A COMPARISON OF TREE-BASED AND TRADITIONAL CLASSIFICATION METHODS

A thesis presented in partial fulfilment of the requirements for the Degree of PhD in Statistics at Massey University.

Robert D Lynn

1994

ABSTRACT

Tree-based discrimination methods provide a way of handling classification and discrimination problems by using decision trees to represent the classification rules. The principal aim of tree-based methods is the segmentation of a data set, in a recursive manner, such that the resulting subgroups are as homogeneous as possible with respect to the categorical response variable. Problems often arise in the real world involving cases with a number of measurements (variables) taken from them. Traditionally, in such circumstances involving two or more groups or populations, researchers have used parametric discrimination methods, namely, linear and quadratic discriminant analysis, as well as the well known non-parametric kernel density estimation and Kth nearest neighbour rules.

In this thesis, all the above types of methods are considered and presented from a methodological point of view. Tree-based methods are summarised in chronological order of introduction, beginning with the Automatic Interaction Detector (AID) method of Morgan and Sonquist (1963) through to the IND method of Buntine (1992).

Given a set of data, the proportion of observations incorrectly classified by a prediction rule is known as the apparent error rate. This error rate is known to underestimate the actual or true error rate associated with the discriminant rule applied to a set of data. Various methods for estimating this actual error rate are considered. Cross-validation is one such method which involves omitting each observation in turn from the data set, calculating a classification rule based on the remaining (n-1) observations and classifying the observation that was omitted. This is carried out n times, that is for each observation in the data set and the total number of misclassified observations is used as the estimate of the error rate.

Simulated continuous explanatory data was used to compare the performance of two traditional discrimination methods, linear and quadratic discriminant analysis, with two tree-based methods, Classification and Regression Trees (CART) and Fast Algorithm for Classification Trees (FACT), using cross-validation error rates. The results showed that linear and/or quadratic discriminant analysis are preferred for normal, less complex data and parallel classification problems while CART is best suited for lognormal, highly complex data and sequential classification problems. Simulation studies using categorical explanatory data also showed linear discriminant analysis to work best for parallel problems and CART for sequential problems while CART was also preferred for smaller sample sizes. FACT was found to perform poorly for both continuous and categorical data. Simulation studies involving the CART method alone provided certain situations where the 0.632 error rate estimate is preferred to cross-validation and the one standard error rule over the zero standard error rule. Studies undertaken using real data sets showed that most of the conclusions drawn from the continuous and categorical simulation studies were valid. Some recommendations are made, both from the literature and personal findings as to what characteristics of tree-based methods are best in particular situations.

Final conclusions are given and some proposals for future research regarding the development of tree-based methods are also discussed. A question worth considering in any future research into this area is the use of non-parametric tests for determining the best splitting variable.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my three supervisors, Associate Professor Dick Brook, Mr Greg Arnold and Dr S Ganesalingam for their constant support and helpful advice throughout my PhD study. I would also like to thank Mum and Dad, Judith and Robin Lynn, for providing me with cheap board and lodgings over the years as well as encouraging me to persevere to the end. I am indebted to Massey University for the use of their computer facilities, and in particular, to the Department of Statistics for providing me with employment over the past five years. Last, and by no means least, I owe a great deal of thanks to Paula McMillan for her efforts in typing this thesis, without her skill in reading my often illegible script this thesis may never have been completed!

ADDITIONAL PUBLICATIONS

Ganesalingam, S and Lynn, R D (1991). Posterior probability based estimator for the overall error rate associated with a linear discriminant function. Occasional Publications in Mathematics and Statistics, **23**, Massey University.

Lynn, R D and Brook, R J (1991). Classification by decision trees and discriminant analysis. New Zealand Statistician, **26**, pp 18-26.

Lynn, R D, Brook, R J and Arnold, G C (1993). A comparison of four classification methods: linear and quadratic discriminant analysis, CART and FACT. Mathematical and Information Sciences Report, Series B: **1**, Massey University.

Table of Contents

1. INTRODUCTION 1

2. TRADITIONAL DATA DISCRIMINATION METHODS 5
2.1 INTRODUCTION 5
2.2 LINEAR DISCRIMINANT ANALYSIS 5
2.2.1 Stepwise Discriminant Analysis 11
2.3 QUADRATIC DISCRIMINANT ANALYSIS 12
2.4 THE ROBUSTNESS OF LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS 12
2.4.1 Modifications to Linear Discriminant Analysis 14
2.5 KERNEL DENSITY ESTIMATION 14
2.6 Kth NEAREST NEIGHBOUR METHODS 17
2.7 CRITIQUES OF KERNEL DENSITY ESTIMATION AND KTH NEAREST NEIGHBOUR METHODS 18

3. A TABULAR COMPARISON ON TEN TREE-BASED METHODS 21
3.1 ORIGINS OF TREE-BASED METHODS 21
3.2 INTRODUCTION 21
3.3 AID 31
3.4 THAID 34
3.5 ID3 36
3.6 CHAID 39
3.7 CART 41
3.8 C4.5 47
3.9 FACT 50
3.10 KnowledgeSeeker 52
3.11 Splus Trees ( ) 55
3.12 IND 59

4. SIMULATION STUDIES INVOLVING CONTINUOUS DATA 65
4.1 INTRODUCTION 65
4.2 ERROR RATES 65
4.3 SIMULATION STUDY I 74
4.3.1 Study Plan 74
4.3.2 Results 76
4.3.3 Summary 82
4.4 SIMULATION STUDY II 83
4.4.1 Study Plan 83
4.4.2 Results 83
4.4.3 Summary and Discussion 91
4.5 THE EFFECTS OF PRIORS ON ERROR RATES 93
4.5.1 Introduction 93
4.5.2 Purpose of this study 93
4.5.3 Study Plan 94
4.5.4 Results 95
4.5.5 Summary 104
4.6 SIMULATION STUDY III 107
4.6.1 Introduction 107
4.6.2 Study Plan 107
4.6.3 Results 108
4.6.4 Summary 112
4.7 CONCLUSIONS 113

5. SIMULATION STUDIES INVOLVING CATEGORICAL DATA 115
5.1 INTRODUCTION 115
5.2 PREVIOUS STUDIES 116
5.3 SIMULATION STUDY I 117
5.3.1 Study Plan 117
5.3.2 Results 118
5.3.3 Summary 121

5.4 SIMULATION STUDY II 122
5.4.1 Introduction 122
5.4.2 Study Plan 122
5.4.3 Results 122
5.4.4 Summary 126
5.5 CONCLUSIONS 127

6. CART SIMULATION STUDY 129
6.1 INTRODUCTION 129
6.2 ERROR RATE ESTIMATION FOR CONTINUOUS DATA IN CART 129
6.2.1 Previous Studies 129
6.2.2 Study Plan 130
6.2.3 Results 132
6.2.4 Summary 143
6.3 ERROR RATE ESTIMATION FOR CATEGORICAL DATA IN CART 144
6.3.1 Study Plan 144
6.3.2 Results 145
6.3.3 Summary 151
6.4 THE STANDARD ERROR RULE IN CART 151
6.4.1 Previous Studies 151
6.4.2 Study Plan 152
6.4.3 Results 152
6.4.4 Summary 158
6.5 TRANSFORMATIONS OF ERROR RATES 158
6.5.1 Study Plan 158
6.5.2 Results 159
6.5.3 Summary 161
6.6 CONCLUSIONS 161

7. CASE STUDIES 165
7.1 INTRODUCTION 165
7.2 PREVIOUS STUDIES 165

7.3 COMPARATIVE STUDIES 166
7.3.1 Methods and Data Sets 166
7.3.2 Cross-Validation Error Rate Results 173
7.3.3 0.632 Error Rate Results 176
7.3.4 Individual Class Error Rates 176
7.3.5 The Standard Error Rule in CART 179
7.3.6 Splus Trees( ) versus CART 179
7.3.7 Summary 181
7.4 ILLUSTRATIVE CASE STUDY 183
7.4.1 Methods and Data 183
7.4.2 Linear Discriminant Analysis 184
7.4.3 CART 186
7.4.4 FACT 189
7.4.5 KnowledgeSeeker 192
7.4.6 Splus Trees( ) 203
7.4.7 Summary 208

8. WHICH CHARACTERISTICS OF TREE-BASED METHODS ARE PREFERRED 209
8.1 INTRODUCTION 209
8.2 WHICH CHARACTERISTICS OF TREE-BASED METHODS ARE PREFERRED? 209
8.2.1 The Method of Splitting 209
8.2.2 Binary versus Multiway Splits 211
8.2.3 Univariate versus Linear Combination Splits 212
8.2.4 Costs and Priors 214
8.2.5 Stopping Rules and Tree Pruning 214
8.3 HUMAN COMPREHENSIBILITY AND USER-FRIENDLINESS OF 216

9. CONCLUSIONS AND PROPOSALS FOR THE FUTURE 221

NOTATION INDEX 235

BIBLIOGRAPHY 239

1. INTRODUCTION

Data often arise in the real world involving many objects with a number of measurements (variables) taken from them. These measurements may be quantitative (continuous or discrete) or qualitative (ordered or unordered categories). The latter may, in some cases, be defined by only two categories and are then binary variables. When more than two categories are involved, instances where the categories can be ordered in a meaningful way are known as ordinal variables, while examples where the categories have no natural ordering are defined as nominal variables. For example, plants may be measured for stem length, stem width and plant height. These measurements are all continuous. A medical study would usually contain information on a patients age, whether he/she smokes or not and whether there is a family history of cancer or not. Age (to the nearest year) is a discrete quantitative variable while the other two variables are binary. A sample survey might ask questions relating to the respondents' educational qualifications, attitudes to race relations and current marital status. Marital status is a nominal variable while the other two variables are ordinal.

Often, the objective of such studies is to distinguish between several groups or populations based on the measurements collected. A botanist may be interested to know which measurements can best distinguish between two related species of plants. A medical practitioner would like to know what variables are best able to predict whether a person will develop cancer or not. A sociologist could be trying to determine if there is any relationship between a person's religious beliefs and various sociological and demographical variables. In such cases involving two or more groups or populations, a large number of methods are available to the botanist/medical practitioner/sociologist to handle the above types of data. The desired intention is that the methods will produce a set of classification or prediction rules, which are both accurate and informative, and serve as a basis for future decisions.

The aim of this thesis is to study and compare the performance of classification methods, both tree-based and more traditional approaches, over a variety of data types with the main goal being to determine in which situations tree-based methods are the preferred approach.

In Chapter 2, the focus is on traditional discrimination and the four most common methods for estimating the conditional density functions of each population in the data set, thereby approximating the Bayes rule. The four methods investigated are linear and quadratic

discriminant analysis, kernel density estimation and Kth nearest neighbour rules. The first two methods are based on parameter estimates while the latter two are wholly non-parametric. A summary table is provided which compares and contrasts each of the above four methods.

In Chapter 3, the focus switches to tree-based classification methods, whose classification rules are portrayed in the form of a decision tree. After surveying the foundations of the tree-based approach to classification, ten tree-based methods are presented from a methodological point of view, examining characteristics such as splitting criteria, stopping rules and interactive and graphical ability among others, as well as critiques of each method from articles in the literature. To conclude the chapter, a summary table is presented comparing all ten tree-based methods.

In Chapter 4, after surveying the various types of error rate estimates that are used in the field of classification, a number of simulation studies are carried out involving continuous explanatory data. In Section 4.2, a comparison is made between two traditional discrimination methods, linear and quadratic discriminant analysis, and two tree-based methods, CART and FACT, in terms of overall accuracy, over every possible combination of five factors involving dimension, sample size, Mahalanobis distance between populations, distribution and priors-covariance structure. In Section 4.4, the same study plan is used except one of the distribution types is changed in order to make comparisons with previous studies. Section 4.5 deals with the estimation of individual class error rates for each of the four methods and how these error rates are affected when the prior probabilities of class membership are altered. The final section of this chapter investigates the reliability of various error rate estimators for three of the methods for predicting the correct class of future observations of the same type.

In Chapter 5, a comparative study is undertaken comparing the four methods used in Chapter 4 for categorical explanatory variables, in particular, five and ten-dimensional binary data. After providing a literature review of previous studies comparing classification methods for such data, a simulation study is carried out using overall accuracy as the measure of classifier performance. In Section 5.4, the reliability of various error rate estimators is determined, as carried out for continuous data in Section 4.6.

Chapter 6 concentrates exclusively on the CART method. Firstly, in Section 6.2, the reliability of various error rate estimation techniques is investigated for continuous data. Data sets are of varying distances between populations. sample sizes and data structure. Four performance criteria are used to evaluate the error rate estimators. In Section 6.3, the same error rate estimators are compared for the categorical data sets used in Chapter 5. In Section 6.4, the so called standard error rule used in CART is analysed while Section 6.5 explores the effects of transforming the error rates.

Chapter 7, firstly, reports the results from a empirical comparison of five classification methods for a number of real world data sets. In Section 7.4, a case study is carried out using some family planning data from India in order to illustrate the approaches taken by linear discriminant analysis and four tree-based methods.

Chapter 8, firstly, compares the approaches taken by tree-based methods to grow a classification tree, through both a survey of the literature and the results of simulation and case studies undertaken in this thesis. Secondly, a subjective comparison of traditional discrimination and tree-based methods is made. A summary of the literature where critical assessment of the interpretability of the two approaches is presented. This is followed by a personal assessment of which method(s) provide the most interpretable and humanly comprehensible models, based on the results of simulation and empirical studies presented in this thesis, as well as personal experience.

In conclusion, a set of recommendations is made, based on the findings of this thesis, as to which methods should be used in which situations. Some proposals for the future development of tree-based programs and research are also presented, after tracing the links and developments of tree-based methods.

2. TRADITIONAL DATA DISCRIMINATION METHODS

2.1 INTRODUCTION

The optimal rule of classification in a p-dimensional, k-class problem is the Bayes rule which is defined to be

$$D_B(\mathbf{x}) = \{\mathbf{x};\ f_i(\mathbf{x})\ \pi_i = \max_j f_j(\mathbf{x})\ \pi_j\} \tag{2.1.1}$$

where $f_i(\mathbf{x})$ is the conditional density of $\mathbf{x}$, given that $\mathbf{x}$ belongs to class i and π_i is the prior probability that $\mathbf{x}$ belongs to class i. The optimal rule for the proportion of observations falsely classified is called the Bayes misclassification error rate. This is calculated as

$$R(B) = 1 - \int \max_i [f_i(\mathbf{x})\ \pi_i]\ d\mathbf{x} \tag{2.1.2}$$

It is very unusual, however, for either the $f_i(\mathbf{x})$ or the π_i to be known. The π_i can easily be estimated by class sample proportions but the $f_i(\mathbf{x})$ are another matter.

This chapter focuses on the four most commonly used methods for estimating the $f_i(\mathbf{x})$, thereby approximating the Bayes rule. The four methods, which attempt to correctly classify a random observation into one of k classes, are linear and quadratic discriminant analysis, kernel density estimation and Kth nearest neighbour rules. The methods are described both algebraically and in words. A table of the assumptions and properties of the four methods is presented in conclusion.

2.2 LINEAR DISCRIMINANT ANALYSIS

Suppose that an object is to be allocated to one of two p-dimensional multivariate ellipsoidal populations on the basis of an observation vector $\mathbf{x}$. Let us assume that observations from the first population, Π_1, occur in a proportion π_1 and the remainder are from Π_2 in the proportion $\pi_2 = (1 - \pi_1)$. Let $f_i(\mathbf{x})$ be the multivariate density of $\mathbf{x}$ in Π_i, with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_i, where

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}\ |\Sigma_i|^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\ \Sigma^{-1}\ (\mathbf{x} - \boldsymbol{\mu}_i)\right] \tag{2.2.1}$$

Suppose that we assign $\mathbf{x}$ to Π_1 if $\mathbf{x}$ is in some region A_1 and to Π_2 if $\mathbf{x}$ is in a region A_2 where A_1 and A_2 form an exhaustive and mutually exclusive partition of the sample space, that is, $Pr(A_1 \cap A_2) = 0$ and $Pr(A_1 \cup A_2) = 1$. Then, the total probability of misclassification, T(A, f), is the proportion of observations from A_1 that are falsely classified as belonging to A_2 and vice-versa. Thus

$$\begin{aligned} T(A, f) &= \pi_1 \int_{A_2} f_1(\mathbf{x})\, d\mathbf{x} + \pi_2 \int_{A_1} f_2(\mathbf{x})\, d\mathbf{x} \\ &= \pi_1[1 - \int_{A_1} f_1(\mathbf{x}) d\mathbf{x}] + \pi_2 \int_{A_1} f_2(\mathbf{x})\, d\mathbf{x} \\ &= \pi_1 + \int_{A_1} [\pi_2 f_2(\mathbf{x}) - \pi_1 f_1(\mathbf{x})]\, d\mathbf{x} \end{aligned} \tag{2.2.2}$$

T(A, f) will be a minimum if $\pi_2 f_2(\mathbf{x}) - \pi_1 f_1(\mathbf{x}) < 0$ for all observations in A_1. That is, the minimum error will occur if the product of the class priors and density functions in A_1 is much larger for Π_1 than for Π_2. With the assumption that $\Sigma_1 = \Sigma_2 = \Sigma$, that is covariance matrices are equal, the optimal rule of allocation $D(\mathbf{x})$ assigns $\mathbf{x}$ to Π_1 if

$$f_1(\mathbf{x}) / f_2(\mathbf{x}) > \pi_2 / \pi_1 \tag{2.2.3}$$

otherwise $\mathbf{x}$ is assigned to Π_2, where the likelihood ratio $f_1(\mathbf{x})/f_2(\mathbf{x})$ is given by

$$f_1(\mathbf{x})/f_2(\mathbf{x}) = \exp[\mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_2) - \tfrac{1}{2}(\mu_1 + \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)] \tag{2.2.4}$$

Taking logarithms produces the rule: assign $\mathbf{x}$ to Π_1 if

$$D(\mathbf{x}) = (\mu_1 - \mu_2)' \Sigma^{-1} [\mathbf{x} - \tfrac{1}{2}(\mu_1 + \mu_2)] > \ln(\pi_2/\pi_1) \tag{2.2.5}$$

otherwise to Π_2.

The above quantity, $D(\mathbf{x})$, is known as the true discriminant function. $D(\mathbf{x})$ is a linear function of $\mathbf{x}$. Now if $\mathbf{x}$ is multivariate ellipsoidal then $D(\mathbf{x})$ will also be multivariate ellipsoidal thus the means and variances of $D(\mathbf{x})$ can also be used to calculate the estimated

error rates from using $D(\mathbf{x})$ as the allocation rule. Now $E[D(\mathbf{x}) \mid \Pi_1]$ is the mean value of $D(\mathbf{x})$ given that $\mathbf{x}$ is from Π_1. Thus

$$\begin{aligned} E[D(\mathbf{x}) \mid \Pi_1] &= [\mu_1 - \tfrac{1}{2}(\mu_1 + \mu_2)]' \Sigma^{-1}(\mu_1 - \mu_2) \\ &= \tfrac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \\ &= \tfrac{1}{2}\delta^2 \end{aligned} \tag{2.2.6}$$

where δ^2 is the square of the true Mahalanobis distance between Π_1 and Π_2.

Similarly

$$E[D(\mathbf{x}) \mid \Pi_2] = -\tfrac{1}{2}\delta^2 \tag{2.2.7}$$

The common variance can be calculated thus:

$$\begin{aligned} &E[D(\mathbf{x}) - D(\mu_i)]^2 = E[(\mathbf{x} - \mu_i)' \Sigma^{-1}(\mu_1 - \mu_2)]^2 \\ &= E[(\mu_1 - \mu_2)' \Sigma^{-1}(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)' \Sigma^{-1}(\mu_1 - \mu_2)] \\ &= (\mu_1 - \mu_2)' \Sigma^{-1} E[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)'] \Sigma^{-1}(\mu_1 - \mu_2)] \\ &= (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \quad \text{as} \quad E[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)'] = \Sigma \\ &= \delta^2 \end{aligned} \tag{2.2.8}$$

Let $R_1(T)$ be the probability of misclassifying an observation from Π_1, so that

$$R_1(T) = \Pr[D(\mathbf{x}) < \ln(\pi_2/\pi_1) \mid \mathbf{x} \in \Pi_1] \tag{2.2.9}$$

Under the assumption of normality (2.2.9) can be expressed by

$$R_1(T) = \Phi\left[\frac{D(\mathbf{x}) - E(D(\mathbf{x}))}{se[D(\mathbf{x})]} < \frac{(\pi_2/\pi_1) - E(D(\mathbf{x}))}{se[D(\mathbf{x})]}\right]$$

$$= \Phi\left[\frac{\ln(\pi_2/\pi_1) - \delta^2/2}{\delta}\right] \tag{2.2.10}$$

and

$$R_2(T) = \Phi\left[\frac{-(\ln(\pi_2/\pi_1) + \delta^2/2)}{\delta}\right] \tag{2.2.11}$$

where $\Phi(.)$ is the cumulative normal distribution function.

If a sample of size n_1 is drawn from Π_1 and size n_2 from Π_2 then μ_i can be replaced by the sample estimate $\bar{\mathbf{x}}_i = \sum_j^{n_i} \mathbf{x}_{ij}/n_i$, (i = 1, 2) and Σ by the estimate of the pooled sample variance, $\mathbf{S}_p$, given by

$$\mathbf{S}_p = [(n_1 - 1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2] / (n_1 + n_2 - 2) \tag{2.2.12}$$

where $\mathbf{S}_i$ are the estimates of Σ_i, (i = 1, 2). If these sample estimates are placed into equation (2.2.5), then the optimal sample allocation rule is to assign a random observation $\mathbf{x}$ to Π_1 if

$$\hat{D}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} [\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)] > \ln(\pi_2/\pi_1) \tag{2.2.13}$$

$\hat{D}(\mathbf{x})$ is the linear discriminant function, the sample estimate of $D(\mathbf{x})$. This discrimination rule assumes that the cost of misclassifying an observation from Π_1 to Π_2, C(2/1), is the same as misclassifying an observation from Π_2 to Π_1, C(1/2). If $C(2/1) \neq C(1/2)$, then (2.2.13) takes the form given so that an observation $\mathbf{x}$ is assigned to Π_1 if

$$\hat{D}(\mathbf{x}) > C(1/2) \ln \pi_2 / C(2/1) \ln \pi_1 \tag{2.2.14}$$

otherwise to Π_2.

In the case of equal a priori probabilities of belonging to a certain class, (2.2.13) simplifies to "classify $\mathbf{x}$ to Π_1 if $\hat{D}(\mathbf{x}) > 0$", otherwise to Π_2.

An alternative way of tackling the problem of classification in the linear discriminant analysis context is to use the group classification functions, $\hat{L}_i(\mathbf{x})$, where

$$\begin{aligned} \hat{L}_i(\mathbf{x}) &= \ln \pi_i + \bar{\mathbf{x}}'_i \mathbf{S}_p^{-1} (\mathbf{x} - \tfrac{1}{2}\bar{\mathbf{x}}_i) \\ &= \ln \pi_i - \tfrac{1}{2}\bar{\mathbf{x}}'_i \mathbf{S}_p^{-1} \bar{\mathbf{x}}_i + \bar{\mathbf{x}}'_i \mathbf{S}_p^{-1} \mathbf{x}_i \qquad (2.2.15) \\ &= a + \mathbf{b}' \mathbf{x}. \qquad (2.2.16) \end{aligned}$$

In the case of k groups, there are k group classification functions, so the rule is to assign $\mathbf{x}$ to Π_m if

$$\hat{L}_m(\mathbf{x}) = \max_i \hat{L}_i(\mathbf{x}), \qquad i = 1, \ldots, k \tag{2.2.17}$$

The above $\hat{L}_i(\mathbf{x})$ can be used to form what has previously been called the linear discriminant function, $\hat{D}(\mathbf{x})$, in the case of k = 2 groups, where

$$\hat{D}(\mathbf{x}) = \hat{L}_1(\mathbf{x}) - \hat{L}_2(\mathbf{x}). \tag{2.2.18}$$

In the case of $k \geq 3$ groups, the situation becomes more complex. A set of linear discriminant functions, $\hat{D}_{ij}(\mathbf{x})$, can be defined as

$$\hat{D}_{ij}(\mathbf{x}) = \hat{L}_i(\mathbf{x}) - \hat{L}_j(\mathbf{x}), \qquad i, j = 1, ..., k \tag{2.2.19}$$

Jennrich (1977) calls these type of linear discriminant functions, group separation functions. In general, there are $C_2^k = k!/2!(k-2)! = k(k-1)/2$ such group separation functions in a sample consisting of k distinct classes. The rule is to assign $\mathbf{x}$ to Π_i if

$$\begin{aligned} &\hat{D}_{ij}(\mathbf{x}) > 0, \quad \forall\, i < j, \text{ and} \\ &\hat{D}_{ij}(\mathbf{x}) < 0, \quad \forall\, i > j \end{aligned} \tag{2.2.20}$$

Otherwise, $\mathbf{x}$ is assigned to one of the other (k-1) classes. For example, in the case of k = 4 classes, the rule is

$$\begin{aligned} &\text{Assign to } \Pi_1 \text{ if } \hat{D}_{1j}(\mathbf{x}) > 0, \quad j = 2, 3, 4 \\ &\text{Assign to } \Pi_2 \text{ if } \hat{D}_{12}(\mathbf{x}) < 0, \quad \hat{D}_{23}(\mathbf{x}) > 0 \quad \text{and} \quad \hat{D}_{24}(\mathbf{x}) > 0 \\ &\text{Assign to } \Pi_3 \text{ if } \hat{D}_{12}(\mathbf{x}) < 0, \quad \hat{D}_{23}(\mathbf{x}) < 0 \quad \text{and} \quad \hat{D}_{34}(\mathbf{x}) > 0 \end{aligned}$$

otherwise assign $\mathbf{x}$ to Π_4.

Figure 2.1 illustrates this procedure for a set of twenty six urinary samples (see data set R from Table 7.1) involving two chemical measurements (androsterone and etiocholanolone) taken from eleven healthy heterosexual and fifteen healthy homosexual males. Linear discriminant analysis is ideally suited to this problem, with the linear discriminant function (LDF), providing perfect discrimination between the two classes, showing that for men with the same values of androsterone, homosexuals have higher values of etiocholanolone than heterosexuals. Therefore, the separation is in a linear combination of the two variables.

Plot of Etiocholanolone (mg/24 hours) against Androsterone (mg/24 hours) from Urinary Samples for 26 Healthy Heterosexuals and Homosexuals with Linear Discriminant Function

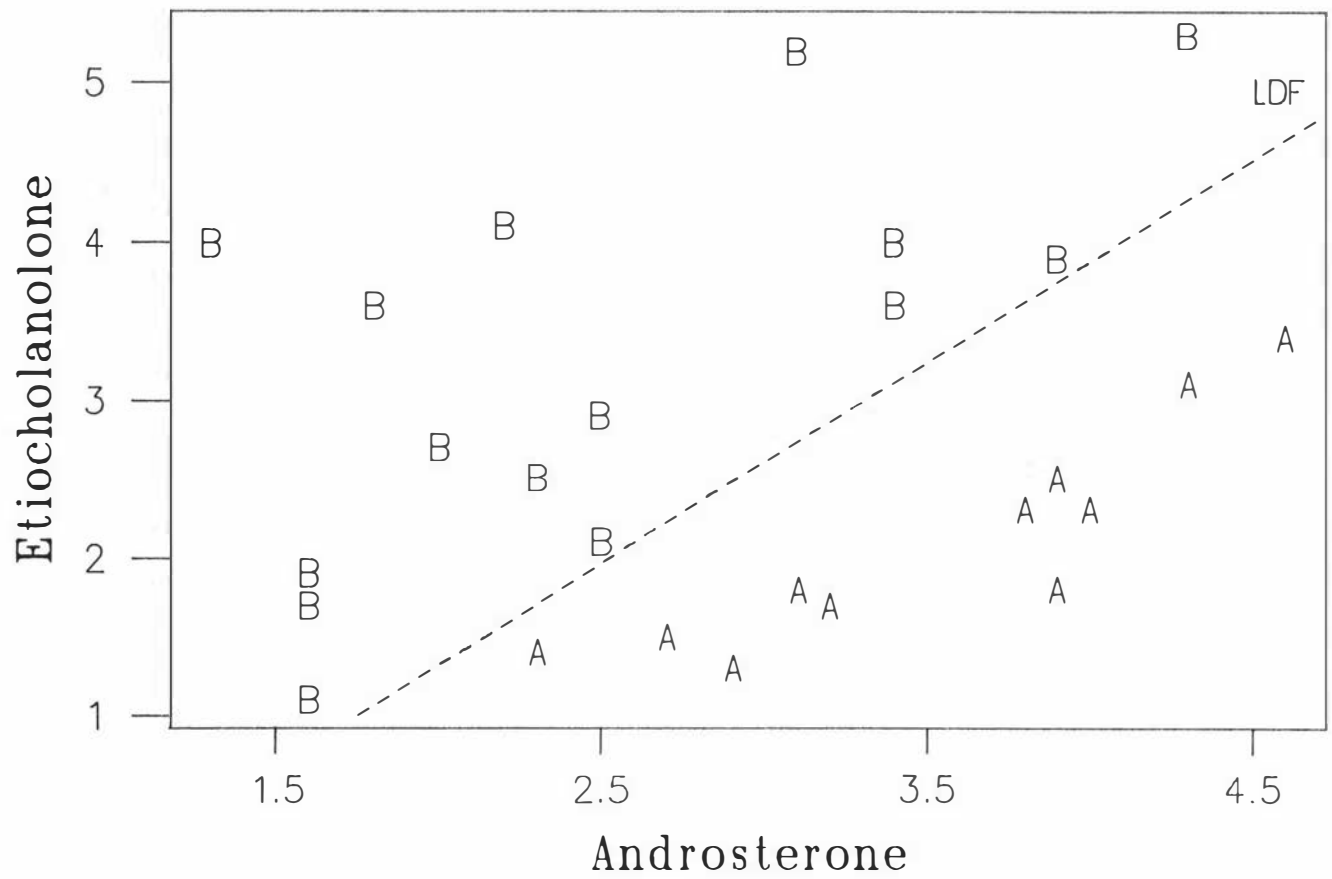


Figure 2.1

2.2.1 Stepwise Discriminant Analysis

A special application of linear discriminant analysis is stepwise discriminant analysis, whereby only a subset of the original p variables is selected to carry out the discriminant analysis. As above, suppose that a sample of dimension p contains n_1 observations from Π_1 and n_2 observations from Π_2. Variables are chosen to either enter or leave the model according to whether the Wilks-Lambda ratio of between to within class variance is greater than or less than a pre-specified significance level, while also taking into account the variables that are already in the model. Alternatively, the partial correlation coefficients between each predictor variable and the class variable can be used to force a variable to either enter or leave the model. In essence, the variables that contribute most to the discriminatory power of the model are selected to carry out the discriminant analysis. However, authors such as Habbema et al (1974) have pointed out that the best q variables selected by stepwise discriminant analysis, may not necessarily be the "best" variables for this type of data, just the best for this particular sample. Snapinn and Knoke (1989) give illustrations where the apparent error rate, the error rate found from resubstituting the original sample, should never be used on a data set which contains only the best q variables, selected by stepwise discriminant analysis, though this has been found to hold for most discrimination methods.

2.3 QUADRATIC DISCRIMINANT ANALYSIS

In real world situations, the requirement of equal covariance matrices is rarely satisfied, though the differences are often too small to cause any deterioration in the performance of linear discriminant analysis. In cases where the covariance matrices are quite different, though, and $\mathbf{x}$ is p-dimensional multivariate ellipsoidal, quadratic discriminant analysis is the appropriate method to use where the discriminant function is

$$Q(\mathbf{x}) = \ln[f_1(\mathbf{x}) / f_2(\mathbf{x})] \tag{2.3.1}$$

and we assign $\mathbf{x}$ to Π_1 if $Q(\mathbf{x}) > \ln(\pi_2/\pi_1)$. If we again replace μ_i and Σ_i by the sample estimates $\bar{\mathbf{x}}_i$ and S_i then the result is the sample estimate of $Q(\mathbf{x})$, the quadratic discriminant function:

$$\hat{Q}(\mathbf{x}) = -\frac{1}{2}\ln\left[\frac{|S_1|}{|S_2|}\right] - \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_1)' S_1^{-1}(\mathbf{x} - \bar{\mathbf{x}}_1) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_2)' S_2^{-1}(\mathbf{x} - \bar{\mathbf{x}}_2) \tag{2.3.2}$$

and $\mathbf{x}$ is assigned to Π_1 if $\hat{Q}(\mathbf{x}) > \ln(\pi_2/\pi_1)$. (2.3.3)

2.4 THE ROBUSTNESS OF LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS

Simulation studies previously undertaken by authors such as Lachenbruch et al (1973), Marks and Dunn (1974), Aitchison et al (1977), Krzanowski (1977) and Wahl and Kronmal (1977), among others, have made many interesting discoveries about the robustness of linear and quadratic discriminant analysis, henceforth called LDA and QDA, respectively.

Seber (1984) has summarised many of these findings, noting in particular that:

(i) LDA and QDA should perform equally well when covariances are roughly equal and the number of variables, p, is small ($p \leq 6$).

(ii) For small samples ($n_1, n_2 < 25$) and small covariance differences and/or p large, LDA is preferred, but when both covariance differences and p are large neither method is recommended.

(iii) QDA is better than LDA when both covariance differences and the number of variables are large, $p > 6$ and when sample sizes are large. It is suggested that $n_1 = n_2 = 25$ and $p = 4$ as a minimum with 25 additional observations per class for every extra two dimensions.

(iv) QDA should not be used in poorly posed situations, that is where the number of variables is not much less than the class sample sizes, resulting in $\mathbf{S}_i$ being a poor estimate of Σ_i. The extreme case occurs where the data is ill-posed, when $p > n_i$ meaning $\mathbf{S}_i$ does not exist.

From the above findings and many applications in case studies, both LDA and QDA should be best when each class is multivariate normal with equal covariances matrices and the ratio of class sample size to dimension is large. LDA is fairly robust to any departures from these conditions, while QDA is only robust to differences in the class covariance matrices.

Morgan and Sonquist (1963), while introducing their Automatic Interaction Detector (AID) program, came to the conclusion that the usual parametric methods of classification were often inadequate in analysing survey data, noting in particular that parametric methods were:

(i) Unable to handle interaction effects, without adding many extra terms to the model, as interactions may be quite complex, affected in different ways by different parts of the data set.

(ii) Variables may not have linear effects, thus there is the need to create many extra terms (for example, quadratic, cubic etc).

(iii) Not good at handling categorical explanatory variables, especially those with many categories. Parametric methods usually treat categorical explanatory variables as a number of binary variables and create linear functions from those variables. As with (i) and (ii) above, the number of variables in the data set could increase dramatically and the data matrix will be sparse.

(iv) Not robust to errors in the variables such as decimal points in the wrong place. As parametric methods make use of all the data at once, any errors in the measurements will lead to false classification rules.

(v) Affected adversely by intercorrelations among the explanatory variables used in the analysis. These correlations interfere with assessing the importance of individual variables.

2.4.1 Modifications to Linear Discriminant Analysis

Friedman (1989) and Raveh (1989) have tried two modifications to LDA in an attempt to solve the problems mentioned in the last section. Friedman developed a method called regularized discriminant analysis (RDA) especially for ill-posed situations, as outlined earlier. He noted that when the sample covariance is singular then the $p - n_i + 1$ smallest eigenvalues are estimated to be 0. The net effect of this biasing phenomenon on discriminant analysis is to, sometimes, dramatically exaggerate the importance associated with the low variance subspace spanned by the eigenvectors corresponding to the eigenvalues near zero.

Friedman tackles the problem using regularization, whereby a reduction in the variance of the sample-based estimates is carried out so as to minimise a potentially increased bias. Two regularization parameters, $0 \leq \lambda \leq 1$ and $0 \leq \gamma \leq 1$, are selected in order to jointly minimise future misclassification errors. The above two parameters are incorporated into a variance function that controls the degree of shrinkage of the individual class covariance matrices that contribute to the pooled estimate. Simulation studies showed that RDA was much better than LDA and QDA in cases where the covariances were spherical. In those cases where the covariance matrices were highly ellipsoidal, and equal, LDA did best but when covariance matrices were unequal, RDA did best.

Raveh developed non-metric discriminant analysis (NDA), a method that requires none of the parametric assumptions required by both LDA and QDA (for example, the assumption of multivariate normality). NDA uses a separation measure so that as many observations as possible from Π_1 are greater than or less than the observations from Π_2. Thus NDA is based solely on the ranks of the individual observations and not the actual values. Through simulation studies, Raveh has shown that NDA is error-free for non-overlapping distributions and that NDA outperforms LDA in cases where the distribution of the data is highly non-normal or where covariance matrices are quite different.

2.5 KERNEL DENSITY ESTIMATION

Often, it occurs that a parametric form cannot be assumed for the $f_i(\mathbf{x})$ so that in order to apply the likelihood-ratio test, the $f_i(\mathbf{x})$ have to be estimated using an unstructured approach. An example where this approach is necessary is in a sample exhibiting gross non-normality and unequal covariance matrices. Such an approach is called non-parametric estimation.

'Kernel density estimation is one form of non-parametric estimation. Hand (1981), Seber (1984) and Fukunaga (1990) all give excellent summaries of how kernel density estimation works.

The basic idea behind kernel density estimation is to use the sample data ($\mathbf{x}_{ij}$, i = 1, ..., k and j = 1, ..., n_i) to estimate each of the $f_i(\mathbf{x})$'s. Hand (1981) first considers the case p = 1. Suppose that $v(x/\Pi_m)$ is the number of sample points belonging to class m, $1 \leq m \leq k$, with values less than or equal to x and $\hat{F}(x/\Pi_m)$ is the estimate of the cumulative distribution and is given by

$$\hat{F}(x/\Pi_m) = \frac{\text{number of class m observations} \leq x}{\text{total number of class m observations}}$$

$$= \frac{v(x/\Pi_m)}{n_m} \qquad (2.5.1)$$

This function cannot be differentiated because the probabilities are not continuous, but an approximation to the derivative of $\hat{F}(x/\Pi_m)$ can be made.

$$\hat{f}(x/\Pi_m) = \frac{\hat{F}(x + h/\Pi_m) - \hat{F}(x - h/\Pi_m)}{2h}$$

$$= \frac{[v(x+h/\Pi_m) - v(x-h/\Pi_m)]/n_m}{2h} \qquad (2.5.2)$$

This can then be rewritten as

$$\hat{f}(x/\Pi_m) = \frac{1}{n_m h} \sum_{i=1}^{n_m} k_0\left(\frac{x - x_i}{h}\right) \qquad (2.5.3)$$

where x_j, i = 1, 2, ..., n_m are the class m sample observations and

$$k_0(z) = \begin{cases} 0, & \text{for } |z| > 1 \\ \frac{1}{2}, & \text{for } |z| \leq 1 \end{cases} \qquad (2.5.4)$$

where $z = (x - x_j) / h$.

The above implies that every point in the interval (x-h, x+h) contributes $1/2n_m h$ to the estimation of the density function at x while any points that lie outside of that interval contribute nothing. It seems wrong that a point near the boundary of (x-h, x+h) carries the same weight as a point very close to x, while a point just outside of the interval contributes nothing. To overcome this problem, a smoothing weighting function is used. For instance, let $k_0(z)$ be from the normal distribution with zero mean so that observations closest to x have the greatest weighting but all observations in the sample contribute to some degree in the calculation of the density function. Another alternative would be to use the uniform distribution as the weighting function so that every observation is equally weighted. "Any other unimodal density could be used as a kernel." (Seber, 1984, p 322.)

Classification is determined by use of the likelihood ratio statistic, $-\ln(f_1(\mathbf{x})/f_2(\mathbf{x}))$, and whether this value is greater than or less than a threshold value. In the case of two populations and $p \geq 2$, the kernel density discrimination function, $\hat{K}(\mathbf{x})$, is given by

$$\hat{K}(\mathbf{x}) = -\ln \frac{\hat{f}_1(\mathbf{x}/\Pi_1)}{\hat{f}_2(\mathbf{x}/\Pi_2)}$$

$$= \frac{(1/n_1) \sum_{j=1}^{n_1} k_1 (\mathbf{x} - \mathbf{x}_{1j})}{(1/n_2) \sum_{j=1}^{n_2} k_2 (\mathbf{x} - \mathbf{x}_{2j})} \qquad (2.5.5)$$

An observation is assigned to Π_1 if

$$\hat{K}(\mathbf{x}) > \ln\left(\frac{\pi_2}{\pi_1}\right) \qquad (2.5.6)$$

otherwise to Π_2. That is, the density estimates of $f_i(\mathbf{x})$ are based on the number of points from Π_i within the region $(\mathbf{x} - \mathbf{h},\ \mathbf{x} + \mathbf{h})$ where $\mathbf{h}$ is a p-dimensional area.

2.6 Kth NEAREST NEIGHBOUR METHODS

The Kth nearest neighbour method (K-NN) is another tool that is used whenever the class density functions, $f_i(\mathbf{x})$, are unknown. In fact, this was the first non-parametric method for classification and was introduced by Fix and Hodges (1951).

The idea behind the method is relatively simple. Cover and Hart (1967) define a random observation $\mathbf{x}_m$, $\mathbf{x}_m \varepsilon\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, as the nearest neighbour to $\mathbf{x}$ if

$$\min d(\mathbf{x}_j, \mathbf{x}) = d(\mathbf{x}_m, \mathbf{x}), \qquad j = 1, 2, ..., n. \tag{2.6.1}$$

where $d(\mathbf{x}_j, \mathbf{x})$ is a distance function. The nearest neighbour rule decides that $\mathbf{x}$ belongs to the class Π_i of its neighbour $\mathbf{x}_m$. The above is the single nearest neighbour rule, that is $K = 1$, and only applies to the single nearest neighbour to $\mathbf{x}$. All other observations are ignored.

The idea is extended naturally to the K nearest neighbours of $\mathbf{x}$. Lachenbruch (1975) describes the general K-NN rule as follows. Suppose there are n_1 and n_2 sample observations from Π_1 and Π_2 respectively. Suppose that the objective is to classify an observation $\mathbf{x}$ to one of Π_1 or Π_2. Using a distance function, $d(\mathbf{x}_{ij}, \mathbf{x})$, order the values, $\mathbf{x}_{ij}$. Let K_i be the number of observations from Π_i among the K closest observations to $\mathbf{x}$. The rule is to assign $\mathbf{x}$ to Π_1 if

$$\frac{K_1}{n_1} > \frac{K_2}{n_2} \tag{2.6.2}$$

otherwise to Π_2. In other words, the procedure involves the relatively simple concept of assigning a random observation $\mathbf{x}$ to the class having the greater proportion of observations closest to $\mathbf{x}$. As $n_i \rightarrow \infty$, it has been found that (2.6.2) tends to the maximum likelihood rule.

2.7 CRITIQUES OF KERNEL DENSITY ESTIMATION AND KTH NEAREST NEIGHBOUR METHODS

Simulation studies by various authors, including Habbema et al (1974), have shown that the kernel density method for classification was just as efficient as LDA in the case of normally distributed data but when non-normality occurred, kernel density estimation was superior.

Feng et al (1993) have shown that the K-NN method were often slow in terms of running time, as was kernel density estimation. For most of the case studies tested in that paper, the K-NN method produced a very low apparent error rate but quite often the test sample error rate found from the classification rules on another set of data that was not used to construct the classifier, was comparatively high. This fact calls into question the reliability of the classification rules proposed by the K-NN method.

Breiman et al (1984), p 17, have criticised both the above non-parametric classification methods on the following grounds:

(i) They are sensitive to the choice of a metric $\|\mathbf{x}\|$ and there is usually no intrinsically preferred definition.

(ii) There is no natural or simple way to handle categorical variables and missing data.

(iii) They are computationally expensive as classifiers. The learning sample must be stored, the inter-point distances and classification rule recomputed for each new observation.

(iv) Most serious, they give very little usable information regarding the structure of the data. That is, neither of the two methods provide a set of simple and intuitive set of classification rules.

2.8 SUMMARY TABLE OF THE ASSUMPTIONS AND PROPERTIES OF TRADITIONAL DATA DISCRIMINATION METHODS

ISSUE	LDA	QDA	KERNEL DENSITY	K-NN
Optimality	Multivariate ellipsoidality and equal covariance matrices within each group.	Multivariate ellipsoidality and equal covariance matrices within each group.	No assumptions about the present distribution of variables.	No assumptions about the present distribution of variables.
Types of Variables	Quantitative	Quantitative.	Quantitative.	Quantitative.
Computations	Computations are based on class sample means and the pooled covariance matrix of the class sample covariances.	Computations are based on class sample means and individual class sample covariances.	Uses the individual data values and a weighting function $k_0(z)$.	Uses the K observations that are closest to **x**.
Discrimination Rule	Assign **x** to Π_1 if $\hat{D}(\mathbf{x}) > \ln(\pi_2/\pi_1)$ where $\hat{D}(\mathbf{x}) = \ln[f_1(\mathbf{x})/f_2(\mathbf{x})]$. Otherwise, assign **x** to Π_2.	Assign **x** to Π_1 if $\hat{Q}(\mathbf{x}) > \ln(\pi_2/\pi_1)$ where $\hat{Q}(\mathbf{x}) = \ln[f_1(\mathbf{x})/f_2(\mathbf{x})]$. Otherwise, assign **x** to Π_2.	Assign **x** to Π_1 if $\hat{K}(\mathbf{x}) > \ln(\pi_2/\pi_1)$ where $\hat{K}(\mathbf{x}) = \left(\frac{n_2}{n_1}\right)\frac{\sum_{j=1}^{n_1} k_1(\mathbf{x}-\mathbf{x}_{1j})}{\sum_{j=1}^{n_2} k_2(\mathbf{x}-\mathbf{x}_{2j})}$. Otherwise, assign to Π_2.	Assign **x** to Π_1 if $K_1/n_1 > K_2/n_2$ where K_i is the number of class i observations among the K nearest neighbours to **x**. Otherwise, assign **x** to Π_2.
Critiques of the Method	• Very fast. • Robust to mild non-normality in the variables. • Unable to properly handle interaction effects. • Not good at handling categorical predictor variables.	• Robust to departures from equal covariance matrices. • Not suited when the ratio of dimension to sample size is small. • Very sensitive to departures from normality in the variables. • Not good at handling categorical predictor variables.	• Not affected by either non-normality or unequal covariance matrices. • Produces reliable classification rules. • Gives very little usable information about the data. • Not good at handling categorical predictor variables.	• Not affected by either non-normality or unequal covariance matrices. • Produces unreliable classification rules. • Gives very little usable information about the data. • Not good at handling categorical predictor variables.

KEY

Optimality	Under what conditions is the method optimal?
Types of Variables	For what type of explanatory variables is the method suited?
Computations	What statistics/values are used in the calculation of the discrimination rules?
Discrimination Rule	How is an observation **x** classified to one of the two populations?
Critiques of the Method	From the literature, what are four key advantages or disadvantages of the method?

3. A TABULAR COMPARISON ON TEN TREE-BASED METHODS

3.1 ORIGINS OF TREE-BASED METHODS

Tree-based methods of classification are children of the computer age. The idea of decision trees could only have been dreamed of before the introduction of the computer as the amount of number crunching required to construct a data-intensive method, such as a decision tree classifier, would have been far too much for the simple adding machine.

The ideas behind decision tree methods were originally developed by Belson (1959). The approach he proposed to take was the binary segmentation of a data set, in a recursive manner, so that each of the subgroups formed would be as homogeneous as possible with respect to the response variable. At each stage of the analysis the predictor variable providing the "best" dichotomous partition would be chosen to partition the subgroup into two further subgroups.

Belson's proposals form the foundations from which all tree-based methods have been built, being the result of a dissatisfaction with standard statistical techniques. In conclusion, Belson states, "[t]he method as I have described it is, it is true, a movement towards a more empirical way of doing things; but it is just as much a movement away from a sophistication which is too often either baffling or misleading" (Belson, 1959, p 75).

3.2 INTRODUCTION

In this chapter, ten tree-based methods are to be summarised in chronological order of introduction. The methods are:

(i)	AID	Automatic Interaction Detector.
(ii)	THAID	THeta AID
(iii)	ID3	
(iv)	CHAID	CHi-squared AID
(v)	CART	Classification And Regression Trees
(vi)	C4.5	
(vii)	FACT	Fast Algorithm for Classification Trees

(viii) KnowledgeSeeker
(ix) Splus Trees()
(x) IND

The ten methods are to be tabulated on the following bases:

Author(s)	Who developed the method?
Introduction	The year the method was introduced and a short summary of how the author(s) describe(s) the method.
Classification/Regression	What type of response variable is handled?
Tree Growth	Is the tree grown on all the data or only on a subset of the data? (a) Splitting Method. What rules are used by each method to partition the data? (b) Type of Splits. How does the method partition the data? Binary/Multiway splits on a single variable (US) or a linear combination of variables? (c) Costs/Priors. Are these incorporated into the splitting algorithm? (d) Stopping Rules. What types of stopping rules are employed? (e) Node Classification/Prediction. How are the nodes classified/predicted?
Tree Pruning	What pruning procedure, if any, exists in the method?
Validation Procedures	Is there validation of the decision rules constructed by means of a test sample or cross-validation?
Interactive Ability	Do facilities exist in the program for the user to easily interact with the tree-growing procedures, so as to automatically change the splitting variable, stop splitting, etc?
Graphical Ability	Can the program display the decision tree graphically?

One-Stage Optimality	Does the method only look for the optimal split of the current node? If not, do facilities exist to examine the effects of splits at the next one or two stages of the tree-growing process?
Missing Values	How are missing values handled?
Criticisms	What is written about the program in the literature? What problems have been identified?
Examples in the literature	A list of important papers using the method.

Finally, a short, summary table comparing all ten tree-based methods is given, over all the attributes described above.

Safavian and Landgrebe (1991) have conducted a survey of a large number of tree-based methods. Their paper includes a summary table comparing each of the methods in terms of the assumptions each approach makes, their performance criterion and some of the specific requirements for each method. The approach adopted in this chapter is intended to be more than a mere enumeration of material or collated bibliography of tree-based methods. Therefore, only ten such methods have been selected and presented in detail with some attempt at critical comparison.

A major difference between the tree-based methods studied in this chapter is the way in which the aims of Belson are carried out, that is, the method of splitting. A decision tree procedure either uses binary splitting where the data is segmented into two groups, or uses multiway splitting, where the data can be split into more than two groups. These splits can either be carried out on a single variable, called univariate splits (US), or on a linear combination of variables.

Figure 3.1 illustrates the method of binary splits for Fisher's Iris data (data set H from Table 7.1) involving three species of iris (I. virginica, I. setosa and I. versicolor) each with 50 cases and measurements taken on four variables (sepal length, sepal width, petal length and petal width). Although the problem involves four variables, only two variables are used to form the tree using the CART algorithm (see Section 3.7). The first split is on petal length and

asks the question as to whether petal length < 1.95, and if so, observations are sent to the right. It would be possible for the next split also to be on petal length, but here the next split is whether or not petal width < 1.75, for those cases where petal length > 1.95.

The tree produced by CART is shown in Figure 3.2. As only two splits were made, there were three terminal nodes, where a node is defined as a subset of the data and a terminal node is a terminal subset of the data which is assigned to one of k classes. The terminal node at top left consisted of 50 class 1 (I. virginica) flowers and was classified as class 1. The terminal node at bottom left was classified as class 2 (I. setosa) and consisted of 49 class 2 and 5 class 3 flowers (I. virginica). The terminal node at bottom right was classified as class 3, consisting of 1 and 45 in classes 2 and 3 respectively. Notice that there were 6 flowers overall misclassified by the classification tree.

Figure 3.3 illustrates the method of multiway splitting for the same data. The FACT algorithm (see Section 3.9) was used to partition the data. In this case, there is only one split carried out, that being on petal width, but it is a three-way split, dividing the data into three subgroups. The first subgroup corresponds to the case where petal width < 0.787 (split 1a), while the second subgroup corresponds to flowers where 0.787 < petal width < 1.677 (split 1b), with the third subgroup corresponding to all those cases where petal width > 1.677. The FACT tree is shown in Figure 3.4. Basically, the FACT tree has split the data into three homogeneous terminal nodes using only one multiway split, compared with the two needed in the binary splits example. As with Figure 3.2, 6 flowers have been misclassified by this classification tree.

The above examples were both carried out using only one variable at a time. Figure 3.5 provides an illustration of a linear combination split, when used with CART. The first split is the same as that of Figure 3.1. The second split, however, involves both petal length and petal width and asks the question, for those cases where petal length > 1.95, as to whether 0.209 * petal length + 0.977 * petal width < 2.51. The CART tree for this example is given in Figure 3.6. The major difference from Figure 3.2 is that no class 3 cases were misclassified, and two fewer cases overall were misclassified.

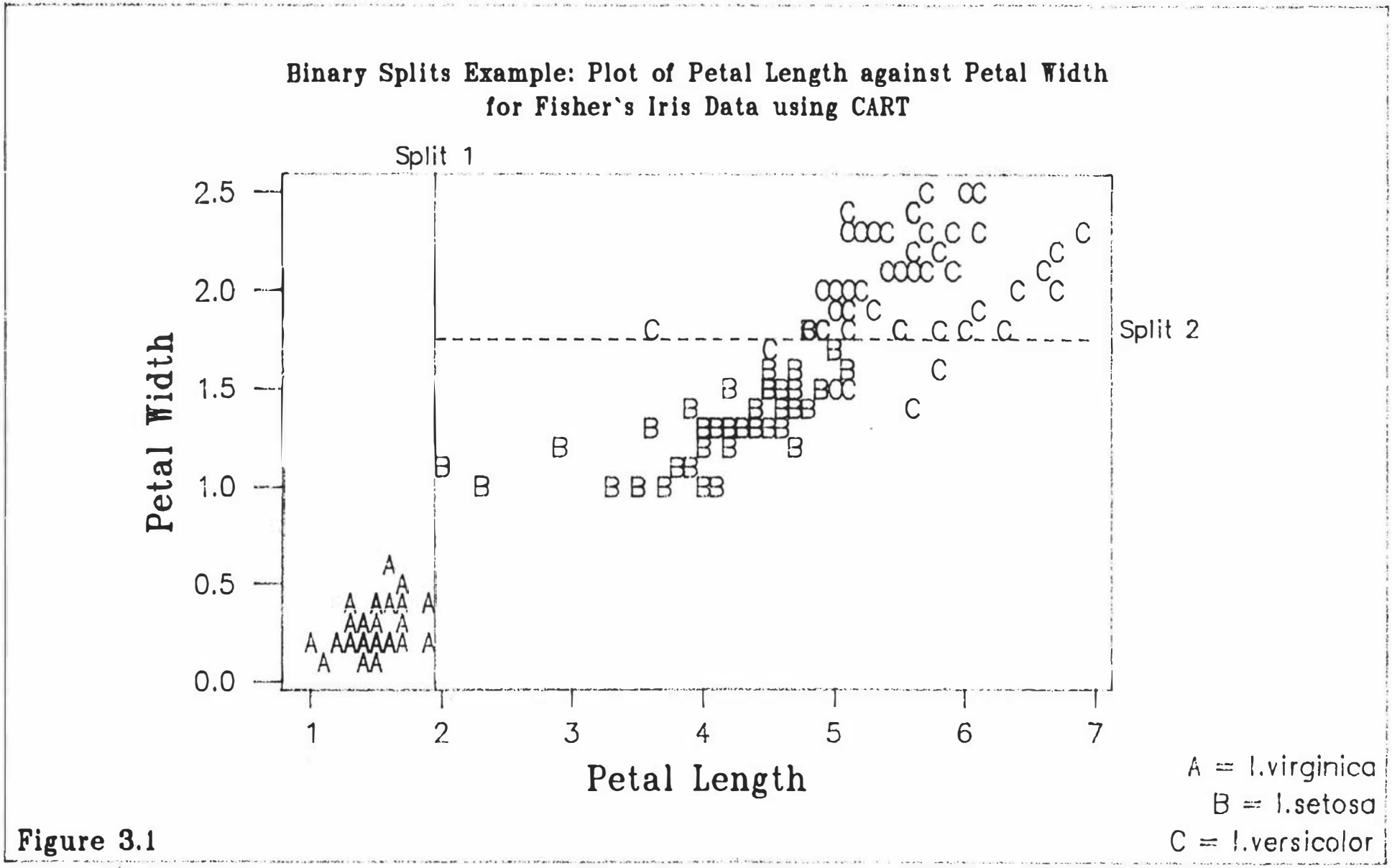
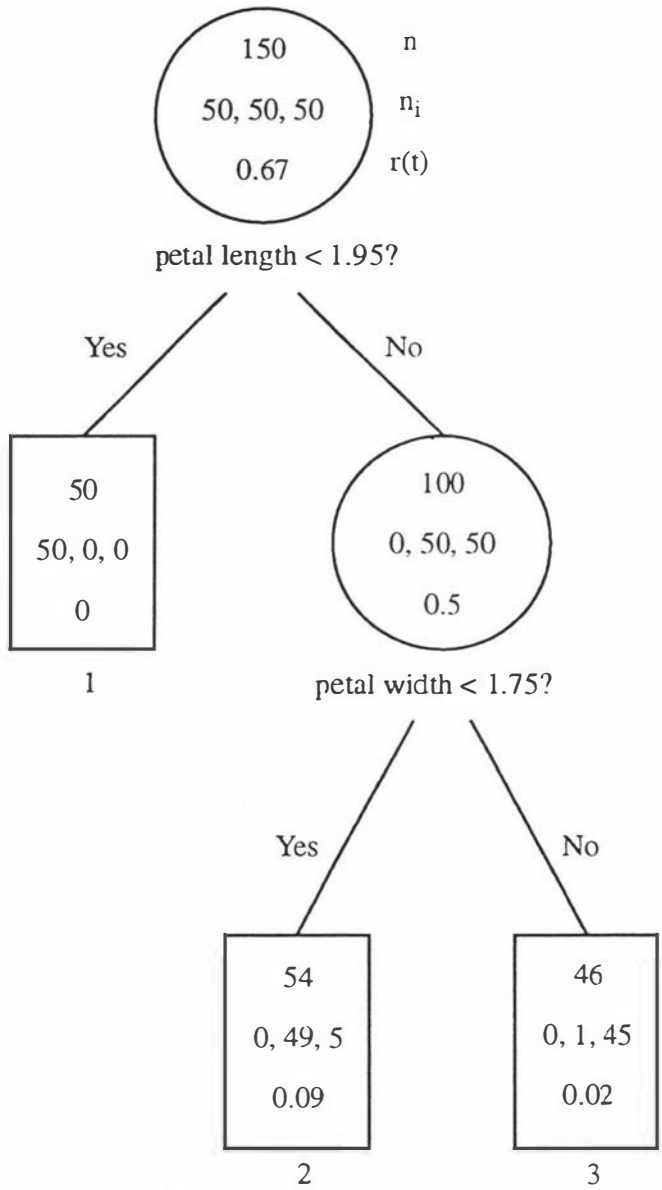

Binary Splits Example: Plot of Petal Length against Petal Width
for Fisher's Iris Data using CART
Split 1
Split 2
Petal Width
Petal Length
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
A = I.virginica
B = I.setosa
C = I.versicolor

Figure 3.1



n = sample size at each node
n_i = sample size for class i at each node
r(t) = purity measure at each node (that is, the proportion of observations not from the class with the largest number of observations at each node)

Circles represent decision nodes which have to be split on while rectangles represent terminal nodes which are assigned to a particular class given below the node.

Figure 3.2: CART Tree for Fisher's Iris Data

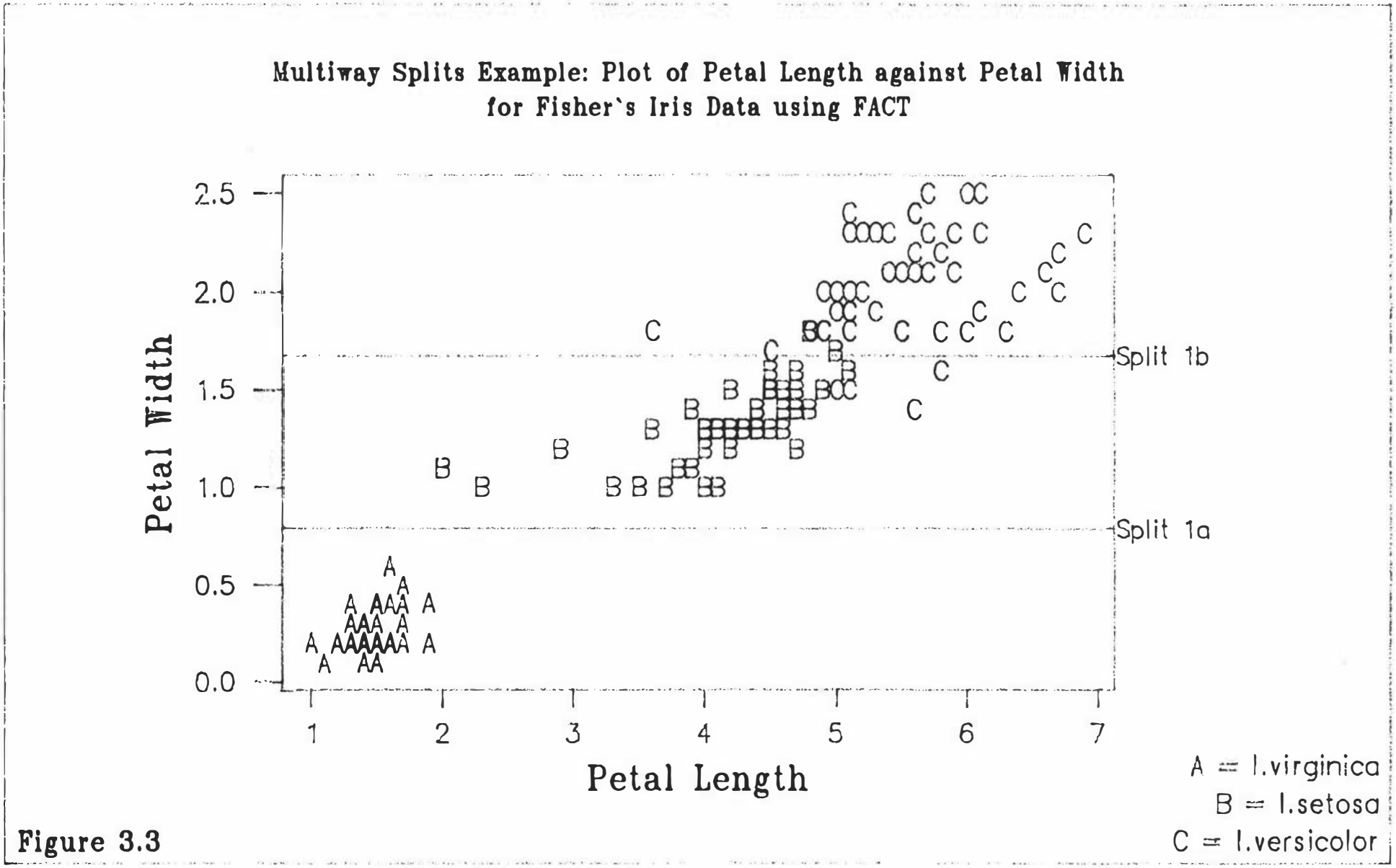

Multiway Splits Example: Plot of Petal Length against Petal Width
for Fisher's Iris Data using FACT
Petal Width
Petal Length
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
Split 1b
Split 1a
A = I.virginica
B = I.setosa
C = I.versicolor

Figure 3.3

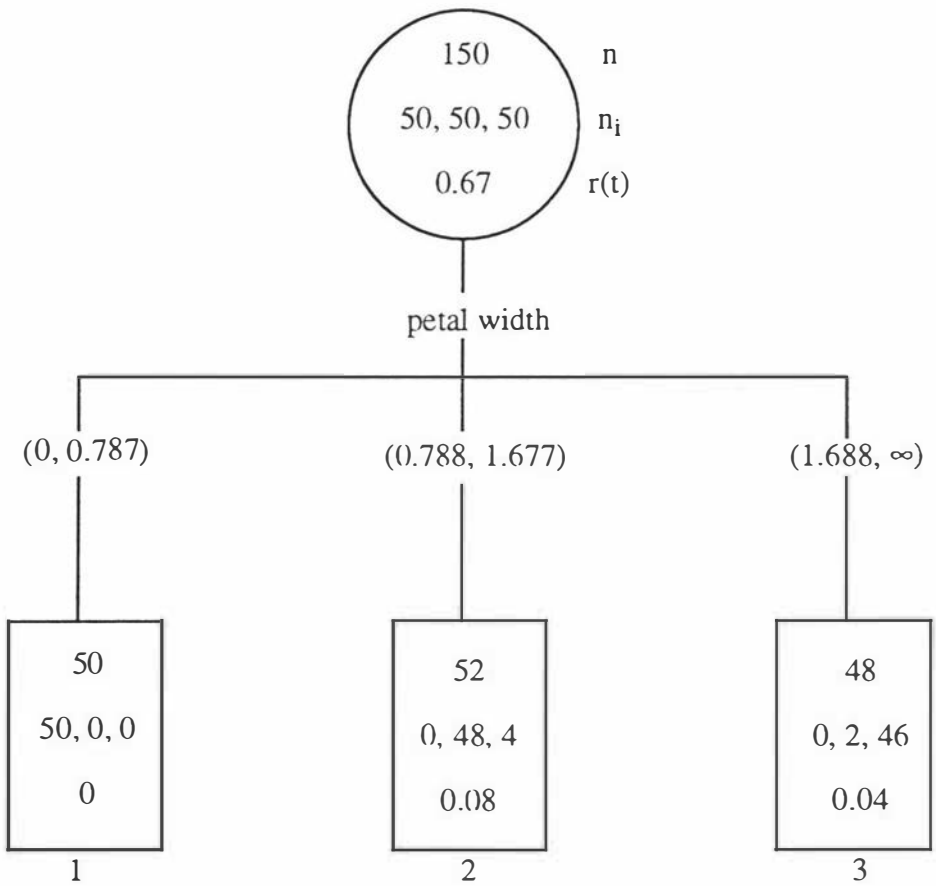


Figure 3.4: FACT Tree for Fisher's Iris Data

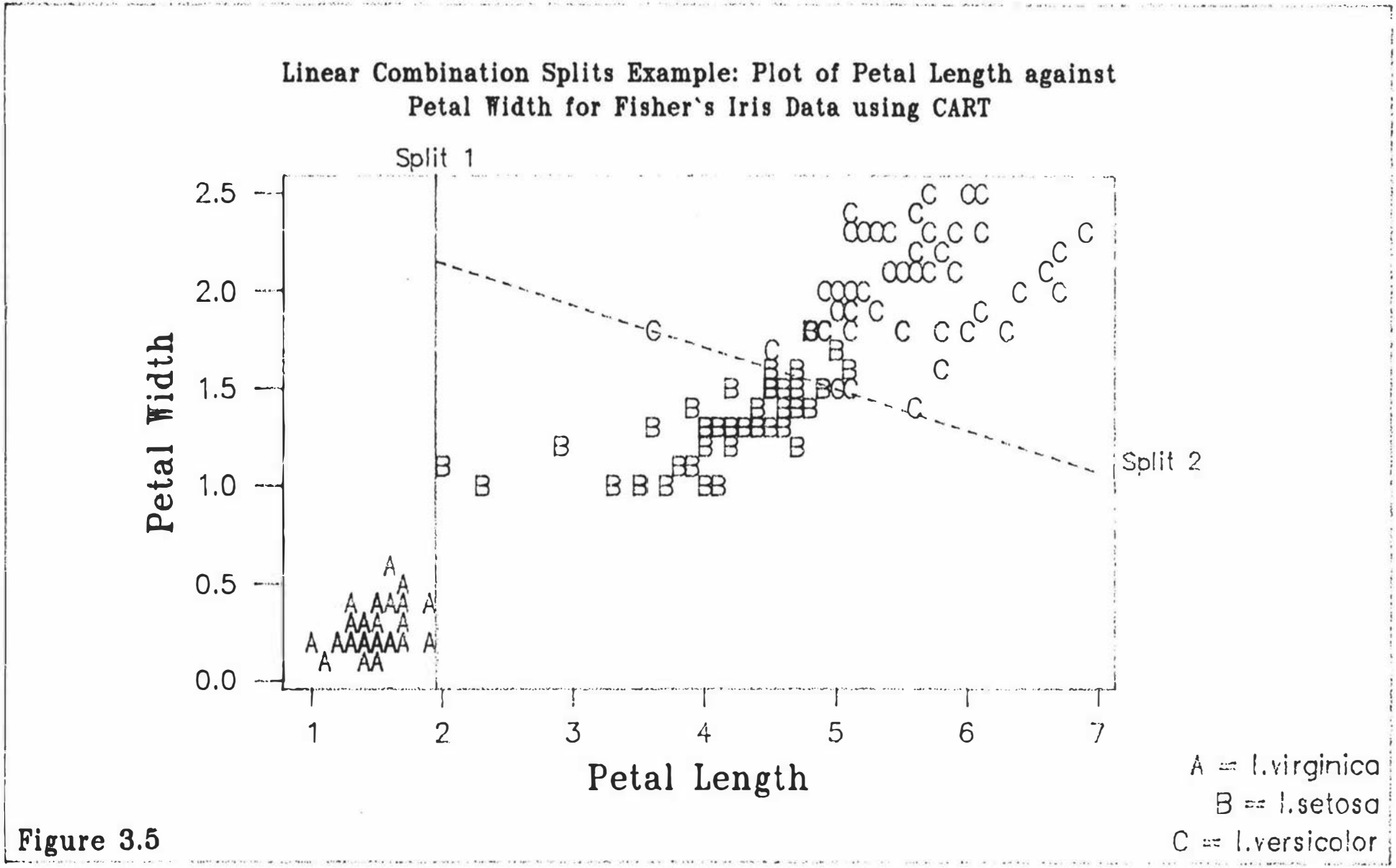

Linear Combination Splits Example: Plot of Petal Length against
Petal Width for Fisher's Iris Data using CART
Split 1
Split 2
Petal Width
Petal Length
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
A = I.virginica
B = I.setosa
C = I.versicolor

Figure 3.5

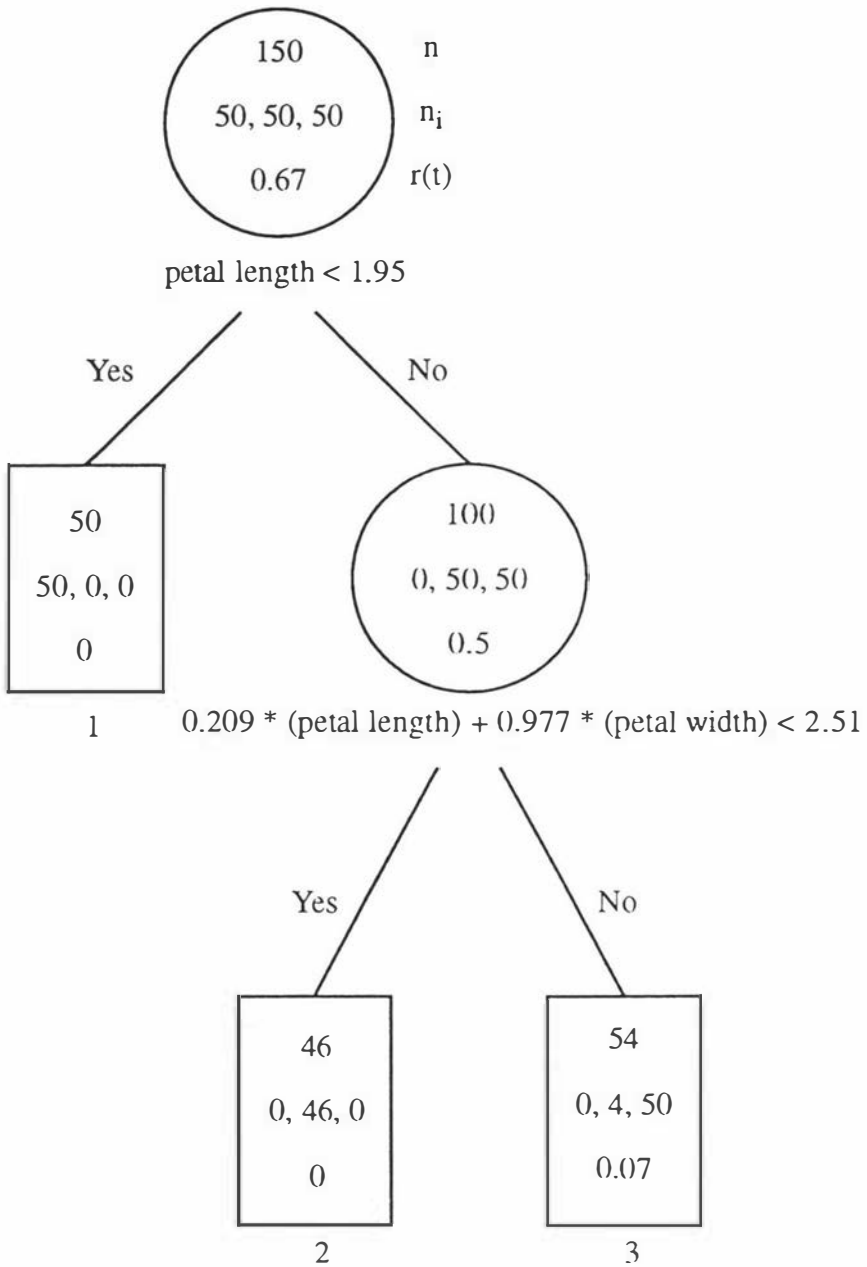


Figure 3.6: CART Tree for Fisher's Iris Data - Linear Combination Split

3.3 AID

Author(s)	J N Morgan and J A Sonquist (USA).
Introduction	The Automatic Interaction Detector was published in 1963. The essence of the algorithm is the sequential application of a one-way analysis of variance model (Morgan and Sonquist, 1963). The purpose of the program was to handle interactions and inter-correlations among the data in a more explicit way.
Classification/Regression	Designed to perform regression using a continuous dependent variable, although dichotomous dependent variables can be handled by transforming one of the two categories into a proportion.
Tree Growth:	The tree is grown on all the data set.
- Splitting Method	Sonquist (1964), summarises the four steps in the tree growing procedure as follows: (i) Choose, for splitting, the node, t, with the largest total sum of squares, $TSS_t = \Sigma y_t^2 - (\Sigma y_t)^2/n$. (ii) Split each variable, x_j, into two subgroups such that this division leads to the biggest decrease in unexplained sum of squares, i.e. maximise $BSS_j = (n_1\bar{y}_1^2 + n_2\bar{y}_2^2) - n_t\bar{y}_t^2$. (iii) Partition variable x_m over node t where BSS_m is $\max_j BSS_j$. (iv) Return to step (i).
- Type of Splits	Binary splits are the only method used and are carried out on only one variable at a time.
- Priors/Costs	No.
- Stopping Rules	Direct stopping rules are used. A number of different criteria exist for stopping tree growth: (i) $TSS_t < R * TSS_1$, where R is a parameter $0 \leq R \leq 1$, and TSS_1 is the total sum of squares for the whole sample. (ii) $BSS_j < Q * TSS_1$ over all x_j, $0 \leq Q \leq 1$. (iii) Number of unsplit nodes > P. (iv) Sample size of each unsplit node < L.

- Node Classification/Prediction

All observations within a node are assigned the average value for the response variable in that node.

Tree Pruning	No.
Validation Procedures	No.
Interactive Ability	No.
Graphical Ability	No.
One-Stage Optimality?	Fielding (1977) describes the r-step lookahead option used by AID III, the then current version of AID. The earlier version of AID grew the tree on a sequential basis so that the prediction error was only minimised at each step of the analysis, that is, "stage by stage optimization". The r-step lookahead option is an attempt to improve on the stage by stage procedure. > For m predictors there will be m tentative splits under consideration. The best splits for each predictor on these subgroups is then obtained. One now has m^2 possible two-stage trees under consideration. This process could be continued for r stages with m^r possible trees. Clearly this lookahead option could involve a tremendous amount of computation and information storage were it not restricted. The current version of AID III limits the lookahead steps to three, including the first split (Ibid, p 249). Morgan (1993), in a personal correspondence, noted that the repeated use of the lookahead feature failed to find any useful applications or examples, and was dropped in later versions of the program. One might think it would find offsetting effects, as when young men and old women are more likely to go to the hospital, but the sequential strategy seems to uncover these too according to Morgan.
Missing Values	Missing values are replaced by class means estimated from non-missing values in the learning sample.

Criticisms

The AID algorithm has been criticised by a number of authors, including Einhorn (1972), Doyle (1973), Kass (1975), Doyle and Fenwick (1975), Kass (1980) and de Ville (1990). The principal reasons for this criticism are:

(i) It requires very large sample sizes, usually > 1000 observations.
(ii) It does not take the intercorrelations among the predictors into account.
(iii) It is not robust to deviations from normality in the variables.
(iv) The tree size is affected too much by noise in the data.
(v) Only binary splits are carried out.
(vi) Most importantly, there is no validation of the prediction rules constructed, either by testing for significance, or using an independent test sample.

Morgan (1993) has responded to these criticisms. He believes that (ii) is wrong, except that once a split is made on one predictor, it may leave groups where a second predictor has lost whatever power it had, but that information is useful to know. For example, in searching for what makes people happy, the program splits first on the quality of the network of friends, then on health, and only then on income! Of criticism (iii), Morgan affirms that this is true of any least squares procedure, though AID alerts the user to isolated cases by splitting these off into a separate subgroup. Problem (v) is irrelevant according to Morgan, since multiple splits on the same predictor are possible, and it is wasteful to start with k subgroups when k-1 will do. The loss of information from grouping data is small, and a very few subgroups contain almost all the information. The last criticism, claims Morgan, is not a function of the program but of the user, who can always grow the tree on three quarters of the sample and see how well the final groups account for the variance in the other quarter. It must be noted, however, that it is wasteful to not use all the data in the tree growing phase and test sample estimates of error are highly varible for small samples. (See Section 4.2 and Breiman et al, 1984.)

Examples in the Literature:

Assael, H (1970). Segmenting markets by groups purchasing behaviour: An application of the AID technique, Journal of Marketing Research, **7**, pp 153-158.

Heald, J I (1972). The application of the automatic interaction detector programme and multiple regression techniques to assessment of store performance and site selection, Operational Research Quarterly, **23**, pp 445-457.

Muxworthy, D T (1972). Review of AID III, British Sociological Association Maths, Statistics and Computing Applications Group Newsletter, **9**.

3.4 THAID

Author(s) J N Morgan and R C Messenger (USA).

Introduction Developed in 1973, THeta AID was designed as an extension of the AID algorithm (Morgan and Sonquist, 1963) to handle categorical dependent variables. It was "... viewed as a simplified version of the present AID". (Messenger and Mandell, 1972, p 18.)

Classification/Regression Designed for classification specifically for use on nominally scaled dependent variables.

Tree Growth: The tree is grown on all the data set.

- Splitting Method Two methods for splitting are used.

(i) Theta criterion or what Messenger and Mandell (ibid, p 12), call "optimal prediction-to-the-mode strategy". The objective is to find the split at the unsplit node t which maximises:

$$\theta_{y/x} = \sum_{i=1}^{2} \left(\frac{n_i}{n_t}\right)\left(\frac{m_i}{n_i}\right) = \frac{1}{n_t}(m_1 + m_2)$$

where n_t = total number of observations in node t
n_i = total number of observations in the ith split group
m_i = total number of misclassified observations in the ith split group.

Or else, the Delta criterion could be used. Messenger and Mandell, (ibid, p 15), define this as "... based on the simple notion that one should find split groups whose probability distributions differ maximally from the original group and hence from each other".

The basic idea is to find the split on the variable for which

$$\delta_{y/x} = n_1 \sum_{j=1}^{k} |p_j - p_{1j}| + n_2 \sum_{j=1}^{k} |p_j - p_{2j}|$$

where p_j = proportion of observations from class j in node t, j = 1, ..., k

and p_{1j} = proportion of observations from class j in split group 1.

Note that the authors of THAID recommend the Delta criterion for splitting if the ratio of sample size of the largest group to the second largest group is greater than 2:1 (Messenger and Mandell, 1973).

- Type of Splits Only binary splits are carried out using only one variable at a time.

- Priors/Costs No.

- Stopping Rules Direct stopping rules are used. Stop if:
(i) $n_s/2 < n_{min}$, where n_{min} is a preset parameter, and
(ii) either $\theta_{y/x} < \theta_{min}$ or $\delta_{y/x} < \delta_{min}$.

- Node Classification/Prediction
Assign a terminal node to the class with the largest number of observations in that node.

Tree Pruning No.

Validation Procedures No.

Interactive Ability No.

Graphical Ability No.

One-Stage Optimality? Yes.

Missing Values Missing values are replaced by class means estimated from non-missing values in the learning sample.

Criticisms	Most of the criticisms levelled at AID, are also valid for THAID. Basically the method does not know when to stop. Kass (1980), p 120, also states that, "[k]nowledge of the theoretical behaviour of the Theta criterion is lacking still". Morgan (1993) has written in, noting that there is a maximum-likelihood χ^2 splitting option available in the new SEARCH program which has replaced THAID. The procedure is designed to maximise stability and remove the chance of erratic results.
Examples in the Literature	Morgan, J N (1990). A conditional analysis of movers' housing responses. Journal of Economic Behaviour and Organisation.

3.5 ID3

Author(s)	J R Quinlan (Australia)
Introduction	Introduced in 1979, this procedure is in the family of recursive partitioning, tree-based algorithms, although it is from the machine learning rather than the statistical literature. Quinlan (1983), describes the method as "... recover[ing] valuable information from large masses of low grade data by a process of inductive inference".
Classification/Regression	Handles classification problems only.
Tree Growth:	A subset of the original learning sample, called a 'window' is chosen at random and a decision tree formed that correctly classifies all observations in the window. All objects in the learning sample, but not in the window are then classified using this tree. If the tree gives the correct classification for all objects then this tree is declared optimal, otherwise some more observations are added to the window with the tree-growing and evaluation process being repeated. The process continues until all cases in the learning sample are correctly classified.
- Splitting Method	Splitting is achieved by means of an information measure. An observation is determined to belong to class 1 with probability $p_1 = m/(n+m)$ and to class 2 with probability $p_2 = n/(m+n)$, where m and n are the number of observations from class 1 and class 2 respectively. The expected information needed to classify an object using a tree is:

$$I(m, n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2.$$

Let a variable, x_j, considered for partitioning, contain v distinct categories $\{A_1, ..., A_v\}$. The node t that is to be considered for splitting will be split into v descendant nodes, $t_1, ..., t_v$, each described by one particular category of x_j. The information required for the subtrees with t_i is $I(m_i, n_i)$, where m_i and n_i are the number of class 1 and 2 observations in the ith node. The expected information required for trees partitioned on x_j at the root node is

$$E(x_j) = \sum_{i=1}^{v} [(m_i + n_i)/(m+n)] * I(m_i, n_i)$$

where the weight of the ith branch is the proportion of objects in t that belong to t_i. Information gained by branching on x_j is

$$gain(x_j) = I(m, n) - E(x_j).$$

ID3 examines all variables, x_j, j = 1, ..., p, and chooses x_j to maximise $gain(x_j)$. This process is continued on the recursively found nodes, $t_1, ..., t_v$. It is known as the **gain criterion**.

- Type of Splits

Multiway splits are used here. In fact, splitting is carried out using every possible value of a variable. If a predictor variable is continuous, some form of clustering of the values is carried out before splitting. Only univariate splits are carried out.

- Priors/Costs

No.

- Stopping Rules

(i) Stop when all cases in the learning sample are correctly classified.

(ii) An alternative stopping rule is: Use the χ^2 statistic to determine if the categories of variable x_j are independent of those class of objects in S. No further testing of the variables (splitting) is done if that variables irrelevance cannot be rejected at a very high confidence level.

- Node Classification/Prediction

Assign a terminal node to the class with the largest number of observations in that node.

Tree Pruning

No.

Validation Procedures

No.

Interactive Ability	No.
Graphical Ability	No.
One-Stage Optimality?	Yes.
Missing Values	Observations can either be discarded from the data set before splitting, or, alternatively, use the ratio of class sample sizes multiplied by what Quinlan calls a 'token' to find a predicted value of the variable for which a particular observation is missing.
Criticisms	Many of the problems inherent in AID have been also found present in this algorithm. deVille (1990) discusses the following problems. (i) Biased towards the selection of variables with many categories though they may not be the best predictor. (ii) Do not know when to stop. ID3 continues splitting on nodes with only a small number of observations. The resultant decision tree would not hold up in the real world, being principally a function of the data at hand. (iii) Overly large trees are too complex and not easy to understand. Quinlan et al (1986), p 164, states that "[e]mpirical investigations have found that trees generated from such sets are usually simpler and more accurate than those constructed from random samples". However, in the next paragraph, it is argued that "... decision trees produced by any top-down approach are more complex than can be justified by the data." (Ibid, p 164).
Examples in the Literature	Schwartz, S, Wiles, J, Gough, I and Phillips, S (1993). Connectionist, rule-based and Bayesian decision aids: an empirical comparison, in "Artificial Intelligence Frontiers in Statistics", D J Hand (ed), London: Chapman & Hall, pp 264-278.

3.6 CHAID

Author(s)	G V Kass (South Africa)
Introduction	Developed in 1980 as an offshoot of AID for use with categorical response variables. It was designed to tackle the criticisms of AID by "... embedding the partitioning problem in a significance testing framework" (Kass, 1980, p 120). Known as CHi-squared AID.
Classification/Regression	Handles only classification problems.
Tree Growth:	The tree is grown on all the data set.
- Splitting Method	According to Kass, the splitting method proceeds as follows: (i) For each predictor variable in turn, with the dependent variable having k classes, cross-tabulate the categories of the predictor with the dependent variable. Go to step (ii). (ii) Find the pair of categories of the 2*k subtable that are least significantly different. Merge the two categories into one compound category if the significance does not exceed a critical value. Repeat this step until no more mergers can be found. (iii) For each compound category consisting of three or more of the original categories, find the most significant binary split into which the merger may be resolved. If the significance exceeds a critical value, implement the split and return to step (ii). (iv) Calculate the significance level of each optimally merged predictor and isolate the most significant one. If this significance is beyond a criterion value, split the data according to the merged categories of the chosen predictor. (iv) Return to step (i) for each, as yet, unsplit node t.
- Type of Splits	Multiway splitting can be used with this method, for example, three-way, four-way or larger splits. Single variable splits only are carried out.
- Costs/Priors	No.

- Stopping Rules	Stop if: (i) $n_s < n_{min}$. (ii) The split on the optimally merged predictor $< \chi^2_{(g-1),\,\alpha}$ on g compound categories for a preset value of α.
Tree Pruning	No.
Validation Procedures	No.
Interactive Ability	A current version of CHAID runs on SPSS for windows. A series of menus with a mouse button allow the user to set values for the parameters used in the tree-growing process, and begin the analysis. The tree-growing process can be interrupted at any point and the values of the parameters altered.
Graphical Ability	The graphical display of the tree works in unison with the tree-building process. As a node is split into a number of sub-nodes, the results are displayed immediately on screen by means of a decision tree. The level of detail about each node and each split carried out can also be altered.
One-Stage Optimality?	Yes.
Missing Values	These are excluded from the tree-growing process.
Criticisms	No criticism of CHAID directly has been discovered in the literature. The weakness of the method, however, would appear to be the lack of any procedure for validating the results. As Einhorn (1972), p 368, stated, eight years before CHAID appeared, "... the results should be subjected to a more rigorous criterion than statistical significance or some other statistical criterion". He noted, "replication is the backbone of science and when techniques capitalise on chance fluctuations in the particular sample at hand, it is imperative to replicate (or cross-validate) the results on a new set of cases" (Ibid, p 368). Another possible criticism of CHAID is the use of a direct stopping rule. The authors of CART, Breiman et al (1984), criticised this type of stopping rule on the following basis. If the significance level is set too high (large p-values), then there is too much splitting so that the tree is too large and just a reflection of the sample data. If the significance level is too

low, then one may cease splitting too early and declare a node as terminal when there still existed splits with large decreases in impurity.

Examples in the Literature

3.7 CART

Author(s)

L Breiman, J H Friedman, R A Olshen and C J Stone (USA).

Introduction

Breiman et al began work on recursive partitioning in the 1970's. Their work was completed with the publication of the CART monograph (Breiman et al, 1984). The purpose of the algorithm had the dual goals of providing a set of accurate decision rules in the form of a tree that were easily intepretable while seeking to solve the problems inherent in some of the earlier methods of the above type, such as AID and THAID.

Classification/Regression

Standing for Classification and Regression Trees, CART handles both numeric and categorical response variables. For comparison with the other methods and simplicity, everything henceforth will be described in the classification context only.

Tree Growth:

The tree is grown on the whole data set. If, however, the data set is overly large, a tree can be grown on only a subsample of the data.

- Splitting Method

At the root node of the tree, the splitting variable is chosen to maximise the class purity, that is, as many observations as possible are from the same class, of the two descendant nodes, these being the two sets of points that went either left or right when the chosen variable was split, as well as aiding the future growth of the tree.

Two different criteria are available in CART to achieve the above two aims, namely the Gini and twoing splitting criteria. Breiman et al (1984) have found that the final classification tree generated is fairly insensitive to the choice of a splitting rule. The Gini splitting criterion works in the following way. Suppose at a node t, an object can be assigned to class i with probability p(i/t) while the estimated probability that the object is in class j is p(j/t), then the estimated probability of misclassification under the Gini index is

$$i(t) = \sum_{i \neq j} p(i/t)\, p(j/t).$$

The Gini criterion seeks to maximise the function

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

where p_L and p_R are the proportion of observations at node t sent left and right respectively by the split.

The twoing criterion seeks to amalgamate the set of k classes into two superclasses, C_1 and C_2. The measure of goodness-of-split, $\Delta i(s, t)$, is computed as if it were a two class problem. "The idea is then, at every node, to select the conglomeration of classes into two superclasses so that considered as a two-class problem, the greatest decrease in node impurity is realised". (Breiman et al, 1984, p 105). The twoing splitting rule thus maximises

$$\Phi(s/t) = p_L \sum_j |p(j/t_L) - p(j/t)| + p_R \sum_j | p(j/t_R) - p(j/t)].$$

In general, Breiman et al state that Gini tends to split into one small pure node and one large impure node, that is, to separate the classes out one at a time. Breiman et al call this end cut preference. In contrast, twoing favours splits that tend to make the two descendant nodes as pure as possible in the two superclasses. "It gives strategic splits and informs the user of class similarities. ... [It] attempts to group together large numbers of classes that are similar in some characteristic [near the top of the tree] ... [and] attempts to isolate single classes [near the bottom of the tree]." (ibid, p 105). The twoing criterion is in fact the same as the delta criterion used in THAID (see Morgan and Messenger, 1973).

- Type of Splits

CART is a **binary** recursive partitioning algorithm that sends a case either left or right. Univariate splits using both ordered and categorical variables can be carried out as well as linear combination splits for ordered or quantitative variables only.

- Costs/Priors

Both priors and misclassification costs can be varied and incorporated into the CART tree building process. Priors can be varied to take account of samples that are not representative of the populations from which they came, that is, the class sample proportions are different from the class proportions in the population. Misclassification costs can also be varied to take into account instances whereby it is more serious to misclassify some class(es) than other class(es). At a node t, the

estimated probability of misclassification using the Gini index is

$$\sum_{j,i} C(i/j)\, p(i/t)\, p(j/t).$$

- Stopping Rules

The process is terminated only when all nodes that have not yet been split are pure, or if the node size for all unsplit nodes falls below a specified value.

- Node Classification/Prediction

A node is assigned to the class with the largest number of observations in that node, in the case of priors proportional to sample size and unit costs. If either priors or costs are varied, then classification of a node must also take into account the values of the costs of misclassification and class priors.

Tree Pruning

It is clear that CART's stopping rule produces a tree that could be very large, with an overly optimistic error rate and many of the splits near the bottom of the tree occurring only because of noise in the data. To guard against this possibility, CART employs a backwards recursive node recombination or pruning algorithm on the completed tree. The algorithm proceeds as follows. For any subtree T of T_{max}, where T_{max} is the fully grown tree, define the cost-complexity measure, $R_\alpha(T)$ as

$$R_\alpha(T) = R(A_T) + \alpha\, L(T)$$

where $R(A_T)$ is the resubstitution estimate of the accuracy of the subtree, L(T) is the number of terminal nodes in T and $\alpha \geq 0$ is the complexity parameter. For each value of α find the subtree T_α that minimises $R_\alpha(T)$ above. In practice, each successive pair of descendant nodes is recombined and an estimate of accuracy is compared for the split/unsplit situations using the cost-complexity function above. A larger penalty is assigned to a larger sized tree. If there is no improvement in accuracy then the two descendant nodes are recombined and tested for accuracy in the same manner. If there is some improvement from splitting, this subtree is retained and the process continues trying to find the next smallest subtree which produces an increase in accuracy through splitting. The end result is a sequence of trees with decreasing size and increasing resubstitution error rate.

A 'honest' sized tree can be obtained by either running an independent test sample down this sequence of subtrees and selecting the tree having the minimum error rate or using g-fold cross-validation, $2 \leq g \leq n$. With cross-validation in the CART

context, a tree of maximum size is grown in each of the sets of size (n - (n/g)) and the pruning algorithm is carried out as for the original sample. Then, each of the g sets of omitted observations (size = n/g) can be used as an independent test sample for the sequence of subtrees created by the pruning algorithm. The sizes of each of the g trees are averaged and the tree from the learning sample that is closest in size to the average of the chosen cross-validated trees is then selected. Breiman et al recommend using a measure of error as well, whereby

$$se(R(\hat{T})) = [R(\hat{T})\,(1 - R(\hat{T}))/n]^{1/2}$$

is the standard error estimate for the test sample or cross-validated error rate $R(\hat{T})$. The idea is to choose the smallest tree within one standard error of $R(\hat{T})$. This was recommended to reduce the size of the decision tree created as it was found that independent error rate estimates were fairly constant over quite a wide range of tree sizes.

Validation Procedures

As seen above, test sample validation and cross-validation are used by CART to select the "right-sized" tree. These two techniques are also used to estimate the true error rate of the prediction rules obtained.

Interactive Ability

CART by Systat provides an enhanced version of CART that is more user-friendly than the original. There is some interaction with a menu system, however a whole tree must be grown before any alterations can be made. CART, though, does not allow you to change the splitting variable at an intermediate stage of the tree-growing process.

Graphical Ability

CART by Systat produces files that can be displayed by an independent graphics program, after the CART analysis has been carried out. This is not possible in the original CART.

One-Stage Optimality?

Yes.

Missing Values

Missing values are handled by what Breiman et al call surrogate splits. This is defined as follows. Suppose that s^* is the optimal partition of a node t into t_L and t_R. If a split, s_j, is carried out on a variable, x_j, then the probability that s_j sends the cases in t the same way as s^* is

$$p(s^*, s_j) = p_{LL}(s^*, s_j) + p_{RR}(s^*, s_j)$$

where

$$p_{LL}(s^*, s_j) = p(t_L \cap t'_L)/p(t)$$

and t'_L is the set of observations sent left by s_j. A surrogate split, $\tilde{s}_j$ on x_j, occurs if

$$p(s^*, \tilde{s}_j) = \max_{s_j} p(s^*, s_j)$$

Breiman et al define a surrogate split as the split on x_j that most accurately predicts the action of s^*.

This then leads to the use of surrogate splits with missing values. If a case has a missing value for the splitting variable, so that s^* is not defined for that case, then for all the non-missing variables for that case, find the best surrogate split, $\tilde{s}_j$, and split the case using $\tilde{s}_j$.

Criticisms

CART could be considered as a 'watershed' in the development of tree-based methods in that it veered away from the direct stopping rules of previous decision tree-based methods and adopted the approach of 'grow an overly large tree, then prune and validate'. In taking this approach it set a benchmark for future methods to build on, as well as being open to criticism.

Loh and Vanichsetakul (1988) criticise CART on the following bases:

(i) Based on sort and search principles.
(ii) Typically no more accurate than LDA.
(iii) Too slow if cross-validation is employed.
(iv) Uses only binary splits.
(v) The cross-validation estimate of error is not genuine as it was also used to select the tree size.
(vi) Produces different results when the variables are transformed.

Breiman and Friedman (1988) answered all these criticisms as well as criticising the FACT program of Loh and Vanichesetakul.

Quinlan (1987) criticises the pruning algorithm employed by CART. First, he believes that there is no valid reason why the cost-complexity model should be favoured over any other. Second, he does not know why the sequence of subtrees produced should be abandoned after selection of the best tree.

Lastly, he feels that the use of cross-validation is computationally expensive.

In an empirical study using a wide variety of data sets, comparing a number of tree-based as well as statistical algorithms and neural networks, Feng et al (1993) found that CART performed rather well over all the data sets. They also found that CART tended to produce the smallest sized trees, hence the simplest trees. Also on the positive side, "... [the introduction of] a cost-handling mechanism in the testing phase (in CART) can make a visible improvement compared to, say, C4.5". (ibid, p 51). Feng et al found that CART produced results much closer to those produced by traditional statistical algorithms than other tree-based methods. They suggested that this could be due to the fact that CART incorporates a cost structure.

On the negative side, they found that CART's pruning algorithm was not all that efficient and wrongly assumed that there was a single global parameter for the amount of pruning to be done. As well, they found that CART can prune too heavily if the one standard error rule is used and there is very little noise in the data. Using the zero standard error rule, however, can mean that trees are too large if there is noise in the data.

Examples in the Literature

Grajski, K A, Breiman, L, Viano Di Prisco, G and Freeman, W J (1986). Classification of EEG spatial patterns with tree-structured methodology: CART, IEEE Transactions on Biomedical Engineering, **33**, pp 1076-1086.

Ildiko, E F and Lanteri, S (1989). Classification models: discriminant analysis, SIMCA, CART, Chemometrics and Intelligent Laboratory Systems, **5**, pp 247-256.

Crawford, S L and Souders, S K (1990). A comparison of two conceptual clustering algorithms, International Journal of Pattern Recognition and Artificial Intelligence, **4**, pp 409-420.

3.8 C4.5

Author(s) J R Quinlan (Australia).

Introduction Published in 1986, C4.5 is a descendant of ID3 (Quinlan, 1979). It is described by Quinlan et al (1986), p 157, as "... a new inductive inference tool that is capable of dealing with large volumes of messy, real-world data". Unlike ID3, though, the tree-growing process is followed by a number of pruning procedures. In addition, a larger range of options and parameter settings is available in C4.5 than ID3.

Classification/Regression Handles classification problems only.

Tree Growth: The first stage of the process is very similar to ID3. According to Quinlan et al (1986), a subset (approximately 10%) of the learning sample is chosen at random. This subset is known as a working set. A decision tree is grown on the working set. The remaining 90% or so of cases in the learning sample are classified using this decision tree. If all the observations from the learning sample are correctly classified, then the process stops and the decision tree is satisfactory. Otherwise, another set of observations from the learning sample is added to the working set and a completely new tree is grown.

- Splitting Method The gain criterion, as used by ID3, can also be used in C4.5. An alternative is the **gain ratio criterion**. If a variable, x_j, has v distinct values then v possible descendant nodes can be found from splitting on x_j. The information measure or 'correctness of the answer', $IV(x_j)$ from splitting on x_j is found by

$$IV(x_j) = - \sum_{i=1}^{v} \frac{m_i + n_i}{m + n} \log_2 \frac{m_i + n_i}{m + n}$$

where m and n are number of observations from class 1 and class 2 respectively, while m_i and n_i are the number of class 1 and class 2 observations in the ith node. Let the expected information content from a split on x_j be defined as

$$E(x_j) = \sum_{i=1}^{v} \frac{m_i + n_i}{m + n} IV(x_j).$$

The gain in splitting on x_j is:

$$gain(x_j) = IV(x_j) - E(x_j).$$

The gain ratio criterion chooses the variable with the maximum ratio of $gain(x_j)/IV(x_j)$ to split on, subject to a number of minor constants.

- Type of Splits

Multiway splitting, as in ID3, is carried out. Splitting is done on every distinct value of the splitting variable. Only univariate splits are carried out.

- Priors/Costs

No.

- Stopping Rules

The process is halted, when after a few iterations of the tree-growing process, no decrease in misclassification error rate has been observed. This would usually occur if there were inconsistencies in the learning sample.

- Node Classification/Prediction

A node is assigned to the class with the maximum number of observations in that node.

Tree Pruning

In contrast with Breiman et al (1984), C4.5 uses **pessimistic pruning**, to decide tree size. Quinlan (1987), defines the method as follows. Let T be a subtree of the tree T_{max}, containing L(T) terminal nodes and letting $\sum K$ and $\sum J$ be the total number of observations and number of misclassified observations respectively in subtree T. A pessimistic view of T is that it will misclassify $L = (\sum J + L(T)/2)$ out of the $\sum K$ unseen cases, with standard error

$$se(L) = \sqrt{\frac{L(\sum K - L)}{\sum K}}.$$

The above involves using the continuity correction as in binomial probabilities. Let E be the number of observations misclassified by the best terminal node within T. The pessimistic pruning algorithm replaces T by the best terminal node whenever $L = \sum J + L(T)/2$ is within the limits of $L + se(L)$. A number of repetitions of the tree growing and pruning process is carried out. As the initial working set is selected purely at random, the same learning sample can give rise to completely different trees, as completely different parts of the learning sample may be chosen. The pruning process selects the best trees based on a combination of low misclassification error and small tree size. Quinlan (1987) states the following as the two prime advantages of pessimistic pruning.

	(i) Faster than other pruning methods. (ii) Does not need a test sample distinct from the learning sample. The validity of the second advantage is questionable as some form of bootstrapping or cross-validation could be used in the pruning process.
Validation Procedures	Validation of the rules is carried out after the creation of the decision trees. Validation by an independent test sample is done by dividing the data into a test sample and a learning sample before analysis.
Interactive Ability	No.
Graphical Ability	No.
One-Stage Optimality?	Yes.
Missing Values	Either omit observations with missing values from the analysis or use class sample sizes multiplied by some parameter to estimate missing values.
Criticisms	Recent studies by Feng et al (1993) and Schwartz et al (1993) have compared C4.5 with other classification methods. Schwartz et al found that C4.5 was very robust to noise in the data, produced sensibly sized trees and provided new insight into a particular set of data by uncovering important relationships among the variables. C4.5, however, due most probably to its creation in the machine learning environment, has no mechanism for incorporating a cost structure. Schwartz et al note that C4.5 produced large differences in group misclassification error rates. Feng et al (1993) carried out a more thorough study than Schwartz et al. Their results tentatively showed that C4.5 produced larger trees than CART, hence produced rules that were biased towards the learning sample. They note "[r]eliability is negatively related to the difference between [learning] and testing accuracy" (Feng et al, 1993, p 48). As a tree increases in size, node sample sizes decrease so rules are being generated from smaller and smaller sized samples. This makes these rules less reliable. They conclude by saying "[as] our tests show, even simply introduc[ing] a cost-handling mechanism in the testing phase (in CART) can make a visible improvement compared to, say, C4.5.

Examples in the Literature

Quinlan, J R, Compton, P J, Horn, K A and Lazarus, L (1986). Inductive knowledge acquisition: a case study, in "Applications of Expert Systems, Volume 1", J R Quinlan (ed), Wokingham: Addison-Wesley, pp 157-183.

Schaffer, C (1993). Selecting a classification method by cross-validation, Personal Communication.

3.9 FACT

Author(s) W Y Loh (USA) and N Vanichsetakul (Thailand).

Introduction Published by Loh and Vanichsetakul in 1988 at the University of Wisconsin, the full title of the program is Fast Algorithm for Classification Trees. The goal of the procedure is an algorithm sharing the best features of LDA and CART, namely the speed of linear techniques and the readily comprehendable structure of decision trees.

Classification/Regression FACT deals with categorical dependent variables only.

Tree Growth: The tree is grown on the whole data set.

- Splitting Method Three splitting algorithms are used by FACT. The first deals with univariate splits. Univariate F-ratios for variable selection are used at each node, to obtain the variable with the highest F-ratio for splitting, and then carrying out LDA on the selected variable to partition the co-ordinate axis. If the largest F-ratio of between to within-class variance is less than a specified threshold, F_0, no split is formed and the node is declared terminal.

Linear combination splits can also be generated by FACT using principal component analysis of the correlation matrix at each node. Then, LDA is carried out on the scores of the m largest principal components with m depending on user input. Loh and Vanichsetakul (1988) prefer linear combination splits over univariate splits.

A third method of splitting can be used whenever spherical symmetry is detected in a node, whereby univariate and linear combinations would be ineffective. Polar coordinate splits solve this problem, which involves transforming the best

splitting variable x_j after subtracting the mean $\bar{y}_i$ from each observation y_{ij}, and splitting on the resulting transformation, where y_{ij} is the ith principal component score for the jth observation.

- Type of Splits Multiway splitting can be used by FACT, dividing a node into two, three or more descendant nodes. As mentioned previously, splitting can be carried out using only one variable at a time, or a linear combination of the variables.

- Priors/Costs As FACT uses LDA to split each node, both different priors and cost matrices can be incorporated into the tree building process.

- Stopping Rules A direct stopping rule is used to determine tree size. Splitting is stopped if the error rate found from resubstituting the original sample does not decrease with splitting or the node size falls below a certain value.

- Node Classification/Prediction
A node is assigned to the class with the largest number of observations in that node, unless priors and/or costs are altered.

Tree Pruning No.

Validation Procedures The final decision tree generated in FACT can be validated by g-fold cross-validation, $2 \leq g \leq 25$.

Interactive Ability No.

Graphical Ability Draws trees using Splus functions.

One-Stage Optimality? Yes.

Missing Values Missing values are replaced by class means estimated from non-missing values in the learning sample.

Criticisms The principal criticisms of FACT appeared in Breiman and Friedman (1988). They were:

The authors of CART, in Breiman and Friedman (1988), regard FACT as a step back in the evolution of binary decision trees. Breiman and Friedman criticise FACT on the following grounds.

(i) The principal motivation for FACT is computational. Running time is sacrificed for accuracy and simplicity.
(ii) Linear combination splits are not better than univariate splits. In most cases where recursive partitioning has performed better than traditional parametric methods it has been through univariate splits.
(iii) Top-down stopping rules, as used in AID, THAID, etc, were one of the main reasons why the above methods were not really recognised. "The optimal-complexity tree pruning algorithm (based on cross-validatory choice) implemented in CART is probably the most important contribution of Breiman et al (1984)." (Breiman and Friedman, 1988, p 726).
(iv) FACT cannot handle categorical variables in a clean and elegant way.
(v) FACT is not invariant under transformations of variables.
(vi) There are no surrogate variables to handle missing values.

Examples in the Literature Wolberg, W H, Tanner, M A, Loh, W Y and Vanichsetakul, N (1987). Statistical approach to fine needle aspiration diagnosis of breast masses, Acta Cytologica, **31**, pp 731-741.

3.10 KnowledgeSeeker

Author(s) B de Ville, E Suen and D Biggs (Canada).

Introduction Released commercially in 1989, KnowledgeSeeker is a decision tree package that according to its authors, "... mine[s] a database for its critical decision-making and problem-solving information" (de Ville, 1990, p 30). The results are presented in a graphical display with an easy-to-use interactive ability which "... provides both end users and specialists with high levels of interaction of accurate, illuminating and reliable decision-making information and knowledge based rules" (Ibid, p 30).

Classification/Regression Both numeric and categorical response variables can be handled. For comparison with the other methods and simplicity, everything henceforth will be described in the classification context only.

Tree Growth: The tree is grown on the whole data set.

- Splitting Method KnowledgeSeeker seeks to overcome the problems inherent in AID and ID3 by using a significance testing approach to splitting as used in the CHAID program (Kass, 1980). Two alternatives exist for splitting using the significance testing approach. The first uses exhaustive partitioning, as in CART, searching over all possible combinations of values of every variable to find the split which maximises the χ^2 statistic with respect to the class variable. This method is guaranteed to find the optimal split for the data at hand based on statistical inference. The second approach is to use a heuristic clustering technique. Values of a particular variable are grouped with one another on the basis of their similarity in the response variable. This merging of values continues until no further merging is significant at a specified level of significance. Once values are merged, they can be split again using a more stringent level of significance. This approach is not optimal, but de Ville (1990) regards it as intuitive and appealing.

- Type of Splits Since the most alike values of a particular predictor are clustered together, multiple branches can accrue from the same node. That is, multiway partitioning is used by KnowledgeSeeker. Splits are carried out on only one variable at a time.

- Costs/Priors No.

- Stopping Rules Splitting is stopped if either:
(i) Node size falls below a certain value.
(ii) The optimal split on a predictor at a particular node does not exceed a specified significance level.

- Node Classification/Prediction
A node is assigned to the class with the largest number of observations in that node.

Tree Pruning According to de Ville (1990), KnowledgeSeeker supports both validation and tree pruning methods, that either verify the decision tree or which rate the quality of new branches on the decision tree and truncate them if its quality fails to pass a certain threshold valve. In the literature of decision tree methods, the above is NOT a pruning method, but rather a set of tests used in the tree growing process. Thus KnowledgeSeeker, like CHAID, its closest ancestor, does not have a tree pruning method.

Validation Procedures	Version 2.0 of KnowledgeSeeker does incorporate a validation procedure whereby part of the data is used to grow the tree while the other part of the data is used to test the rules created. It is relatively quick and easy to divide the data in two and use each half in turn as a learning sample and a test sample. Unfortunately, the latest version (2.1) of KnowledgeSeeker **does not** contain any validation procedure.
Interactive Ability	With the click of a mouse button, the user can investigate other possible partitionings of the decision tree at every step of the tree growing process. KnowledgeSeeker automatically calculates the best alternate splits at every node, so the user can examine the effects on the tree by changing from the best split to the best alternate/second best alternate split etc, so as to "... correspondingly mould the creation of the decision tree or rule base to support their understanding of the problem area and decision making task at hand" (de Ville, 1990, p 30). In addition, the user has the choice of either growing the tree automatically or growing it on a node-by-node (stepwise) basis. All operations are started by the use of a pull-down menu.
Graphical Ability	The graphical display of the tree works in unison with the tree building process. As a node is split into a number of sub-nodes, the results are displayed immediately on the screen by means of a decision tree. The user can interrupt the process to investigate the current state of the tree and then continue at the point where splitting was ceased. The level of detail about each node and each split carried out can also be altered.
One-Stage Optimality?	KnowledgeSeeker is, in theory, one-stage optimal. In practice, however, it is relatively quick and easy to investigate the effects on the tree of changing the partition of a particular node to one of the other significant partitions.
Missing Values	KnowledgeSeeker handles missing values in two different ways: (i) They are excluded from the decision tree growing process. (ii) They are treated as an additional category of a variable, and so can be combined with the categories that they most resemble.

Criticisms

Very few, if any, reviews of KnowledgeSeeker have appeared in the literature. Biggs et al (1991) conducted some simulation studies with different significance levels. They found that KnowledgeSeeker can confidently be used with either small or large data sets involving categorical predictors and a response. They also found that the same confidence applied with continuous responses, provided that the response was approximately normally distributed with roughly equal variances.

One major criticism that could be made of KnowledgeSeeker is that it moves away from any form of validation of the results, by independent test samples. As Breiman et al (1984) state, the use of the resubstitution estimate of error rate as an estimate of the true error rate can give an overly optimistic picture of the set of decision rules constructed. The omission of a validation procedure goes against current statistical practice in the decision tree field and even contradicts what was written in de Ville (1990). In a personal communication, de Ville (1994) affirms that the forthcoming version of KnowledgeSeeker does support a hold back sample and validation facility.

Examples in the Literature

3.11 Splus Trees ( )

Author(s)

L Clark (USA) and D Pregibon (Canada).

Introduction

Developed in 1991 using the Splus language to carry out a CART-like decision tree modelling method. Of all the decision tree-based methods, the Splus tree routines are the closest to those used by CART with binary recursive partitioning, pruning and cross-validation.

Classification/Regression

Splus trees() handles both categorical and numeric dependent variables hence can be used for classification and regression problems. For comparison with the other methods and simplicity, everything henceforth will be described in the classification context only.

Tree Growth: The tree is grown on the whole data set.

- Splitting Method The deviance function for an observation y_i is defined as (Clark and Pregibon, 1992).

$$D(\mu_i, y_i) = -2 \sum_{j=1}^{k} y_{ij} \log(\gamma_{ij})$$

that is, negative two times the log-likelihood function, where γ_{ij} denotes the probability that the ith response falls in jth class.

At a given node, the mean parameter μ is constant for all observations. The maximum likelihood (or minimum deviance) estimate of μ is given by the node proportions. The deviance of a node is defined as the sum of the deviances of all the observations in the node, $D(\hat{\mu}; y) = \sum D(\hat{\mu}; y_i)$. A node where all the observations belong to the same class will have a deviance of zero.

Splitting is achieved by comparing the deviance of the current node to that of the two descendant nodes, where the combined deviance of the two descendant nodes is

$$D(\hat{\mu}_L, \hat{\mu}_R; y) = \sum_L D(\hat{\mu}_L; y) + \sum_R D(\hat{\mu}_R; y)$$

and the split that maximises

$$\Delta D = D(\hat{\mu}; y) - D(\hat{\mu}_L, \hat{\mu}_R; y)$$

is the split used at the given node.

- Type of Splits Only binary splits are used, and only on one variable at a time.

- Priors/Costs No.

- Stopping Rules Splitting is stopped by one or two different rules. The first sets a minimum node size below which splitting cannot be done while the second stops splitting if the ratio of deviances between a tree with r terminal nodes and the root node is less than some threshold value.

- Node Classification/Prediction
A node is assigned to the class with the largest number of observations in that node.

Tree Pruning

As with CART, an overly large tree, biased towards the learning sample, can be grown by Splus trees(). The next step is to apply a pruning procedure that determines a nested sequence of subtrees of the original tree by cutting off branches containing relatively unimportant splits. This is achieved by means of a cost-complexity measure

$$D_\alpha(T) = D(T) + \alpha\, L(T)$$

where $D_\alpha(T)$ is deviance of the subtree T, size L(T) is the number of terminal nodes contained in T and α is the cost-complexity parameter. By default, the procedure produces a sequence of subtrees that minimise the cost-complexity measure. Note that this algorithm is very similar to the one used by CART, except that CART uses a measure of misclassification error rate rather than deviance.

A similar procedure used by Splus trees() is the shrink-tree() function which determines a sequence of subtrees from the original tree that differ in their fitted values. The function uses the recursion relation

$$\hat{y}(\text{node}) = \alpha(\bar{y}(\text{node})) + (1-\alpha)\, \hat{y}(\text{parent})$$

where $\bar{y}$(node) is the usual fitted value for each node and $\hat{y}$(parent) is the shrunken fitted value for the node's parent, which was in turn obtained in the same way. The technique basically uses a parameterization of α that optimally shrinks the descendant nodes to their parent nodes based on the magnitude of the difference between $\bar{y}$(node) and $\bar{y}$(parent). The result of a plot of deviance against size of the subtrees found by shrinking is a smooth decreasing curve which flattens out as the size of the subtrees increase.

Validation Procedures

A second approach Splus trees() uses to test the sequence of subtrees produced by either pruning or shrinking, is to use g-fold cross-validation, $2 \leq g \leq n$, to select the tree with the minimum cross-validated **deviance** rather than the minimum error rate. No standard error rule is used by Splus trees(). Thus, like CART, the decision tree that is produced should be relatively robust, giving a set of rules that are valid when applied to another set of data from the same population.

Interactive Ability

Splus trees() has an option to allow the user to examine the goodness of split for each variable at a particular node. This information is conveyed by means of a scatter plot for categorical variables and a high density bar graph for numeric predictors. The user is able to see what variable (and value(s) of that variable) is the best discriminator of the class variable. In order to change the splitting variable, however, from that chosen by the splitting algorithm, the edit.tree() function is called with the variable to be split and its value explicitly stated.

Graphical Ability

Trees can be drawn and labelled through Splus graphics. Either the use of a mouse or Splus commands can edit the tree, examine splits and examine the distribution of the classes in the terminal nodes.

One-Stage Optimality?

The method is one-stage optimal examining only the best splits at a current node. As seen above, though, the splitting variable at a particular node can easily be changed but a new tree cannot be grown after changing the splitting variable.

Missing Values

The function na.tree.replace() is used to handle missing values in Splus trees(). The function creates a new level for any variable containing missing values, coded as 'NA'. Numeric predictors are first grouped into c categories. Clark and Pregibon (1992) describe how missing values are predicted as follows:

> The approach we adopt is that once an NA is detected while dropping a (new) observation down a fitted tree, the observation 'stops' at that point where the observation is required to continue the path down the tree. This is equivalent to sending an observation down both sides of any split requiring the missing value and taking the weighted average of the vector of predictions in the resulting set of terminal nodes.

Criticisms

Many of the criticisms levelled at CART would also apply to Splus trees(). However, "([t]he S computing language ... is currently one of the most developed interactive programming environments for data analysis and graphics". (Le Blanc and Crowley, 1993, p 466). The interactive facilities that Splus trees() has appears to give it a distinct advantage over CART. As Clark and Pregibon (1992), p 415 state "[o]ur recommended approach to tree building is far less automatic than that provided by other software for the same purpose, as the unbundling of procedures for growing, displaying and challenging trees requires user initiation in all phases". Perhaps one other

criticism of Splus trees() is that the method requires an adequate knowledge of the Splus language, which is not menu-driven nor user-friendly.

Examples in the Literature

Bradford, E (1993). Tree-based models in S, New Zealand Statistician, **28**, pp 36-51.

Morton, S C (1992). Personal crunching: new advances in statistical dendrology, Chance, **5**, pp 76-79.

3.12 IND

Author(s) W Buntine (Australia).

Introduction Introduced in 1991, the technique tries to combine the simplicity of decision tree rules with the power of Bayesian methods. Bayesian methods are used for splitting, smoothing and tree averaging. "IND provides a potentially bewildering number of options to allow the user to precisely control how data is interpreted, how trees are grown and tested, and how results are displayed" (Buntine and Caruana, 1993, p 1-4). As well, IND has the ability to simulate the CART and C4.5 tree-based methods, or follow a minimum message length idea such as that used in Wallace and Patrick (1993), or indeed the newly developed decision graph approach of Oliver (1993).

Classification/Regression Used only for classification.

Tree Growth: The tree is grown on the whole data set.

- Splitting Method IND can choose from a number of different criteria when evaluating the quality of different splits or tests. For example the Gini and twoing splitting criterion (see Section 3.7) can be used or the gain ratio criterion, as used with C4.5 (see Section 3.8). Buntine recommends the use of Bayesian splitting which evaluates each possible split of a particular node into several sub-nodes.

The Bayesian estimate of the posterior probability of each possible split being correct is then evaluated, with the split producing the maximum posterior probability being carried out.

- Type of Splits

Multiway splitting is used, splitting on each distinct value of a variable. Only univariate splits are carried out.

- Priors/Costs

In the Bayesian mode, "class priors" in the CART sense of the word, do not exist. The algorithm incorporates priors but these are Bayesian prior probabilities of the class probabilities in each terminal node prior to seeing any data.

IND has cost structures. As the technique is Bayesian, the incorporation of priors is trivial. The tree returns a class probability vector. The cost vector is then combined with that so that a minimum cost decision can be made. It's a simple add-on to the tree interpretation routine.

- Stopping Rules

Splitting stops when the quality measure above for the best split at a particular node fails to exceed a prespecified criterion.

- Node Classification/Prediction

If classification is not determined by cost, then each terminal node is assigned to the class with the maximum probability within the terminal node. If costs are incorporated into the tree growing process then the terminal node is assigned to the minimum cost class.

Tree Pruning

IND adopts a Bayesian approach to pruning, which uses a smoothing technique. The usual approach is to find the class probabilities of observations in the terminal nodes. The smoothing approach also takes into account the class probability vectors for all the intermediate (decision) nodes en route to the terminal nodes. The resultant final tree may have widely differing class probabilities for the terminal nodes. If the class probabilities of two or more terminal nodes are similar, and come from the same parent, they can be pruned upwards.

Validation Procedures

One has to divide the data into a learning sample and a test sample before the tree growing process is begun. A routine exists in IND to carry this out. In a personal communication, Buntine (1993) recommends growing the tree on the full data set and believes that if you want to produce the best predictions possible, all of the data should be used to grow the tree, as per standard Bayesian theory. As several different class probability trees are grown for each data set, the weighted average of the class probability vectors each tree assigns to an observation can be taken. This is the Bayesian averaging approach which Buntine favours instead of learning/test samples.

Interactive Ability	Advanced features allow the user to interactively search for the best splits and control the tree-growing process.
Graphical Ability	IND contains a graphical display routine that is used to display the tree classifiers in various forms.
One-Stage Optimality?	No. The method encompasses an N-ply lookahead facility, whereby not only are the best splits examined at the current node, but also how the resulting descendant nodes and their sons should be split? For example, a 2-ply lookahead scheme would search for the best split at the current node as well as the best split at the resulting son nodes. This may lead the user to find that the so-called 'best' split of the current node was not the optimal split in terms of future tree development, as this variable may not interact with any other variables in the data set. A 'lesser' split may in fact lead to a greater reduction in misclassification in the next stage of tree development. The N-ply lookahead facility allows the user to uncover such a structure in the data.
Missing Values	Missing values are handled using Quinlan's preferred strategy. That is, IND sends a case down each branch with the proportion found in the learning sample at that node. In effect, each case with missing variables is split into a number of parts, with the largest part going down the branch where most other cases have gone. Otherwise, a routine exists to send a case down the branch of the tree most commonly taken by other examples.
Criticisms	IND is a relatively new method so there has yet to be an article either criticising or praising IND in the literature. It would perhaps be criticised for having no mechanism to incorporate "class priors" in the CART sense of the word. Buntine (1993) feels that the main criticism of IND is that he hasn't optimised to handle all those important real-world things like real-valued splits, which can be done, but without much thought.
Examples in the Literature	

3.13 SUMMARY TABLE COMPARING THE TEN TREE-BASED METHODS

Attribute	AID	THAID	ID3	CHAID	CART
Author(s)	J N Morgan J A Sonquist	J N Morgan R C Messenger	J R Quinlan	G V Kass	L Breiman J H Friedman R A Olshen and C J Stone
Introduction	1963	1973	1979	1980	1984
Classification/Regression	Regression	Classification	Classification	Classification	Both
Tree Growth	Uses all the data.	Uses all the data.	Uses a subsample of the data.	Uses all the data.	Uses either all or a subsample of the data.
- Splitting Method	Maximum between to within-group variance.	Theta or delta splitting criterion.	Maximisation of the gain criterion.	Maximising the χ^2 statistic of grouped categories.	Gini or twoing splitting criterion.
- Type of Splits	Binary/US.	Binary/US.	Multiway/US.	Multiway/US.	Binary/US or LC.
- Costs/Priors	No.	No.	No.	No.	Yes.
- Stopping Rules	Direct stopping.	Direct stopping.	All cases in the learning sample are correctly classified.	No significant splits.	All nodes are pure to one class.
- Node Classification/Prediction	Average value of the cases in the terminal node.	Class with largest number of observations in the node.	As for THAID.	As for THAID.	As for THAID after accounting for costs and priors.
Tree Pruning	No	No	No	No	Cost-Complexity
Validation Procedures	No	No	No	No	Yes
Interactive Ability	No	No	No	Yes	Yes/No
Graphical Ability	No	No	No	Yes	Yes
One-Stage Optimal?	Yes	Yes	Yes	Yes	Yes
Missing Values	Estimated using class means.	Estimated using class means.	Estimated using class proportions	Omitted	Estimated using surrogate splits
Criticisms	• Produces overly large trees. • Too dependent on sizes of stopping rules. • No validation of results.	• Produces overly large trees. • Too dependent on sizes of stopping rules. • No validation of results.	• Does not know when to stop. • No validation of the results.	• No validation of the results.	• Instability of cost-complexity pruning. • Tree size affected too much by the standard error rule.

TABLE 3.13 (cont'd)

Attribute	C4.5	FACT	Knowledge Seeker	Splus Tree(s)	IND
Author(s)	J R Quinlan	W Y Loh N Vanichsetakul	B de Ville E Suen D Biggs	L Clark D Pregibon	W Buntine
Introduction	1986	1988	1989	1991	1991
Classification/ Regression	Classification	Classification	Both	Both	Classification
Tree Growth	Uses a sub-sample of the data.	Uses all the data.	Uses all the data.	Uses all the data.	Uses all the data.
- Splitting Method	Maximisation of the gain ratio criterion.	Discriminant analysis.	Maximising the χ^2 statistic of grouped categories.	Likelihood ratio statistic.	Quality measure.
- Type of Splits	Multiway/US.	Multiway/US or LC.	Multiway/US.	Binary/US.	Binary/US.
- Costs/Priors	No.	Yes.	No.	No.	Yes/No.
- Stopping Rules	Direct stopping.	Direct stopping.	No significant splits.	Deviance below a certain value.	Quality measure below a certain value.
- Node Classification/Prediction	As for THAID.	As for CART.	As for THAID.	As for THAID.	Maximum class probabilities after accounting for cost.
Tree Pruning	Pessimistic	No	No	Cost-complexity with deviances	Bayesian
Validation Procedures	Yes	Yes	No	Yes	Yes
Interactive Ability	No	No	Yes	Yes	Yes
Graphical Ability	No	Yes	Yes	Yes	Yes
One-Stage Optimal?	Yes	Yes	Yes/No	Yes	No
Missing Values	Estimated using class proportions	Estimated using class means	Creation of a new category	Creation of a new category	Estimated using class proportions
Criticisms	• No mechanism for incorporating a cost structure. • Tends to produce overly large trees.	• Decision rules not simple and accurate. • Direct stopping rules are used. • Not robust to non-normality.	• No validation of the results.	• Instability of cost-complexity pruning.	• No mechanism for incorporating priors structure • Not designed for real-valued splits.

4. SIMULATION STUDIES INVOLVING CONTINUOUS DATA

4.1 INTRODUCTION

Data sets involving distinct groups of populations arise in many disciplines, including the social sciences, business and medicine. Multivariate data sets are often not easy to analyse, so it is important that the method should be both powerful and easy to understand. In the classification context, the prediction rule should be accurate, that is have a low, unbiased error rate, yet be as easy to interpret as possible.

In Section 4.2, through a literature survey, an investigation is carried out into the various types of error rate estimates that are used in the field of classification.

Next, in Sections 4.3 and 4.4, a comparison of four classification methods from the domains of traditional discrimination and tree-based methods is done. LDA and QDA are the two most commonly used classification methods in statistics to handle the above type of data. These two parametric techniques are compared and contrasted with two tree-based methods, CART and FACT. (See Sections 2.2, 2.3, 3.7 and 3.9 for details on LDA, QDA, CART and FACT respectively.)

This chapter compares both the accuracy and reliability of these four classification methods in classifying individuals into two multivariate populations under certain combinations of parameters. Three types of data distributions will be investigated, involving normal, lognormal and standardised lognormal. The robustness of each method to a change in the value of the a priori probabilities of class membership will also be determined.

4.2 ERROR RATES

A great deal has been written on the subject of the error rates in classification analysis over the past three or more decades. In this section, a review of the literature on error rates is given as well as a formal definition of each of the error rate estimators, including those to be used later in this chapter. Extensive reviews of error rate estimation may be found in Kanal (1974), Toussaint (1974), Lachenbruch (1975), Efron (1982, 1983), Hand (1986) and McLachlan (1986, 1987) among others.

In any classification problem, the object is to assign a random observation **x** to one of k populations, $\Pi_1, ..., \Pi_k$. Paraphrasing Toussaint (1974), one of the most important problems with the use of any classification method is estimating the probability of misclassification.

The optimal error rate of any classifier is the Bayes error rate, which is defined as

$$R(B) = 1 - \int \max_i [f_i(\mathbf{x})\, \pi_i]\, d\mathbf{x} \qquad (4.2.1)$$

where $f_i(\mathbf{x})$ is the class conditional density function for Π_i and π_i is the a priori probability of belonging to Π_i. According to Hand (1986), p 335, "[this] is the minimum possible error rate given a set of [variables]." This is the error rate that would result if the class conditional density functions were known.

The actual or true error rate, R(T), is defined as the expected probability of misclassification when the class conditional density functions are known and the Bayes rule is not used. In practice, this is the error rate that would result when applying the classifier to an infinite test sample. In notation form, and assuming that $\pi_1 = \pi_2$

$$R(T) = \frac{1}{2}\int_{\hat{A}_2} f_1(\mathbf{x})\, d\mathbf{x} + \frac{1}{2}\int_{\hat{A}_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (4.2.2)$$

where $\hat{A}_1 = \{\mathbf{x}: D(\mathbf{x}) = [\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)]' S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > 0\}$ and with,

$$\hat{A}_1 \cap \hat{A}_2 = \phi.$$

The probability that **x** falls in $\hat{A}_1$ given that $\mathbf{x} \in \Pi_2$ is

$$\Pr[D(\mathbf{x}) > 0 \mid \Pi_2] = \Pr\left[\frac{D(\mathbf{x}) - D(\mu_2)}{\sqrt{V}} > \frac{-D(\mu_2)}{\sqrt{V}}\right] = \Phi\left[\frac{D(\mu_2)}{\sqrt{V}}\right] \qquad (4.2.3)$$

where $D(\mathbf{x})$ is as defined in (2.2.5) and

$$V = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S^{-1} \Sigma S^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

It follows that

$$\Pr(D(\mathbf{x}) < 0 \mid \Pi_1) = \Phi\left[\frac{-D(\boldsymbol{\mu}_1)}{\sqrt{V}}\right] \tag{4.2.4}$$

so that

$$R(T) = \frac{1}{2}\Phi\left[\frac{-D(\boldsymbol{\mu}_1)}{\sqrt{V}}\right] + \frac{1}{2}\Phi\left[\frac{D(\boldsymbol{\mu}_2)}{\sqrt{V}}\right] \tag{4.2.5}$$

Realistically, the samples at hand are always finite and the class conditional density functions are often not known. Therefore, another measure of classification error is needed to be used for a classification rule. One such measure is the expected error

$$R(E) = E[R(T)] \tag{4.2.6}$$

for learning samples of a given size. In practice, however, the amount of data at hand does not allow for the estimation of R(T). Hence, another estimate of R(T) is needed.

Two approaches are available for the estimation of R(T). The first uses functions which combine the sample estimates of the class means and covariances with the number of variables and sample size in a data set to estimate the error rate. These methods are known as **parametric error rate estimators**. Examples of these types of estimators are the L estimator of Lachenbruch (1967), the M estimator of McLachlan (1974), and the NS and NS* smoothing estimators of Snapinn and Knoke (1985, 1988). These methods have been shown to provide accurate, unbiased estimates of the actual error rate as seen in their use with LDA (Prada Sanchez and Otero Cepeda, 1989 and Ganeshanandam and Krzanowski, 1990) and with stepwise discriminant analysis (Snapinn and Knoke, 1989) in the case of normal or other symmetric distributions, such as the uniform distribution. It has been noted that "caution should be exercised with the use of the parametric estimators of the error rates as they may not be reliable under departures from the parametric model adopted" (McLachlan, 1986, p 271.) Indeed, recent studies by Snapinn and Knoke (1989) and Konishi and Honda (1990) have shown that these parametric estimators do not perform well with skewed distributions. Konishi and Honda found that parametric estimators, in particular, the M estimator above, should be applied with extreme care in the case of skewed distributions such as the lognormal. Snapinn and Knoke found that the smoothed bootstrap estimator deteriorated markedly from the normal case, when used in stepwise discriminant analysis for lognormal

distributions. Since simulation studies in this chapter involve lognormal data, parametric error rate estimators will not be used.

The second approach available for the estimation of the actual error rate are **non-parametric error rate estimators**, where no assumptions about $f_i(\mathbf{x})$ are required. All such methods can be called error count estimates in that they involve counting the number of falsely classified observations determined by the classification rules.

Hand (1986), describes error rate estimation as a relatively simple process if in addition to the data set $[(\mathbf{x}_1, ..., \mathbf{x}_n)]$ from which the classifier was designed, there is available an independent test set of observations $[(\mathbf{x}_1, ..., \mathbf{x}_m)]$ sampled from the same distribution. This error rate estimator is defined as the test sample estimate of R(T), R(TS). In practice however, there is very seldom any additional data available. If there is, then the test set is often small and although providing an unbiased estimate of R(T), it has large variance. That is, if many test samples of small size are obtained and a classifier is trained on these then the individual test sample error rates will vary between being pessimistically and optimistically biased, though most probably average out to zero.

A second, more feasible approach is to train the classifier on the learning sample $L = (\mathbf{x}_1, ..., \mathbf{x}_n)$ from which the classifier was formed. This estimator, R(A), is known as the apparent or resubstitution error rate estimator. Authors such as Lachenbruch and Mickey (1968), among others, have shown that this method is optimistically biased in LDA in that it almost always underestimates R(T). Efron (1983), offers an explanation for this phenomenon. R(A) is the error rate for points zero distance from the learning set while R(T) is the expected error rate for a new observation, $\mathbf{x}_0$, which may be some distance away from the learning set. If the error rate of the prediction rule increases as the point being predicted moves away from L then R(A) will underestimate R(T). In other words, it is unlikely that all future observations will lie within the range of values spanned by L or be distributed in the same manner as L. The classification rules are designed to optimise the error rate for all observations in L. Therefore, a test sample which either has values outside the range of the values in L or not distributed the same way as L will have a larger error rate than R(A).

A third non-parametric method of estimating R(T) is the hold out error rate estimator, R(H). This is found by dividing the data set into two and using one part as the learning sample to construct the classifier and the other part as the test sample. As with R(TS), the test sample estimate of the error rate, R(H) has large variability when n is small. As well "... it is an inefficient use of data - one would like to use all the available data to design the best possible classifier" (Hand, 1986, p 336). Toussaint (1974), quotes numerous studies which have found R(H) to have a pessimistic bias in estimating R(T).

The rotation method makes more efficient use of the data. The idea is to divide the data into two halves and use each half in turn as a learning sample and a test sample. By averaging the two test sample error rate estimates the rotation error rate estimate, R(ROT), is obtained. Although making a more efficient use of the data, Toussaint (1974) still found this method to be pessimistically biased.

The above idea was extended to divide the data into g mutually and randomly chosen sets of data of size n/g. The method of g-fold cross-validation omits each of the g groups in turn from the data set, calculates a classification rule based on the remaining (n - (n/g)) observations and classifies the omitted group of observations. Then it counts the total number of misclassified observations divided by the size of the data set to get an estimate of the error rate. Toussaint (1974), refers to this as the Π method. When g = 2, that is, two-fold cross-validation, this is the rotation method. When g = n, this is the n-fold cross-validation error estimator, R(CV), attributed to Lachenbruch (1967), where, in the case of two populations

$$R(CV) = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij}/n_i \qquad (4.2.7)$$

This method is also known as the 'leave-one-out' or U estimate. Studies undertaken by numerous authors including Efron (1983) have shown that n-fold cross-validation has large variance. Thus, although R(CV) may be an unbiased estimate, the confidence with which the user can expect R(CV) for his/her sample to approach R(T) is not great.

The jackknife error rate estimate is due to Quenouille (1949). The method involves omitting each observation in turn from the learning sample and to obtain the apparent error rate for the learning sample with the jth observation omitted, $R_j^*(A)$, so that

$$R_J^*(A) = \frac{1}{n} \sum_{j=1}^{n} R_j^*(A) \tag{4.2.8}$$

so that $\hat{w}_J$, the jackknife estimate of the bias of R(A), is

$$\hat{w}_J = (n\text{-}1)\,[R_J^*(A) - R(A)] \tag{4.2.9}$$

leading to the jackknife estimate of the error rate

$$R(J) = n\,R(A) - (n\text{-}1)\,R_J^*(A) \tag{4.2.10}$$

Another approach is to use the estimated posterior probabilities of class membership, $\theta_i(\mathbf{x})$, where

$$\theta_i(\mathbf{x}) = \pi_i\, f_i(\mathbf{x}) \Big/ \left[\sum_{m=1}^{k} \pi_m\, f_m(\mathbf{x}) \right] \tag{4.2.11}$$

is the posterior probability that $\mathbf{x}$ belongs to Π_i. An observation, $\mathbf{x}$, is assigned to Π_i if

$$\theta_i(\mathbf{x}) = \max_m \theta_m(\mathbf{x}). \tag{4.2.12}$$

This naturally leads to the posterior probability error rate estimator, R(PP), where each observation is not assigned outright to a population; rather it is given an estimated probability of membership to each population. This estimator takes the form

$$R(PP) = \sum_{j=1}^{n} \min \theta_m(\mathbf{x}_j)/n \tag{4.2.13}$$

Glick (1978) has shown this estimator to be optimistically biased, though with smaller variability than R(A). Ganesalingam and Lynn (1991) have considered posterior probability error rate estimation in the context of a mixture of two normal populations. They also found that R(PP) generally underestimated R(T).

A recent development in the field of error rate estimation is the bootstrap error rate estimator, R(B), due to Efron (1979) and developed further in Efron (1982, 1983). The idea is as follows. Let $\mathbf{x}_j$ be a random observation from C_j, and let $\mathbf{x}_0$ be a new observation that is to be classified, belonging to C_0. Let the $\mathbf{x}_j$'s and $\mathbf{x}_0$ be from the entire population mixture distribution, $F(\mathbf{x})$. Letting $\hat{C}_0$ be the predicted class of $\mathbf{x}_0$ using the classifier constructed from L. Efron (1983) defines a loss function

$$Q(C_0, \hat{C}_0) = \begin{cases} 0 & \text{if } \hat{C}_0 = C_0 \\ 1 & \text{else} \end{cases} \tag{4.2.14}$$

Therefore, the actual error rate of the classification rule is

$$R(T) = E[Q(C_0, \hat{C}_0)] \tag{4.2.15}$$

while the apparent error rate is

$$R(A) = \frac{1}{n} \sum_{j=1}^{n} Q(C_j, \hat{C}_j) \tag{4.2.136}$$

The true bias involved in using R(A) as an estimate of R(T) is

$$w = E[R(T) - R(A)] \tag{4.2.17}$$

If w was known, then an accurate estimate of R(T) could be obtained. In theory, there is no knowledge of w. Bootstrapping estimation is an attempt to approximate w by calculating $\hat{w}_B$, the bootstrap estimate of the bias involved in using R(A) as an estimate of R(T).

The basic sampling procedure behind bootstrapping is a clever, yet simple idea. In the univariate case, let $(x_1^*, x_2^*, ..., x_n^*)$ be a random sample of observations drawn with replacement from L, with weight 1/n placed on each observation in L. This is known as the bootstrap sample. A classification rule is constructed from the bootstrap sample and the apparent error rate estimate, $R^*(A)$, for the classifier is found. In addition, the cases from L are classified using the rules generated from $(x_1^*, ..., x_n^*)$. Then

$$\hat{w}_b = \sum_{j=1}^{n} \left(\frac{1}{n} - p^*_{jb} \right) Q(C_j, \hat{C}_j) \qquad (4.2.18)$$

where p^*_{jb} is the resampled proportion of observations in $(x^*_1, ..., x^*_n)$. Therefore, only observations that are not in $(x^*_1, ..., x^*_n)$ will contribute positively to $\hat{w}_b$. B bootstrap samples are generated in the same way and the $\hat{w}_b$ are calculated in the same way for each bootstrap sample. These are then averaged over the B samples to get the bootstrap estimate of the bias of R(A), that is

$$\hat{w}_B = \sum_{b=1}^{B} \hat{w}_b/B = E[R^*(T) - R^*(A)] \qquad (4.2.19)$$

so that, R(BOOT), the boostrap bias-corrected estimate of the actual error rate is

$$R(BOOT) = R(A) + \hat{w}_B \qquad (4.2.20)$$

Variants on the bootstrap have also been proposed by Efron (1982). These include the randomised bootstrap whereby empirical bootstrap samples are drawn from L with the proportion of observations drawn from Π_i a preset value. For example, the numbers drawn from each class could be set to be proportional to class sample sizes. Then if $\pi_1 = 0.6$, 60% of each bootstrap sample would be taken from Π_1. Another variant is the double bootstrap which was designed to correct the observed optimistic bias of the ordinary bootstrap. The process involves taking another lot of bootstrap samples to correct the above-mentioned bias of the ordinary bootstrap.

A third variant is the 0.632 estimator, defined as

$$R(0.632) = 0.368 * R(A) + 0.632 * R(\varepsilon) \qquad (4.2.21)$$

where $R(\varepsilon)$ is the average error rate for all observations not in a bootstrap sample over all B bootstrap samples. The 0.632 estimator was developed by considering the distribution of the distance δ between the point where the classification rule is applied and the closest point in the learning sample. It was noted that observations in the bootstrap sample have a high probability of being instance $\delta = 0$ away from observations in L, whereas the reverse would

occur if another independent sample of data from the same distribution was taken. "Their probability is equal to the probability that the point at which the rule is applied is included in the bootstrap sample, which is $1 - (1-1/n)^n$ and tends to 0.632 as $n \rightarrow \infty$." (McLachlan, 1987, p 234.) The bias of R(A) estimated by the 0.632 estimator is

$$\hat{w}_{0.632} = 0.632\,(R(A) - R(\varepsilon)) \tag{4.2.22}$$

so that

$$R(0.632) = R(A) - \hat{w}_{0.632} = 0.368 * R(A) + 0.632 * R(\varepsilon).$$

Efron (1983) showed that the asymptotic expansion of $R(\varepsilon)$ was very similar to the asymptotic expansion of R(ROT) and that the correlations between $R(\varepsilon)$ and R(ROT) in the simulation studies undertaken by him were very high (range 0.86 - 0.98). This implies that the estimator

$$R(0.632) = 0.368 * R(A) + 0.632 * R(ROT) \tag{4.2.23}$$

is almost the same as (4.2.21). McLachlan (1977), Wernecke, Kalb and Sturzebecher (1980) and Wernecke and Kalb (1983) have considered similar estimators to (4.2.23). McLachlan (1977) tried to find the parameter τ which lead to the greatest reduction in the bias of the apparent error rate, whereby

$$R(\tau) = \tau * R(GCV) + (1-\tau) * R(A) \tag{4.2.24}$$

where $0 \leq \tau \leq 1$ and R(GCV) is the g-fold cross-validation error rate estimate. McLachlan found that very little weight should be given to R(A), that is, τ close to 1, unless G is set to 2, where τ varied from 0.3 to 0.4. Simulation studies undertaken by Chernick, Murthy and Nealy (1985, 1986) have shown that $R(\varepsilon)$ is overly pessimistic, as is R(ROT). Therefore, a weighting function of a pessimistic estimator ($R(\varepsilon)$ or R(ROT)) and an optimistic estimator (R(A)) seems a very logical step.

Simulation studies undertaken by Efron (1983), Chernick et al (1985) and Fitzmaurice et al (1991) with LDA, Gong (1986) with logistic regression and Crawford (1989) with CART have found that the bootstrap and more particularly the 0.632 estimator are unbiased as well as having low variability. Rather contradictory results from those reported above were

obtained by Ganeshanandam and Krzanowski (1990). In a study involving error rate estimation in two group discriminant analysis, they found that the 0.632 error rate always estimated the actual error rate in the vicinity of the 0.3 to 0.4 range. This meant that the method was best when either the Mahalanobis distance between populations or sample size was small, while estimation of the actual error rate was overly pessimistic for large samples and/or large Mahalanobis distance between populations. They also found that contrary to Efron (1983) and others, the n-fold cross-validation estimator, R(CV), performed especially well though its relative variability did increase as sample size decreased.

4.3 SIMULATION STUDY I

4.3.1 Study Plan

Eighty different bimodal probability models were generated by using every possible combination of five different factors. The five factors used were; the number of variables (p), total sample size (n), Mahalanobis distance (δ), type of distribution (f($\mathbf{x}$)) and priors-covariance structure (e). The values of the first four factors, each at two levels, were

$$\begin{aligned} &p = 2, 6 \\ &n = 60, 300 \\ &\delta = 2, 3 \\ &f(.) = \text{normal, lognormal} \end{aligned} \qquad (4.3.1)$$

However, for this study, the lognormal data is transformed to have mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_i for Π_i, by letting

$$z_{ij} = \frac{x_{ij} - E(x_{ij})}{s.d.(x_{ij})}, \qquad i = 1, 2, \qquad \text{and } x_{ij} \text{ is lognormal}$$

which is lognormal (0, 1).

For Π_2,

$$x_{2j} = z_{2j} + \mu_{2j}.$$

Hence, the distribution is standardized lognormal rather than pure lognormal.

The fifth factor, e, with five levels, had values:

$$\begin{aligned}
&\pi_1 = 0.5: \ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I} && (1)\\
&\pi_1 = 0.5: \ \boldsymbol{\Sigma}_1 = \mathbf{I}, \ \boldsymbol{\Sigma}_2 = 3\boldsymbol{\Sigma}_1 && (2)\\
&\pi_1 = 0.25: \ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I} && (3)\\
&\pi_1 = 0.25: \ \boldsymbol{\Sigma}_1 = \mathbf{I}, \ \boldsymbol{\Sigma}_2 = 3\boldsymbol{\Sigma}_1 && (4)\\
&\pi_1 = 0.75: \ \boldsymbol{\Sigma}_1 = \mathbf{I}, \ \boldsymbol{\Sigma}_2 = 3\boldsymbol{\Sigma}_1 && (5)
\end{aligned} \qquad (4.3.2)$$

The values of the first three factors were carefully chosen, both from examples in the literature and to approximate real world data situations. The extreme situations encountered in this study were at one end, small, bivariate, moderately separated populations while at the other extreme had large, six-dimensional, well separated populations.

Other authors have used the pure lognormal to compare with the normal in simulation studies. However, the use of the pure lognormal means that if observations in Π_2 have a larger mean, the range of values will be much larger, resulting in a much higher covariance for Π_2 than for Π_1. Therefore, methods are not only being compared across degree of skewness of a distribution, but also across covariance structures, which are now different. Lachenbruch et al (1973) noted this problem but took no action to correct it. The use of the standardised lognormal distribution, on the other hand, preserves covariance structure but maintains the degree of skewness.

The values of the fifth factor were based on those used in other authors' simulation studies. More extreme covariance differences between populations were considered, for example, $\boldsymbol{\Sigma}_2 = 10\boldsymbol{\Sigma}_1$, but the present values were chosen to reflect real-world situations.

The effects of the experimental factors on the four classification methods, LDA, QDA, CART and FACT were investigated, using a split-plot design as employed by Ganeshanandam and Krzanowski (1990). The experimental factors were given in the experimental factor (main plot) stratum while the main effect of classification method (R), together with all first and second-order interactions involving R were contained in the method (sub-plot) stratum. All

second order interactions and above in the experimental factor stratum and all third order interactions and above in the method stratum were pooled with the respective error variances. Individual analyses of the main effects and first order interactions for the four methods were also carried out in order to compare the effects of the different experimental factors within each method separately, and also because the mean square errors for each method were found to be unequal.

In order to obtain an unbiased estimate of the error rate for each run, g-fold cross-validation was carried out. Ideally it would be desirable to set the value of g = n for all methods, but using this caused the FACT program to crash on many occasions, due to either a floating point error or cross-validation samples being too unbalanced or too similar at some node. On those occasions where n-fold cross-validation was successful, however, the ten-fold and n-fold cross-validation error rates were found to be nearly identical. Hence, as suggested by Loh (1988), the number of cross-validations for FACT was set to ten.

G-fold cross-validation was used by Ildiko and Lanteri (1989), Feng et al (1993) and Schaffer (1993) as a means of comparing various classification methods using real data sets. Although, as seen in Section 4.2, it was noted by many authors that cross-validation had large variance for small samples, most of the reported results were for smaller sized samples than occur here.

4.3.2 Results

The results from the experimental factor stratum of the ANOVA are not of interest in this study, indicating only if the experimental factors had any effect on the error rates for all methods considered as a whole. The method stratum, however, gave more relevant results showing that the R (method) main effect ($F = 72.13$) and the R * f(**x**) (method by distribution) interaction ($F = 54.98$) were extremely significant although all other first order interactions were also highly significant with smaller F-ratios (range: $F = 9.37$ to $F = 17.04$), except R * e (the interaction of method by priors-covariance structure) which was not significant. Five second order interactions were also found to be significant, but all except two were only just significant, those being R * p * f(.), where p is the number of variables ($F = 60.01$) and R * δ * f(.), where δ is the distance between groups ($F = 4.24$). These results not only showed that there were differences in error rates between the four methods but also that a comparison of error rates between the four methods depended on the factors p, n (sample size), δ and f(.) as well as the p * f(.) and δ * f(.) interactions.

Tables 4.1 to 4.7 give the means and standard errors of the differences in means for the computed error rates of the main effects and two most significant first order interactions for each method, with the method producing the lowest error rate for each factor or interaction given in bold. Three standard errors for the differences between the means are given below each table. They are:

The standard error of the difference between the QDA means.
The standard error of the difference between the LDA, CART and FACT means.
The standard error of the difference between QDA and the other three methods.

This was carried out because LDA, CART and FACT had roughly equal mean square errors from the individual ANOVA's while QDA had a mean square error roughly twice that of the other three methods.

It was found that increasing the number of variables increased the error rate for all of the methods, with the largest effect for the individual ANOVA's being for FACT (F = 168.22), with the error rate being 9.9 points larger when p = 6 compared to p = 2. Table 4.1 shows that CART produced the lowest average error rate, no matter what the value of p, though the average error estimates for LDA were both within 2.5% of the CART error rates. QDA, however, was influenced by distribution as well. This will be discussed in a later paragraph on interactions.

Table 4.1: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the dimension (p)

Level	LDA	QDA	CART	FACT
p = 2	0.089	0.096	**0.068**	0.114
p = 6	0.115	0.144	**0.102**	0.212
Standard error of the difference between the QDA means = 0.011				
Standard error of the difference between the LDA, CART and FACT means = 0.008				
Standard error of the difference between QDA and the other three methods = 0.010				

Increasing sample size had the effect of reducing the error rates for all the methods with LDA having the largest effect for increasing n (F = 94.67). Table 4.2 shows that CART produced

the lowest average error rate, no matter what the value of n, though as for p, the average error estimates for LDA were within 2.5% of the CART error rates.

Table 4.2: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the sample size (n)

Level	LDA	QDA	CART	FACT
n = 60	0.111	0.136	**0.101**	0.200
n = 300	0.093	0.104	**0.069**	0.126
Standard error of the difference between the QDA means = 0.011				
Standard error of the difference between the LDA, CART and FACT means = 0.008				
Standard error of the difference between QDA and the other three methods = 0.010				

Increasing the distance between groups also had the effect of reducing the error rates for all the methods with LDA having the largest effect for increasing δ ($F = 82.63$). Table 4.3 shows that if $\delta = 3$, LDA did slightly better than CART, but when $\delta = 2$, the average error rate for LDA was 4% more than that for CART. CART, and to a lesser extent, QDA and FACT, were also influenced by the distribution. This will be discussed in a later paragraph on interactions.

Table 4.3: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the distance between groups (δ)

Level	LDA	QDA	CART	FACT
$\delta = 2$	0.139	0.161	**0.097**	0.193
$\delta = 3$	**0.065**	0.079	0.073	0.133
Standard error of the difference between the QDA means = 0.011				
Standard error of the difference between the LDA, CART and FACT means = 0.008				
Standard error of the difference between QDA and the other three methods = 0.010				

The most interesting finding however, was with respect to the distribution of the data set. If f(.) was lognormal rather than normal the error rate would decrease for LDA, CART and

FACT, whereas for QDA, the error rate would significantly increase (F = 6.35). The LDA finding does not support Lachenbruch et al (1973), but in that study the pure lognormal distribution was used. CART was found to be most sensitive to changes in f(.). Table 4.4 shows that when f(.) was normal, QDA produced the lowest error rate, slightly lower than LDA and moderately lower than CART, but when f(.) was lognormal the average error rate for CART was at least 5% less than that for LDA, which in turn had a mean error rate at least 5% less than that for QDA. The error rates for some of the methods though were affected by the dimension and/or distance between the groups.

Table 4.4: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the type of distribution, (f(.))

Level	LDA	QDA	CART	FACT
f(**x**) = normal	0.121	**0.106**	0.140	0.200
f(**x**) = lognormal	0.083	0.134	**0.030**	0.126
Standard error of the difference between the QDA means = 0.011				
Standard error of the difference between the LDA, CART and FACT means = 0.008				
Standard error of the difference between QDA and the other three methods = 0.010				

Table 4.5 shows that the error rates for QDA when f(.) was normal were not greatly affected by the size of p, but when f(.) was lognormal, increasing p almost doubled the error rate for QDA. QDA did best on all occasions where f(.) was normally distributed, while CART did appreciably better where f(.) was lognormal.

Table 4.5: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the dimension-distribution interaction (p * f(.))

Interaction	LDA	QDA	CART	FACT
p = 2, f(.) = normal	0.112	**0.100**	0.111	0.151
p = 2, f(.) = lognormal	0.066	0.091	**0.025**	0.077
p = 6, f(.) = normal	0.130	**0.112**	0.169	0.249
p = 6, f(.) = lognormal	0.100	0.176	**0.034**	0.176
Standard error of the difference between the QDA means = 0.016				
Standard error of the difference between the LDA, CART and FACT means = 0.011				
Standard error of the difference between QDA and the other three methods = 0.013				

Table 4.6 shows that CART, and to a lesser extent, QDA and FACT, were affected more by f(.) for $\delta = 2$ than for $\delta = 3$. As for the p * f(.) interaction, QDA did best on all occasions where f(.) was normally distributed, while CART did appreciably better where f(.) was lognormal.

Table 4.6: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the distance-distribution interaction (δ * f(.))

Interaction	LDA	QDA	CART	FACT
$\delta = 2$, f(.) = normal	0.160	**0.139**	0.162	0.238
$\delta = 2$, f(.) = lognormal	0.118	0.183	**0.031**	0.148
$\delta = 3$, f(.) = normal	0.082	**0.073**	0.118	0.162
$\delta = 3$, f(.) = lognormal	0.048	0.085	**0.028**	0.105
Standard error of the difference between the QDA means = 0.016				
Standard error of the difference between the LDA, CART and FACT means = 0.011				
Standard error of the difference between QDA and the other three methods = 0.013				

From Tables 4.1 through 4.6 it can also be seen that the average error rate for FACT exceeded all others except where f(.) was lognormal and $\delta \neq 3$.

Of the five-level factor, only the fifth level was found to be significant for any method, and that for LDA only, implying that using LDA on samples which had more observations in the group with the smallest variance will increase the error rate from the ideal equal priors, equal variance instance.

Table 4.7 gives the results for the five level factor for completeness, showing that CART produced the lowest error rate every time, but the differences between the methods only mirrored the overall differences between the methods, taken over all 80 data sets.

Table 4.7: Means and standard errors of the differences in means of the cross-validation error rate estimates for each classification method with respect to the priors-covariance structure (e)

Level	LDA	QDA	CART	FACT
$\pi_1 = 0.5$: $\Sigma_1 = \Sigma_2 = I$	0.106	0.127	**0.099**	0.171
$\pi_1 = 0.5$: $\Sigma_1 = I$, $\Sigma_2 = 3\Sigma_1$	0.100	0.128	**0.075**	0.166
$\pi_1 = 0.25$: $\Sigma_1 = \Sigma_2 = I$	0.086	0.107	**0.079**	0.160
$\pi_1 = 0.25$: $\Sigma_1 = I$, $\Sigma_2 = 3\Sigma_1$	0.090	0.114	**0.072**	0.157
$\pi_1 = 0.75$: $\Sigma_1 = I$, $\Sigma_2 = 3\Sigma_1$	0.128	0.123	**0.100**	0.161
Standard error of the difference between the QDA means = 0.017				
Standard error of the difference between the LDA, CART and FACT means = 0.012				
Standard error of the difference between QDA and the other three methods = 0.015				

ANOVA's were also calculated for the size of the decision trees from CART and FACT on the experimental factors. On average, a CART tree contained 4.35 terminal nodes with only increasing p and/or a lognormal data set having any significant influence on the size of the tree. The average size of a FACT tree also contained 4.35 terminal nodes, but depended on many factors. For both CART and FACT, various rules were used to limit the size of a tree to less than ten terminal nodes involving either increasing the size below which a node cannot be split for both CART and FACT, or selecting the smallest tree within b standard errors, $0 \leq b \leq 2$, of the tree with the smallest cross-validated error rate, where the standard error of the misclassification cost is

$$se_{R(CV)} = [R(CV)\,(1 - R(CV))\,/\,n]^{0.5}$$

where n is the size of the data set and R(CV) is as defined earlier. This last rule only applies for CART.

4.3.3 Summary

In this study, it was found that increasing the dimension significantly increased the error rate while increasing either sample size or distance between groups both significantly decreased the error rate for each of the four methods. Using a data set that was lognormally rather than normally distributed significantly reduced the error rate for each of LDA, CART and FACT, but increased the error rate when QDA was used. The size of a FACT tree depends on many different criteria, but CART is only influenced by the distribution, normal or lognormal, and/or the dimension of the data set.

This study has also shown that CART performs better (on average) than either LDA or QDA no matter what the sample size or number of variables. LDA performed slightly better on average than CART when the distance between groups was large and when the data set was normally distributed. In the latter situation, QDA even outperformed LDA. CART, on the other hand, performed much better than both LDA and QDA when the distance between groups was not so large, and moderately better than LDA and considerably better than QDA when the data set was lognormal.

The differences in error rates between CART and LDA or QDA for the data sets with different priors and/or covariance structures, was found to be negligible, whereas for the two most significant first-order interactions, the dimension * distribution and the distance * distribution interactions, CART performs best on average except where the data set was normally distributed for which QDA did best. QDA performed poorly when the distribution was lognormal though, while FACT had the largest error rate for almost every run, performing especially badly when there were six variables in the data set.

It can thus be concluded that CART, a tree-based, non-parametric method, will in many cases perform as well as if not better than the usual parametric methods of classification, LDA and QDA, constructing a tree that is usually not too large. Only when the distribution of the data

set is normal or the distance between the popuations is large does either LDA and/or QDA perform better than CART. From a predictive point of view then, CART is a narrow winner.

Quinlan (1993) noted the distinction between **parallel** classification problems whereby all the variables have equal weightings, so that the classification rules depend on all the variables, and **sequential** classification problems whereby only a few of the variables contribute to the classification rules generated. He suggested that connectionist methods such as LDA and QDA are preferred for parallel classification problems while symbolic methods such as tree-based procedures are best for sequential problems. In this study, all the data sets were examples of parallel classification problems, and not suited to CART (and FACT). Hence, CART has been shown to perform relatively well for classification problems for which it is not particularly suited.

4.4 SIMULATION STUDY II

4.4.1 Study Plan

It was decided, in retrospect, to compare normally distributed data with the pure lognormal, so that comparisons could be made with the work of other authors (for example, Lachenbruch et al, 1973 and Chinganda and Subrahmaniam, 1979). This is, in fact, a monotonic transformation of the variables so the results for CART are invariant under either type of distribution. In the previous section, each observation was divided by the original class standard deviations. As the standard deviation for Π_2 was greater than that for Π_1, in the lognormal case, the rankings of the x_{ij}'s would not remain the same after transformation.

Therefore, another simulation study was carried out using exactly the same probability models that were used in Section 4.3 except that f(.) was either normal or true lognormal (without standardization) so that log[f(.)] was normal.

4.4.2 Results

As in Section 4.3, the results from the probability model stratum of the ANOVA are not useful for this study, indicating only if the probability models had any effect on the error rates for all methods considered as a whole. The method stratum, however, gave more relevant results showing that the R(method) main effect (F = 74.57) and the R * f(.) (method by distribution) interaction (F = 52.45) were extremely large while all other first order

interactions were highly significant (range: F = 5.66 to 26.19). Five second order interactions were also found to be significant at the 5% level with four of those interactions also being significant at the 1% level, those being the interactions of R * p * δ where p is the number of variables and δ is the distance between groups, R * p * f(.), R * δ * f(.) and R * f(.) * e, where e is the prior-covariance structure of the data. Residual mean square for the analysis was 1.35×10^{-3}. These results showed that a comparison of error rates between the four methods depended on the factors p, n (sample size), δ, f(.) and e as well as the p * δ, p * f(.), δ * f(.) and f(.) * e interactions.

Tables 4.8 to 4.16 give the means and standard errors of the differences in means for the computed error rates of the main effects and the four most significant first order interactions for each method, with the method producing the lowest error rate for each level of the factor or interaction given in bold. Three standard errors for the differences between means are given below each table. They are:

The standard error of the difference between the FACT means.
The standard error of the difference between the LDA, QDA and CART means.
The standard error of the difference between FACT and the other three methods.

This was carried out because LDA, QDA and FACT had roughly equal mean square errors from the individual ANOVA's while FACT had a mean square error roughly five times that of the other methods.

It was found that increasing the number of variables increased the error rates for CART and FACT, with the largest effect being for CART (F = 78.02), whereas for LDA, the error rate decreased (F = 16.56). There was no real effect on the QDA error rate when increasing p. Table 4.8 shows that CART produced the lowest average error rate when p = 2 but when p = 6, QDA did best. Note though that CART and FACT were influenced by distance as well while LDA and FACT were also influenced by the distribution.

Table 4.8: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the dimension (p)

Level	LDA	QDA	CART	FACT
p = 2	0.198	0.128	**0.111**	0.174
p = 6	0.173	**0.129**	0.169	0.231
Standard error of the difference between the FACT means = 0.014				
Standard error of the difference between the LDA, QDA and CART means = 0.006				
Standard error of the difference between FACT and the other three methods = 0.011				

Increasing sample size had the effect of reducing error rates for CART and FACT, with CART having the largest effect (F = 22.37), but had no real effect on the error rates for LDA and QDA. Table 4.9 shows that QDA produced the lowest average error rate when n = 60 while the difference between CART and QDA was negligible for n = 300.

Table 4.9: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the sample size (n)

Level	LDA	QDA	CART	FACT
n = 60	0.191	**0.133**	0.156	0.235
n = 300	0.180	0.125	**0.124**	0.170
Standard error of the difference between the FACT means = 0.014				
Standard error of the difference between the LDA, QDA and CART means = 0.006				
Standard error of the difference between FACT and the other three methods = 0.011				

Increasing the distance between groups had the effect of reducing error rates for all the methods with QDA having the largest effect for increasing δ (F = 210.37). Table 4.10 shows that if δ = 2, CART did slightly better than QDA but when δ = 3, the average error rate for CART was 3% more than that for QDA. Note though that CART and FACT were also influenced by dimension while LDA, QDA and FACT were also influenced by distribution.

Table 4.10: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the distance between groups (δ)

Level	LDA	QDA	CART	FACT
$\delta = 2$	0.215	0.173	**0.162**	0.228
$\delta = 3$	0.156	**0.085**	0.118	0.177
Standard error of the difference between the FACT means = 0.014				
Standard error of the difference between the LDA, QDA and CART means = 0.006				
Standard error of the difference between FACT and the other three methods = 0.011				

It was found that if f(.) was lognormal rather than normal, the error rate would increase for LDA and QDA but have no real effect for CART and FACT. For CART, there was no change at all to the results due to the fact that CART is invariant under all monotone transformations of the variables. LDA was found to be most sensitive to changes in f(.) (F = 408.96) supporting the findings of Lachenbruch et al (1973). Table 4.11 shows that QDA did best when f(.) was normal while CART did best when f(.) was lognormal. The error rates for LDA and FACT were also influenced by dimension while all other methods except CART were influenced by distance and priors-covariance structure.

Table 4.11: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the distribution of the data set (f(.))

Level	LDA	QDA	CART	FACT
f(.) = normal	0.121	**0.106**	0.140	0.200
f(.) = lognormal	0.250	0.151	**0.140**	0.205
Standard error of the difference between the FACT means = 0.014				
Standard error of the difference between the LDA, QDA and CART means = 0.006				
Standard error of the difference between FACT and the other three methods = 0.011				

All methods were affected by changing the priors-covariance structure of the data with the largest effect being for LDA (F = 48.79), which had a large increase in error rate when $\pi_1 = 0.25$ and $\Sigma_2 = 3\Sigma_1$. CART was the least affected by any change in priors-covariance structure. Table 4.12 shows that QDA did best on all occasions where $\Sigma_1 \neq \Sigma_2$. Note, though, that all methods except CART were also affected by the distribution.

Table 4.12: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the priors-covariance structure (e)

Level	LDA	QDA	CART	FACT
$\pi_1 = 0.5$: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$	**0.154**	0.163	0.178	0.225
$\pi_1 = 0.5$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$	0.185	**0.092**	0.124	0.189
$\pi_1 = 0.25$: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$	0.190	0.191	**0.148**	0.230
$\pi_1 = 0.25$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$	0.264	**0.100**	0.118	0.219
$\pi_1 = 0.75$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$	0.134	**0.098**	0.132	0.149
Standard error of the difference between the FACT means = 0.023				
Standard error of the difference between the LDA, QDA and CART means = 0.010				
Standard error of the difference between FACT and the other three methods = 0.018				

Table 4.13 shows that the error rates for CART when p = 2, were more affected by an increase in δ than when p = 6, while for FACT, the opposite effect occurred. CART did best on both occasions when p = 2 while QDA produced the lowest average error rates when p = 6.

Table 4.13: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the dimension-distance interaction ($p * \delta$)

Level	LDA	QDA	CART	FACT
$p = 2, \quad \delta = 2$	0.222	0.172	**0.142**	0.188
$p = 2, \quad \delta = 3$	0.175	0.085	**0.079**	0.160
$p = 6, \quad \delta = 2$	0.207	**0.174**	0.181	0.268
$p = 6, \quad \delta = 3$	0.138	**0.084**	0.157	0.194
Standard error of the difference between the FACT means = 0.020				
Standard error of the difference between the LDA, QDA and CART means = 0.009				
Standard error of the difference between FACT and the other three methods = 0.016				

Table 4.14 shows that the error rates for LDA were affected more by f(.) when $p = 2$ than for $p = 6$, whereas for FACT when $p = 2$, the error rate increased if f(.) was lognormal rather than normal, but when $p = 6$, the error rate decreased if f(.) was lognormal. QDA did best except when $p = 2$ and f(.) was lognormal where CART produced the lowest mean error rate.

Table 4.14: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the dimension-distribution interaction ($p*f(.)$)

Level	LDA	QDA	CART	FACT
$p = 2$, f(.) = normal	0.112	**0.100**	0.111	0.151
$p = 2$, f(.) = lognormal	0.285	0.156	**0.111**	0.197
$p = 6$, f(.) = normal	0.130	**0.112**	0.169	0.249
$p = 6$, f(.) = lognormal	0.215	**0.146**	0.169	0.213
Standard error of the difference between the FACT means = 0.020				
Standard error of the difference between the LDA, QDA and CART means = 0.009				
Standard error of the difference between FACT and the other three methods = 0.016				

Table 4.15 shows that f(.) had a different effect on LDA, QDA and FACT when $\delta = 2$ compared to $\delta = 3$. The most interesting finding was for FACT where the error rates for $\delta = 2$ were lower when f(.) was lognormal rather than normal, but when $\delta = 3$ the reverse effect occurred. QDA produced the lowest error rate except when $\delta = 2$ and f(.) was lognormal where CART did best.

Table 4.15: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the the distance-distribution interaction (δ * f(.))

Level	LDA	QDA	CART	FACT
$\delta = 2$, f(.) = normal	0.160	**0.139**	0.162	0.238
$\delta = 2$, f(.) = lognormal	0.269	0.207	**0.162**	0.219
$\delta = 3$, f(.) = normal	0.082	**0.073**	0.118	0.162
$\delta = 3$, f(.) = lognormal	0.230	**0.096**	0.118	0.192
Standard error of the difference between the FACT means = 0.020				
Standard error of the difference between the LDA, QDA and CART means = 0.009				
Standard error of the difference between FACT and the other three methods = 0.016				

Table 4.16 shows that all methods except CART were affected by the f(.) * e interaction. QDA did best on all occasions where $\Sigma_1 \neq \Sigma_2$, while LDA did best when f(.) was normal and $\Sigma_1 = \Sigma_2$ and CART did best when f(.) was lognormal and $\Sigma_1 = \Sigma_2$.

Table 4.16: Means and standard errors of the difference in means of the cross-validation error rate estimates for each classification method with respect to the distribution-priors-covariance interaction (f(.) * e)

Level	LDA	QDA	CART	FACT
f(.) = normal, e = 1	**0.120**	0.127	0.178	0.224
f(.) = normal, e = 2	0.112	**0.087**	0.124	0.201
f(.) = normal, e = 3	**0.120**	0.132	0.148	0.211
f(.) = normal, e = 4	0.113	**0.084**	0.118	0.188
f(.) = normal, e = 5	0.141	**0.101**	0.132	0.175
f(.) = lognormal, e = 1	0.187	0.199	**0.178**	0.226
f(.) = lognormal, e = 2	0.258	**0.097**	0.124	0.177
f(.) = lognormal, e = 3	0.261	0.250	**0.148**	0.250
f(.) = lognormal, e = 4	0.416	**0.116**	0.118	0.250
f(.) = lognormal, e = 5	0.128	**0.095**	0.132	0.123

Standard error of the difference between the FACT means = 0.032

Standard error of the difference between the LDA, QDA and CART means = 0.014

Standard error of the difference between FACT and the other three methods = 0.025

Legend:

e = 1: $\pi_1 = 0.5$: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$

e = 2: $\pi_1 = 0.5$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$

e = 3: $\pi_1 = 0.25$: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$

e = 4: $\pi_1 = 0.25$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$

e = 5: $\pi_1 = 0.75$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$

From Tables 4.8 to 4.16 it can be seen that on most occasions the average error rate for FACT exceeded all others.

ANOVA's were also calculated for the size of the decision trees from CART and FACT on the experimental factors. On average, a CART tree contained 5.18 terminal nodes with only increasing p and/or the p * e interaction having any significant influence on the size of the tree. The average FACT tree contained 4.41 terminal nodes, but depended on many factors. For both CART and FACT, various rules were used to limit the size of a tree to less than ten terminal nodes involving either increasing the size below which a node cannot be split for

both CART and FACT, or selecting the smallest tree within b standard errors, $0 \leq b \leq 2$, of the tree with the smallest cross-validated error rate, where the standard error of the misclassification cost is as given in Section 4.3.

4.4.3 Summary and Discussion

In this study, it was found that either increasing dimension or decreasing sample size had the effect of increasing the error rate for CART and FACT while increasing dimension decreased the error rate for LDA. Neither dimension nor sample size had any real effect on the error rates for QDA nor did sample size for LDA. This was most probably because neither the sample size nor ratio of dimension to sample size were set low enough to seriously affect the error rates from the above two methods. Increasing the distance between groups was found to significantly decrease the error rate for all methods while using a data set that was lognormally rather than normally distributed significantly increased the error rate for all methods except CART where it had no effect at all. The error rates for LDA were affected most by the changes in the five-level factor while CART was affected least. Note, however, that only for sample size were there no significant interactions. It was also found that the size of a FACT tree depended on many different criteria in contrast to a CART tree.

This study has also shown that CART performs better on average than the other three methods when either the distribution was lognormal, dimension was small or the distance between groups was small, as well as when there were equal covariance matrices but unequal priors. On all other occasions, QDA produced the lowest average error rate except in the equal priors, equal covariance case, where LDA did best and when sample size was large where the differences between QDA and CART were negligible.

For the four most significant first order interactions, QDA did best on average when the distribution was normal, except when the covariance matrices were equal. CART did best on most occasions where the distribution of the data set was lognormal.

At this stage, it would be desirable to tie together the results from both Sections 4.3 and 4.4, in order to provide some reasons for the differences in results. In Section 4.3, normally distributed data was compared with lognormal data with the means and covariances (thus δ values) being exactly as specified in (4.3.1) and (4.3.2) by standardising the data. In Section 4.4, however, no standardisation was undertaken resulting in means and covariances grossly

different to the values used in Section 4.3. Consider the simple case of p = 1. If $X \sim N(\mu, \sigma^2)$ and $Y = e^x$ (that is, lognormally distributed) the values of E[Y] and V[Y], the mean and covariance of Y respectively, are defined as (Aitchison and Brown, 1957):

$$E[Y] = \exp(\mu + \sigma^2/2) \tag{4.4.1}$$

and

$$V[Y] = \exp(2\mu + \sigma^2)\,[\exp(\sigma^2) - 1] \tag{4.4.2}$$

Suppose that $X_1 \sim N(0, 1)$ and $X_2 \sim N(2, 1)$. Let $Y_1 = e^{x_1}$ and $Y_2 = e^{x_2}$, then

$$E[Y_1] = e^{0.5} \quad \text{and} \quad V[Y_1] = e(e - 1)$$

with

$$E[Y_2] = e^{2.5} \quad \text{and} \quad V[Y_2] = e^5(e - 1)$$

giving $\delta = 0.924$.

Therefore, the distance between populations has been reduced by 53.8% from the case of normally distributed data. When the distance between populations is increased in the case of normally distributed data, the relative reduction in δ caused by the exponential transformation increases correspondingly.

From the results, it is obvious that methods which use linear discriminant functions to form the classification rules (LDA and FACT) have lower error rates for lognormally shaped data when the covariance structure of the data is unaltered (see Tables 4.4, 4.5 and 4.6). When the data is transformed to be true lognormal, the means and covariances are drastically altered so that the distance between populations is reduced. This implies that the markedly increased error rates for LDA in this situation have been caused by a reduction in the separation between populations rather than the lognormally shaped data. For FACT, the increase in error rates is minimal. QDA, which models the individual class means and covariances separately, was affected in the same way by lognormally shaped data (increased error rates), no matter what the means and covariance structure of the data were. This also provides an explanation why there were no significant priors-covariance structure main effects or interactions in Section 4.3 but that there were numerous such occurrences in Section 4.4.

4.5 THE EFFECTS OF PRIORS ON ERROR RATES

4.5.1 Introduction

This section gives details of a simulation study which was carried out using the data sets that were used in Section 4.4. The pure lognormal was used rather than the standardised lognormal so that comparisons could be made with the work of other authors (for example, Lachenbruch et al, 1973). The purpose of the study was to compare the group misclassification error rates for the two parametric discrimination methods, LDA and QDA, and those of the two tree-based methods, CART and FACT, using both priors proportional to sample size (PPSS) and equal priors. A comparison of the overall error rates using PPSS and equal priors was also done. The results are presented followed by a discussion of the implications of the study.

4.5.2 Purpose of this study

Tests using several data sets have found that although the set of decision rules produced by CART were about as accurate as those produced by LDA, the individual group misclassification error rates tended to be more variable for CART using PPSS. The most noticeable trend observed was that CART favoured the group(s) with the larger sample size(s) to the detriment of the group(s) with the smaller sample size(s). This occurred despite the fact that a larger proportion of observations from a class with a smaller sample size were sent one way than a class with a larger sample size. However, because there were more observations from the class with the larger sample size sent the same way, the node was assigned to that class. This resulted in all the observations from the smaller class(es) being misclassified. Using LDA, however, in the case of PPSS, did not lead to such extremes of group misclassifications that were encountered above for CART. This is because the group separation functions used in LDA to discriminate between two groups

$$\hat{D}_{ij}(\mathbf{x}) = \ln(\pi_i/\pi_j) + (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)\, S^{-1}[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_j)]$$

are only changed by one term, that is $\ln(\pi_i/\pi_j)$, when sample sizes are different. The rules tend to favour the class with the larger size but not to the same extent as those of CART, but if an observation is much closer to one sample than another then that observation should still be allocated to that sample no matter what the ratio of class sample sizes.

Making the priors equal had a vastly different effect on CART than LDA for many of the examples carried out. The net result was that when the observations in the small sample were even slightly more homogeneous than those in the larger sample, in relation to the discriminatory variables, then the splits were biased in favour of the small samples. This often resulted in all the observations from a larger class being completely misclassified, while most of those in the smaller samples were correctly classified, leading to a large overall error rate and hence an inaccurate set of decision rules. As well, the decision rules created were totally different from those found using PPSS, with the variables strongly associated with the smaller sample(s) being split on. It could be stated that altering the priors had somewhat reversed the group misclassification error rates for CART. With LDA, however, $\ln(\pi_i/\pi_j) = 0$ so there was no bias due to sample size with discrimination based solely on the distance measure of one observation to a pair of sample means.

This study was carried out to assess the performance of LDA and CART, as well as QDA and FACT, in correctly classifying observations in the simple case of just two groups. The study was intended to test whether CART is more susceptible to changes in the structure of the prior probabilities of group membership. Note, though, that a failure to show this trend will not necessarily invalidate what was conjectured before about CART's sensitivity to changes in the structure of the priors, but instead that CART works adequately in the case of two samples.

4.5.3 Study Plan

This study uses the eighty different bimodal probability models that were generated in Section 4.4. The prior probabilities for each of the two classes will be set to either equal or PPSS.

As in the estimation of overall error rates in Sections 4.3 and 4.4, the group misclassification error rates R(i/j), i, j = 1, 2, $i \neq j$, were estimated using n-fold cross-validation. The resubstitution or apparent error rates were also calculated in the case of both PPSS and equal priors for LDA and CART only.

Two measures of comparison of the group misclassification error rates were carried out. The first compares the difference between the two group error rates

$$R(1/2) - R(2/1)$$

It could be argued that the above is not a true measure of the performance of the individual methods over the possible range of error rates. For example, suppose that R(1/2) = 0.4 and R(2/1) = 0.35. This gives a difference of 0.05. In another set, R(1/2) = 0.05 while R(2/1) = 0 which also gives a difference of 0.05. However, in the high error case the two group error rates are relatively similar, while those in the low error case are not. The straight difference between error rates is heavily weighted towards the data sets with the lowest error rates. Therefore, a second proposed measure of performance is the ratio of the two group error rates adjusted to avoid the possibility of invalid values in the cases where R(2/1) = 0.

$$(R(1/2) + 0.01) / (R(2/1) + 0.01)$$

4.5.4 Results

The results of this study were analysed by a series of split-plot ANOVA's. (See Section 4.3 for details.) The first involved a comparison of the differences between group misclassification error rates using n-fold cross-validation for LDA, QDA, CART and tenfold cross-validation for FACT with PPSS. The results from the experimental factor stratum of the ANOVA are not useful for any of the analyses in this study indicating only if the experimental factors had any effect on either the difference between group error rates or the ratio of group error rates. The results of interest appear in the method stratum.

When comparing the differences between group error rates, it was found that the method (R) main effect (F = 393.57) was by far the most important effect. Seven method * factor(s) first and second order interactions were also significant at less than the 0.01% significance level of the F-distribution with the R * f(.) (method by distribution) interaction (F = 130.71) and the R * e (the method by priors-covariance structure) interaction (F = 91.4) being the largest. The results could not be summarised in terms of either method by distribution or method by priors-covariance structure as the second order interaction of R * f(.) * e (F = 9.91) was also highly significant. As a check on the assumptions for carrying out the ANOVA a plot of residuals against fitted values revealed no dramatic trends with just a small number of unusual observations. Boxplots of residuals for each method showed variation to be relatively similar for LDA, QDA and CART although the variability for FACT did appear to be much larger than the other three methods. A weighting using the inverses of the mean square errors from the ANOVA's of the differences in error rates on the experimental factors was employed.

Previous analyses where this weighting was carried out had shown that the results were not drastically affected.

Figures 4.1a and 4.1b give the mean absolute differences between group error rates for the five levels of e, using the data sets that were normally and lognormally distributed respectively, using PPSS. The ideal situation is where R(1/2) - R(2/1) = 0. Figure 4.1a shows that all methods produced relatively similar error rates in the ideal equal priors, equal variance case. When the covariance structure of the data was changed, it was LDA that suffered the most, with more observations from the class with the larger variance being misclassified. When one looks at the three cases where priors were not equal then it is noticeable that the two parametric methods were least affected while the two tree-based methods, FACT more so than CART, were most affected, by misclassifying a larger number of observations from the class with the smaller sample size.

Figure 4.1b shows in the case of lognormal data that the mean differences for CART were the same as in Figure 4.1a, as would be expected since one of the properties of CART is its invariance to monotone transformations of the variables. LDA has done better than CART in the unequal priors, equal covariance case but worse when covariances were unequal. The mean values for R(1/2) - R(2/1) using LDA for e = 1 and e = 3 were very similar as were those for e = 2, e = 4 and e = 5 implying that it was the difference in covariance values that caused the large disparity between group error rates when using LDA rather than the fact the data sets were lognormally distributed. FACT was the most affected by lognormal data when e = 3 and e = 4, that is when the class with the smallest sample size had the largest variance, but was less affected when e = 1, 2 or 5. As in Figure 4.1a, QDA was least affected by changes in the prior-covariance structure of the data. In general, QDA produced differences in misclassification errors which were usually larger when f(.) was lognormally rather than normally distributed but those differences were not as great as those using LDA. These results confirm those of Lachenbruch et al (1973) where a very similar set of parameters were used.

Comparison of the Difference between Group Misclassification Error Rates using LDA, QDA, CART and FACT with Priors Proportional to Sample Size: (a) Normally Distributed Data

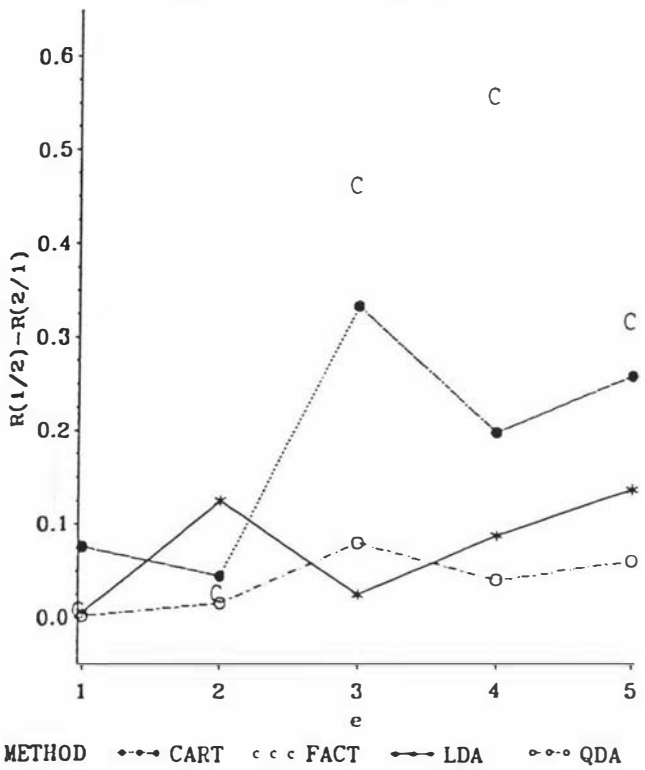


Figure 4.1a

(b) Lognormally Distributed Data

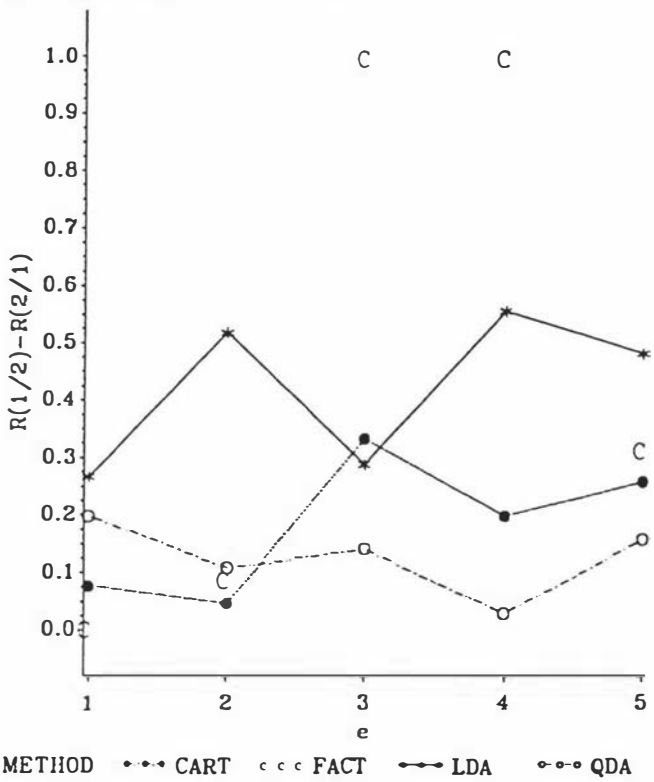


Figure 4.1b

Comparison of the Ratio of Group Misclassification Error Rates using LDA, QDA, CART and FACT with Priors Proportional to Sample Size: (a) Normally Distributed Data

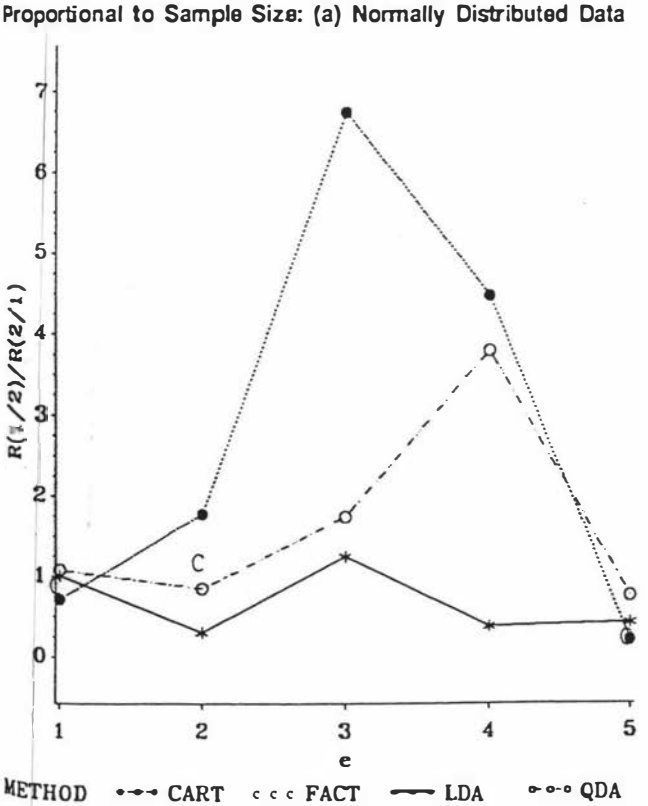


Figure 4.2a

(b) Lognormally distributed Data

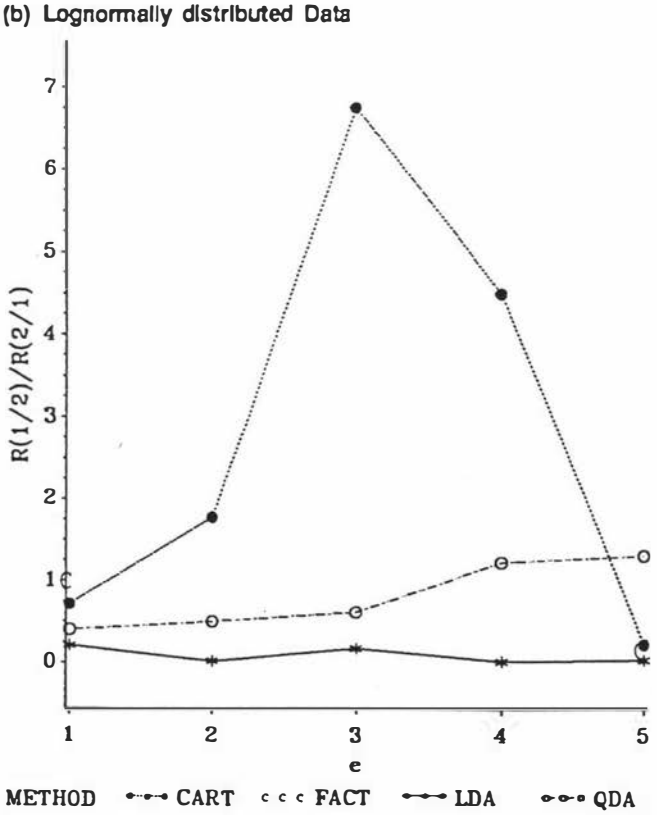


Figure 4.2b

The second analysis compared LDA, QDA, CART and FACT using the ratio of the two group error rates. The results of the split-plot ANOVA show that as in the previous analysis, the R main effect was by far the most important effect (F = 164.12) with the R * f(.) interaction (F = 71.8) also being important. As with the previous analysis, the R * f(.) * e (F = 21.14) interaction was significant so results will be presented in terms of these three factors. Analysis of the residuals shows that there was rather a funnel-like pattern among them, implying that the variation of residuals was not constant. Increasing the fitted value increased the variation of the residuals. Boxplots of the residuals for each method showed, as in the previous analysis, that the mean square error for FACT was much larger than that for the other three methods.

Figures 4.2a and 4.2b give the mean ratios of the group error rates using the adjustment factor mentioned earlier. The ideal situation in this case is where R(1/2)/R(2/1) = 1. It is noticeable that the trends observed are similar to, yet somewhat different from those in Figures 4.1a and 4.1b. Figure 4.2a shows that in the case of f(.) being normally distributed both the LDA and QDA error rates were the most stable of all the methods over the given priors-covariance levels. CART performed very badly in the unequal priors case while the results for FACT when e = 3 and e = 4 were omitted from the graph due to the excessively high values recorded.

Figure 4.2b shows that in the case of lognormal data, QDA was the best method. It must be remembered, as noted in Section 4.3.1, the variance for the second class where $\bar{\mathbf{x}}_2 > \mathbf{0}$ will be substantially larger than that for the first class where $\bar{\mathbf{x}}_1 = \mathbf{0}$. Therefore, the quadratic rules should be expected to work better than the linear rules. LDA has performed badly again but as in Figures 4.1a and 4.1b, LDA has performed consistently poorly, whereas the ratio of priors for CART was affected mostly by sample size. The results for FACT were not really worth quoting due to the excessively high ratios of error rates.

In the third analysis, it was decided to compare the mean difference between both the cross-validation and apparent group error rates using CART and LDA, but this time using equal priors. The results, from the method stratum again show the R main effect (F = 222.46) to be most important with the R * f(.) interaction (F = 130.8) also being very important.

The fourth analysis compared the CART and LDA apparent and cross-validation group error rates using the ratio of group error rates as the measure of performance. The results from the ANOVA show rather different trends to those exhibited in the previous analysis with the magnitude of the F-ratios having dramatically decreased. For instance, for the R main effect, the F-ratio was only 2.43, which not significant at even the 5% level of significance. This indicates that differences between the group misclassification error rates did indeed increase with the total error rate so that the transformations were not really necessary here though the analysis is included for completeness.

Both analyses showed that the R * f(.) * e second order interaction was highly significant so that results are presented in terms of these three factors as in the first two analyses. Figure 4.3a shows that in the case of the data sets that were normally distributed, the apparent group error rates were very similar but there was a large difference between the cross-validated group error rates, with the class having the larger sample size having the substantially larger error rate. In contrast, the apparent and cross-validated error rates for LDA were very close together, being affected most by the change in covariance structure rather than sample size. This is illustrated by the error rate for class 2 being larger than that for class 1 when both e = 4 and 5.

Figure 4.3b shows a similar pattern to Figure 4.3a except that the differences in group error rates for LDA have greatly increased using lognormally rather than normally distributed data sets. These patterns are further exemplified in Figures 4.4a and 4.4b. It has thus been shown, in the case of equal priors, that class sample size is the main factor in determining group error rates for CART when cross-validation is done, though this problem does not manifest itself when used to calculate the group error rates from the learning sample. Sample size, however, as expected, had no influence on the group error rates for LDA, in the case of equals priors which were instead influenced by the covariance structure and distribution of the data set.

Comparison of the Difference between Cross-Validation and Apparent Group Misclassification Error Rates using LDA and CART with Equal Priors: (a) Normally Distributed Data

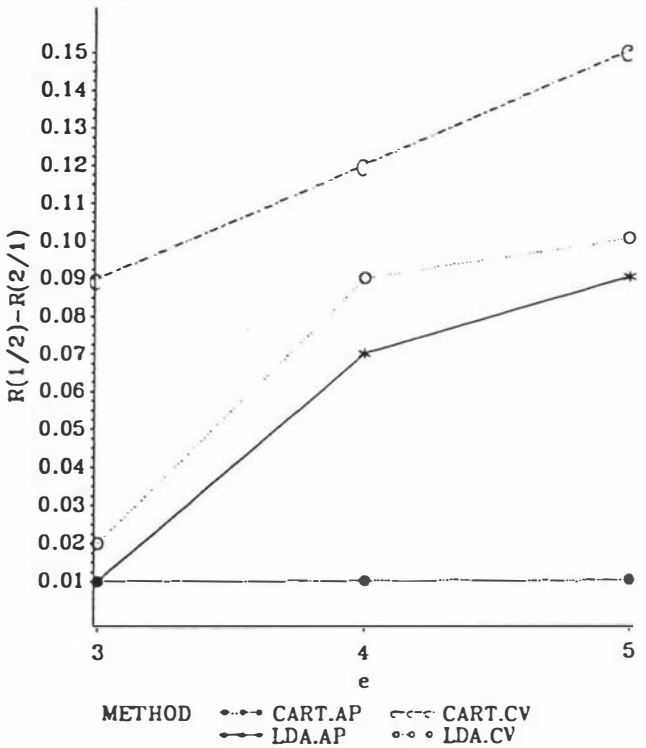


Figure 4.3a

(b) Lognormally Distributed Data

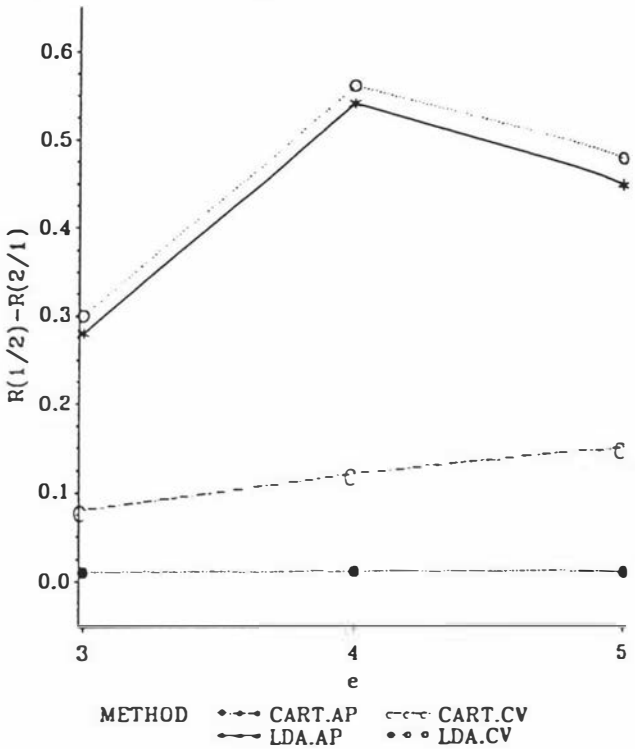


Figure 4.3b

Comparison of the Ratio of Cross-Validation and Apparent Group Misclassification Error Rates using LDA and CART with Equal Priors: (a) Normally Distributed Data

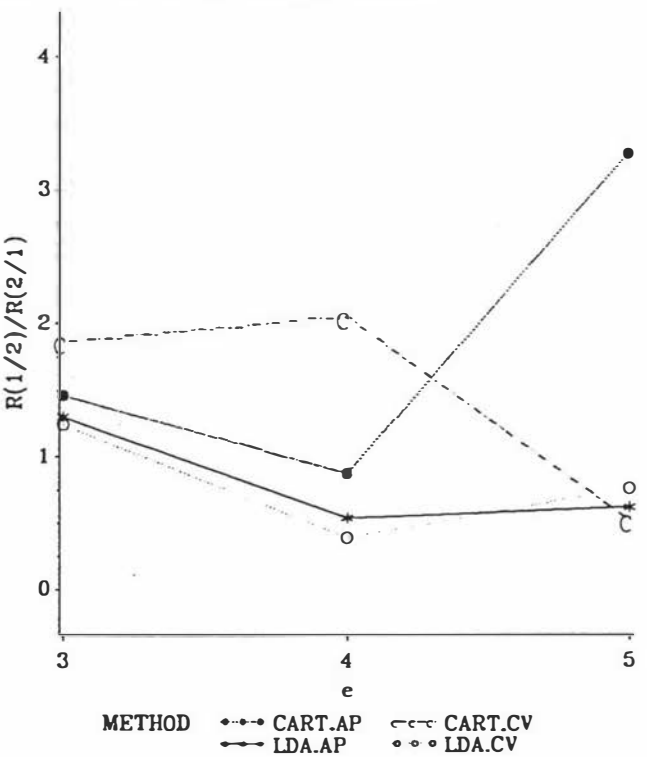


Figure 4.4a

(b) Lognormally Distributed Data

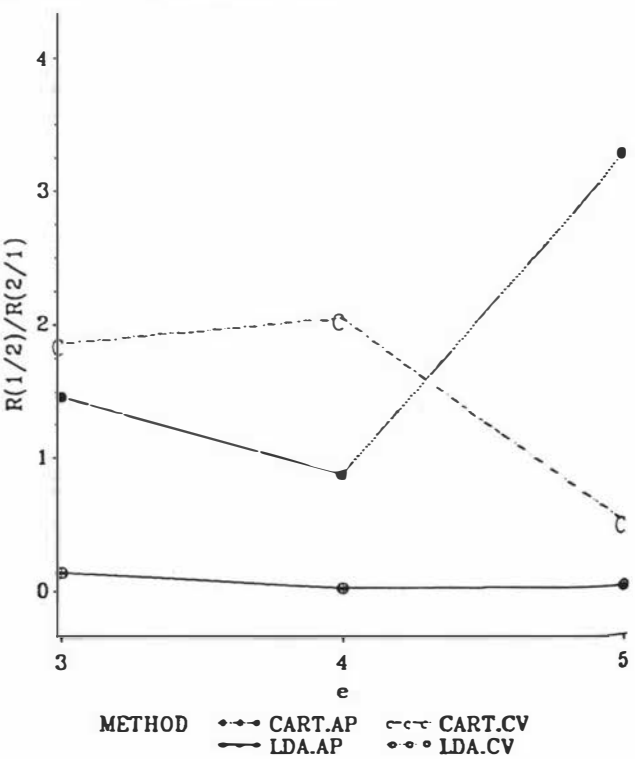


Figure 4.4b

The fifth analysis compared the CART and LDA cross-validated group error rates using the difference between error rates as the measure of performance, while the sixth analysis used the ratio of error rates, in the case of both equal priors and PPSS. As in the previous analyses, the R * f(.) * e interaction was very important so the results will be presented in terms of those three factors. Figures 4.5a and 4.5b show that choice of priors did not affect the absolute difference in group error rates, no matter what the distribution of the data set. As in the previous analyses, the group error rates for LDA were most affected by change in the covariance structure of the data set. With CART, a different pattern emerges. The use of equal priors rather than priors proportional to sample size has meant a reduction in the difference between group error rates. The greatest influence on individual group error rates for CART was sample size. Figures 4.6a and 4.6b confirm the trends mentioned above.

A next step in the analysis was to compare the overall LDA and CART error rates using equal priors. As with the group misclassification error rates, the results were analysed using a split-plot ANOVA. The ANOVA showed that the R main effect ($F = 111.8$) was highly significant as were all the R * f(.) interactions and a number of second order interactions. The two most important of these were the R * p * n ($F = 8.43$) and R * p * f(.) ($F = 6.47$) interactions.

Figures 4.7a and 4.7b compare the n-fold cross-validation and apparent error rates over all combinations of dimension and sample size using equal priors. The two graphs suggest that there was a large difference between the R(CV) and R(A) error rates in CART for smaller samples, indicating the bias of the latter estimator in such situations. This bias was only minimal for LDA. Considering the cross-validation estimates only, CART did best when $p = 2$ and LDA when $p = 6$.

Figures 4.8a and 4.8b illustrate the situation of the p * f(.) interaction. As in Figures 4.7a and 4.7b, there was a large discrepancy between the R(CV) and R(A) error rates for CART, which did not arise for LDA. Looking at the R(CV) estimates shows that LDA did best for normal data and lognormal data when $p = 6$, while CART had the lowest error rate when $p = 2$ and the data was lognormal, as distinct from the situation of PPSS where CART did best for lognormal data no matter what the number of variables in the data set.

Comparison of the Difference between Cross-Validation Group Misclassification Error Rates using LDA and CART with both Equal Priors and PPSS: (a) Normally Distributed Data

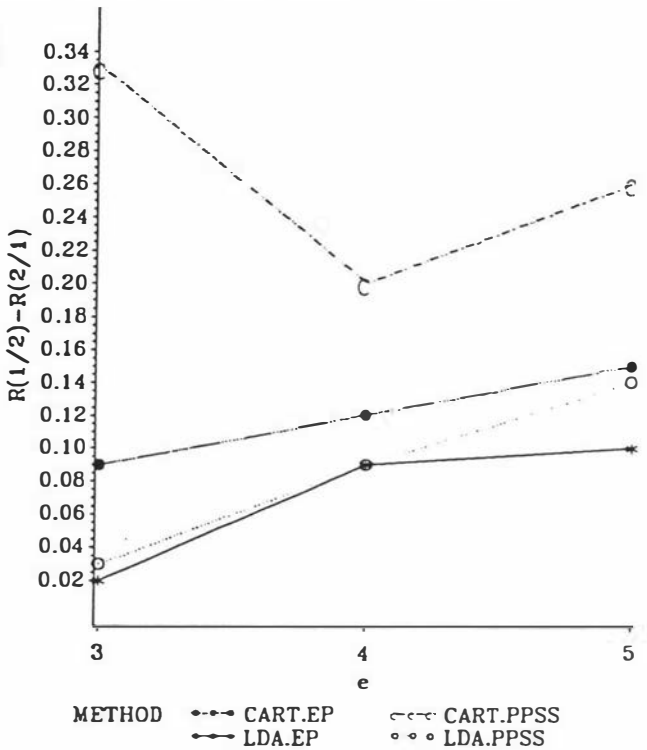


Figure 4.5a

(b) Lognormally Distributed Data

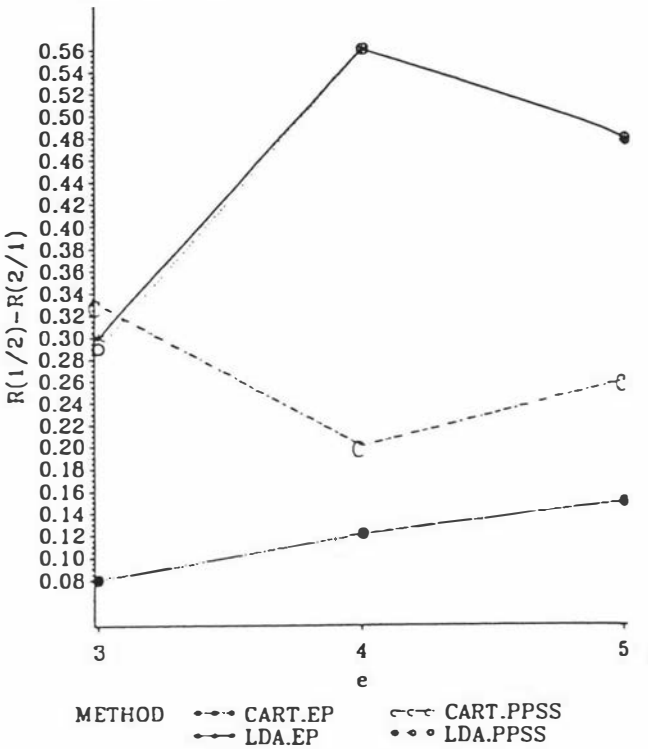


Figure 4.5b

Comparison of the Ratio of Cross-Validation Group Misclassification Error Rates using LDA and CART with both Equal Priors and PPSS: (a) Normally Distributed Data

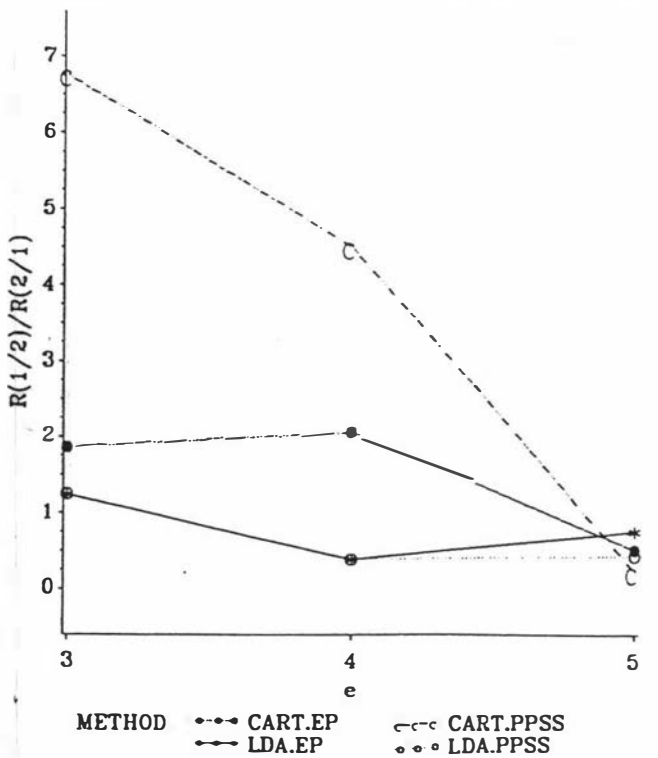


Figure 4.6a

(b) Lognormally Distributed Data

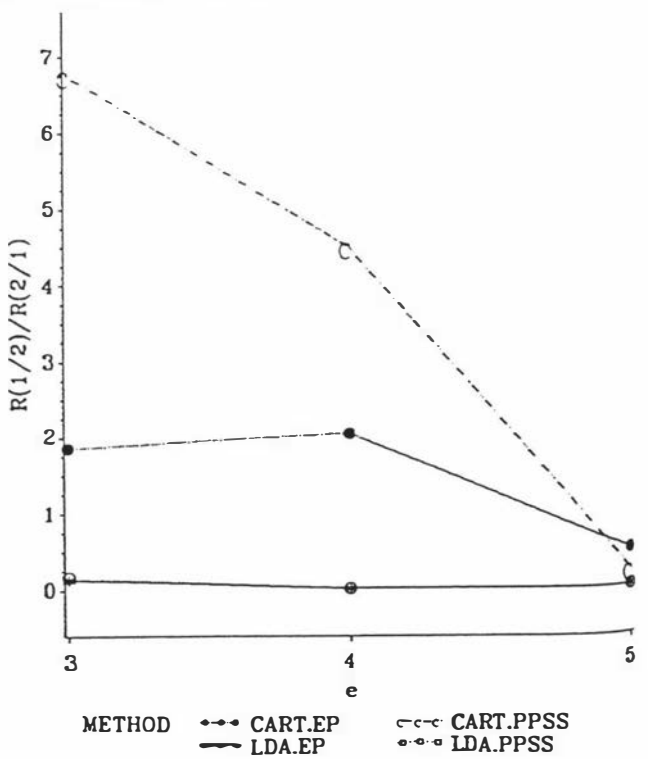


Figure 4.6b

Comparison of the Cross-Validation and Apparent Misclassification Error Rates using LDA and CART with Equal Priors: (a) Dimension = Two

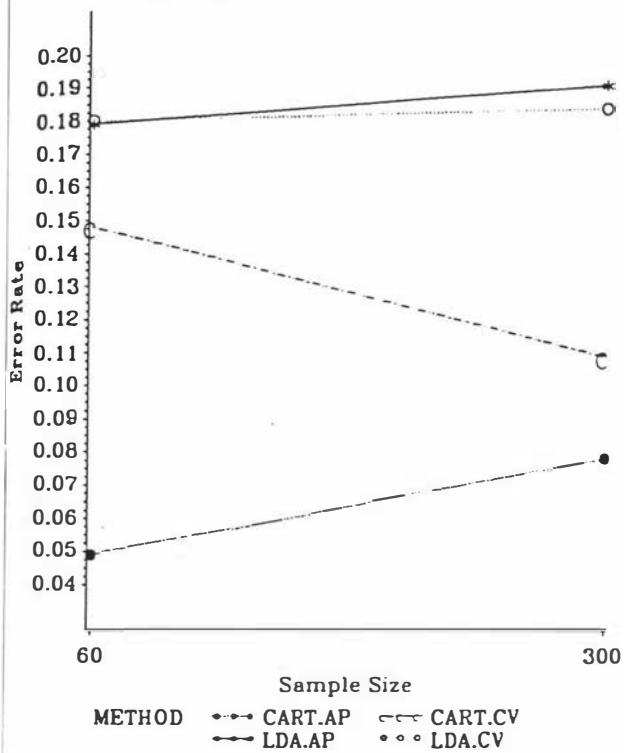


Figure 4.7a

(b) Dimension = Six

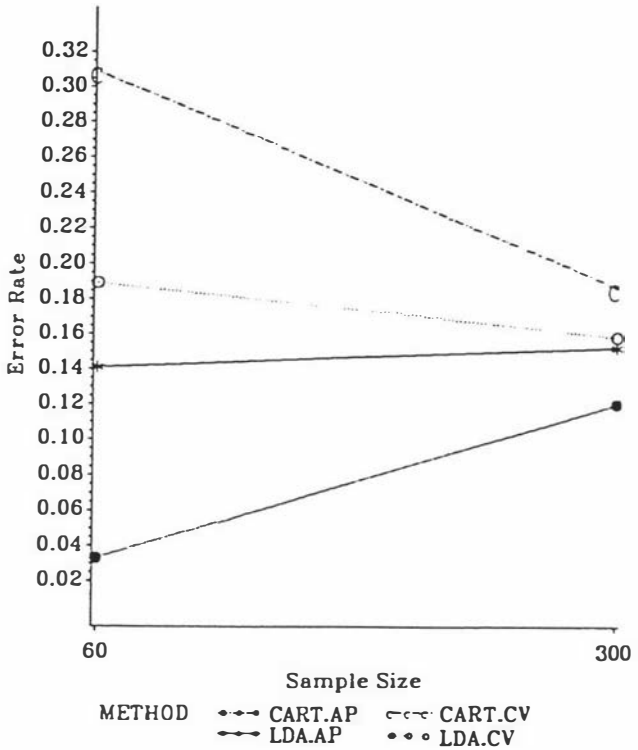


Figure 4.7b

Comparison of the Cross-Validation and Apparent Misclassification Error Rates using LDA and CART with Equal Priors and Different Dimension: (a) Normally Distributed Data

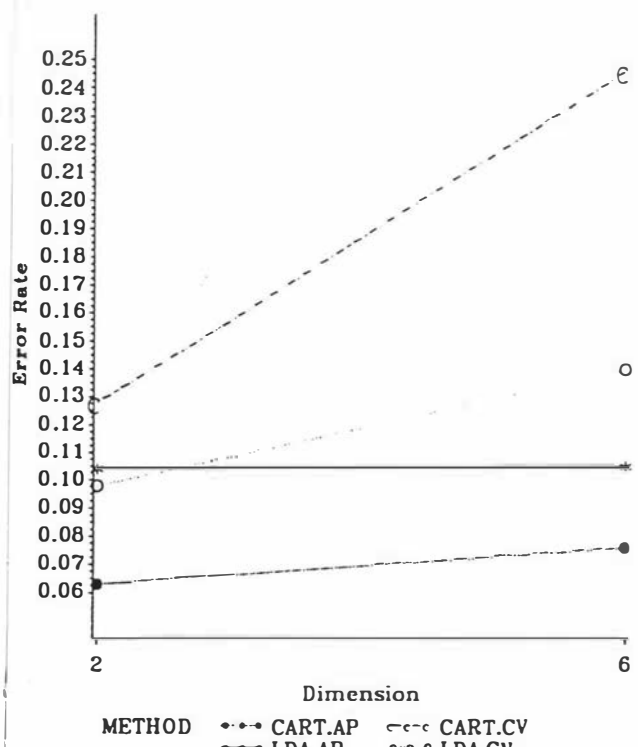


Figure 4.8a

(b) Lognormally Distributed Data

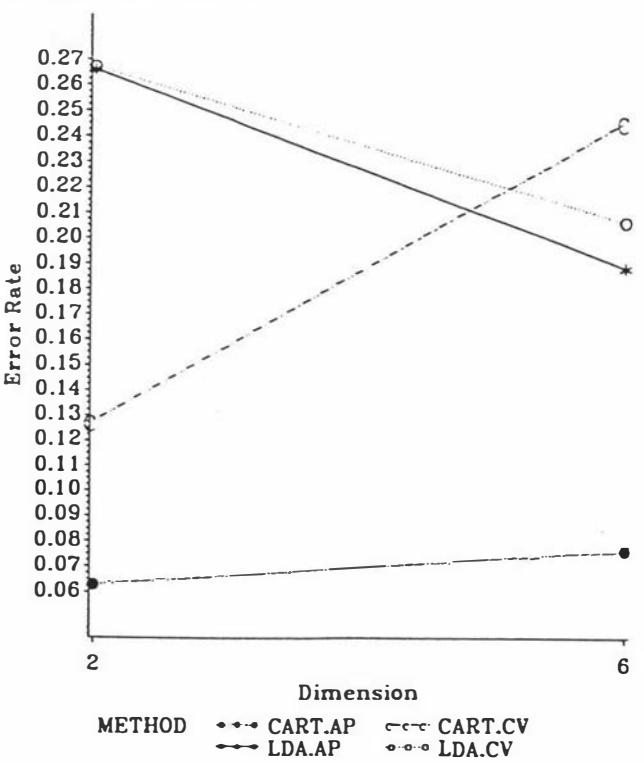


Figure 4.8b

A final analysis was done to compare the overall LDA and CART R(CV) estimates using both PPSS and equal priors. A split-plot ANOVA showed that the R main effect (F = 18.75) was not the most significant effect present. More important were the R * p (F = 39.37), R * f(.) (F = 82.69) and R * e (F = 22.04) interactions, as well as the R * f(.) * e second order interaction (F = 24.15). The R * p * f(.) interaction was also found to be highly significant (F = 6.49). Therefore, the differences between some of the methods/priors * factor interactions were more important than the differences between the methods/priors themselves.

Figure 4.9a shows there was a relatively small difference in the LDA error rates when p = 2 and no real difference when p = 6. For CART when p = 2, there was very little difference in the error rates but when p = 6, the use of equal priors resulted in an increase of approximately 0.09 from using PPSS. Figure 4.9b shows for lognormal data, that there was a larger difference in the LDA error rates, with the equal priors case resulting in the lower error rate. Note though, that even in circumstances unfavourable to LDA, this difference was smaller than that between the CART error rates.

From Figure 4.10a, it can be seen that using equal priors for LDA produced the lowest error rate except when $\pi_1 = 0.75$ and $\Sigma_2 = 3\Sigma_1$ though the differences were relatively minor. Figure 4.10b illustrates that for lognormal data, the same situation as above occurred though the differences were much larger. For CART, using PPSS produced the lowest error rate, no matter what the priors-covariance structure of the data.

4.5.5 Summary

In this section, the group misclassification error rates for LDA, QDA, CART and FACT were compared using both priors proportional to sample size and equal priors. Overall, the results showed that the individual error rates for QDA were least affected by changes in the priors and covariance structure of the data with LDA being the second least affected by the above mentioned alterations. CART was the least affected by using data that was lognormally rather than normally distributed, but was severely affected by changes in the priors-covariance structure of the data, misclassifying fewer observations from the class with the largest sample size, as suggested by Breiman et al (1984). This trend was even more apparent for FACT. Lognormal data severely affected both parametric discrimination methods, LDA more so than QDA which supports the results of Lachenbruch et al (1973). As noted previously, a different set of parameters was used in that study. It was stated in

Comparison of the Cross-Validation Misclassification Error Rates using LDA and CART with both Equal Priors and PPSS and Different Dimension: (a) Normally Distributed Data

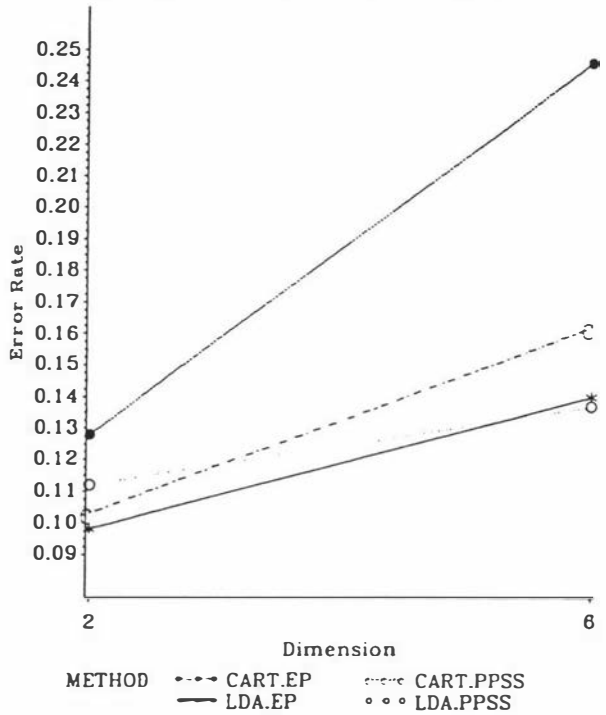


Figure 4.9a

(b) Lognormally Distributed Data

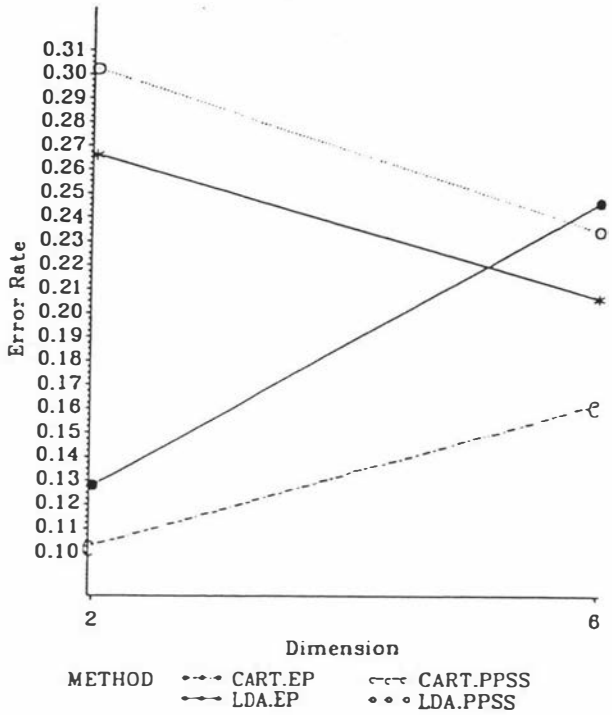
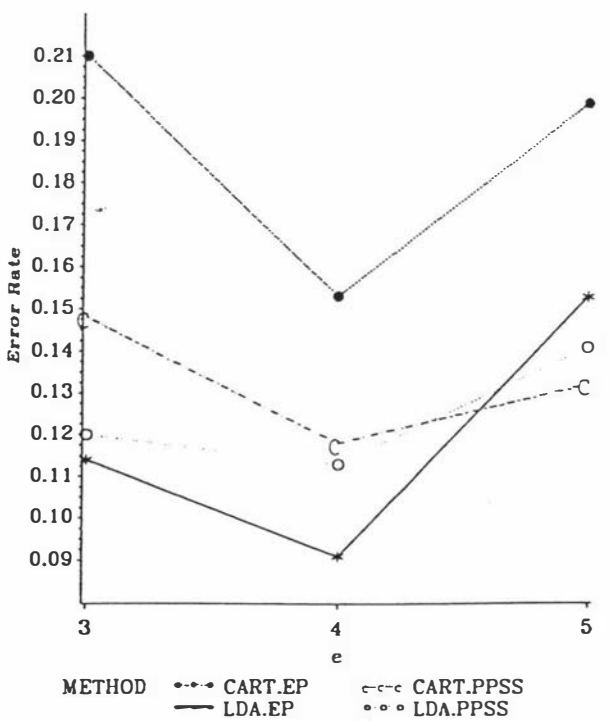


Figure 4.9b

Comparison of the Cross-Validation Misclassification Error Rates using LDA and CART with both Equal Priors and PPSS: (a) Normally Distributed Data



re 4.10a

(b) Lognormally Distributed Data

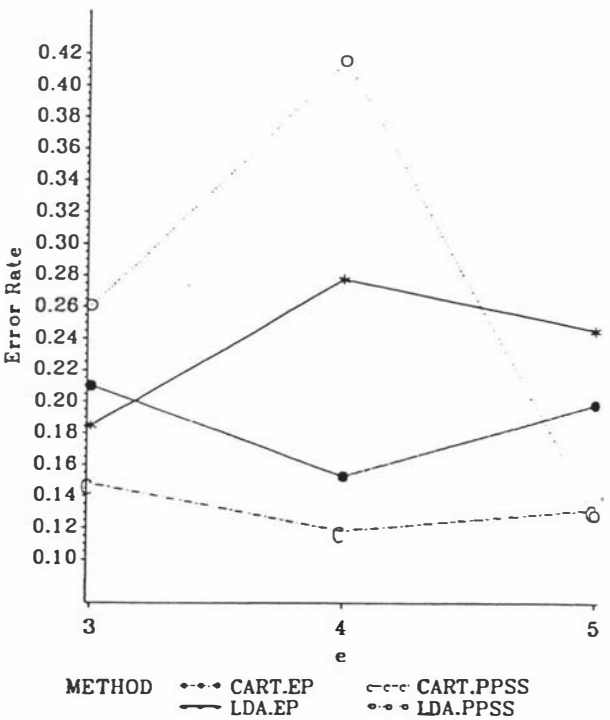


Figure 4.10b

Section 2.3 that if the ratio of class sample sizes to dimension is large, QDA works particularly well. In the majority of cases studied here, the above ratio was large so that QDA should be expected to work well.

Using FACT on the lognormal data sets had little effect on the group error rates when sample sizes were equal, but, when one class was larger than the other, all observations from the class with the smallest number of observations were misclassified, except when that class had unit covariance.

It was also found that there were large differences between the apparent and cross-validated group error rates for CART with the apparent error rates for each group being very similar, but the cross-validated error rates exhibited wide differences. With LDA, there were negligible differences between the apparent and cross-validated error rates. Using equal priors rather than priors proportional to sample size with CART decreased the difference between (or ratio of) group error rates but the differences were still larger than those using LDA on normally distributed data sets.

It could be recommended, based on these simulations, that CART would be the preferred method when the data is lognormal, if the criterion used to judge a method performance is the group misclassification error rates.

When comparing overall error rates using equal priors, it was found that in comparing the n-fold cross-validation and apparent error rates for LDA and CART, the differences were very dependent on the interaction of dimension and sample size as well as dimension and distribution of the data. LDA did best in all the above situations, except when there was only a low number of variables with lognormal data.

In comparing the error rates found from using priors proportional to sample size and equal priors for LDA, the differences between the two error rates were minimal for normal data, though rather large for lognormal data. The choice of priors affected overall error rates more for CART except when the data was lognormal.

The results from this section, even in the simple case of two populations, point to CART being very sensitive to the proportion of observations from each class in the sample. The results have shown that the objective of CART is to optimise the overall error rate at the expense of the respective group misclassification error rates. When $\pi_1 = \pi_2$, there is no such problem, but when one of the class sample sizes is small, then CART will tend to correctly classify as many observations as possible from the largest class at the cost of misclassifying many or most observations from the smallest class.

It could thus be recommended based on these simulation results, that LDA (or QDA) is preferred in the case of disparate sample sizes. If CART is used, some caution should be shown when interpreting the results.

4.6 SIMULATION STUDY III

4.6.1 Introduction

Sections 4.3, 4.4 and 4.5 compared the accuracy of the four classification methods in the setting of multivariate continuous data. This section compares the reliability of n-fold cross-validation error rate estimators for each of LDA, QDA and CART in estimating the actual error rate. The predictive ability of FACT was shown to be particularly poor, especially in non-ideal situations, hence further consideration of this method will not be done.

Numerous papers including Efron (1983), Hand (1986) and McLachlan (1986, 1987) have noted the n-fold cross-validation estimator of error rate has a large variance, especially when n is small, when used with LDA. These results were echoed by Crawford (1989) when used with CART. Hence, the reliability of the n-fold cross-validation error rate estimators is compared to a number of other error rate estimators that were introduced in Section 4.2, both within and across the three methods.

4.6.2 Study Plan

The same probability models that were used in Section 4.4 will be used here to compare and assess the reliability of each error rate estimator, across and within the three methods, in approximating the actual error rate, R(T). The error rate estimators that were considered for LDA and QDA are the n-fold cross-validation estimator, R(CV), the apparent estimator, R(A), the rotation estimator, R(ROT), and the 0.632 estimator, R(0.632), using

$$R(0.632) = 0.368 * R(A) + 0.632 * R(ROT)$$

as this has been shown to be very similar to using $R(\varepsilon)$, instead of R(ROT), the average error rate for all observations not included in the bootstrap sample.

For CART, the R(CV) and R(ROT) estimators were used as for LDA and QDA, while a ten-fold cross-validation error rate estimator, R(TEN), was also calculated. Since the tree chosen for each of the probability models by the above three error estimation techniques may be of different size and hence have different apparent error rates, three different apparent error rates were calculated. These were associated with each of the above three error estimation techniques and given by, R(ACV), R(AR) and R(AT) corresponding to n-fold cross-validation, rotation and ten-fold cross-validation respectively. In addition, the R(0.632) estimator was calculated in the same manner as for LDA and QDA with tree size chosen by R(ROT).

For all methods, a test sample of size 5000 was used to give an accurate value of R(T) and in the case of CART, to choose the right-sized tree. The performance of each method was determined by the mean square error criterion

$$MSE = E[R(\hat{T}) - R(T)]^2$$

where $R(\hat{T})$ is the particular error rate estimator. The MSE criterion provides a measure of both bias and variability of an error rate estimator. As in Sections 4.3, 4.4 and 4.5, a split-plot ANOVA was used to analyse the results of the simulations.

4.6.3 Results

The first analysis looked at the error rate estimators for only LDA and QDA. The R main effect ($F = 8.71$) and six R * factor interactions were significant at the $\alpha = 1\%$ level (and indeed at the $\alpha = 0.1\%$ level). The largest among these were the R * f(.) ($F = 12.07$), R * e ($F = 7.55$) and R * f(.) * e ($F = 6.00$) interactions. Of the first order interactions, only the R * δ interaction was not significant showing that for both LDA and QDA, all error rate estimators produced fairly similar results, no matter what the distance between populations. Note, though, that when $\delta = 2$, $R(B) = 0.159$ and when $\delta = 3$, $R(B) = 0.067$, hence the probability models studied here were for fairly well separated populations. Ganeshanandam

and Krzanowski (1990) found the R * δ interaction to be highly significant although they used δ = 1.01 and δ = 2.53 (R(B) = 0.291 and R(B) = 0.103), thus based their results over a much wider range of Bayes error rates. The results given here, though, were comparable with those of Fitzmaurice et al (1991) for R(B) = 0.05 and 0.15.

When comparing the seven error rate estimators for CART, the R main effect (F = 50.76) and seven R * factor interactions were significant at the α = 1% level. Of the first order interactions, only the R * f(.) interaction was not significant, which was as expected, given that CART is robust to non-normality of the variables. The results showed though that the CART error rate estimators were very sensitive to the choice of the number of variables, sample size, Mahalanobis distance between populations and the priors-covariance structure of the data.

A final analysis was done to compare all the error rate estimators over the three methods. As expected, with so many error rate estimators over different methods, almost all the R* factor interactions were significant at the α = 1% level. A personal correspondence from David Hand suggested that this approach may be infeasible if there were different residual variances among the method-error rate estimator combinations. A plot of the residuals against each of the method - error rate estimator combinations showed that the assumption of equal residual variances was not really valid. As suggested by Hand, a weighting of the MSE's for each method error rate estimator combination was carried out, using $1/s_i^2$ as the weights, where s_i^2 is the variance of the MSE's for the ith method-error rate estimator combination. The results, however, showed a number of differences from the unweighted analysis, in that a few less of the R* factor interactions were significant, but in the main, the important effects identified in the unweighted analysis showed up in the weighted analysis. However, the magnitude of the F-ratios can still be used to indicate which were the most important effects influencing the performance of the various error rate estimators across the three methods.

Tables 4.17 to 4.19 show the mean values for the most important second order interactions, those being R * n * δ (F = 5.79) and R * f(.) * e (F = 4.37). Table 4.17 shows that most estimators were more reliable (that is, had lower mean square error) when δ = 3 than when δ = 2 for smaller sample size. A notable exception to the rule was R(0.632) for CART which confirms what was shown in Crawford (1989). R(0.632) for LDA had lower mean square error for δ = 3 than for δ = 2, which confirms a trend shown in Fitzmaurice et al (1991).

When n = 300, some of the CART error rate estimators were more reliable when $\delta = 2$ than when $\delta = 3$. Of the R(CV) estimators, QDA and CART were the most reliable while those for LDA were rather unreliable, due most probably to the large variability of the estimator. Overall, R(0.632) for CART did best when $\delta = 2$, no matter what the sample size. When $\delta = 3$, R(CV) for CART did best when n = 60 and R(TEN) for CART did best when n = 300.

Table 4.17: Average mean square errors (MSE's) for different error rate estimators using LDA, QDA and CART with respect to the sample size-distance interaction (n * δ) (* 10^{-5})

		n = 60		n = 300	
Method	**Error Rate Estimator**	**δ = 2**	**δ = 3**	**δ = 2**	**δ = 3**
LDA	R(CV)	708	344	504	424
	R(A)	564	271	492	417
	R(ROT)	427	525	484	366
	R(0.632)	400	375	483	381
QDA	R(CV)	189	111	84	39
	R(A)	278	86	70	33
	R(ROT)	773	533	57	34
	R(0.632)	366	174	49	31
CART	R(CV)	409	**51**	57	101
	R(ACV)	2869	763	36	293
	R(ROT)	185	234	58	56
	R(AR)	1238	687	221	224
	R(0.632)	**106**	118	**23**	74
	R(TEN)	483	477	66	**24**
	R(AT)	1532	850	262	184

Table 4.18 shows that the error rate estimators using LDA and QDA were generally more reliable than those for CART, for normally distributed data. Of the R(CV) estimates, LDA and QDA were the most reliable except when $\pi_1 = 0.75$ and $\Sigma_2 = 3\Sigma_1$ (e = 5) where CART did best and LDA especially fell down. The R(0.632) estimator for CART performed uniformly well over all five levels of factor e while the other estimators for CART were more

variable. The R(ROT) estimate for CART did best overall in the equal covariance, unequal priors case (e = 3).

Table 4.18: Average mean square errors (MSE's) for different error rate estimators using LDA, QDA and CART with respect to priors-covariance structure (e) and normal data (* 10^{-5}).

Method	Error Rate Estimator	e = 1	e = 2	e = 3	e = 4	e = 5
LDA	R(CV)	60	49	109	152	440
	R(A)	108	**20**	64	112	268
	R(ROT)	**18**	97	331	140	374
	R(0.632)	29	52	179	91	315
QDA	R(CV)	87	55	173	**39**	194
	R(A)	411	133	83	87	**39**
	R(ROT)	232	294	704	197	940
	R(0.632)	237	104	237	56	372
CART	R(CV)	121	182	192	204	74
	R(ACV)	1422	1203	970	1064	699
	R(ROT)	114	175	**56**	177	145
	R(AR)	1027	421	482	605	427
	R(0.632)	102	47	79	101	72
	R(TEN)	434	297	187	268	126
	R(AT)	1247	651	409	710	519

Legend:

e = 1: $\pi_1 = 0.5$: $\Sigma_1 = \Sigma_2 = \mathbf{I}$

e = 2: $\pi_1 = 0.5$: $\Sigma_1 = \mathbf{I}$, $\Sigma_2 = 3\Sigma_1$

e = 3: $\pi_1 = 0.25$: $\Sigma_1 = \Sigma_2 = \mathbf{I}$

e = 4: $\pi_1 = 0.25$: $\Sigma_1 = \mathbf{I}$, $\Sigma_2 = 3\Sigma_1$

e = 5: $\pi_1 = 0.75$: $\Sigma_1 = \mathbf{I}$, $\Sigma_2 = 3\Sigma_1$

Table 4.19 shows the error rate estimators for the lognormally distributed data. The general trend observed was that the estimators for LDA deteriorated markedly in the case of unequal variances, but in the equal covariance case, the estimates closely approximated the actual

error rate. The estimators for QDA remained relatively constant for all levels of e and not too dissimilar from the normal data situation. Naturally, the CART estimators were exactly the same as in the normal data situation. Of the R(CV) estimators, LDA did best in the case of both equal priors and covariances, CART for unequal priors but equal covariances and QDA elsewhere. The R(0.632) estimate for CART had consistently low mean square error for all levels of e.

Table 4.19: Average mean square errors (MSE's) for different error rate estimators using LDA, QDA and CART with respect to priors-covariance structure (e) and lognormal data.

Method	Error Rate Estimator	e = 1	e = 2	e = 3	e = 4	e = 5
LDA	R(CV)	**16**	89	652	3231	150
	R(A)	50	100	499	2959	182
	R(ROT)	32	125	1178	1971	240
	R(0.632)	**16**	70	873	2264	208
QDA	R(CV)	43	84	267	62	**52**
	R(A)	62	75	210	**20**	56
	R(ROT)	191	269	178	184	305
	R(0.632)	71	89	160	67	157
		121	182	192	204	74
CART	R(CV)					
	R(ACV)	1422	1203	970	1064	699
	R(ROT)	114	175	**56**	177	145
	R(AR)	1027	421	482	605	427
	R(0.632)	102	**47**	79	101	72
	R(TEN)	434	297	187	268	126
	R(AT)	1247	651	409	710	519

Legend:

e = 1: $\pi_1 = 0.5$: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$

e = 2: $\pi_1 = 0.5$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$

e = 3: $\pi_1 = 0.25$: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$

e = 4: $\pi_1 = 0.25$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$

e = 5: $\pi_1 = 0.75$: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = 3\mathbf{\Sigma}_1$

4.6.4 Summary

The results of this study have found that the differences between the error rate estimators for CART were most affected by dimension, sample size, distance between populations and the priors-covariance structure of the data. The differences in error rate estimators for LDA and QDA were affected most by all the above factors, except distance between populations, as well as being affected by the distribution of the data. Considering the error rate estimators over all methods, it was found that estimation of the actual error rate depended very much on the sample size - distance interaction as well as the distribution - priors - covariance structure interaction.

Overall, the QDA and CART error rate estimates most closely approximated the actual error rate, excluding the apparent error rates for CART, which were unreliable in most situations.

Of the n-fold cross-validation estimates, those for QDA and CART were the best, usually having the lowest mean square error. The n-fold cross-validation estimates for LDA, away from the ideal situations, were found to be rather poor.

The 0.632 error rate estimate for CART was found to be very reliable throughout and not influenced to a great extent by any of the factors. This was confirmed by a separate ANOVA explaining the effects on the 0.632 estimator alone. It was noted that this estimator was particularly good for less well separated populations. On the other hand, the 0.632 estimate for LDA (and QDA) was sensitive to changes in the priors-covariance structure of the data and lognormal data.

4.7 CONCLUSIONS

In this chapter, a comparative study was undertaken for four classification methods, namely LDA, QDA, CART and FACT on the basis of predictive accuracy. The four methods were compared over different dimensions, sample sizes, distances between populations, distributions and priors-covariance structures. The results showed that LDA and QDA performed best for normal data, higher dimension and well separated populations, while CART performed well for lognormal data, lower dimension, less well separated populations and equal covariances with unequal priors. QDA was found to be the preferred method in the case of unequal covariances. In most situations, FACT's prediction rules were a poor fourth.

In Section 4.5, a study was undertaken, based on the findings from real data sets, to determine the effect on the individual group error rates for each of the four classification methods. These studies found that CART was affected quite drastically by the ratio of class sample sizes used, though not to the same extent as FACT. The individual error rates for LDA and especially QDA were least affected by unequal class sample sizes. It was recommended from that study that caution should be shown when using CART on data sets with grossly unequal class sample sizes.

In the final section of this chapter, an investigation was carried out into the reliability of each n-fold cross-validation error rate estimate for LDA, QDA and CART over the different probability models studied. The results showed that the n-fold cross-validation estimates were fairly reliable for QDA and CART in most situations, but that for LDA, were unreliable in situations where there were unequal covariance matrices. A deal of promise was shown by the 0.632 estimator for CART, in that it performed uniformly well over all situations studied. The reliability of this and other error rate estimates will be investigated in more detail in Chapter 6.

It could therefore be concluded that LDA and QDA would be preferred over CART in many situations, and CART as the preferred method in others, if accuracy were the sole measure of the performance of a method.

5. SIMULATION STUDIES INVOLVING CATEGORICAL DATA

5.1 INTRODUCTION

Often in multivariate data settings, problems do not involve continuous variables. Rather, the problem may involve ordered categorical variables such as the ratings of a certain product (bad, average, good) or number of years education. These variables can be treated as continuous although the requirement of multivariate ellipsoidality may not always be met. In other situations, the problem may involve unordered categorical or nominal variables, whereby there is no natural ordering of the categories. Race and marital status are two examples of nominal variables.

The general approach taken by traditional discrimination methods for such variables is to code the c categories into (c-1) binary variables where

$$x_i = \begin{cases} 1 & \text{if } c_i \text{ present} \\ 0 & \text{else} \end{cases}$$

As seen earlier, CART and many other tree-based procedures do not require the use of binary variables to handle nominal variables. Instead, most tree-based procedures attempt to find the grouping of categories that leads to the least overall misclassification error. FACT, in contrast, is one tree-based method which takes the LDA approach to handling categorical variables.

This chapter focuses on a comparison of LDA, QDA, CART and FACT for the above type of data. Firstly, the four methods are compared from an accuracy point of view in classifying observations into two distinct populations, Π_1 and Π_2. For the sake of a direct comparison to be made between CART and the other three methods, only p-variate binary data is used. As in Chapter 4, the reliability of these classification methods is investigated as well as different error rate estimates for LDA and QDA.

5.2 PREVIOUS STUDIES

Moore (1973) carried out some simulation studies using six-dimensional, bimodal, binary data sets comparing LDA and QDA with various multinomial procedures. Two factors were varied, those being

p_{ij} = probability of getting a response $x_j = 1$ for Π_i

and r_{ijk} = correlation between x_j and x_k for Π_i.

Moore's results showed that LDA performs very well except when there is a "reversal" in the log-likelihoods for each population. Moore illustrates by using the following example. Let x_1 and x_2 be given as

$$x_1 = \begin{cases} 1 & \text{if birthweight is high} \\ 0 & \text{if birthweight is low} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if gestation length is long} \\ 0 & \text{if gestation length is short} \end{cases}$$

A baby would be classified as normal when $x_1 = 0$ and $x_2 = 0$ or $x_1 = 1$ and $x_2 = 1$, otherwise it is abnormal.

The optimal linear rule would be to assign $\mathbf{x}$ to Π_1 if

$$a(\mathbf{x}) = \beta_0 + \sum_{j=1}^{2} \beta_j x_j > c.$$

Now $a(1, 1) = a(0, 1) + a(1, 0) - a(0, 0)$.

As $a(1, 1) > \max\{a(1, 0),\ a(0, 1)\}$

$\Rightarrow$ $a(0, 0) < \min\{a(1, 0),\ a(0, 1)\}$

then, if (1, 1) is assigned to Π_1, (0, 0) has to be assigned to Π_2. This leads to gross errors in misclassification. The problem is in using a monotonically increasing function of x_1 and x_2 to approximate the log-likelihood, $L(\mathbf{x})$, which is not monotonic. In the two-dimensional

case, the problem can be solved quite simply by including an interaction term in the model. When there are a large number of variables, however, this approach is infeasible.

Others have shown that not only different correlation structures in the two populations lead to these reversals, but so will moderate and large positive correlations. Krzanowski (1977) considered a mixture of both binary and continuous random variables with various values of p_{ij} used and r_{ijk} set to either 0 or 0.375 (no, or moderate, positive correlation). The results showed that LDA performed well when there was no correlation between the binary variables, but performed poorly if there was a moderate positive correlation among all the binary variables or the correlations between the binary and continuous variables differ markedly between the two groups. For both types of data, QDA has been found rarely to perform as well as LDA.

Ganeshanandam and Krzanowski (1990) compared a number of different error rate estimators for LDA as well as the n-fold cross-validation estimate for QDA, for multivariate binary data. Their simulation results showed that the p_{ij}'s and sample size all had significant effects on the estimation of the actual error rate though the r_{ijk} factor did not.

.3 SIMULATION STUDY I

5.3.1 Study Plan

The factors used in this study were the same as those employed in Ganeshanandam and Krzanowski (1990) in order to be able to check the results for LDA and QDA against theirs.. The factor p had settings of 5 and 10. A separate analysis was done for each of the two dimension levels. Factor n had three settings, those being "small, medium and large" relative to the number of variables. In the case of p = 5, n = 20, 60 and 100, while for p = 10, n = 40, 120 and 200. In all cases, $\pi_1 = \pi_2 = 0.5$ implying that class sample sizes were equal. When p = 5, factor r_{ijk} had two levels, those being, all $r_{ijk} = 0$ and all $r_{ijk} = 0.25$. When p = 10, all r_{ijk} were set to zero for the first five variables and to 0.25 for the second block of five. The last factor was the values of the p_{ij}'s. The levels of the p_{ij}'s are shown in Table 5.1 with level 1 corresponding to wide separation between groups with increasing levels leading to narrower separation. Level 5 corresponds to identically distributed populations. For p = 10, the p_{ij}'s for the first block of five variables were repeated for the second block of five. This gives 45 multinomial learning samples for p = 5 and 10 combined. Three replicates for each multinomial situation were conducted for p = 5 and six replicates for p = 10, giving 180 data

sets in total (90 for each dimension size). Generation of the data was a straightforward process using MINITAB macros.

Table 5.1: Values of p_{ij} for each set of five binary variables.

Level	Π_1					Π_2				
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
1	0.20	0.20	0.20	0.20	0.20	0.80	0.80	0.80	0.80	0.80
2	0.25	0.30	0.35	0.40	0.45	0.75	0.70	0.65	0.60	0.55
3	0.40	0.45	0.50	0.55	0.60	0.60	0.55	0.50	0.45	0.40
4	0.25	0.30	0.35	0.40	0.45	0.45	0.40	0.35	0.30	0.25
5	0.30	0.40	0.50	0.60	0.70	0.30	0.40	0.50	0.60	0.70

The methods were compared by means of the n-fold cross-validation error rates, R(CV), except for FACT where ten fold cross-validation was used as outlined in Section 4.3. For both CART and FACT, the minimum size below which a node will not be split was set to five, while for CART alone, the one standard error rule was used. A split-plot ANOVA was used to identify which factors lead to differences between the methods, as in Chapter 4. Tables of means and the standard deviations of the differences between means are presented for each significant effect.

5.3.2 Results

For the case p = 5, the results showed that the R(method) main effect (F = 10.19) was highly significant and dominated the variation in error. All interactions involving R had F-values less than 1.1. This meant that there were differences between the methods when using p = 5 binary variables, and these differences were not influenced by other factors.

In both the analyses for p = 5 and p = 10, the plot of residuals against fitted values showed no real trends, in contrast with the results for the continuous data. The plots of residuals against each method showed there to be roughly equal residual variances for each method. Individual ANOVA's were constructed for each method separately. These results confirmed the above finding of equal residual variances. Therefore, the results of the split-plot ANOVA are strictly valid.

The average values for each method, when p = 5, are given in Table 5.2, as well as the standard error of difference between the methods. The results showed that CART was the best method by some distance from LDA, FACT and QDA.

Table 5.2: Means and standard error of the differences in means of the cross-validation error rate estimates for each method.

	p = 5			
Level	**LDA**	**QDA**	**CART**	**FACT**
	0.343	0.368	**.304**	.358

Standard error of the difference between means = 0.013.

In the case of p = 10, the split-plot ANOVA showed that the R(method) main effect (F = 20.78) was highly significant, though the R * p_{ij} (method by probability pattern) interaction (F = 4.2) and R * n (method by sample size) interaction (F = 4.91) were also highly significant ($\alpha < 0.001$). This showed that for p = 10, there were not only differences between the methods but that these differences were affected by both sample sizes, and to a slightly lesser extent, the pattern of probabilities in the parent populations.

Table 5.3 shows that increasing sample size had the effect of reducing the error rates for all methods, except CART, where sample size had no real effect. The greatest reduction in error rate occurred when going from n = 40 to n = 120, for LDA, QDA and FACT. CART did best notably when n = 40 and slightly better than LDA when n = 120. When n = 200, however, LDA did better than CART. FACT was a poor fourth except when n = 40, where, because of the small ratio of dimension to class sample sizes, QDA did worst.

Table 5.3: Means and standard error of the differences in means of the cross-validation error rate estimates for each classification method with respect to sample size (n).

	p = 10			
Level	**LDA**	**QDA**	**CART**	**FACT**
n = 40	0.329	0.401	**0.284**	0.366
n = 120	0.289	0.307	**0.287**	0.345
n = 200	**0.274**	0.292	0.284	0.333

Standard error of the difference in means = 0.016.

Increasing the level of the probability patterns, p_{ij}, effectively narrowing the distance between populations, had the effect of increasing the error rates for all methods, with the two tree-based methods being less affected than both LDA and QDA. Table 5.4 shows that LDA did best for $p_{ij} = 1$ and narrowly better than CART when $p_{ij} = 2$. For all other levels of p_{ij}, CART had the lowest average error rate. An explanation why LDA did best for $p_{ij} = 1$ is that this is an example of a parallel classification problem (see Section 4.3), in that all variables are equally important in determining the classification rules. Levels 2 to 4 for p_{ij} are examples of a sequential classification problem, in that only a subset of the variables are ever used to determine the classification rules. Methods such as LDA and QDA are suited to the former type of problems while CART is designed for the latter. The good performance of CART for less well separated populations mirrors what was observed for continuous data in Chapter 4. A noteworthy trend observed here was that the average error for CART in the case of identical populations ($p_{ij} = 5$) was 0.43, so that CART was managing to build a tree from noise. All other methods for $p_{ij} = 5$ contained error rates in the vicinity of 0.5.

Table 5.4: Means and standard error of the differences in means of the cross-validation error rate estimates for each classification method with respect to probability patterns (p_{ij}).

	p = 10			
Level	**LDA**	**QDA**	**CART**	**FACT**
$p_{ij} = 1$	**0.073**	0.097	0.132	0.175
$p_{ij} = 2$	**0.177**	0.230	0.189	0.249
$p_{ij} = 3$	0.346	0.389	**0.325**	0.410
$p_{ij} = 4$	0.365	0.426	**0.349**	0.415
$p_{ij} = 5$	0.524	0.525	**0.430**	0.492

Standard error of the difference in means = 0.021.

5.3.3 Summary

For five-dimensional categorical data, it was found that there were differences in the cross-validated error rates of the four methods, but these differences were due only to the methods and not to any other factor such as sample size or probability patterns. It was found that CART was clearly the best method followed by LDA.

For ten-dimensional categorical data, the differences in error rates were due not only to method, but also to probability patterns and sample size. CART was found to be the least affected of all methods by changes in either sample size or the pattern of probabilities and had the lowest error rate for smaller sized samples and poorly separated populations. These results echo very much what was observed for CART in Chapter 4, for continuous data. In all other situations, LDA did best. In accordance with other studies, QDA performed poorly except when sample size was large or the two populations were well separated. As for the continuous data, FACT was a poor fourth in almost all situations.

5.4 SIMULATION STUDY II

5.4.1 Introduction

Section 5.3 compared the accuracy of four classification methods in the setting of categorical data, specifically with p-variate binary data. This section compares the reliability of the cross-validation error rate estimates produced by three of the four methods as well as a comparison of different error rate estimators using both LDA and QDA. Further analyses for FACT were not done due to the poor predictive capability exhibited by the method through the simulation studies.

5.4.2 Study Plan

The first analysis compared the n-fold cross-validation, R(CV), and apparent, R(A), estimators for both LDA and QDA, as well as the rotation, R(ROT), and 0.632, R(0.632), estimators for LDA alone. The latter was calculated as in (4.2.23). For both $p = 5$ and $p = 10$, when n was small, the R(ROT) estimates could not be calculated for QDA as $n_i \leq p$ for each class. Therefore, the R(ROT) and R(0.632) estimators for QDA were not included in the analysis.

The R(CV) estimators for LDA, QDA and CART were then compared to test their reliability in estimating the actual error rate, R(T), for each data set used in Section 5.3. Test samples of size 5000 were used throughout to calculate the values of R(T). A comparison of different error rate estimators for CART, using categorical data, will be undertaken in a latter chapter.

5.4.3 Results

The results are presented in the same format as Section 5.3. Comparing the error rate estimators for LDA and QDA first, in the case of $p = 5$, the split-plot ANOVA showed that there was a definite difference between the error rate estimators, R, ($F = 34.76$) and that those differences depended to a large extent on sample size, R * n ($F = 7.64$) and, to a much lesser degree, on the pattern of probabilities in the Π_i's, R * p_{ij} ($F = 2.37$). These results very closely follow the results exhibited in Ganeshanandam and Krzanowski (1990) where the R * n effect was also more important than the R * p_{ij} effect. Any effect involving r_{ijk} had very little effect on the estimation of R(T).

For p = 10, it was observed that, in addition to the R, R * n and R * p_{ij} effects being highly significant, the R * p_{ij} * n interaction (F = 3.96) was also significant at the 0.1% level. Ganeshanandam and Krzanowski gave no indication of the importance of this second order interaction, but, it must be taken into account in any analysis of means implying that the error rate estimators for both LDA and QDA were influenced by sample size in conjunction with the pattern or probabilities.

It is clear from Table 5.5 that the mean square errors (MSE's) for all estimators decreased as sample size increased, that is, the estimators were more precise for larger rather than smaller sized samples. The R(CV) estimator was the least affected by altering sample size while the R(A) estimators were most affected. The R(CV) estimator for LDA did best for the smallest sized samples, the R(ROT) estimator for n = 60 and the R(CV) estimator for QDA for the largest samples. These results were in relatively close agreement to those of Ganeshanandam and Krzanowski (1990), although their results showed sample size to have no effect on the MSE for the R(0.632) estimator. Note that the standard error of the difference was not calculated from either of the apparent estimators. This occurred, because the variation in MSE's for these two estimators was four to five times larger than that for the other estimators.

Table 5.5 Means and standard error of the differences in means of the MSE's for each error rate estimator with respect to sample size (n) (* 10^{-4})

	p = 5		
Level	**n = 20**	**n = 60**	**n = 100**
LDA, R(CV)	**139**	56	46
LDA, R(A)	443	107	41
LDA, R(ROT)	181	**47**	35
LDA, R(0.632)	200	49	25
QDA, R(CV)	216	71	**23**
QDA, R(A)	685	252	155

Standard error of the difference between means = 37.2.

Table 5.6 shows that the error rate estimators exhibited different behaviour over the various levels of p_{ij}. The R(CV) estimator, for LDA, had lowest MSE for $p_{ij} = 1$ and highest for $p_{ij} = 3$, while the R(0.632) estimator had lowest MSE for $p_{ij} = 3$ and almost the highest at $p_{ij} = 1$. R(CV) did best overall for $p_{ij} = 1$, 4 and 5, the R(0.632) estimator for $p_{ij} = 2$ and the R(ROT) estimator for $p_{ij} = 3$. In the case of p = 5, these results fairly closely matched those of Ganeshanandam and Krzanowski (1990) where R(0.632) was found to perform best for less well separated populations.

Table 5.6: Means and standard error of the differences in means of the MSE's for each error rate estimator with respect to probability patterns (p_{ij}) (* 10^{-4})

	p = 5				
Level	$p_{ij} = 1$	$p_{ij} = 2$	$p_{ij} = 3$	$p_{ij} = 4$	$p_{ij} = 5$
LDA, R(CV)	**22**	131	85	**85**	**75**
LDA, R(A)	74	184	214	268	245
LDA, R(ROT)	86	96	**48**	120	89
LDA, R(0.632)	118	**74**	63	118	84
QDA, R(CV)	56	114	109	117	120
QDA, R(A)	172	219	481	485	462

Standard error of the difference between means = 48.

Figures 5.1 to 5.3 show the MSE's for the different error rate estimators over the different levels of p_{ij}, for sample sizes of 40, 120 and 200 respectively. The apparent estimators for both LDA and QDA are not shown due to their exceedingly high MSE's in most cases. The general trend for the other estimators was an increase in MSE as sample size decreased while decreasing the distance between populations generally increased the MSE, though there were some exceptions. The graphs show that the R(CV) estimator was the most consistent, and thus reliable over all combinations of probability pattern and sample size. The R(CV) estimator for QDA was sometimes the most reliable estimator, but in other situations was the least reliable, especially for larger sample sizes. The R(0.632) estimator did best for poorly separated populations and n = 120 or 200, though not for populations which were the same ($p_{ij} = 5$).

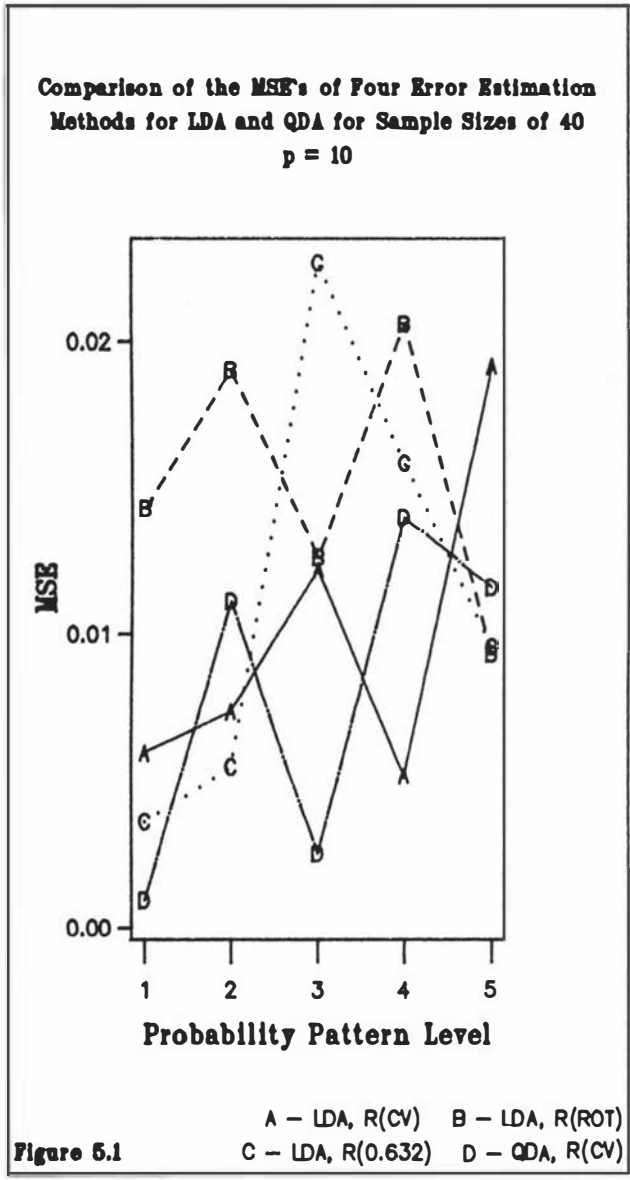


Figure 5.1

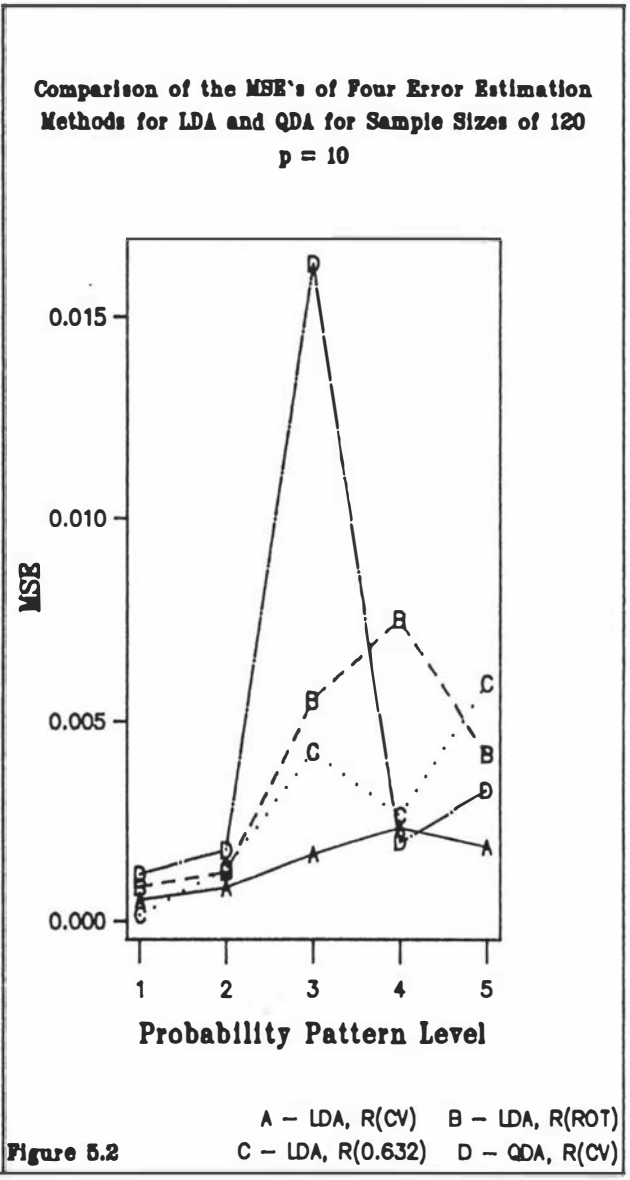


Figure 5.2

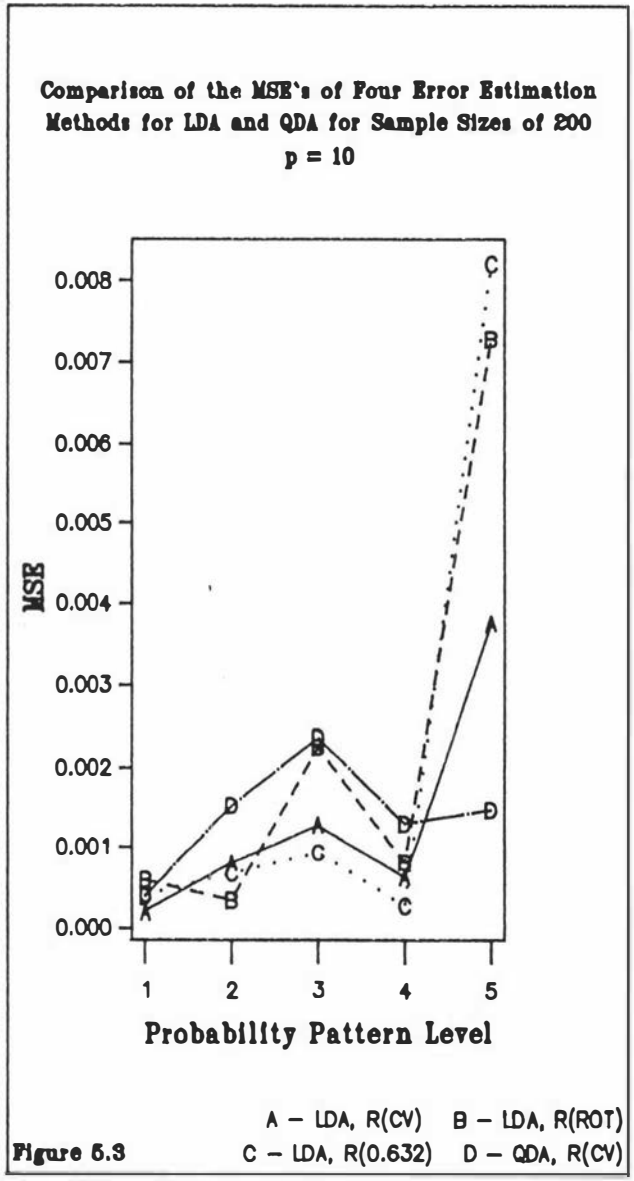


Figure 5.3

A comparison of the reliability of the n-fold cross-validation error rates for LDA, QDA and CART showed that when p = 5, the R main effect was not significant. That is, there were no differences in the reliability of the R(CV) estimators between the three methods.

For p = 10, the R main effect (F = 3.45) was the only significant result, and that only at the α = 5% level. Table 5.7 shows that both the LDA and QDA R(CV) estimators were roughly equally reliable while that for CART was comparatively high (roughly twice the magnitude of the LDA estimate). This discrepancy was due mainly to the overoptimistic estimates for R(T) when $p_{ij} = 5$ produced by cross-validation.

Table 5.7: Means and standard error of the difference in means of the MSE's for the n-fold cross-validation error rate estimates for each classification method (* 10^{-4})

	p = 10		
Level	**LDA**	**QDA**	**CART**
	43	48	82

Standard error of the difference between means = 17.

5.4.4 Summary

For five-dimensional categorical data, it was found when comparing the reliability of various error rate estimators for LDA and QDA that sample size and probability patterns for each population were important in determining differences between error rate estimators. The R(CV) estimator for LDA did best for either small samples or large differences between populations while the R(ROT) and R(0.632) estimators were better for larger samples or smaller differences between populations. For ten-dimensional categorical data, it was found that the interaction of sample size and probability patterns was important in differentiating the estimators, while R(CV) was the most reliable estimator overall.

In comparing the n-fold cross-validation estimators for LDA, QDA and CART, it was found that there were no differences present in the five-dimensional case. However in the ten-dimensional case, LDA and QDA had the most reliable estimates with CART some distance behind.

5.5 CONCLUSIONS

In this chapter, four classification methods were compared in the setting of p-variate categorical data. In the case of five-dimensional categorical data, the only significant effect was the overall difference between the methods, where CART was found to be the best method. For the ten-dimensional data, the results followed a very similar pattern to those for the continuous data in that LDA did best when every variable counted an equal amount for the classification rules or where there was fairly good separation between groups. CART, on the other hand, did better for less well separated groups or where only a few variables were important to the creation of the classification rules. As well, CART did better for smaller samples and LDA for larger samples, thus appears particularly useful for categorical data.

A comparison of different error rate estimators for LDA and QDA showed that the n-fold cross-validation estimator for LDA was the better estimator of the actual error rate. A comparison of the n-fold cross-vlidation error rate estimators for LDA, QDA and CART showed that the CART estimator was the least reliable of the three.

The last finding led to a comparison of the various error rate estimators for CART using categorical data, the results of which are reported in Chapter 6 along with a comparison of the various estimators for continuous data.

Another interesting finding was that there was only a minor difference between the third and fourth levels of the probability pattern variable (see Table 5.1). Ganeshanandam and Krzanowski (1990) reported that the error rates for the third level were twelve to seventeen points higher than those for the fourth level. The results here have shown that difference to be approximately only a few points. It is possible that Ganeshanandam and Krzanowski (1990) actually used different probability patterns than those stated, because theoretically, the error rates for the third level should be closer to those in level 4 rather than level 2 as occurred in their studies.

6. CART SIMULATION STUDY

6.1 INTRODUCTION

In Section 4.6, an investigation was carried out into the performance of various techniques for determining tree size and estimating the actual error rate in CART. Recent studies have suggested, however, that the range of Bayes error rates directly affects the performances of the error rate estimators, especially the 0.632 estimator for CART in the case of continuous data.

The objective of this chapter was firstly to compare the various error rate estimators over a wider range of Bayes error rates, and reduced sample sizes from those studied in Chapter 4. Numerous studies (Efron, 1983, Gong, 1986 and Crawford, 1989 for instance) have shown that sample size is a crucial factor in determining the performance of an error rate estimator.

A second objective was to compare the various error rate estimation techniques for CART over the categorical data sets that were used in Chapter 5, to determine if similar patterns as were observed for continuous data could be seen.

Thirdly, based on the comments of Feng et al (1993), a comparison between the zero and one standard error rules for selecting the right sized tree, was carried out, in order to determine in which situations, if any, one should/should not use the one standard error rule.

A final objective was to brought about by a study of Fitzmaurice et al (1991). They affirmed that the untransformed error rate scale, bounded by 0 and 1, may not be appropriate for the comparison of different methods. Thus, a number of transformations were carried out on the error rates and the effects of the transformations analysed.

6.2 ERROR RATE ESTIMATION FOR CONTINUOUS DATA IN CART

6.2.1 Previous Studies

This study was motivated by the work of Breiman et al (1984) and Crawford (1989). Breiman et al in Section 11.7, suggest that on the basis of tests on both real and simulated data sets, the bootstrap error rate estimate had lower variance than the cross-validated error rate estimate, but was highly overoptimistic when compared with those based on g-fold cross-

validation. When the learning sample is large, they state that the bias effect dominates the variance so that the g-fold cross-validation estimator is superior to bootstrapping.

Breiman et al (1984) suggested that perhaps a modified bootstrap estimator could be used to determine both optimal tree size and provide an estimate of the actual error rate. Crawford (1989) tested these assertions by comparing the performance of cross-validation, the bootstrap and the 0.632 bootstrap, using

$$R(0.632) = 0.368 * R(A) + 0.632 * R(\varepsilon).$$

He found for small data sets (n = 20) that the 0.632 estimator, R(0.632), was the best in terms of having the lowest mean square error (MSE), while for larger samples (n = 100), the cross-validation estimator, R(CV), was best for high values of R(B) but R(CV) for low values of R(B). Crawford suggested the use of a combined strategy whereby n-fold cross-validation was used to select the right sized tree and R(0.632) to estimate R(T) on the selected tree. Crawford concluded that this combined approach minimised the chance of poor performance when faced with either a high or low value of R(B).

In Section 11.5, Breiman et al, p 85, affirmed that "... we have not come across any situations where taking [g] larger than 10 gave a significant improvement in accuracy for the tree selected." They suggest that the use of ten fold cross-validation gives adequate accuracy in most situations, and indeed, this is the default value used within the CART program. As yet, no results have appeared in the literature validating these assertions.

6.2.2 Study Plan

The aim of this study was to use CART to compare the performance of the n-fold and ten-fold cross-validation, rotation and 0.632 estimators, along with the associated apparent error rates in approximating the actual error rate, R(T), of the sample. As in Section 4.6, R(T) was found by using an independent test sample of size 5000. The objective was to expand on the work of Breiman et al and Crawford (1989) in order to find out which method was the best in selecting the most "honest" sized tree. The error rate estimators were as used in Section 4.6. The 0.632 estimator, using R(ROT) in the equation instead of $R(\varepsilon)$ provides a simple alternative to the combined strategy proposed by Crawford (1989). His proposal involved a double calculation, hence a large increase in processing time, in that n-fold cross-validation

was needed to select the right-sized tree, then B bootstrap samples had to be generated in order to estimate R(T). This version uses the rotation method to calculate the right-sized tree and then uses that error rate in the equation for R(0.632).

In addition, a comparison of the sizes of the trees produced by each method was made to determine if there were any differences between methods.

The data were generated from two multivariate normal populations, as it is known that CART is invariant under monotone transformations of the variables. Three factors were varied in a full factorial design; R(B), the Bayes error rate; n, the sample size with $\pi_1 = \pi_2$ for all cases; and q, a combination of dimension (p), means and correlations between variables.

The values of the first and second factors were:

(i) R(B) = 0.05, 0.15, 0.25 and 0.35.

(ii) n = 20, 100.

The third factor, q, had levels whereby $\mu_{1j} = 0$ for all j, where μ_{1j} is the mean of the first population for variable j; and μ_{2j} is the mean of the second population for variable j.

1: p = 2, $\mu_{21} = \mu_{22}$, $\rho = 0$

2: p = 2, $\mu_{21} = \mu_{22}$, $\rho = 0.5$

3: p = 4 whereby $\mu_{21} = 2\mu_{22} = 6\mu_{23}$; $\mu_{24} = 0$

and $P_1 = \begin{bmatrix} 1 & 0 & -0.5 & 0 \\ 0 & 1 & 0 & 0 \\ -0.5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $P_2 = \begin{bmatrix} 1 & 0 & 0.75 & 0 \\ 0 & 1 & 0 & 0 \\ 0.75 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

where $P_i = [(\rho_{ijk})]$ is the population correlation matrix for Π_i.

The values for the first and second factors were similar to those used by Crawford (1989) and Fitzmaurice (1991), except that no studies were done for R(B) = 0.45, where, as noted by Fitzmaurice et al, any classification rule which produced an error rate in the region of 0.45 would probably not be widely used.

The levels of the third factor were chosen after some conclusions by Quinlan (1993) about parallel and sequential classification problems and summarised in Chapter 4. Therefore, in this study, the first two levels of factor q correspond to situations which are less favourable to CART while the third level corresponds to a situation more favourable to CART.

Each of the 24 factor combinations was used for 25 simulations. The number of simulations was chosen so as to be able to adequately depict the true trends. The effects of a few 'bad' samples will be minimised by the large number of 'good' samples. Four criteria of performance were used to compare the various error rate estimators, namely, the bias of each technique in estimating R(T), where

$$\text{bias} = R(\hat{T}) - R(T), \quad \hat{T} = CV, A, ROT, 0.632 \text{ or } TEN$$

the standard deviation of the bias; thirdly, the MSE, where

$$MSE = E[(R(\hat{T}) - R(T))^2].$$

A fourth measure used was the COUNT criterion, corresponding to the proportion of samples for each factor combination in which the estimated error rate was less than the actual error rate, and is therefore a measure of the optimism involved in using each estimator. A large value for COUNT, say > 75%, would correspond to an overoptimistic estimation whereas a low value for COUNT, say < 25%, would correspond to a pessimistic or under optimistic estimation.

For all the data sets in this section, the zero standard error rule was employed while the size below which a node will not be split was set to five. Independent test samples of size 5000 were used throughout to determine R(T).

6.2.3 Results

As in Chapters 4 and 5, a split-plot ANOVA was used to assess the relative importance of the experimental factors in influencing the MSE's for the various error estimation techniques. A large number of replicates were carried out in this study, in contrast to Chapters 4 and 5, hence the F-ratio from the ANOVA should not be regarded as a true measure of the significance of each result.

The R (method) main effect (F = 87.68) and R * R(B) (method by Bayes error) interaction (F = 20.18) dominated the other R * factor interactions. All other effects were very small though, rather surprisingly, the R * R(B) * n * q interaction was the next largest (F = 2.57). Therefore, it was decided to compare the four error estimation techniques over all possible combinations of R(B), n and q.

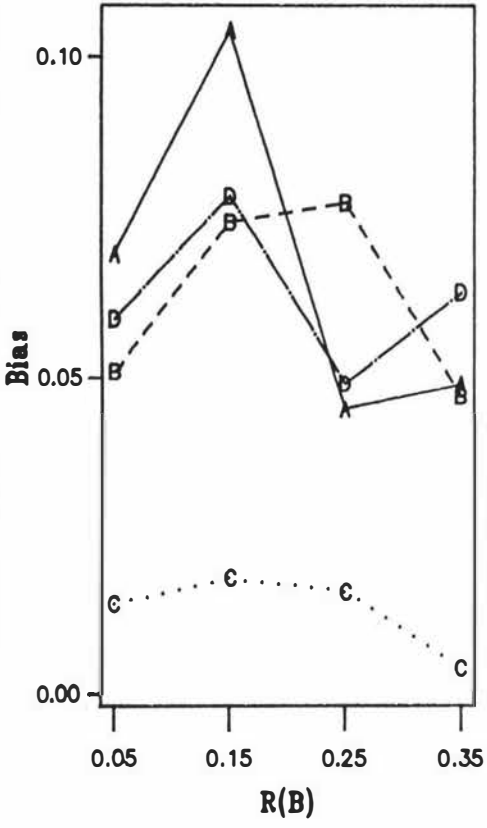
The results of the average bias, standard deviation, MSE's and COUNT's are presented graphically in Figures 6.1 to 6.24. Only the R(CV), R(ROT), R(0.632) and R(TEN) estimators are shown as the values for the respective apparent estimators were on most occasions highly overoptimistic, leading to extreme values of the above four measures.

Figures 6.1 to 6.6 show the average bias values for each estimator over the ranges of the factors used. The results show that for almost all values of R(B), n = 20 and q = 1 and 2, that R(0.632) had the lowest bias. All other methods were markedly overly pessimistic in the estimation of R(T). The exception to the rule was when q = 2 and R(B) = 0.15. When q = 3, however, a different picture emerged. The R(CV), R(0.632) and R(TEN) estimators all had similar bias with this bias increasing pessimistically as R(B) increased. The R(ROT) estimator was consistently more pessimistic than the other three estimators.

For larger samples (n = 100), it is noticeable that the performance of the R(0.632) estimator deteriorated as R(B) increased in that the bias became overly optimistic. The R(CV) estimator was usually the least biased for R(B) = 0.05 and 0.15, but deteriorated for higher R(B). In those situations, R(TEN) produced the lowest errors. For highly separated populations (R(B) = 0.05), the R(0.632) estimate was comparable to or better than the R(CV) estimate. As with the smaller samples, the R(ROT) estimator was consistently pessimistic.

Turning to the standard deviations of the estimates, it can be seen that for smaller samples (Figures 6.7 to 6.9), for q = 1 and 3, that all estimators exhibit a distinctive inverted U shaped pattern in that the lowest standard deviations occurred for either the lowest or highest R(B). For q = 2, the trends were different, for all estimators. In terms of performance, the R(0.632) estimator had lowest variability when R(B) = 0.05 and 0.15, though highest variability when R(B) = 0.25 and 0.35. The large variability of the R(CV) and R(TEN) estimators is clearly evident.

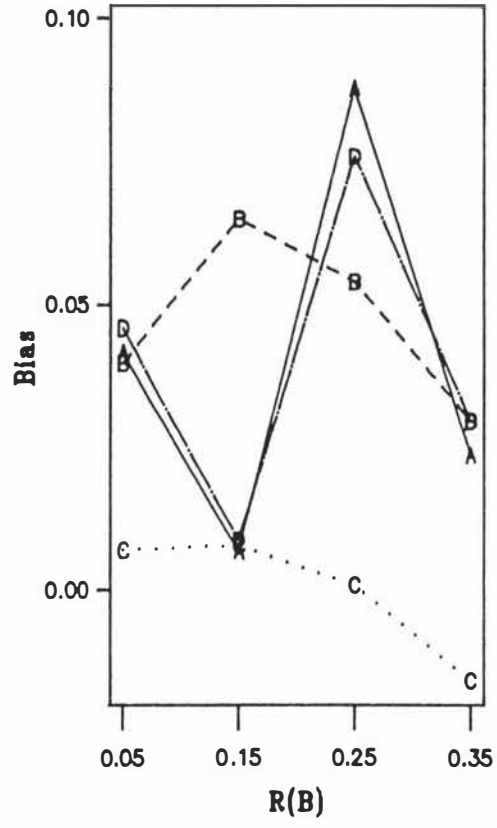
Comparison of the Bias of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.1

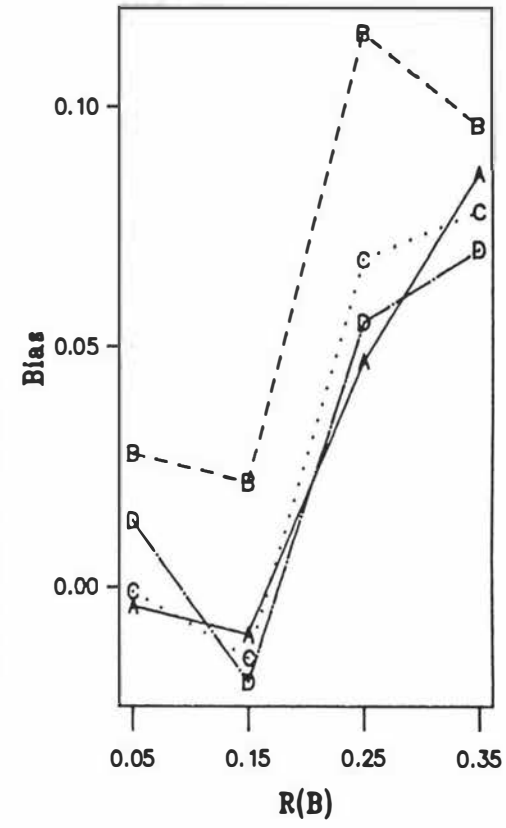
Comparison of the Bias of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.2

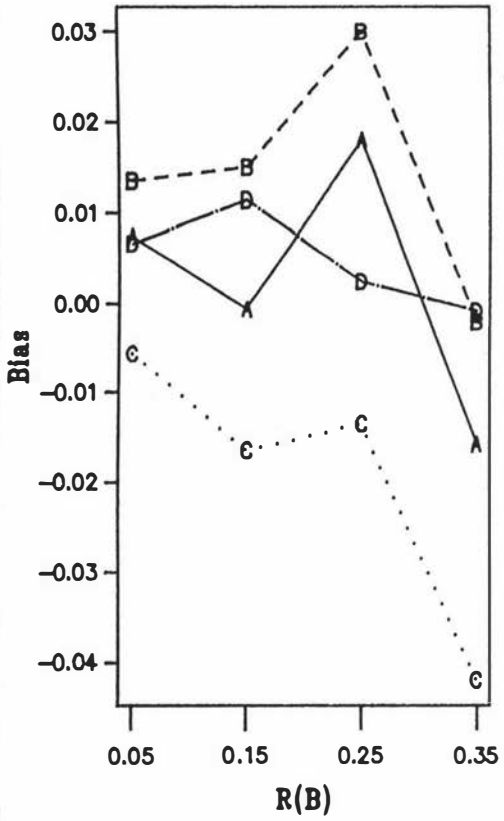
Comparison of the Bias of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.3

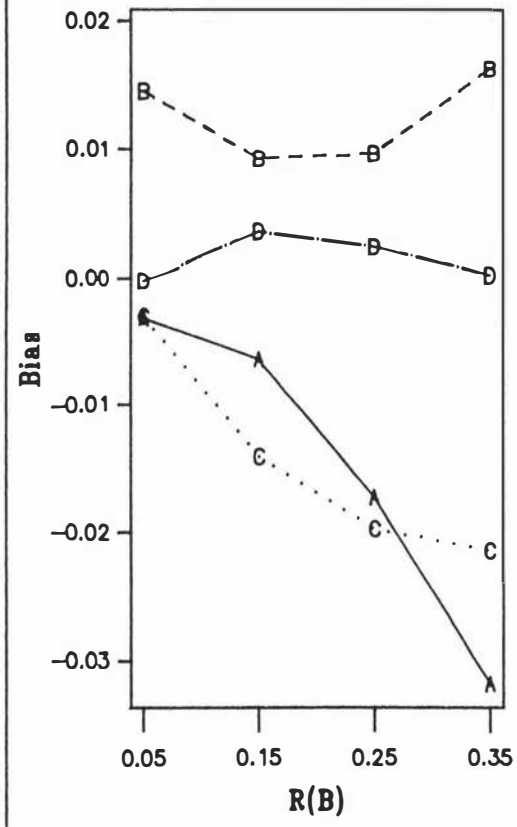
Comparison of the Bias of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.4

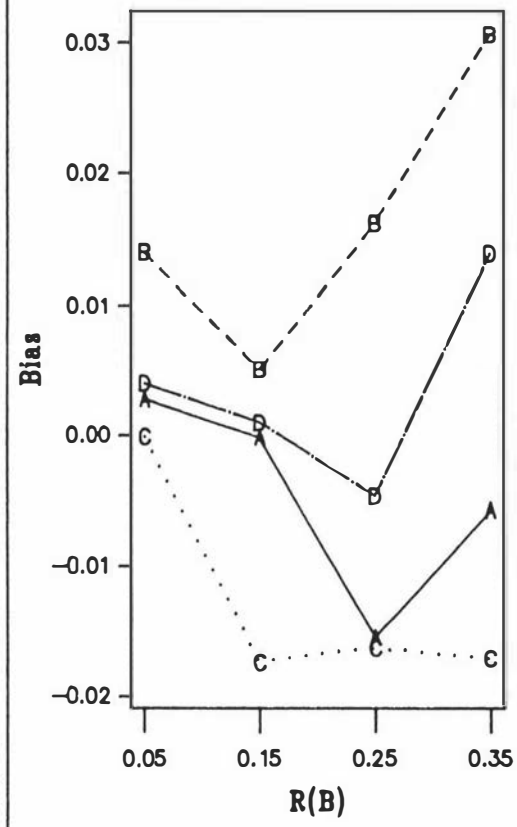
Comparison of the Bias of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.5

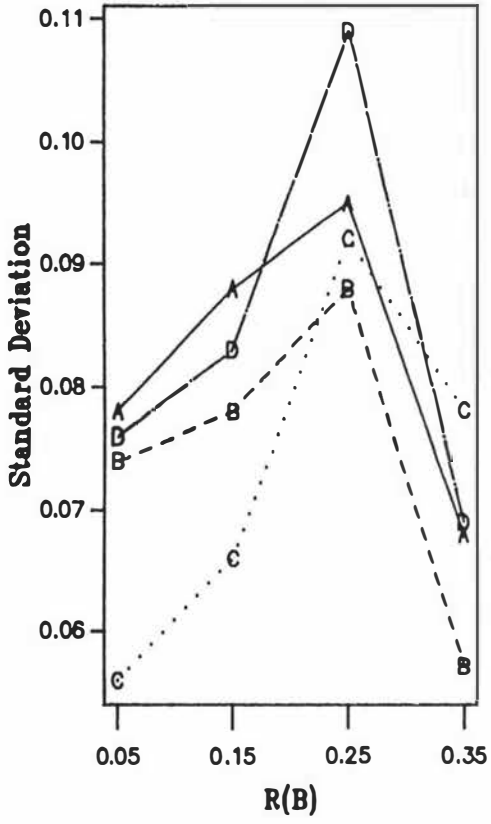
Comparison of the Bias of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.6

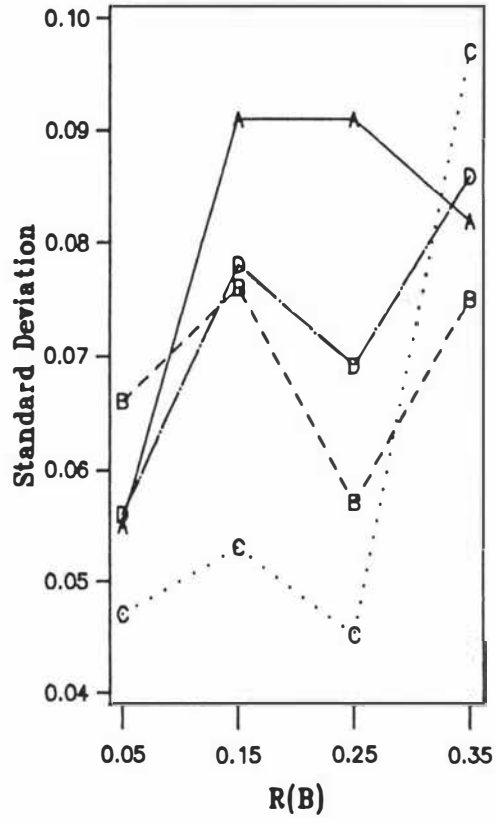
Comparison of the Standard Deviations of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.7

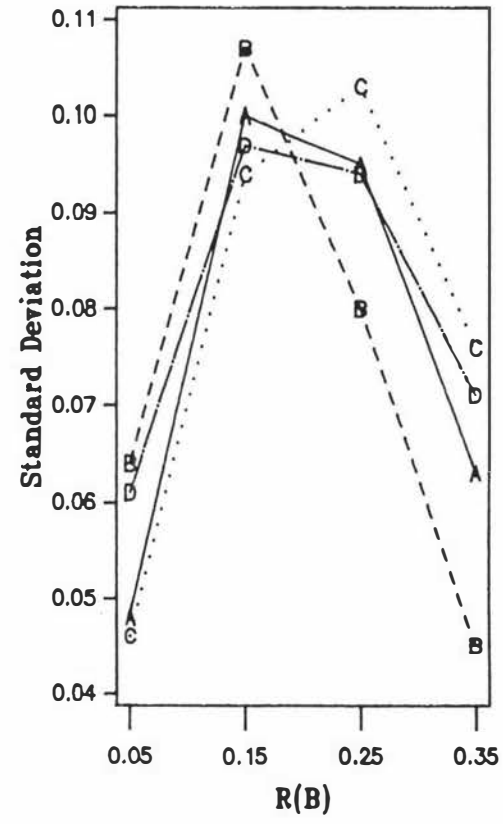
Comparison of the Standard Deviations of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.8

Comparison of the Standard Deviations of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.9

For larger samples (Figures 6.10 to 6.12), a more linear trend is apparent for all estimators in that variability increased as R(B) increased. For q = 1 and 2, the R(0.632) estimator was often the least variable estimator for higher R(B) with roughly similar variability to the cross-validation estimators for lower R(B), but for q = 3, it was the best estimator for low R(B) and worst for high R(B).

In the next analysis, the bias and variation of the estimators were combined into the MSE criterion. For small samples (Figures 6.13 to 6.15), it is clear that for q = 1 and 2, the R(0.632) estimator did best except when R(B) = 0.35 and that the best results occurred at low R(B) and the worst at moderate R(B). It should noted, though, that the R(0.632) estimator was the least affected of all estimators by changes in the values of R(B). In accordance with the results of Crawford (1989), the R(CV) estimate had high MSE, due mostly to the large variability, as shown in Figures 6.7 and 6.8. For q = 3, a slightly different picture emerged. All estimators had roughly similar MSE except R(ROT) when $R(B) \leq 0.25$.

In the case of n = 100, similar trends can be seen in all three graphs (Figures 6.16 to 6.18). Generally, the performance of each estimator deteriorated as R(B) increased. The R(CV) and R(TEN) estimates performed very much the same. The R(0.632) estimator did best overall for q = 1 and 2 while the R(CV) and R(TEN) estimators had lowest MSE for q = 3.

Figures 6.19 to 6.24 provide another measure of performance giving the values of the COUNT's of optimism for each estimator. Values closest to 0.5 were the most ideal. For small samples (Figures 6.19 to 6.21), the trends exhibited are very similar to those exhibited for bias. For q = 1 and 2, the R(0.632) estimator produced the most unbiased estimates of error while the other estimators were highly pessimistic. For q = 3, all methods, except R(ROT), had similar COUNT's. For R(B) = 0.05 and 0.15, these estimates were around 0.5 but for R(B) = 0.25 and 0.35, the estimates were highly pessimistic.

For larger samples (Figures 6.22 to 6.24), it is clear that for q = 1 and 2, the R(0.632) estimator was consistently optimistic. For q = 3, this also occurred when $R(B) \geq 0.15$, but for R(B) = 0.05, the proportion of samples where R(T) was either over or underestimated was roughly 0.5. For q = 1, R(CV) did best while for q = 2, R(TEN) did best with R(CV) tending to be rather optimistic. For q = 3, both R(CV) and R(TEN) performed equally well. The overly pessimistic nature of the R(ROT) estimator is reinforced by these results. In other words, R(ROT) was much higher, on average, than the actual error rate, R(T).

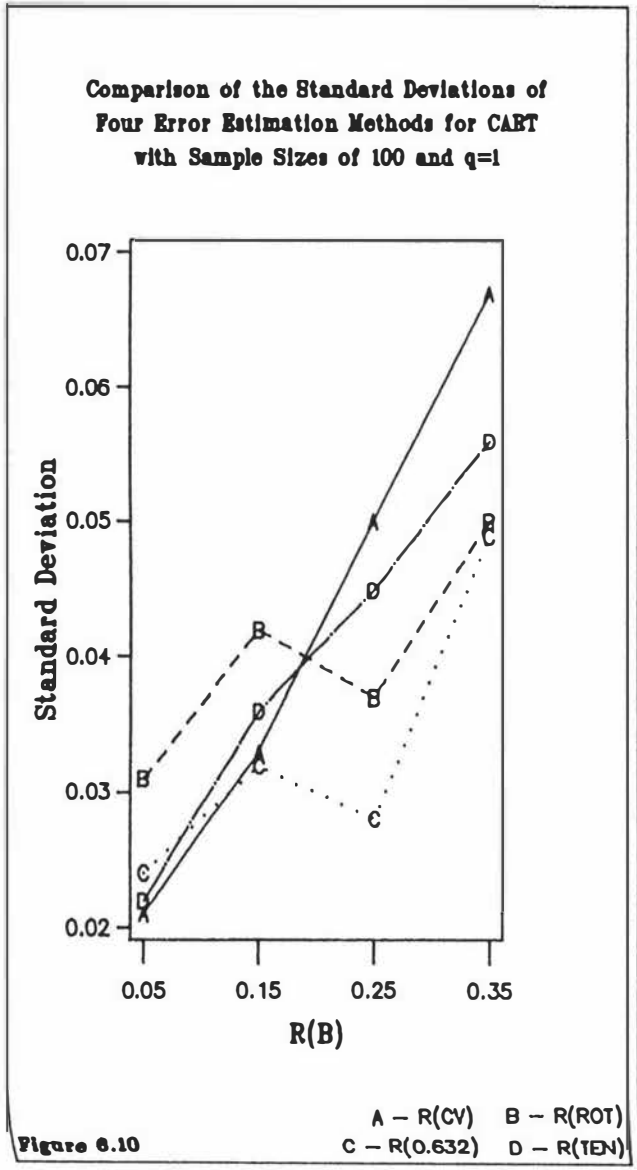


Figure 6.10

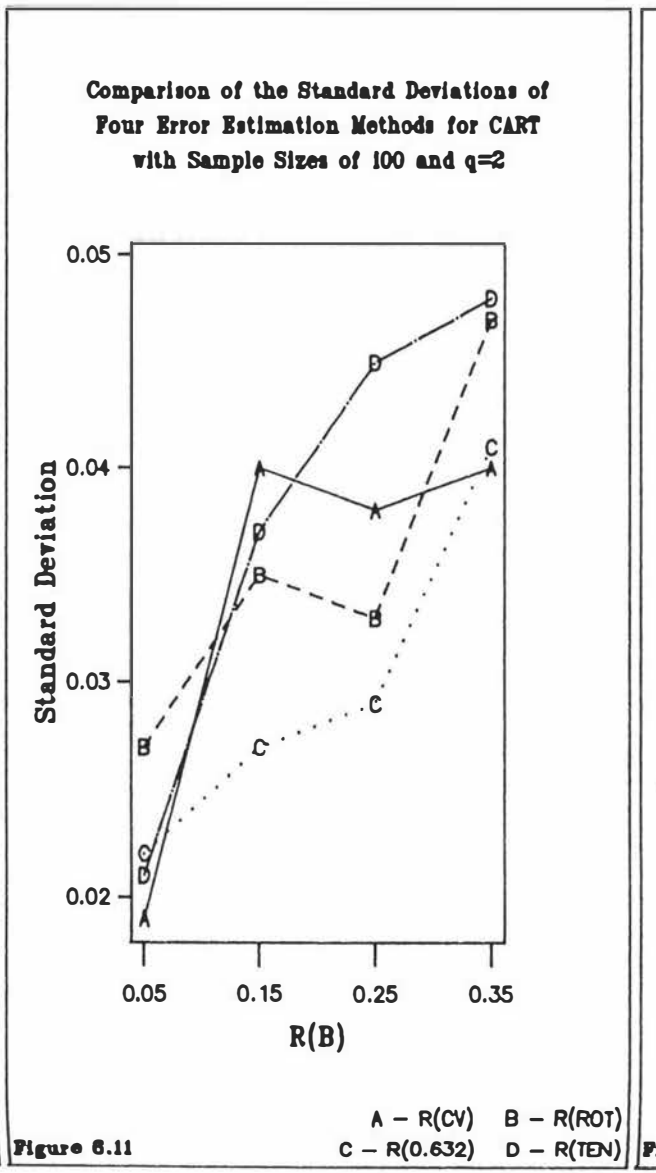


Figure 6.11

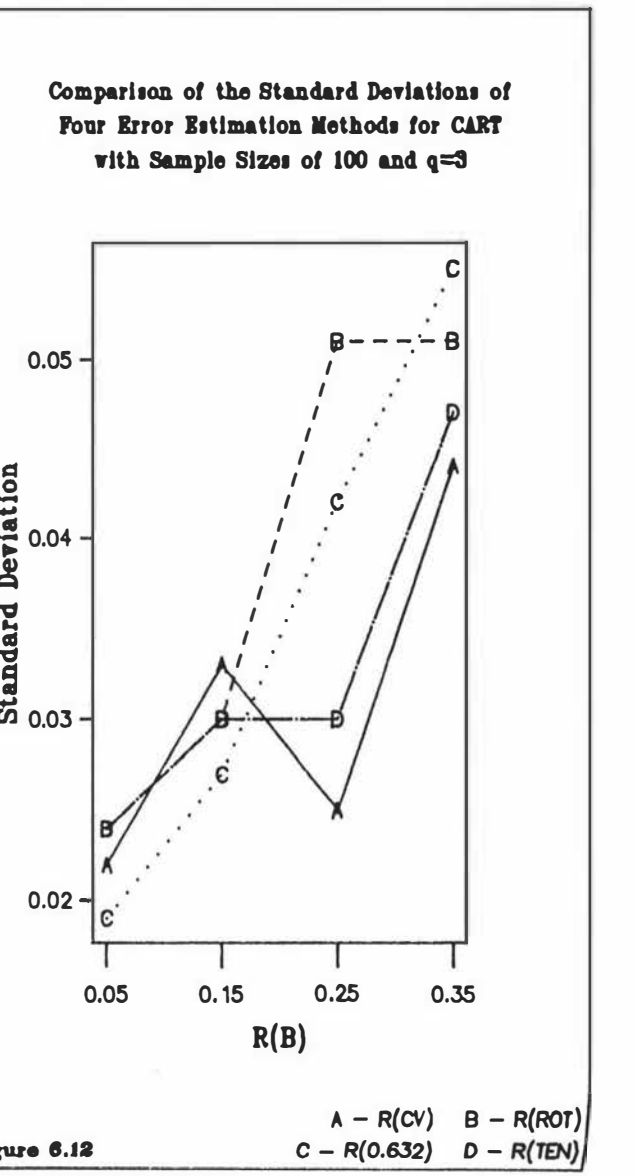
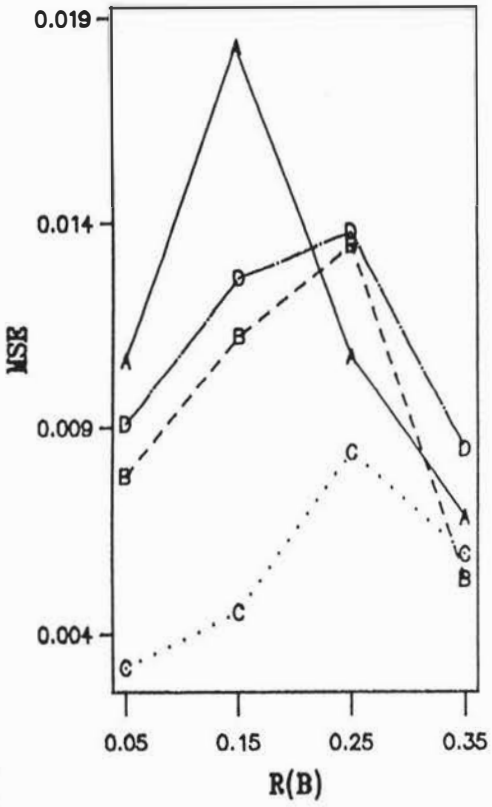


Figure 6.12

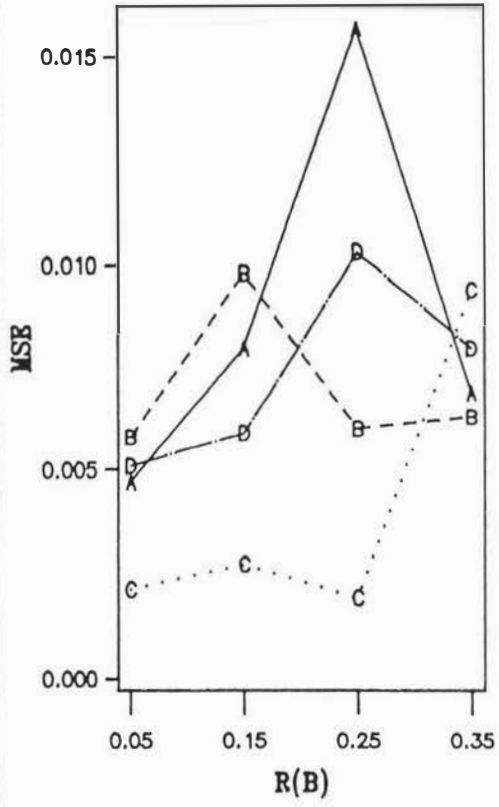
Comparison of the MSE's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.13

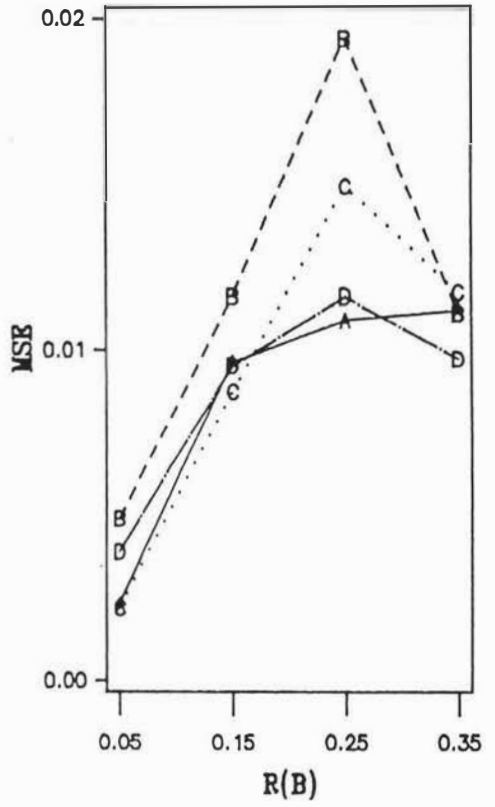
Comparison of the MSE's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.14

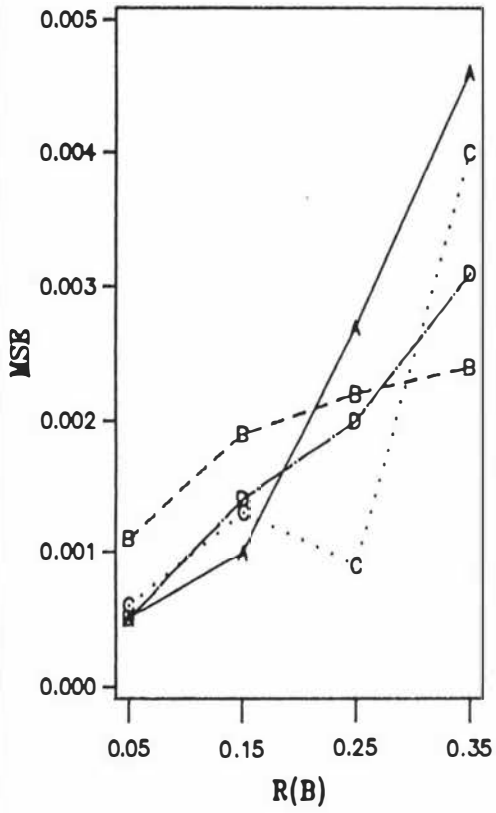
Comparison of the MSE's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.15

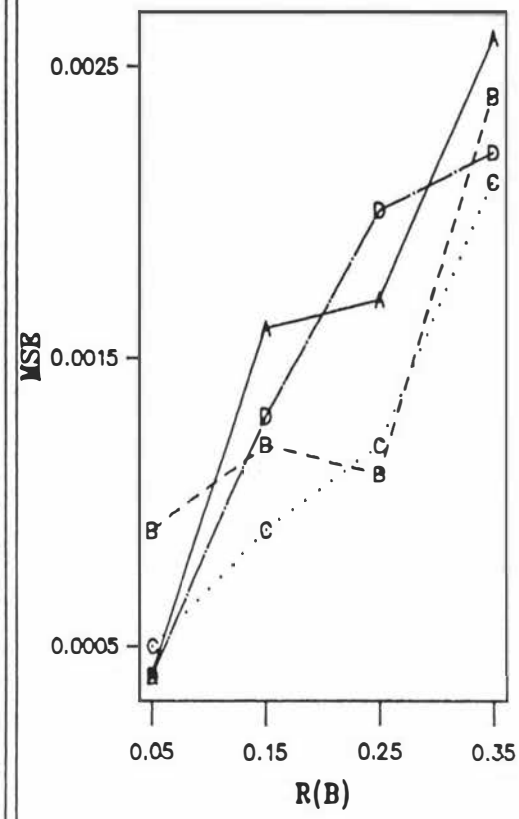
Comparison of the MSE's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.16

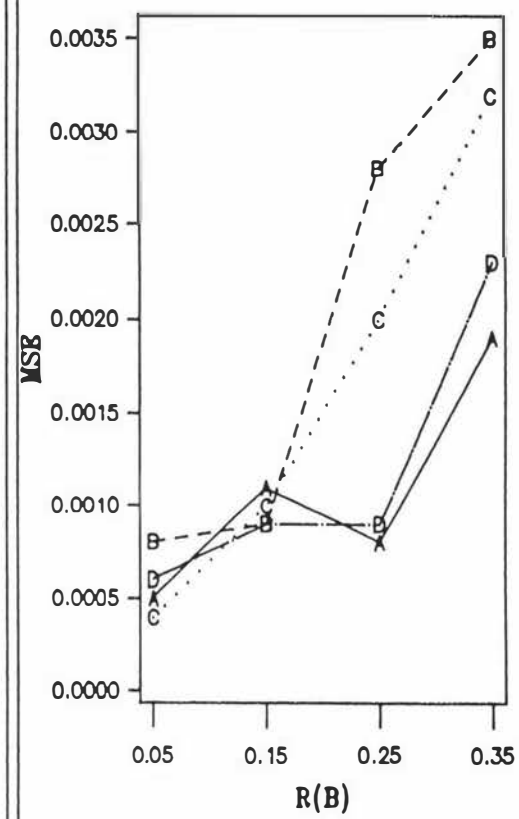
Comparison of the MSE's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.17

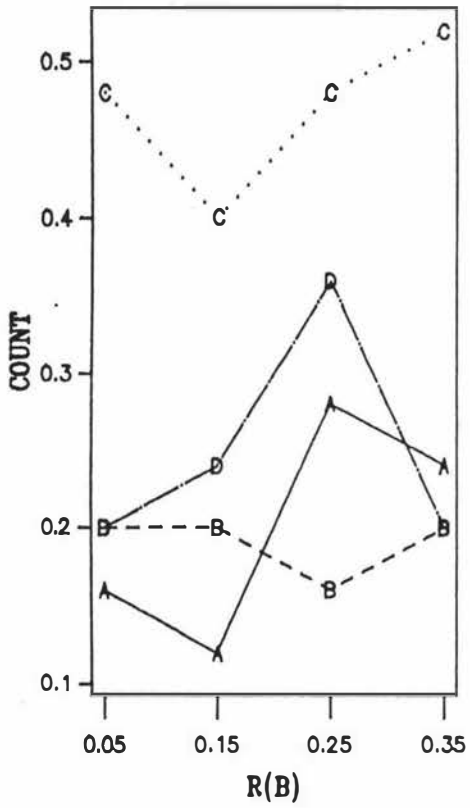
Comparison of the MSE's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.18

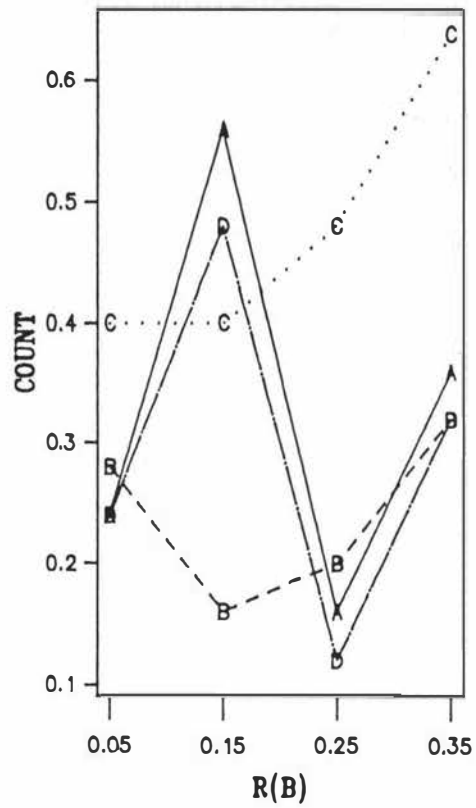
Comparison of the COUNT's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.19

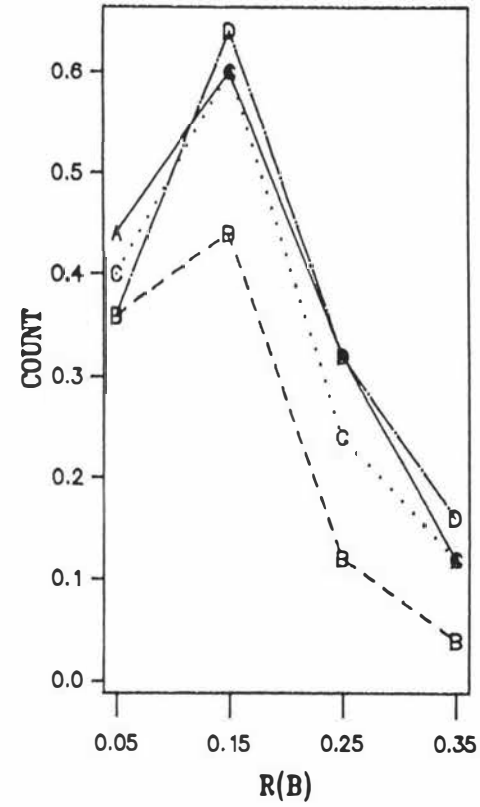
Comparison of the COUNT's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.20

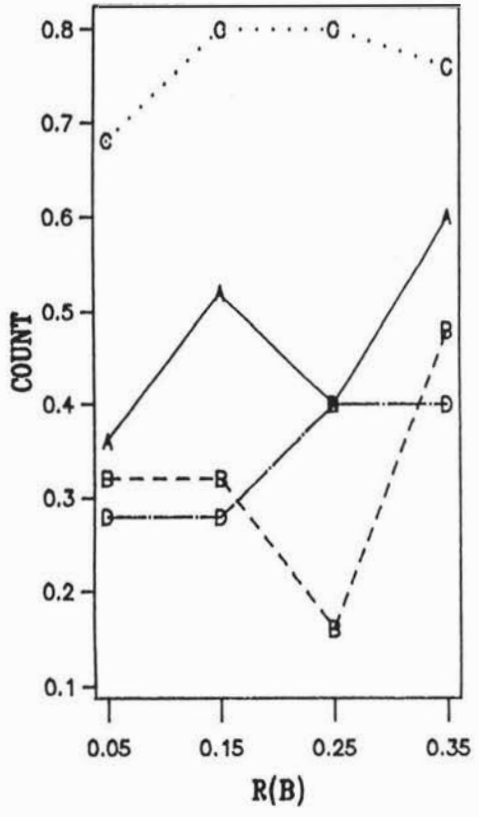
Comparison of the COUNT's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.21

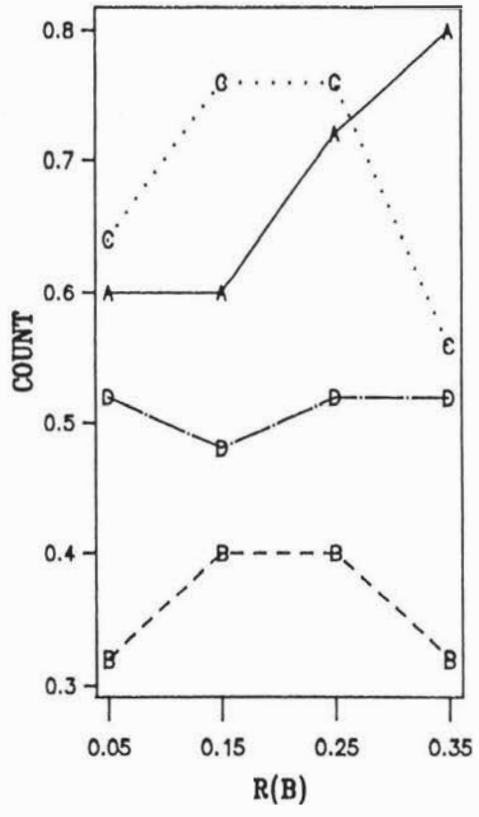
Comparison of the COUNT's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=1



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.22

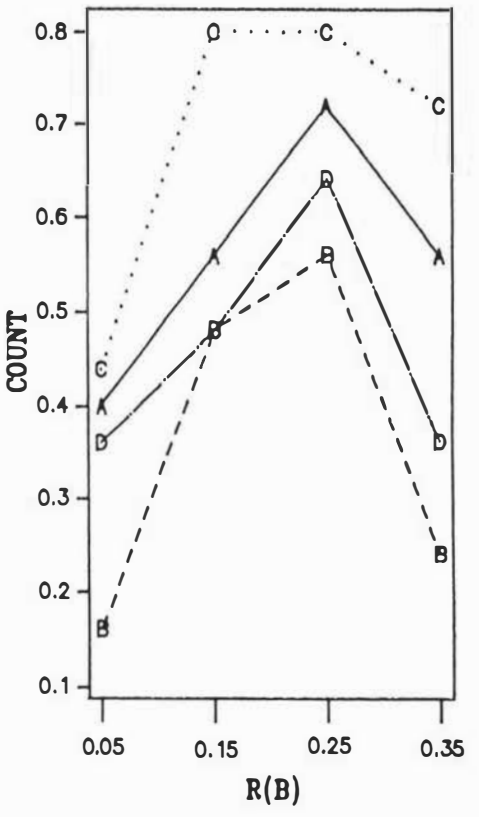
Comparison of the COUNT's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=2



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.23

Comparison of the COUNT's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=3



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.24

These results were fairly similar to those of Crawford (1989), though he used a different MSE criterion to the one used here. Generally, these results have shown that the R(0.632) estimator for CART was clearly the most reliable estimator for smaller samples and marginally the best for larger samples in the case of conditions less favourable to CART. The R(CV) and R(TEN) estimators were best for larger samples in situations most favourable to CART and for larger R(B). For smaller samples and situations favourable to CART, there was little to choose between the estimators. The R(ROT) estimator was found to be the worst due to the often large pessimistic bias.

As an extension to the studies undertaken here, it was decided to compare the performance of the holdout estimator, R(H), for n = 100 only. The R(H) estimator was calculated by using two thirds of the original data as the learning sample to grow a classification tree and the other third as a test sample to select the tree size and estimate the error rate of the chosen tree. In summary, the results showed that the R(H) estimator was unbiased but that the variability of the estimator was very large, leading to a consistently higher MSE than the other estimators. As recommended by Breiman et al (1984), the holdout method should not be used with CART unless the data set is very large. (They recommend a combined data set of at least 1000 cases.)

A final analysis in this study was carried out to compare the tree sizes generated by each of the three error estimation methods. Naturally the tree chosen by R(ROT) was the same as that for R(0.632). The results showed that the method effect (F = 14.08) dominated all others, and that on average, the trees produced by the rotation method were the simplest, containing 3.5 terminal nodes compared with 3.63 for tenfold cross-validation and 4.35 for n-fold cross-validation.

6.2.4 Summary

This study has shown that trees produced by using the rotation method for tree selection, then using the 0.632 method to estimate R(T), the actual error rate, for the selected tree, were in the main fairly unbiased, or if biased optimistically, generally exhibited low variability. The R(0.632) estimator was found to be the most reliable of all error estimation techniques for small samples in non-ideal situations (parallel classification problems) and marginally more reliable than other methods for smaller samples and ideal situations (sequential classification problems), when R(B), the Bayes error rate, was low to moderate, as well as for larger samples and non-ideal situations. Only when the classification problem was sequential and

R(B) was moderate to high did R(CV) and R(TEN) clearly outperform R(0.632). The R(CV) and R(TEN) estimates were found to be extremely variable for small samples, although as claimed by Breiman et al (1984), R(TEN) did no worse than R(CV). The R(ROT) estimator was found to be overly pessimistic in nearly all situations.

The general trend observed for all estimators was that for small samples, reliability was best for either lowest or highest R(B), but for large samples, reliability decreased with increasing R(B).

6.3 ERROR RATE ESTIMATION FOR CATEGORICAL DATA IN CART

6.3.1 Study Plan

In this section, the various error rate estimation techniques were calculated using the same data sets that were employed in Chapter 5. However, the fourth level of the probability patterns, p_{ij}, was omitted as the results for that level were found to be very close to those of the third level. This meant that there was unnecessary replication of the same type of classification problem. For simplicity, level 5 was recoded as level 4.

The aim of this study was the same as that of Section 6.2, except that the conclusions apply to categorical rather than continuous data. Comparisons of the error estimation techniques were made by means of the twin criteria of bias and MSE. In addition, a comparison of the tree sizes by each method was made to determine if there were any differences.

For all the data sets in this section, the one standard error rule was used while the size below which a node will not be split was set to five. Independent test samples of size 5000 were used throughout to determine R(T).

6.3.2 Results

For p = 5 variables, the split-plot ANOVA showed that the method main effect (R, F = 4.36) and the interactions of method with probability pattern (R * p_{ij}, F = 2.74), method with sample size (R * n, F = 2.66) and method with probability pattern and sample size (R * p_{ij} * n, F = 2.44) were the only significant effects, hence the results are presented in terms of the latter.

As in Section 6.2, the results for the three apparent error rate estimators are not presented in order to preserve display resolution, as in most situations, the bias and MSE's of these estimators was much larger than that of the other estimators. Figures 6.25 to 6.27 give the values of the bias for the other four estimators when p = 5. For n = 20, the R(0.632) estimator was least biased for the first three levels of p_{ij} as well as either R(ROT) or R(TEN). For all four levels of p_{ij}, R(CV) was highly optimistic, and most disturbingly, underestimating R(T) by almost 10% in the case of identical populations. For n = 60, R(CV) did best for the most highly separated populations (p_{ij} = 1) but for other levels of p_{ij} was consistently the worst (overoptimistic). The R(0.632) and R(ROT) estimators exhibited fairly similar trends, being fairly unbiased. For n = 100, the R(CV) estimator did equally well for p_{ij} = 1 and 2, but was the worst for the other two levels of p_{ij}. The other estimators all had very similar error rates.

In terms of MSE, Figure 6.28 shows that when n = 20, the performance of all estimators varied across the levels of p_{ij}, with R(0.632) doing best for p_{ij} = 1 and 3, along with R(ROT) for p_{ij} = 3, R(CV) for p_{ij} = 2 and R(TEN) for p_{ij} = 4, where R(CV) did worst. For n = 60, Figure 6.29 shows that R(ROT) did marginally better than R(0.632) except for p_{ij} = 1 where R(0.632) did narrowly better than R(CV). The poor performance of R(CV) for p_{ij} = 2, 3 and 4 can also be seen. For n = 100, Figure 6.30 shows a similar pattern to that for the smallest samples in that the R(0.632) estimator had the lowest MSE for p_{ij} = 1 and 3, R(CV) for p_{ij} = 2 and R(TEN) for p_{ij} = 4.

For p = 10, the split-plot ANOVA showed that the R main effect (F = 14.58) and R * p_{ij} (F = 2.48) and R * n (F = 2.32) interactions were all significant at the α = 1% level. Hence, the results are presented in terms of p_{ij} and n in turn. Figure 6.31 shows that in terms of bias, the R(0.632) estimator was best for p_{ij} = 2 and 4. For p_{ij} = 3, R(ROT) did best, otherwise it was overly pessimistic. The R(TEN) estimator was narrowly better than R(0.632) for p_{ij} = 1 but was overly optimistic for other levels.

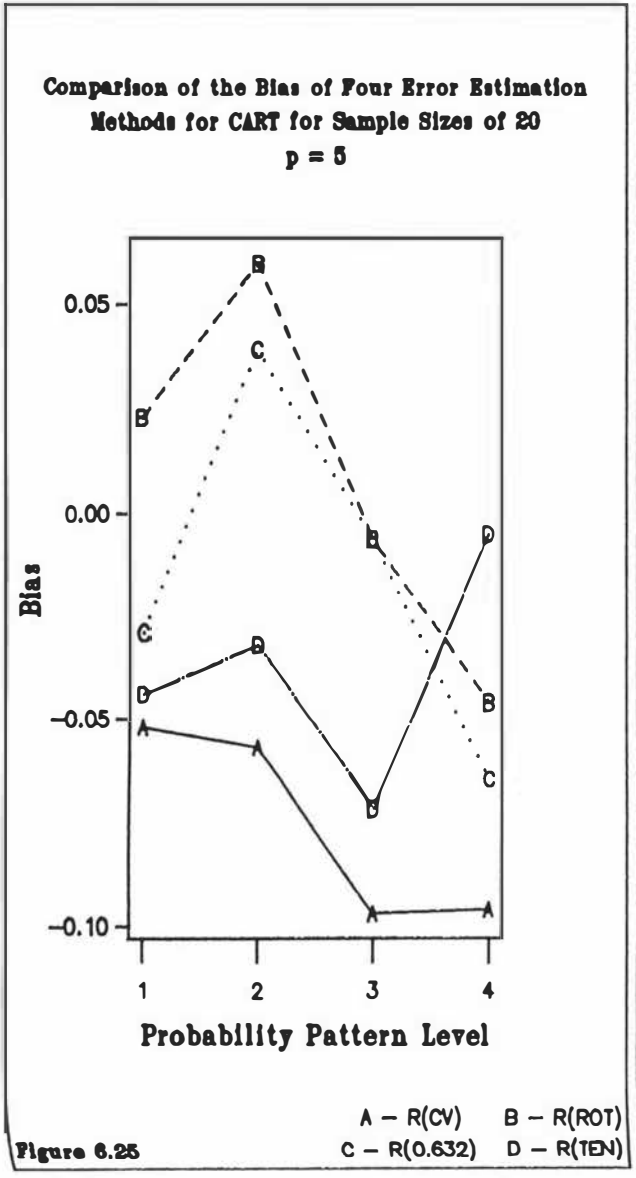


Figure 6.25

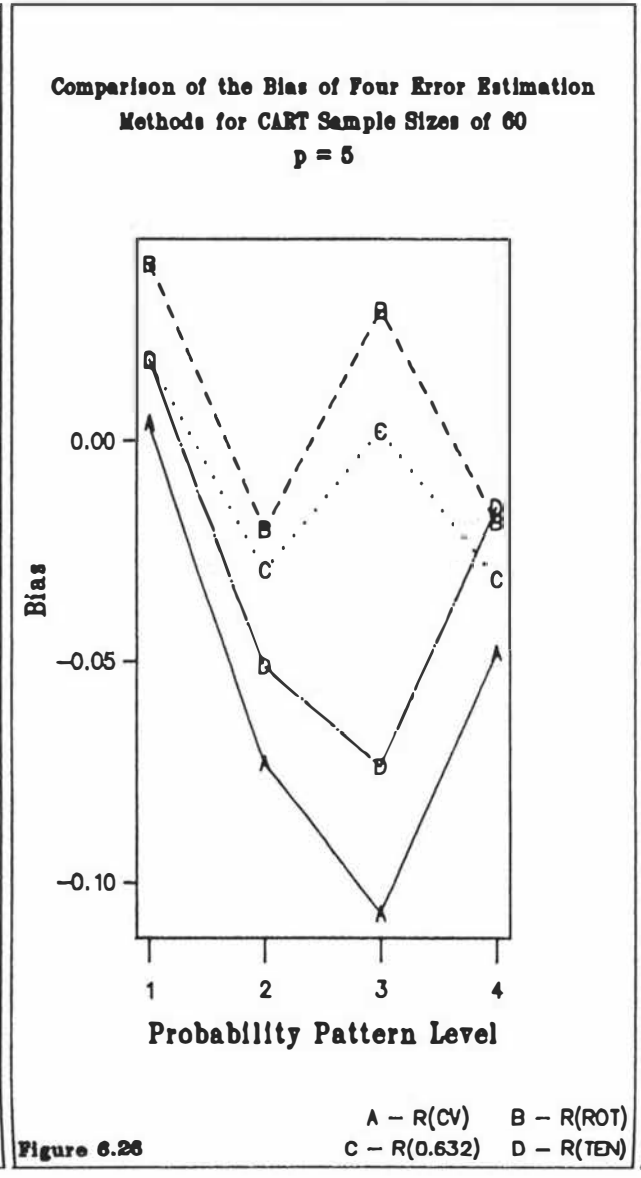


Figure 6.26

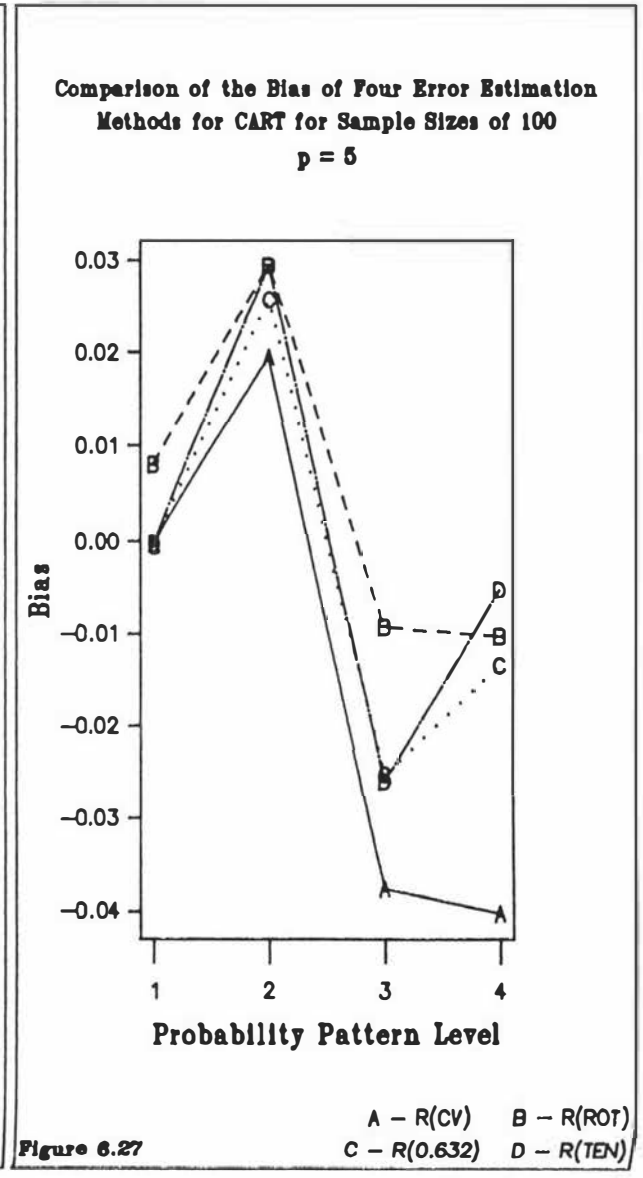


Figure 6.27

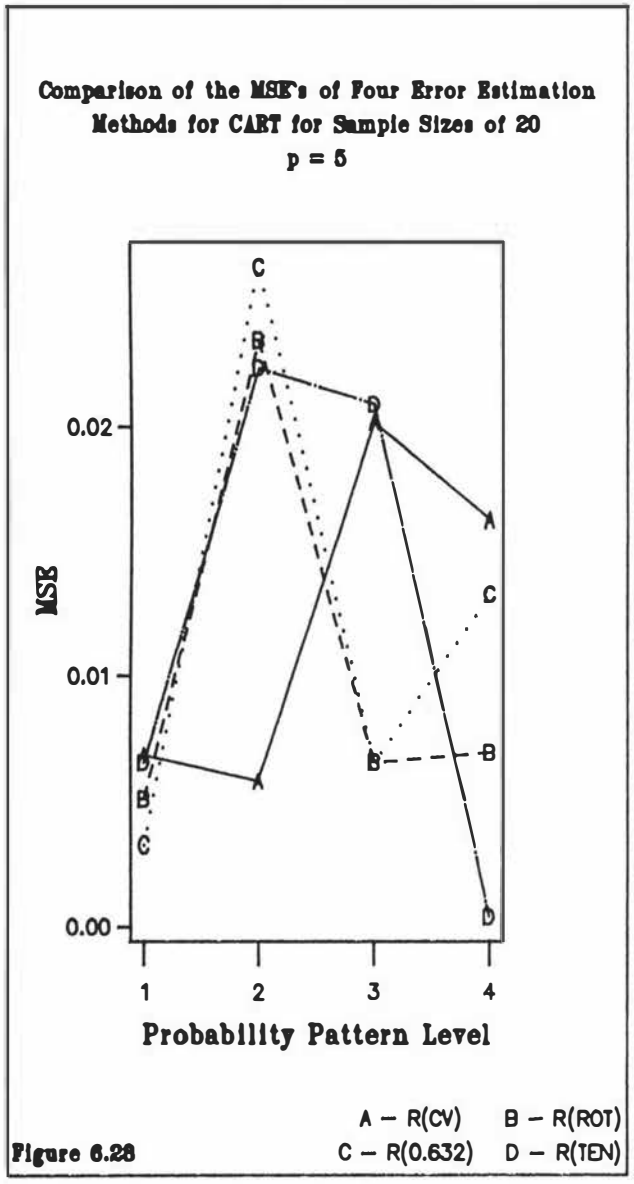


Figure 6.28

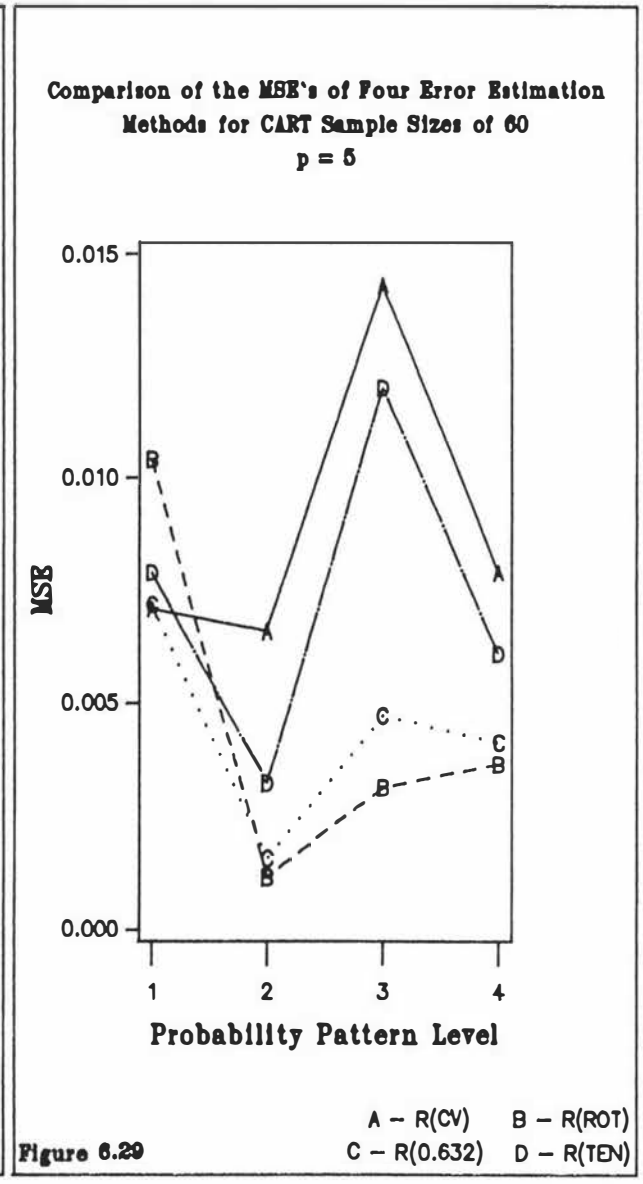


Figure 6.29

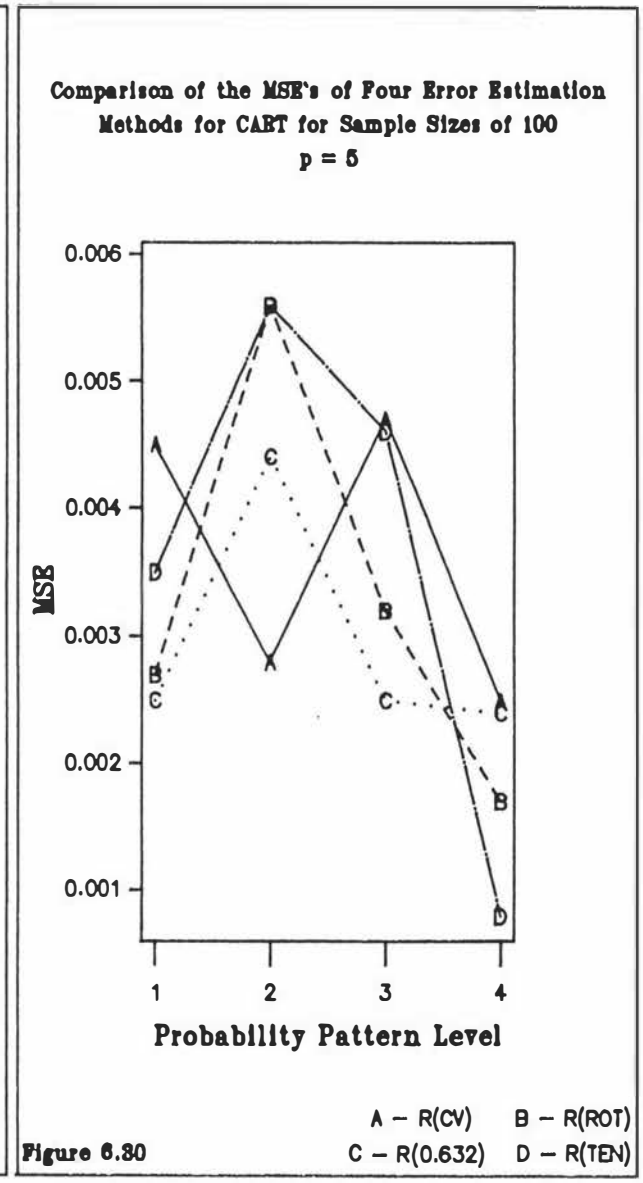


Figure 6.30

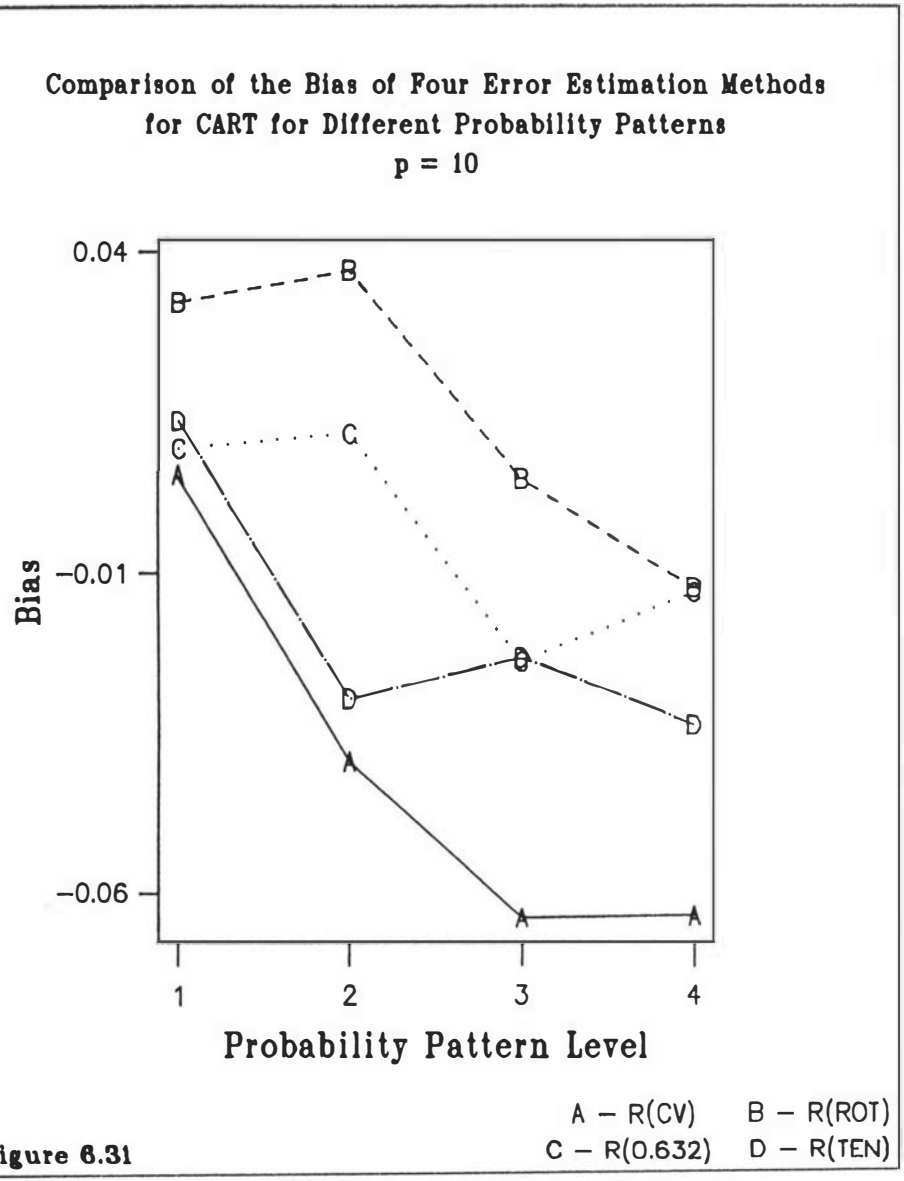


Figure 6.31

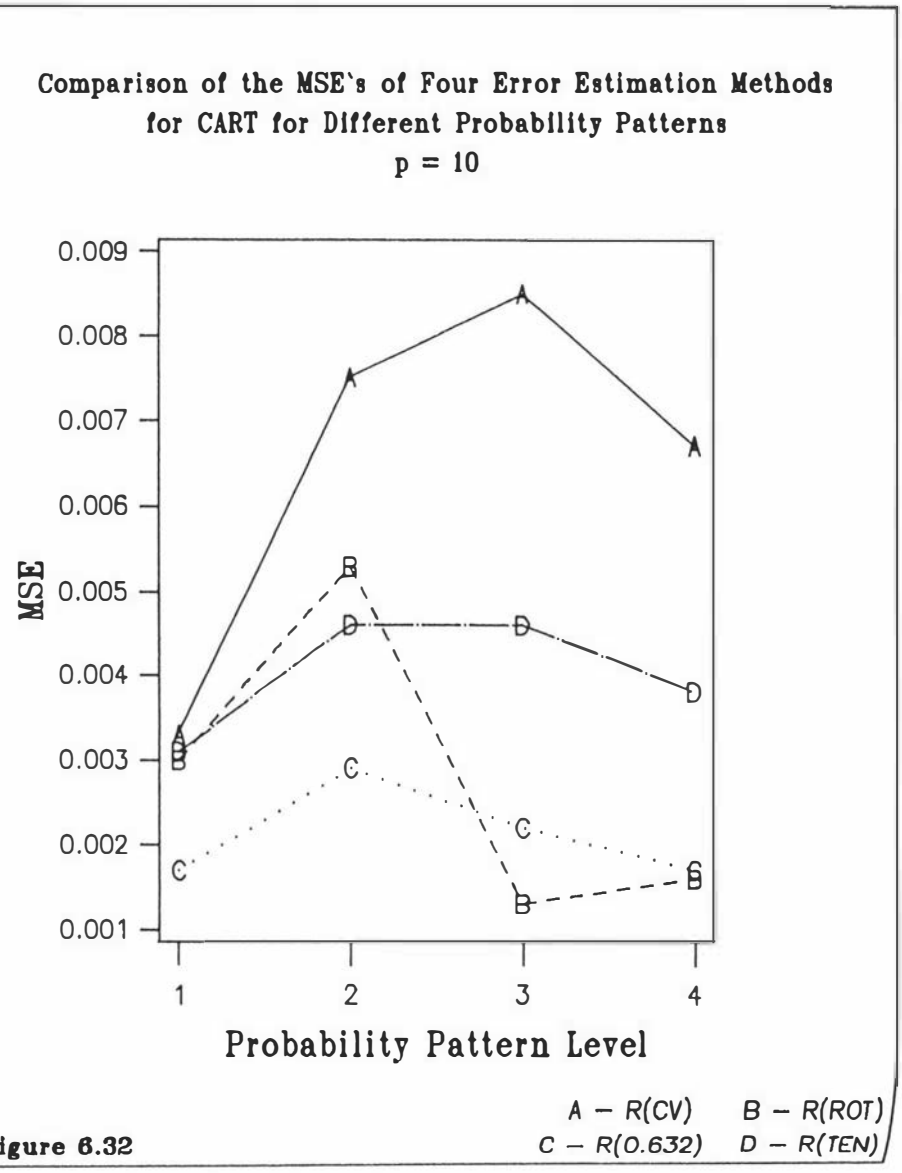


Figure 6.32

Figure 6.32 shows that in terms of MSE, R(0.632) was clearly best for p_{ij} = 1 and 2, while R(ROT) was the most reliable for p_{ij} = 3 and 4. R(CV) was found to be the least reliable estimator in all situations. Interestingly, these results mirror those found for continuous data in that R(0.632) was shown to be the best estimator for parallel classification problems which were not suited to CART as well as for the best separated populations (p_{ij} = 1 and 2). Note, though, that the reliability of the R(0.632) estimate exhibited a distinctive U shape pattern, doing best for either well-separated or identical populations.

Figure 6.33 shows that R(0.632) was the least affected by changing sample size and was the least biased for small samples. The high optimistic bias of R(CV) for small samples is clearly evident and remains optimistic for larger samples.

In terms of MSE, Figure 6.34 shows that R(0.632) did best for n = 40. For n = 120, R(0.632) was slightly better than R(ROT) while for n = 200, R(0.632), R(ROT) and R(TEN) did equally well. The R(CV) estimator for all sample sizes performed consistently the worst.

A final analysis in this section compared the sizes of the trees resulting from each of the three error estimation methods for both p = 5 and p = 10 binary variables. For p = 5 variables, the split-plot ANOVA showed there to be a difference between methods (F = 7.69) with R(ROT) producing the smallest sized trees (2.03 terminal nodes) and R(CV) the largest (2.64 terminal nodes). For p = 10, again the method effect (F = 7.7) dominated all others with the R(ROT) estimator producing the smallest sized trees (2.9 terminal nodes) on average. The average sized trees produced by R(CV) and R(TEN) contained 5.07 and 4.44 terminal nodes respectively implying that the rotation method produced trees with 1.5 less terminal nodes than any other method.

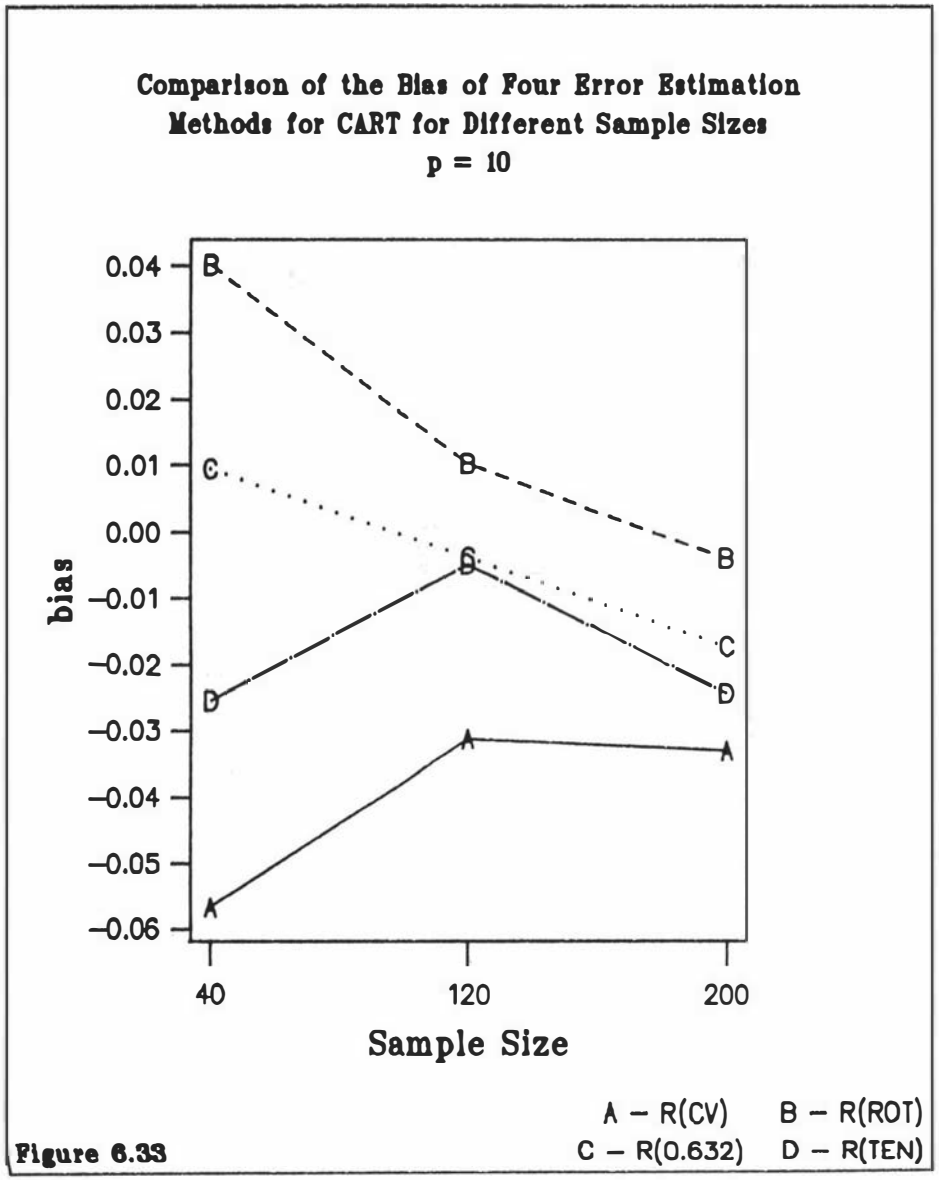


Figure 6.33

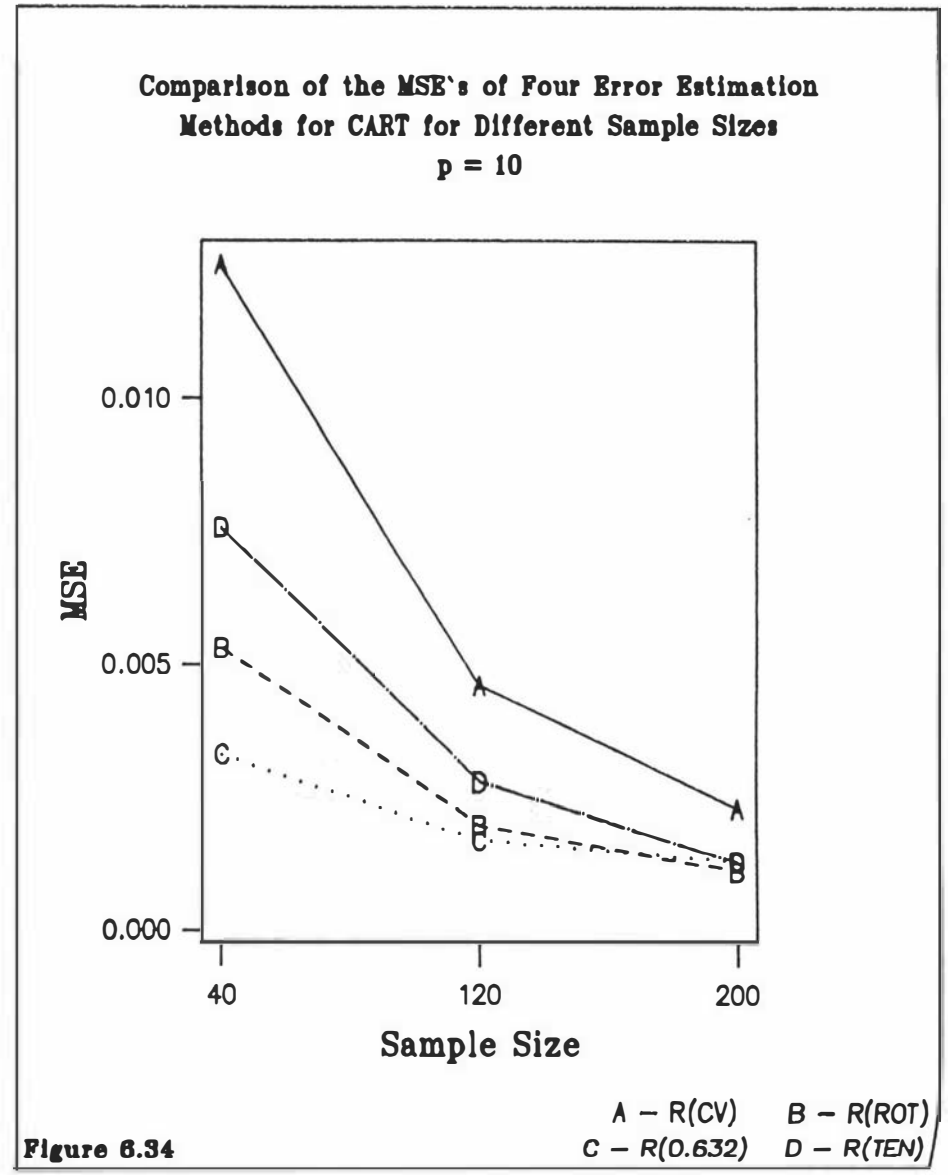


Figure 6.34

6.3.3 Summary

This study has shown that for categorical data, the R(0.632) and R(ROT) estimators produced the most reliable classification trees as well as being the simplest. It was found that, as for the continuous data in Section 6.2, the R(0.632) estimator was the most reliable for small samples, parallel classification problems and well separated populations. As with the continuous data, and in agreement with other studies, the R(CV) estimator was found to be a poor estimator for small samples and highly optimistic, especially for poorly separated populations. Based on the results of this study, the use of n-fold cross-validation with categorical data is not recommended.

6.4 THE STANDARD ERROR RULE IN CART

6.4.1 Previous Studies

The motivation for doing this study was provided by Breiman et al (1984). They recommended the use of the one standard error (1-SE) rule for, firstly, the sake of accuracy, noting that in most cases, the cross-validation estimate of error was over optimistic. Secondly, they stated that a plot of R(CV) against tree size had the characteristics of an initial sharp decrease followed by a long, flat valley across a wide range of tree sizes and then an increase for very small trees. Inside the long valley, most error rates were found to be within the $\pm$ 1-SE range and that the position of the minimum may be unstable. The 1-SE rule was used to reduce that instability as well as produce trees which are as simple as possible.

Feng et al (1993) carried out a small-scale empirical study comparing the zero standard error (0-SE) and 1-SE rules using various data sets. They found that trees produced by the 0-SE rule were between two and ten times larger than those constructed using the 1-SE rule, so that the latter were biased towards simplicity. In determining which rule was better they were rather inconclusive. "We believe that there is no single best rule, instead it depends on how much "noise" there is in the data. If there is little noise in the data, then the 0-SE rule should be used. If there is a lot of noise then ... the 1-SE rule should be used." (ibid, p 49.)

6.4.2 Study Plan

In Section 6.2, the performance of various error rate estimators was compared in estimation of the actual error rate using the 0-SE rule. In this section, only two estimators, R(CV) and R(0.632), were used, one of which worked best in any one of the factor combinations studied in Section 6.4. The two error estimation methods were compared over the factor combinations studied in Section 6.2 involving the Bayes error rate, R(B), sample size, n, and the third factor q, using both the 0-SE and 1-SE rules, with the objective of determining in which situations either of the above two rules should be used. Comparisons were made using both the bias and MSE performance criteria. In addition, a comparison of the decrease in tree size produced by using the 1-SE rather than the 0-SE rule was made for q = 3 only.

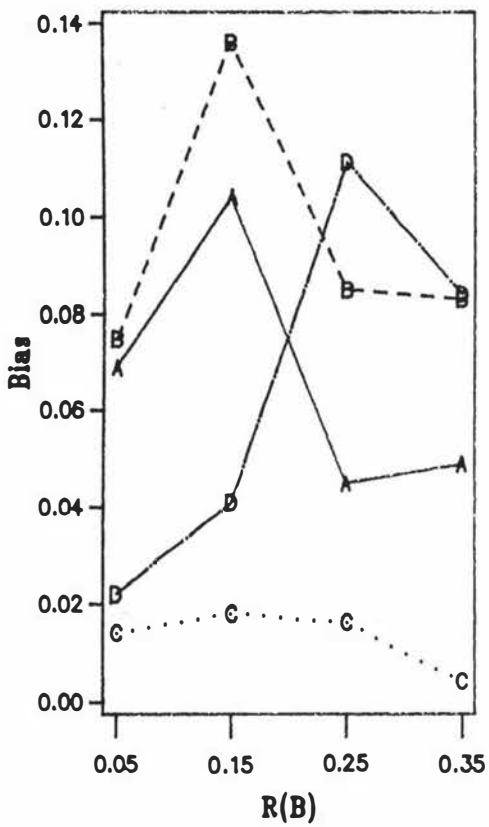
6.4.3 Results

Figures 6.35 to 6.37 compare the average bias for n = 20. For q = 1 and 2, it is clear that both the R(CV) and R(0.632) estimators using the 0-SE rule were less biased than the corresponding estimates using the 1-SE rule except for R(B) = 0.05. For q = 3, there was little to choose between the use of the 0-SE or 1-SE rules, except the R(0.632) estimate for larger R(B) which was excessively pessimistic. Note too that the two R(CV) estimators exhibited very similar trends as functions of R(B) while the two R(0.632) estimates did not. For q = 1 and 2, the average bias decreased as R(B) increased using the 0-SE rule while the bias increased as R(B) increased using the 1-SE rule.

Figures 6.38 to 6.40 illustrate the cases of n = 100. For q = 1, 2 and 3, the estimates using the 1-SE rule were generally less biased than those using the 0-SE rule, with the latter tending to be over optimistic. For q = 3, the disparity in bias between the 0-SE and 1-SE rule estimates was less marked for low R(B) than high R(B). As with n = 20, the R(CV) estimates followed similar patterns while the R(0.632) estimates behaved rather differently.

In terms of MSE, Figures 6.41 to 6.43 illustrate the cases of n = 20. The trends shown are very similar to those for bias. For q = 1 and 2, the 0-SE estimates produced the lowest MSE except when R(B) = 0.05. For q = 3, the difference between methods was marginal.

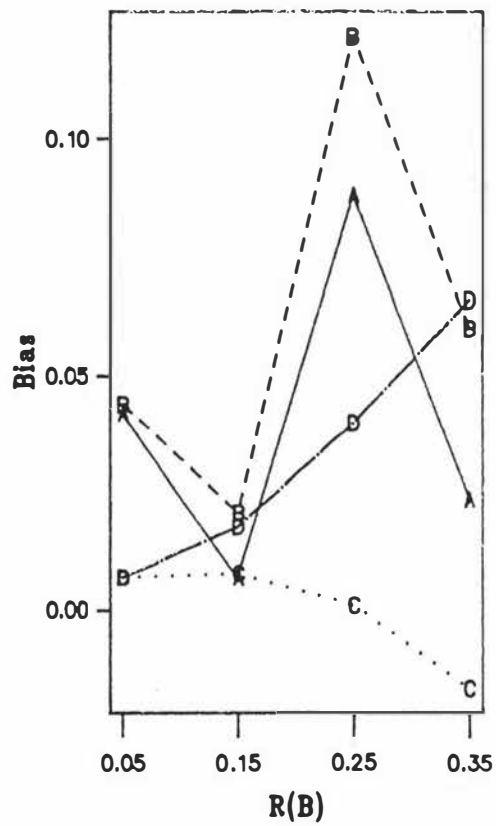
Comparison of the Bias of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 20 and q=1



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.35

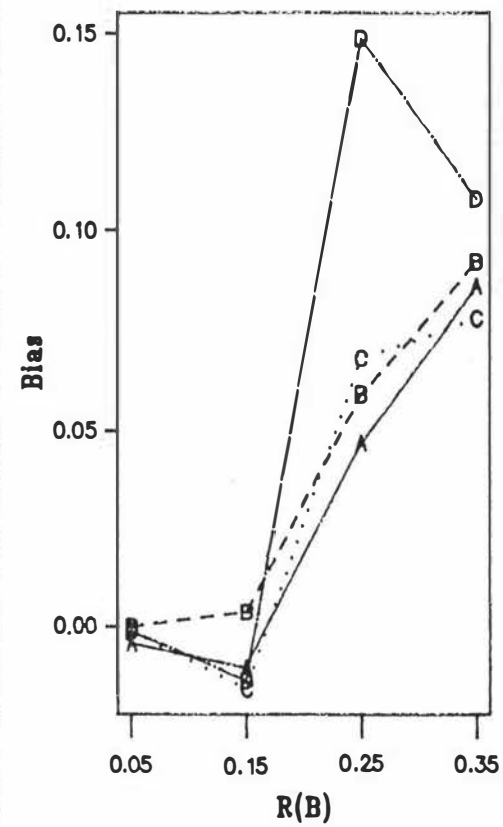
Comparison of the Bias of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 20 and q=2



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.36

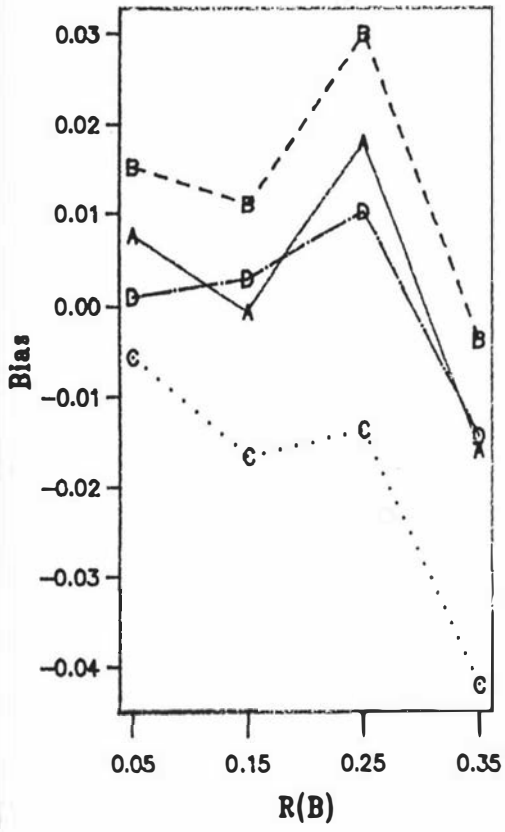
Comparison of the Bias of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 20 and q=3



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.37

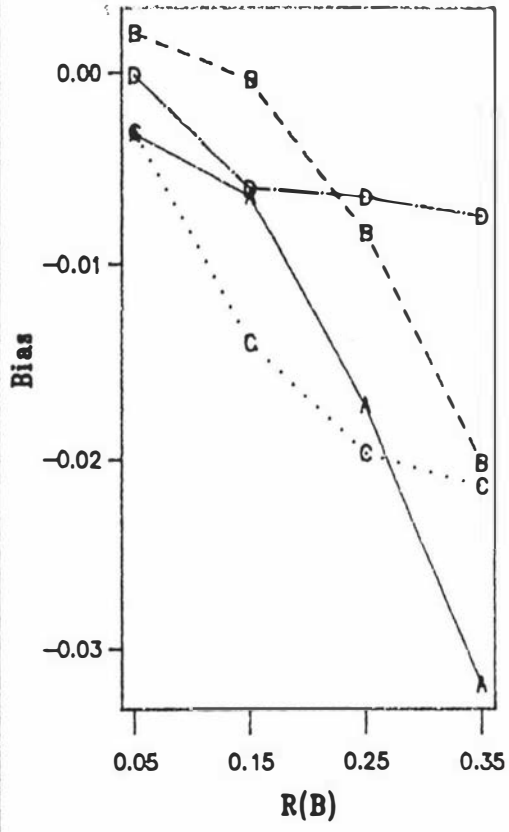
Comparison of the Bias of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 100 and q=1



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.38

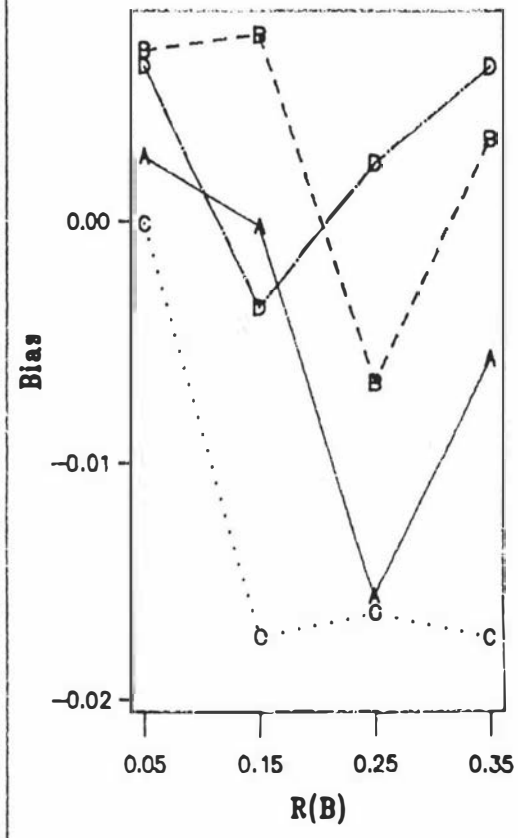
Comparison of the Bias of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 100 and q=2



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.39

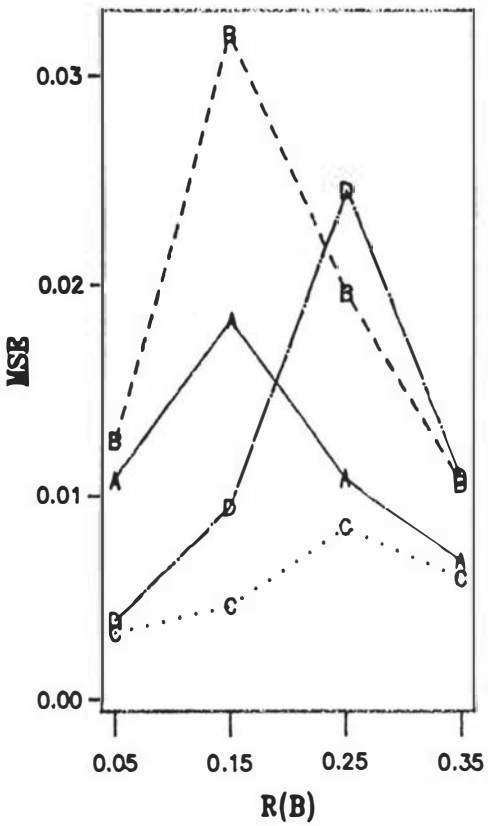
Comparison of the Bias of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 100 and q=3



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.40

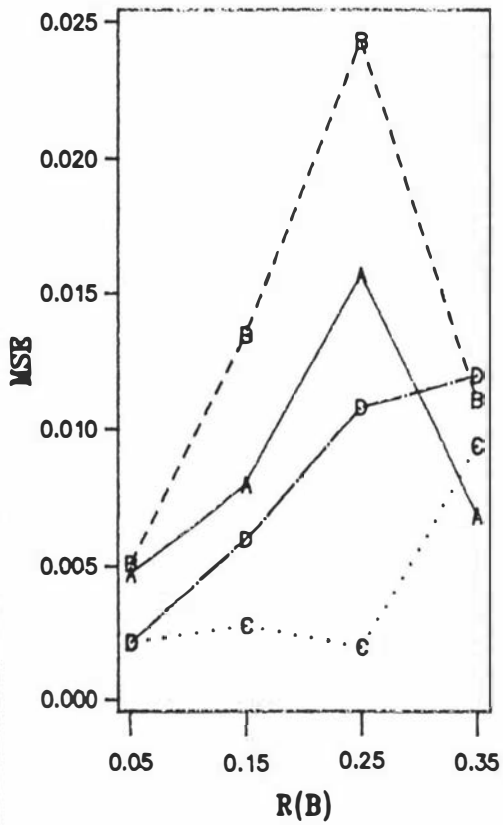
Comparison of the MSE's of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 20 and q=1



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.41

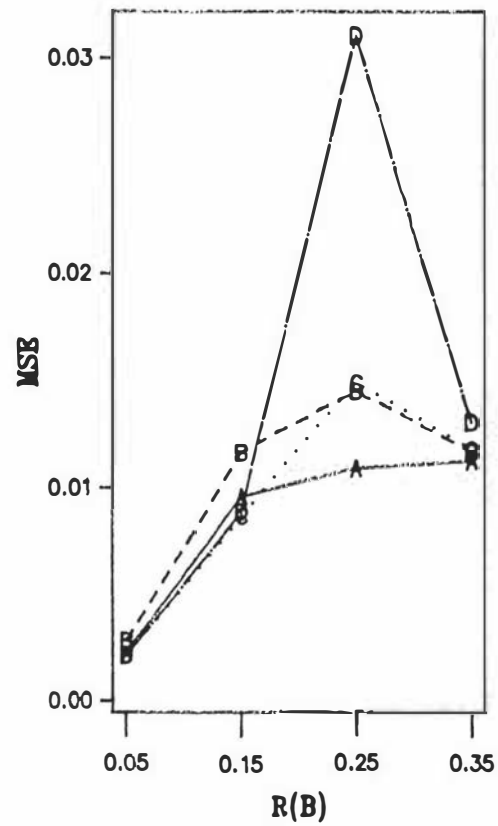
Comparison of the MSE's of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 20 and q=2



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.42

Comparison of the MSE's of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 20 and q=3



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.43

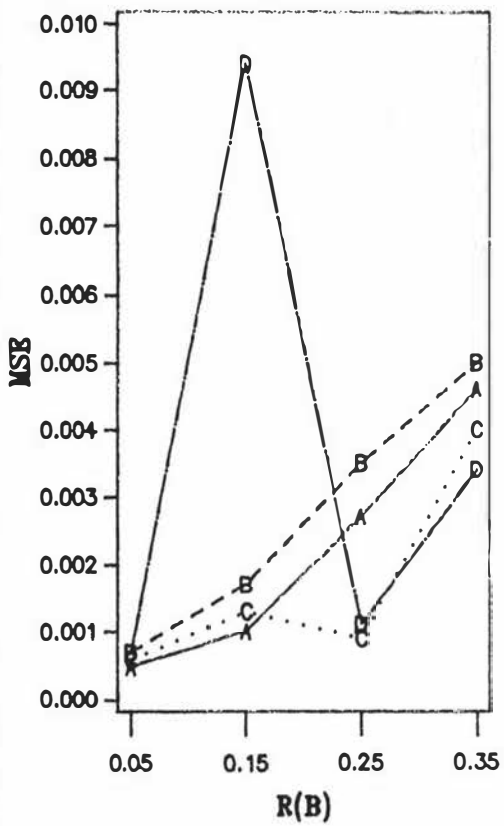
For larger samples, Figures 6.44 to 6.46 show trends slightly different to those for bias. For q = 1 and 2, one of the 0-SE estimates was lowest in terms of MSE for higher R(B), with R(0.632) using the 1-SE rule having a very high MSE, while for higher R(B) (R(B) ≥ 0.25), the 1-SE estimates were more reliable. For q = 3, the 1-SE estimates were equivalent to those using the 0-SE rule for R(B) = 0.05 and better than the 0-SE for R(B) = 0.15 and 0.25. For R(B) = 0.35, the 0-SE estimates performed best.

These results would tend to suggest that sample size plays an important part in determining the choice between the 0-SE and 1-SE rules for CART, as well as if there is any noise in the data or not. Based on these results, the recommendation is to use the 0-SE rule with very small data sets for parallel classification problems involving little or no noise, while for sequential classification problems and some noise in the data, the 1-SE rule is preferred on the grounds of simplicity. For larger samples, the 0-SE rule should be used for well separated populations in cases involving parallel classification problems. For sequential classification problems, the 1-SE rule should be used unless the populations are not well separated.

Comparing the tree sizes obtained by using the 1-SE rule instead of the 0-SE rule, showed that overall, for both methods, n had a very large effect (F = 12.94) compared with all other effects and interactions. This implies that sample size was a major factor in determining if tree size decreased or not with the use of 1-SE rule. In fact, the overall increase in tree size was one terminal node larger for large n than for small n.

Investigation of the probability model stratum of the ANOVA showed there to be no significant method effect or method by factor interactions, therefore the decrease in tree size resulting from using the 1-SE rule was no different for either of the two error estimation methods. On the evidence here it would appear that tree size was little affected by using the 0-SE rule instead of the 1-SE rule, certainly less than suggested by Feng et al (1993), though for larger samples with more variables, the increase may be much greater.

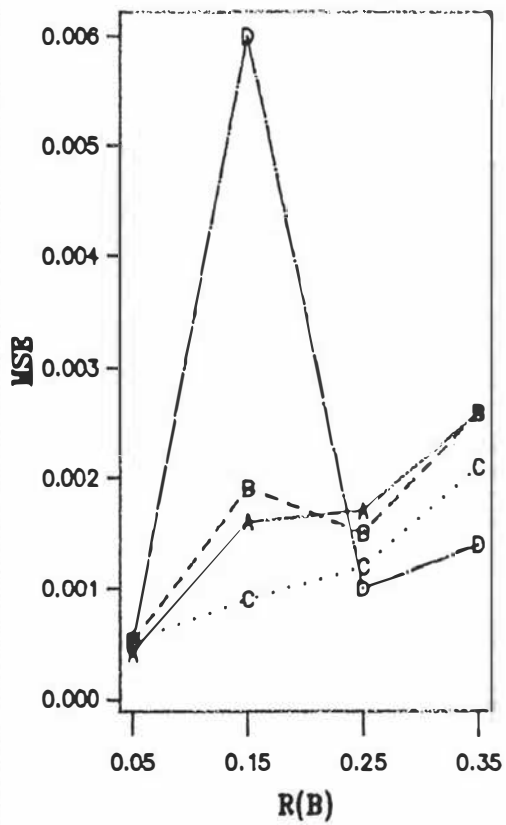
Comparison of the MSE's of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 100 and q=1



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.44

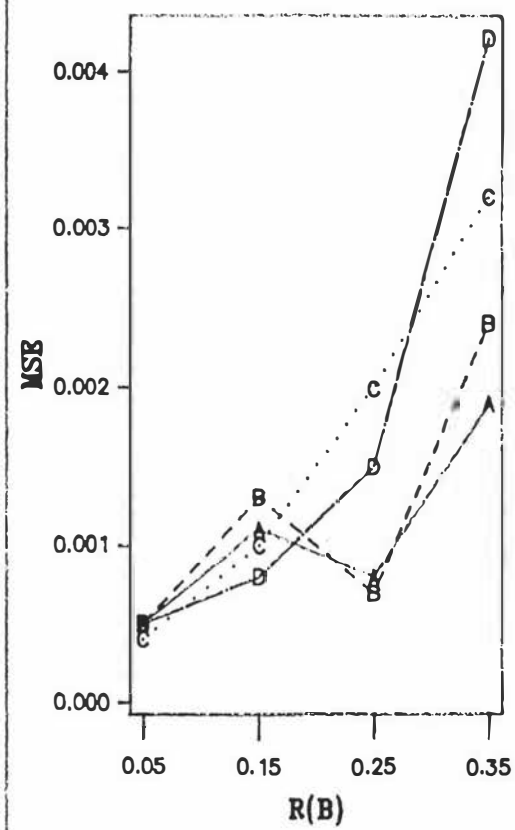
Comparison of the MSE's of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 100 and q=2



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.45

Comparison of the MSE's of Two Error Estimation Methods for CART with and without the One Standard Error Rule for Sample Sizes of 100 and q=3



A – R(CV) B – R(CV) – 1SE
C – R(0.632) D – R(0.632) – 1SE

Figure 6.46

6.4.4 Summary

The results from this particular study have indicated that when using either the R(CV) or R(0.632) estimates to both select tree size and calculate an "honest" estimate of R(T), the 1-SE rule should not be used for either small samples or when there is little noise in the data, unless the populations are not well separated. For situations where there exists a lot of noise in the data, the 1-SE rule is preferred unless the populations are not well separated.

6.5 TRANSFORMATIONS OF ERROR RATES

6.5.1 Study Plan

The previous sections have dealt with the analyses of untransformed error rates so a difference of 0.1 from R(T) = 0.05 was treated the same as a difference of 0.1 from 0.35. This seems a somewhat unfair and inappropriate comparison, as suggested by Fitzmaurice et al (1991), in that the former difference should receive more weight than the latter (see Section 4.5.3).

In this section, two transformations of the error rates were tried to try and right this imbalance, namely the logit and proportion transformations. For the logit transformation, R(T) was replaced by LR(T), where

$$LR(T) = \ln[R(T) / (1 - R(T))]$$

and its estimate, $R(\hat{T})$, by

$$LR(\hat{T}) = \ln[R(\hat{T}) / (1 - R(\hat{T}))]$$

while for the proportion transformation, R(T) was replaced by PR(T), where

$$PR(T) = (R(T) - R(T)) / R(T) = 0$$

and its estimate, $R(\hat{T})$, by

$$PR(\hat{T}) = (R(\hat{T}) - R(T)) / R(T).$$

For example, if $R(\hat{T}) = 0.05$ and $R(\hat{T}) = 0.2$

$$LR(\hat{T}) - LR(T) = \ln[0.2 / (1-0.2)] - \ln[0.05 / (1-0.05)] = 1.558$$
$$PR(\hat{T}) - PR(T) = (0.2 - 0.05) / 0.05 = 3$$

while if $R(T) = 0.35$ and $R(\hat{T}) = 0.5$

$$LR(\hat{T}) - LR(T) = \ln[0.5/0.5] - \ln[0.35/0.65] = 0.619$$
$$PR(\hat{T}) - PR(T) = (0.5 - 0.35) / 0.35 = 0.429.$$

From these results it can be seen that using the proportion transformation has the greatest effect on the error rates, as the magnitude of differences between (0.2 - 0.05) and (0.5 - 0.35) is 7 and 2.517 for the proportion and logit transformations respectively.

The two transformations were used on the R(CV), R(ROT), R(0.632) and R(TEN) error rate estimates calculated in Section 6.2 with the intention of determining what differences, if any, appeared in the MSE's for all four estimators.

6.5.2 Results

As with Fitzmaurice et al (1991), the two transformations had very similar effects on the patterns of MSE's. Therefore, only the results for the proportion transformation are demonstrated here. The results for n = 20 appear in Figures 6.47 to 6.49 and differ from the untransformed results, given in Figures 6.13 to 6.15, in a number of respects. Firstly, all estimators now exhibit the general trend of an initial sharp decrease in MSE going from R(B) = 0.05 to 0.15 then a gradual decrease from R(B) = 0.15 to 0.35. Note, though, that the MSE's for the R(0.632) estimator were least affected by changes in the values of R(B). For q = 1 and 2, the differences between the R(0.632) and other estimators for R(B) = 0.05 were accentuated. As with the untransformed scale, the R(0.632) estimator did best when q = 1 and 2, except for high R(B), while no single estimator was best for q = 3.

For larger samples, a slightly different trend than appeared with smaller samples is highlighted in Figures 6.50 to 6.52, with the initial decrease in MSE for all estimators, except R(ROT), being not as large as that for small samples and increasing MSE for R(B) = 0.35. As with smaller samples, though, the performance of each estimator is clearly defined for low R(B). R(0.632) did best, in the cases of q = 1 and 2, for moderate R(B), and high R(B) for q = 3.

6.5.3 Summary

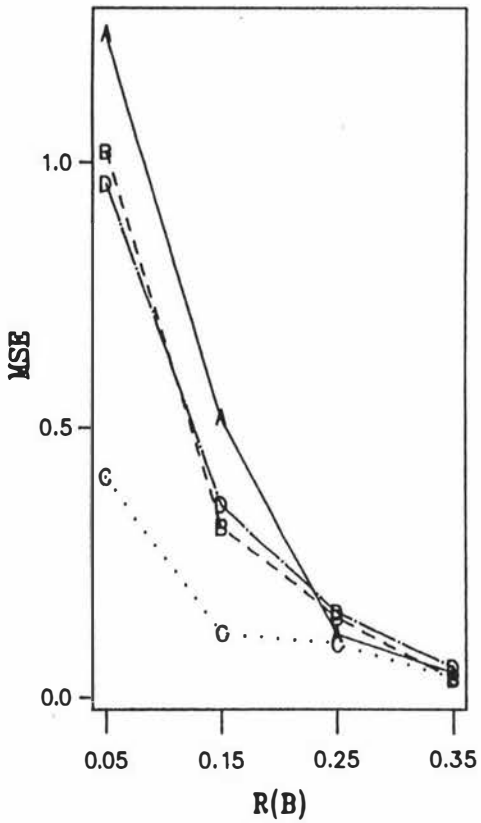
The results reported here were very similar to those given by Fitzmaurice et al (1991) using LDA. A comparison of the MSE's of the four error methods was not greatly affected by the transformations . However, as recorded by Fitzmaurice et al, the methods now performed best for high R(B) and worst for low R(B) in contrast with the untransformed results where for small samples, MSE was highest for moderate R(B), while for larger samples, it was largest for high R(B).

6 CONCLUSIONS

In this chapter, simulation study results have shown that the R(0.632) method for estimating the actual error rate when using CART performed well in most situations for both continuous and categorical data. For continuous data, the R(0.632) clearly had the lowest MSE for smaller samples and parallel classification problems and marginally lower MSE for smaller samples and sequential classification problems as well as larger samples and parallel classification problems. Only when the classification problem was sequential and the distance between populations was moderate to large did other techniques outperform R(0.632).

For categorical data, most of the trends noted above were also observed. The R(CV) estimator, as for continuous data, was found to be a poor estimator for small samples and highly optimistic for poorly separated populations. For both continuous and categorical data, the R(0.632) estimator (R(ROT)) was found to produce the smallest sized trees.

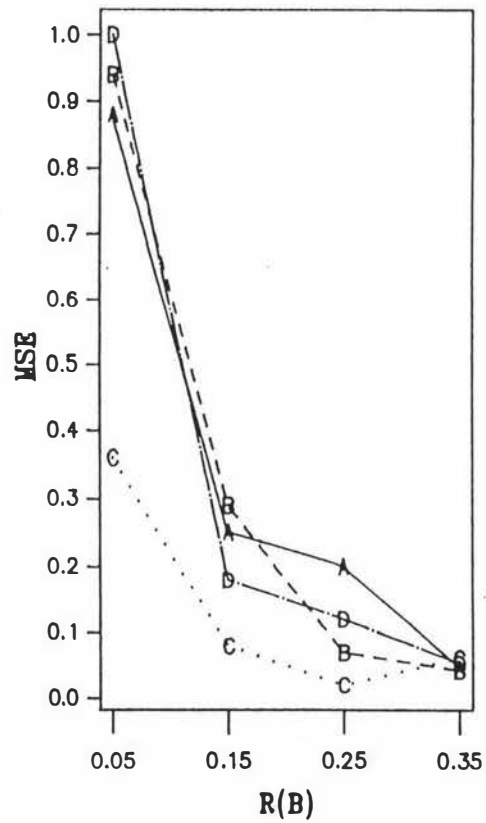
Comparison of the Transformed MSE's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=1: Proportion Transformation



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.47

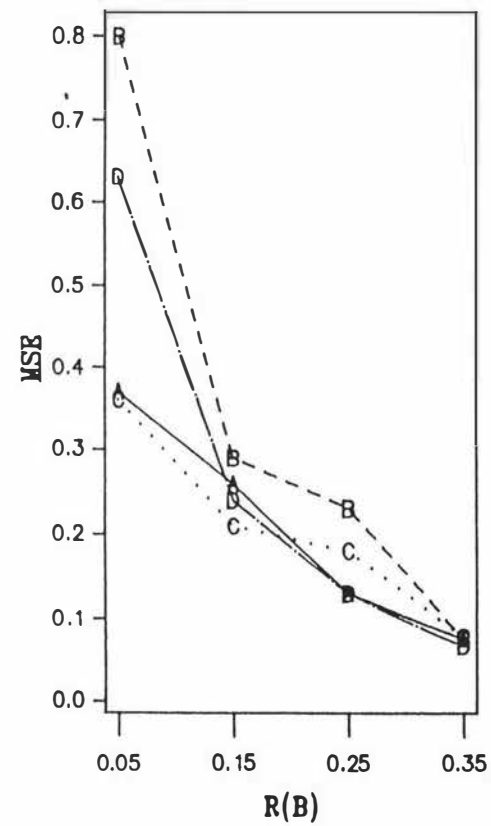
Comparison of the Transformed MSE's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=2: Proportion Transformation



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.48

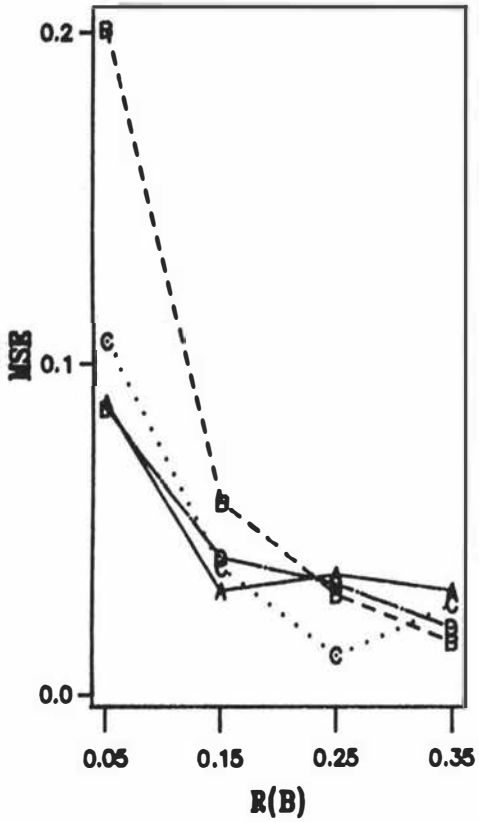
Comparison of the Transformed MSE's of Four Error Estimation Methods for CART with Sample Sizes of 20 and q=3: Proportion Transformation



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.49

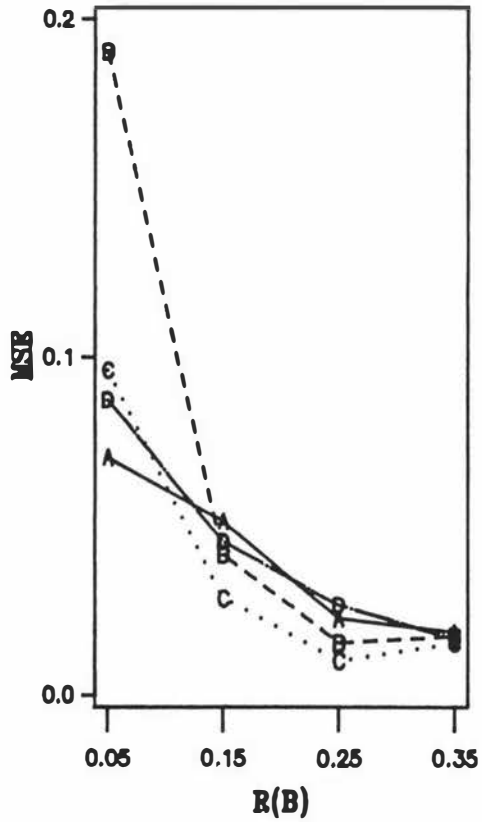
Comparison of the Transformed MSE's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=1: Proportion Transformation



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.50

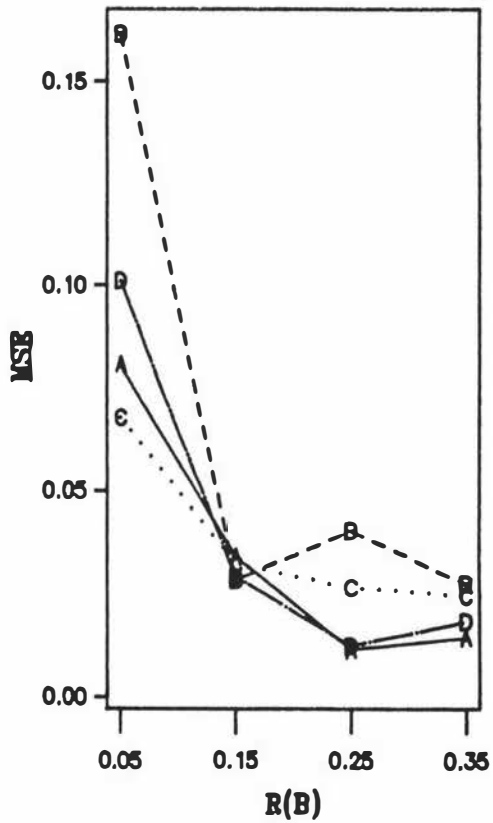
Comparison of the Transformed MSE's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=2: Proportion Transformation



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.51

Comparison of the Transformed MSE's of Four Error Estimation Methods for CART with Sample Sizes of 100 and q=3: Proportion Transformation



A – R(CV) B – R(ROT)
C – R(0.632) D – R(TEN)

Figure 6.52

In studies comparing the use of the zero and one standard error rules in CART, it was found that the one standard error rule should not be used for either small samples for when there is little noise in the data, unless the populations are poorly separated. In all other situations, the one standard error rule is the preferred method.

Finally, a transformation of the error rate scale produced results which were not unexpected. For large differences between populations, that is low Bayes error rates, the differences between error estimation techniques were accentuated from the case of untransformed error rates.

Therefore, the technique of using the rotation method to select tree size then using the 0.632 method to estimate the actual error rate of a data set is recommended as a quick, easy and reliable technique when used with CART decision trees. However, as mentioned by Crawford (1989), the user should not be constrained to using one method to select the right-sized tree, but instead, with a mixture of common sense and prior knowledge of the domain, make a sensible tree selection.

'. CASE STUDIES

.1 INTRODUCTION

In this chapter, a number of the classification methods outlined in Chapters 2 and 3, are applied to 24 real-world data sets and compared by means of a number of criteria to be outlined later in this chapter. These data sets are used to either validate or not some of the conclusions reached after the simulation studies undertaken in Chapters 4, 5 and 6.

Later in this chapter, a comparison of various tree-based methods is made for one particular data set.

.2 PREVIOUS STUDIES

Ildiko and Lanteri (1989) compared LDA, QDA, SIMCA (a form of QDA) and CART on four data sets selected from various fields of chemistry. They concluded that no overall method was superior in terms of prediction error. They also recommend that the type of data structure involved should be explored and then to choose the optimal rule for that particular type of data. If in doubt, several different methods should be used and compared.

Lynn and Brook (1991) undertook an empirical study comparing the performance of traditional discrimination methods with CART on twelve predominantly multivariate normal classification problems, differing in sample size, dimension and modality. Subsequently, it was found that for only three of the data sets was the assumption of equality of variances valid, hence, it was decided to use only LDA to compare with CART. For the other nine data sets investigated both QDA and kernel density estimation were carried out. In all cases, comparisons were made by means of n-fold cross-validation. The findings of this paper suggested that CART does not perform as well as discriminant analyses in cases where the data set is small and/or simple but does perform at least as well as discriminant analysis in most cases where the data set is larger and/or complex (multi-modal, non-normally distributed and/or high-dimensional), especially where the covariance structure is heterogeneous.

Feng et al (1993) reported a number of papers from the literature which have compared various classification methods. However, they noted a number of problems common to the papers referred to above, such as applying different methods to data sets which were not the same, and using old versions of some methods while using the latest versions of others. In their study, Feng et al as mentioned earlier, compared a large number of classification methods for eight data sets involving industrial applications. Generally, the data sets were of much larger sample size and dimension than those used by Lynn and Brook (1991). "In conclusion, it seems that there is no one particular algorithm or one particular method superior to the others on all the data sets. There is indication from our results that which algorithm performs best depends on the characteristics of the data sets. Our work is, however, incomplete in the analysis of such dependent relationships". (Feng et al, 1993, p 51).

Brown et al (1993) compared CART with neural networks, a method which uses multiple layers of processes. Each processor produces a weighted non-linear function of the variables. Their comparative studies were carried out for several multi-modal classification problems and found that the two methods produced classification rules with comparable error rates, but CART is preferred for data sets with a large number of irrelevant or noisy variables and when the ratio of sample size to dimension is small.

7.3 COMPARATIVE STUDIES

7.3.1 Methods and Data Sets

Five classification methods were used in this study, involving two categories of methods:

(1) Traditional discrimination methods, which include LDA, QDA and kernel density estimation.

(2) Decision tree-based methods, which include CART and FACT.

Twenty four data sets, described in Tables 7.1 and 7.2, were chosen for the purpose of comparison. All the data sets were a convenient selection of published data. Twelve of the data sets were used previously in a comparative study undertaken by Lynn and Brook (1991). Those data sets, however, all contained continuous variables, nine of which were approximately normally distributed. The additional twelve data sets used here contain a wider variety of data types, including some data sets involving binary and ordinal categorical variables.

Table 7.1: Description of data sets (block 1)

A. **Mammo**: This problem involves an attempt to discriminate between women's experiences with mammography (three levels) based on five variables, describing their knowledge, attitude and behaviour towards mammography.
Source: R J Zapka and Ms D Spotts, University of Massachussets, Division of Public Health.

B. **Marks1**: This involves discriminating between males and females based on their Grade Point Average at university and five pre-university academic variables.
Source: Moore and McCabe (1989).

C. **Marks2**: This involves discriminating between three groups of students with different majors on the basis of the same six variables in B.
Source: Moore and McCabe (1989).

D. **Marks3**: This involves discriminating between six groups of students with different sex and/or majoring subject on the basis of the same six variables in B.
Source: Moore and McCabe (1989).

E. **Digit**: In this example, the data are generated from a faulty calculator. Each of the seven lights (X_1, ..., X_7) of the digit display has 0.1 probability of not doing what it is supposed to do. The problem is an attempt to distinguish between the values 1 to 10, which occur with equal probability.
Source: Breiman et al (1984).

F. **Birth**: For this set of data, an attempt was made to discriminate between overweight and underweight babies based on various medical and demographic variables relating to the mother.
Source: Hosmer and Lemeschow (1989).

G. **Family**: This problem was analysed by Kumar (1993), containing information on the type of contraceptive device used by 174 Indian couples. An attempt was made to discriminate between the four types of contraceptive device based on the values of twelve variables collected from each couple.
Source: Family Planning Association of India.

H. **Iris**: This is the classic problem posed by Fisher involving discrimination between three allied species of iris based on four measurements relating to the size of the iris.
Source: Fisher (1936).

I. **Enures**: One method of treatment of enuretic children involves an alarm buzzer which wakes the child whenever a bed becomes wet. It was proposed to investigate whether the outcome of the treatment could be predicted from seven measurements, where the possible outcomes are 1 = fail, 2 = relapse after apparent cure and, 3 = long term cure.
Source: Dr Sylvia Dische (from Hand (1981)).

J. **Blood1**: In the context of genetic counselling, the question of discriminating between normal and haemophilia A carrying women was considered on the basis of two variables.
Source: Habbema, Hermans and Van Den Broek (1974).

K. **Pinetree**: This data consists of the measurements, in centimetres of 60 pinetrees which were felled in three different areas of the forest. For each tree, measurements were taken on four positions. The problem involves distinguishing between trees grown in each of the three areas.
Source: NZ Forestry Department.

L. **Wheat**: This problem involves discriminating between two varieties of wheat on the basis of six measurements taken from a sample of the two species.
Source: Indian Agricultural Research Institute, India (1972).

M. **Biomass1**: This problem involves discriminating between three different islands on the basis of the growth of spartina biomass and four different chemicals from each of the three islands.
Source: Rawlings (1988).

N. **Biomass2**: This involves distinguishing between three different types of vegetation cover on the basis of the same five variables in M.
Source: Rawlings (1988).

O. **Biomass3**: This involves distinguishing between nine different location-vegetation types on the basis of the same five variables in M.
Source: Rawlings (1988).

P. **Comp1**: Users of the University of London Computer Centre are divided into non-medical and medical users. An attempt is made to distinguish bertween the two based on the numbers of units of computing used under two different operating systems.
Source: Hand (1981).

Q. **Employ**: This problem involves discriminating between three groups of countries (North-Western, Southern and Eastern Europe respectively) on the basis of the percentages of the labour force employed in nine different types of industry.
Source: Euromonitor (1979).

R. **Urinary**: This problem involves discriminating between homosexual and heterosexual males on the basis of two chemical measurements taken from urinary samples.
Source: Margolese (1970).

Table 7.2: Description of data sets (block 2)

S. **Blood2:** This is the same problem as J except that the two discriminating variables are logged (base 10).

T. **Lingual:** This problem involves discrimination between a set of 27 children who had an inborn error of metabolism known as transient neonatal tyrosinemia (TNT) and a control group of 27 normal children based on the scores of ten psycholingual variables.
Source: Peter Mullins.

U. **Sparrow:** This problem involves discriminating between sparrows that did or did not survive a severe storm off Rhode Island on 1 February 1889, on the basis of five measurements taken on each bird.
Source: Bumpus (1898).

V. **Comp2:** This is the same problem as P except that the two discriminatory variables are logged (natural log).

W. **Beetle:** In this case, an attempt is made to distinguish between two allied species of flea beetles that were long confused with one another, on the basis of two joint measurements.
Source: Lubischew (1962).

X. **Nuclear:** This data involves two measurements (population and area) on each of the fifteen largest British cities, excluding London. The fifteen cities used are divided into two classes; those with an estimated fatality rate of 70% or more resulting from a nuclear strike and those with an estimated rate of less than 70%.
Source: Laurie (1979).

The data sets were first sorted into two blocks after a Chi-squared test for heterogeneity of variance within groups was carried out for each data set, and where heterogeneity was present to a significant amount, those data sets were assigned to the first block (Table 7.1), otherwise to the second (Table 7.2). Within each of the two blocks of comparison methods, the data sets were ordered by sample size. As some data sets within each block were of similar size, those data sets were ordered by dimension.

Table 7.3 lists some details about each of the data sets, including sample size, dimension, number of classes, Chi-squared test for the equality of class covariance matrices, equal priors or not, variable types and data structure. Variable type refers to whether the variables in the data set are continuous (normal (N)/skewed (S)), ordinal (O), nominal (C), binary (B) or a mixture of the above five types. Data structure refers to how many of the variables in the data were important for the classification process and how many were irrelevant. A data set could be described as either a parallel classification problem, whereby all of the variables have approximately equal weighting, or as a sequential classification problem where relatively few variables are important. As outlined in Section 4.3, traditional discrimination methods should perform best for parallel classification problems with tree based methods doing better for sequential problems. A third category "mixed" was also used for problems where a particular data set did not fit neatly into any of the above two categories. Both stepwise discriminant analysis and CART's varible ranking technique were employed to determine into which category each of the twenty four data sets should be classed.

For each data set, priors proportional to class sample sizes were used, and for all methods, with the exception of FACT, models were obtained which minimised the misclassification error rate by n-fold cross-validation, although LDA and CART were also compared using the 0.632 error rate (see further on in this section). Tenfold cross-validation was used with FACT for reasons outlined in Section 4.3. With kernel density estimation, a normal kernel was used with smoothing parameter, $h = 0.5$. For both CART and FACT, the size below which a node would not be split on was set to 5. It was decided to use the one standard error rule throughout for CART for the purpose of consistency, although simulation results in Section 6.4 had shown that the use of the zero standard error rule would be a more reliable estimate of the actual error rate for smaller data sets.

Table 7.3: Data sets and their various characteristics

Data Set	Sample Size	Dimension	Classes	χ^2	Equal priors?	Variable type(s)	Data structure	Best Method
A	374	5	3	155.94**	No	O, B	parallel[a]	LDA
B	234	6	2	58.80**	Yes	S	parallel	CART
C	234	6	3	107.62**	Yes	S	parallel	CART
D	234	6	6	294.29**	Yes	S	mixed[c]	**KDE**
E	200	7	10	2105.08**	No	B	parallel	LDA
F	188	8	2	42.76*	No	N, S, C B	sequential[b]	CART
G	174	12	4	1108.47**	No	N, O, C, B	mixed	LDA
H	150	4	3	154.42**	Yes	N	parallel	**KDE/QDA**
I	112	7	3	2363.75**	No	B	sequential	LDA
J	75	2	2	15.90*	No	S	parallel	**KDE/QDA**
K	60	4	3	92.12**	Yes	N	parallel	**LDA**
L	54	6	2	43.23**	Yes	N	mixed	**LDA**
M	45	5	3	135.50**	Yes	N	mixed	**KDE**
N	45	5	3	107.81**	Yes	N	parallel	CART
O	45	5	9	400.55**	Yes	N	parallel	KDE
P	49	2	2	101.51**	No	S	mixed	FACT
Q	26	9	3	193.35**	No	N	mixed	**KDE**
R	26	2	2	11.22*	No	N	parallel	**LDA**
S	75	2	2	5.24	No	N	parallel	**KDE**
T	54	10	2	52.27	Yes	N	sequential	FACT
U	49	5	2	0.69	No	N	parallel	CART
V	49	2	2	3.92	No	N	mixed	QDA
W	36	2	2	2.15	Yes	N	parallel	**LDA/KDE**
X	15	2	2	9.30	No	N, S	mixed	**KDE**

χ^2 Chi-squared test for homogeneity of the within class covariances
N Normally distributed variables
S Skewed variables
O Ordinal variables
C Nominal variables
B Binary variables
** $\alpha < 0.01$
* $\alpha < 0.05$
a Refers to a problem where most of the variables are important in forming the classifier.
b. Refers to a problem where relatively few variables are important.
c. Refers to a problem not fitting neatly into one of the above two categories.
d. The method producing the lowest error rate in Table 7.4. Where one of {LDA, QDA, kernel density estimation (KDE)} had an error rate at least 33% less than one of {CART, FACT}, or vice-versa, that method is shown in bold.

7.3.2 Cross-Validation Error Rate Results

The results for the twenty four data sets analysed by each of the five methods are shown in Table 7.4. Relating the performance of each method to the characteristics of the data sets provides some explanations for the results.

In terms of covariance structure, it can be seen that CART produced the lowest cross-validated error rate on four of the data sets with heterogeneous covariance structures, but only in one out of the six data sets with homogeneous covariance structures, and that being the sparrow data set (U), which is a trivial case since no tree was formed and all class 1 objects were classified as belonging to class 2. For most of the other fourteen variables with heterogeneous covariance structures, CART also performed relatively well with respect to the best classification method, though with a number of exceptions.

There appears to be little common pattern in the results when related to either dimension, sample size or modality. Nor did whether class sample sizes were equal or not have any real influence on the comparison between methods, or if they did, the effects were tied up with other factors. More important were the types of variables in the data set, how well separated were the classes (error structure) and the structure of the data.

With respect to the types of variables first, the results show that CART did well relative to the other methods when at least some of the variables were skewed, categorical or a mixture of data types, with the relative performance declining for normally distributed data. In accord with simulation studies in Sections 4.3 and 4.4 and/or critiques from the literature in Section 2.7, QDA, and to a lesser extent, kernel density estimation were not suited to categorical data, though handling skewed continuous data fairly well. The performance of LDA markedly deteriorated for skewed data, as did FACT, whose error rates were on the whole higher than those of CART as expected from the simulation study results.

With reference to the inherent error structure in the data, it is apparent that CART's performance relative to other methods was best for high error models (greater than 0.2), but for lower error models, worked much better than both CART and FACT. These trends support the conclusions made from the simulation studies for both continuous and categorical data in Chapters 4 and 5 respectively.

Table 7.4: A summary of cross-validation error rates for data sets A-X

Data Set	Method				
	LDA	QDA	KDE[a]	CART	FACT
A	**0.37**	0.41	0.39	0.38(8)[b]	0.42(4)
B	0.30	0.31	0.23	**0.22**(31)	0.34(2)
C	0.53	0.52	0.30	**0.25**(40)	0.49(7)
D	0.55	0.61	**0.35**	0.45(62)	0.56(21)
E	**0.27**	0.60	0.29	0.34(10)	0.34(10)
F	0.29	0.34	0.30	**0.28**(2)	0.36(2)
G	**0.20**	0.27	0.26	0.21(4)	0.21(3)
H	0.07	**0.03**	**0.03**	0.05(3)	0.04(3)
I	**0.37**	0.48	0.43	0.41(5)	0.46(9)
J	0.16	**0.13**	**0.13**	0.20(3)	0.30(3)
K	**0.08**	0.12	0.12	0.41(4)	0.45(7)
L	**0.07**	0.11	0.15	0.13(3)	0.17(6)
M	0.11	0.09	**0.07**	0.15(6)	0.13(9)
N	0.29	0.29	0.31	**0.22**(3)	0.33(6)
O	0.38	N/A	**0.31**	0.34(12)	0.47(22)
P	0.24	0.24	0.24	0.30(2)	**0.22**(2)
Q	0.27	N/A	**0.04**	0.08(3)	0.19(5)
R	**0**	0.04	0.04	0.16(3)	0.23(3)
S	0.17	0.17	**0.12**	0.20(3)	0.16(4)
T	0.48	0.48	0.48	0.50(1)[c]	**0.43**(2)
U	0.45	0.57	0.57	**0.43**(1)[c]	**0.43**(1)[c]
V	0.31	**0.20**	0.22	0.30(2)	0.24(4)
W	**0.03**	0.06	**0.03**	0.14(2)	0.11(4)
X	0.20	0.27	**0.13**	0.40(2)	0.47(1)[c]

a KDE is kernel density estimation

b The number in parenthesis indicates the number of terminal nodes in the decison tree

c No trees were created in these cases

N/A QDA was not able to be carried out as at least one of the class covariance matrices was not of full rank

Finally, in reference to the data structure, one of the tree-based methods has produced the lowest error rate for two (F, T) of the three data sets which were described as sequential. For the other sequential classification problem (I), CART had the second lowest error rate. Interestingly, for most parallel and mixed classification problems, either LDA, QDA or kernel density estimation produced the lowest error rates, although CART did best for three such data sets. One could draw the conclusion from these empirical results that tree based methods are preferred for sequential classification problems. Unfortunately, there were not more data sets of this type to lend more weight behind this assertion.

In comparing the decision tree sizes of the two tree-based methods, some interesting results are noticeable from Table 7.4. As Breiman et al (1984) point out, tree size is negatively related to the reliability of the classification model, in that smaller trees are heavily biased in favour of the learning sample. The general trend exhibited here is that CART tended to produce the larger trees for larger data sets (I-X), but for larger data sets (A-H), the FACT trees were never larger than those produced by CART, and in some cases were considerably smaller. The most striking examples of the latter situation were for the marks data sets (B-D) where CART produced excessively large trees, which, in two cases at least, led to the most accurate set of prediction rules. With such large trees, however, one should not have too much confidence in the resulting set of classification rules.

In general, it appears that CART should best be used for problems involving either skewed or categorical data, where the classes are not well separated and only a few out of many variables are important in the classification process. In other situations, LDA and/or QDA/kernel density estimation are preferred. It must be noted, in passing, that there are always exceptions to the rule and the above recommendations should not be regarded as "set in stone". The biomass2 data set (N) involving discrimination between three types of vegetation provides an example of normally distributed data, where all the variables are important and the classes are at least moderately separated, and in which CART has produced the most accurate set of classification rules.

7.3.3 0.632 Error Rate Results

The 0.632 error rates were also calculated for the twenty four data sets using equation (4.2.20), for both LDA and CART. The results of these analyses are shown in Table 7.5. Overall, the results are not drastically different from those found using n-fold cross-validation, in that CART worked well relative to LDA for data sets with heterogeneous covariance structures for lognormal data and higher error models. LDA had the lowest error rate for two of the three sequential classification problems, but some other characteristics of these data sets were perhaps influencing these results. When compared with the cross-validation error rates for LDA and CART in Table 7.4, it appears that the cross-validation error rates for both methods were higher than the corresponding 0.632 error rates for a majority of the data sets, with a relatively large proportion of cross-validated error rates being higher for smaller sized data sets. Otherwise, there appears to be no patterns in the data which determine what error rate will be lower than the other. In comparing tree sizes, the trees produced using the 0.632 error rate were smaller than those produced by using the cross-validation error rate, in the main, which tends to suggest that the classification trees produced by using the former estimate were more reliable than the latter. The results for tree size correspond to the findings of the simulation study in Section 6.2.

7.3.4 Individual Class Error Rates

In Section 4.5, simulation results suggested that CART was more sensitive to unequal class sample sizes than LDA (and indeed QDA) for continuous data. Table 7.6 gives the individual class error rates for the thirteen data sets with unequal class sample sizes in an attempt to verify the above assumptions. The results show that the above findings hold for nine of the thirteen data sets. The instances where the class error rates for CART were less variable than those for LDA, occurred for either categorical variables (A and E) or for skewed variables (P and X), with simulation results showing that the error rates for LDA suffered under the latter situation. In addition, it should be noted that the ratios of class sample sizes were not too dissimilar, except A, that is $n_i < 2n_j$ where n_i is the sample size for the class with the largest number of observations and n_j is the sample size for the class with the smallest number of observations.

Table 7.5: 0.632 error rates for LDA and CART

Data Set	Method	
	LDA	CART
A	**0.37**	0.39(7)[a]
B	0.28	**0.20**(15)
C	0.49	**0.28**(26)
D	0.62	**0.41**(36)
E	**0.30**	0.32(10)
F	0.31	**0.29**(2)
G	**0.20**	**0.20**(2)
H	**0.05**	**0.05**(3)
I	**0.37**	0.48(1)[b]
J	**0.15**	0.23(2)
K	**0.08**	0.28(8)
L	**0.08**	0.12(3)
M	**0.13**	0.14(5)
N	0.26	**0.24**(3)
O	**0.22**	0.30(7)
P	0.21	**0.20**(2)
Q	0.29	**0.22**(2)
R	0.23	**0.19**(3)
S	**0.16**	0.23(2)
T	**0.39**	0.50(1)[b]
U	0.45	**0.43**(1)[b]
V	0.31	**0.20**(2)
W	**0.03**	0.17(2)
X	**0.13**	0.33(3)

a The number in parenthesis indicates the number of terminal nodes in the decision tree

b No trees were created in these cases

Table 7.6: Class error rates for data sets with unequal class sample sizes

Data Set	Method	R(i/j)									
		1	2	3	4	5	6	7	8	9	10
A	LDA	0.49	1	0.12							
	CART	0.53	0.92	0.16							
E	LDA	0.06	0.31	0.30	0.25	0.23	0.29	0.21	0.41	0.42	0.10
	CART	0.17	0.42	0.25	0.42	0.38	0.29	0.26	0.41	0.46	0.30
F	LDA	0.73	0.15								
	CART	0.69	0.09								
G	LDA	0.14	0.13	0.86	1						
	CART	0.27	0.68	1	1						
I	LDA	0.44	0.47	0.28							
	CART	0.50	0.53	0.31							
J	LDA	0.20	0.13								
	CART	0.13	0.31								
P	LDA	0.39	0.05								
	CART	0.43	0.14								
Q	LDA	0.15	0.60	0.13							
	CART	0	0.4	0.13							
R	LDA	0	0								
	CART	0.27	0.13								
S	LDA	0.17	0.16								
	CART	0.17	0.31								
U	LDA	0.76	0.21								
	CART	1	0								
V	LDA	0.25	0.38								
	CART	0.43	0.14								
X	LDA	0.38	0								
	CART	0.50	0.29								

R(i/j) = proportion of class j observations misclassified as class i, $i \neq j$.

7.3.5 The Standard Error Rule in CART

In the comparative study with other methods in Section 7.3.2, the 1-SE rule was used in CART to select the right sized tree. In Section 6.4, it was found through simulation studies that the 1-SE rule is inappropriate for smaller samples or when there is little noise in the data, unless the populations are not well separated. For situations where there exists a large amount of noise in the data, the 1-SE rule should be used. Hence, it was decided to analyse the error rates and tree sizes of the classification rules produced by using both the 0-SE and 1-SE rules for CART over all twenty four data sets. The 0-SE results are given in Table 7.7. Comparing these results with the 1-SE results in Tables 7.4 and 7.5, the empirical evidence suggests that tree sizes were not greatly affected by either rule, though, as expected, tree sizes for the larger data sets were somewhat reduced, while those for the smaller data sets remained basically unchanged, as evidenced in Section 6.4. The small number of sequential classification problems encountered here, situations in which tree size should be greatly reduced, makes it very difficult to reach any firm conclusions about the differences between the two rules.

7.3.6 Splus Trees( ) versus CART

Since Splus Trees() is basically the incorporation of the CART method into the Splus programming environment, the trees produced by Splus should be of roughly comparable size to those produced by CART, with similar error rates. The most obvious differences between the two methods are that Splus Trees() uses deviances as a measure of goodness of split in contrast with the misclassification error rate criterion used by CART, in addition to having a shrinking algorithm as well as a pruning algorithm, both of which are based on deviances. In order to test whether there are any differences between the two methods, Splus Trees() was carried out on all 24 data sets, using both optimal shrinking and cost-complexity pruning. Both n-fold cross validation error rates as well as tree sizes were recorded and compared with CART. The CART results were the same as those used in Table 7.4, that is, for trees using the 1-SE rule.

Bradford (1993) has compared Splus Trees() with CART using the data sets from Lynn and Brook (1991). Bradford, however, has only constructed trees using cost-complexity pruning and has used tree size as the sole measure of comparison between the two methods.

Table 7.7: Error rates for CART using the zero standard error rule

Data Set	Error Rate	
	Cross-Validation	0.632
A	0.38(8)[a]	0.39(7)
B	0.22(31)	0.17(31)
C	0.25(40)	0.28(26)
D	0.45(62)	0.41(36)
E	0.34(21)	0.32(10)
F	0.28(2)	0.29(2)
G	0.18(6)	0.17(4)
H	0.05(3)	0.04(4)
I	0.41(5)	0.45(3)
J	0.25(2)	0.23(2)
K	0.38(15)	0.28(8)
L	0.11(6)	0.09(6)
M	0.13(7)	0.09(6)
N	0.22(3)	0.22(4)
O	0.31(14)	0.30(7)
P	0.30(2)	0.20(2)
Q	0.08(3)	0.13(3)
R	0.19(3)	0.19(3)
S	0.25(2)	0.23(2)
T	0.48(2)	0.50(1)[b]
U	0.43(1)[b]	0.43(1)[b]
V	0.30(2)	0.20(2)
W	0.14(2)	0.17(2)
X	0.33(3)	0.33(3)

a The number in parenthesis indicates the number of terminal nodes in the decision tree

b No trees were created in these cases

The results are given in Table 7.8. The empirical evidence suggests that the final decision trees created by Splus Trees() were not always the same as those built by CART. In fact, for only 13 out of the 24 data sets were the error rates the same. For those examples where there were differences between the two methods, the CART trees were generally larger and had lower error rates than the Splus trees.

The evidence also suggests that sample size and covariance structure are major factors in determining whether CART and Splus Trees() produce the same set of decision rules or not. CART's trees were noticeably less succinct than Splus for larger samples, as well as for data sets where the covariances were not equal. Therefore, it could be concluded that Splus provides shorter trees than CART with more conservative error rates.

Comparing the shrunken and pruned trees created by Splus, there seems to be only minor differences in tree sizes and error rates. Clark and Pregibon (1991) believe that optimally shrunk trees have lower error rates than pruned trees, but are correspondingly larger. The case studies given here produce a number of counter examples to this assertion, most notably for data sets B, C and E. This shows that the choice of either optimal shrinking or cost-complexity pruning should not influence the final Splus tree to a large extent.

7.3.7 Summary

In this section, five classification methods from the fields of both traditional discrimination and tree-based methods were compared over twenty four real data sets. The cross-validation results showed that, to a considerable degree, which method performed best depended on the characteristics of each data set. CART worked well for either categorical or skewed data, poorly separated classes and where only a small proportion of the variables in the data set were important in the classification process. Otherwise, traditional data discrimination methods worked best. These findings are in general agreement with the simulation study results of Chapter 4.

On most occasions, CART trees were found to be smaller than FACT trees for smaller data sets while FACT trees were never larger than those of CART for larger data sets.

When using 0.632 error rates, it was discovered that very little differences occurred from the cross-validation results when using LDA and CART, though the cross-validation error rates tended to be higher than the corresponding 0.632 error rates, especially for smaller data sets.

Table 7.8: Cross-validation error rates and tree sizes for CART and Splus Trees()

Data Set	Method		
	CART	ST-OS[a]	ST-TP[b]
A	**0.38**(8)[c]	0.43(2)	0.43(2)
B	**0.22**(31)	0.28(12)	0.30(3)
C	**0.25**(40)	0.42(14)	0.34(20)
D	**0.45**(62)	0.65(15)	0.58(23)
E	**0.34**(10)	**0.34**(10)	**0.34**(15)
F	**0.28**(2)	0.31(1)[d]	0.31(1)[d]
G	**0.21**(4)	**0.21**(4)	**0.21**(4)
H	**0.05**(3)	0.06(5)	0.06(4)
I	**0.41**(5)	0.48(3)	0.48(2)
J	**0.20**(3)	**0.20**(3)	0.25(2)
K	**0.41**(4)	0.67(3)	0.67(3)
L	**0.13**(3)	**0.13**(3)	**0.13**(5)
M	**0.15**(6)	0.36(4)	0.25(5)
N	**0.22**(3)	**0.22**(3)	**0.22**(3)
O	**0.34**(12)	0.79(5)	0.59(7)
P	**0.30**(2)	**0.30**(2)	**0.30**(2)
Q	**0.08**(3)	**0.08**(3)	0.19(2)
R	**0.16**(3)	**0.16**(3)	**0.16**(3)
S	**0.20**(3)	**0.20**(3)	**0.20**(2)
T	**0.50**(1)[d]	**0.50**(1)[d]	**0.50**(1)[d]
U	**0.43**(1)[d]	**0.43**(1)[d]	**0.43**(1)[d]
V	**0.30**(2)	**0.30**(2)	**0.30**(2)
W	**0.14**(2)	0.17(4)	0.17(4)
X	**0.40**(3)	**0.40**(3)	**0.40**(3)

a ST-OS is Splus Trees() using optimal shrinking.

b. ST-TP is Splus Trees() using cost-complexity tree pruning.

c. The number in parenthesis indicates the number of terminal nodes in the decision tree.

d. No trees were created in these cases.

Trees produced by using the 0.632 error rate were generally smaller than those constructed using the cross-validation error rate, supporting simulation results.

Also in accord with simulation results, CART was found to be more affected by unequal class sample sizes than LDA, except perhaps for skewed and categorical data.

The empirical results given here appear rather inconclusive as to the choice of standard error rule to use in CART. Simulation study results suggested that sample size and the amount of noise in the data were two determining criteria in such a choice, though only some evidence appears here to support both those assertions.

In comparing CART with Splus Trees(), it was found that the two methods produce fairly similar sized trees with not dissimilar error rates, though CART's trees were noticeably larger with lower error rates for both larger samples and where the class covariance matrices were not equal. The empirical evidence also pointed to trees produced by using either cost-complexity pruning or optimal shrinking being generally of a similar size with comparable error rates.

4 ILLUSTRATIVE CASE STUDY

7.4.1 Methods and Data

The data in this study was collected from the Family Planning Association of India (FPAI), Lucknow (UP) branch. This data (used as data set G in Section 7.3) contains the information on all the family planning cases done at FPAI during 1990. Kumar (1993) hopes that the data, especially the analysis done on it, will be of some use to the motivators and policy makers of India and help in developing the promotional strategies for various family planning devices, and hence, manpower and resources can be allocated accordingly.

There are four types of family planning devices measured in this study; IUD (56), Tubectomy (103), foam tablets (7) and oral pills (8), with Tubectomy being the only terminal device in nature. The figures in parenthesis represent the number of couples that used each family planning device. Information on twelve socio-economic and demographic variables for each of the 174 couples who accepted the use of the one of the four devices, was also collected.

The twelve variables were:

1.	Wife_Age	Age of wife
2.	Husb_Age	Age of husband
3.	Husb_Edu	Education of husband, levels 0-7
4.	Wife_Edu	Education of wife, levels 0-7
5.	Occupn	Occupation of husband (service, business, farming, labour)
6.	Income	Household income
7.	No_Child	Total number of living children
8.	No_Males	Total number of male living children
9.	No_Femal	Total number of female living children
10.	Age_Baby	Age (in months) of youngest child
11.	Urb/Rur	Whether an individual belongs to an urban or a rural background
12.	Religion	Hindu, Muslim, Christian, Sikh

Thus there was a mixture of continuous, ordinal and nominal categorical variables. Note, too, that a previous study by Kumar and Srivastava (1989) had found all these variables to be significant while analysing the profile of those couples that accept family planning.

7.4.2 Linear Discriminant Analysis

In order to be able to carry out LDA on this data set, it was first necessary to transform the two nominal categorical variables, each having four levels, into two batches of three binary variables. Considering the relatively large number of variables involved in this problem ($p = 16$, with ten untransformed and six binary variables), it was decided to use stepwise LDA (SDA), with stepwise selection of the best q variables, using an $\alpha = 0.15$ significance level to enter variables to or delete from the model. A summary table of the order they were entered in the model is given in Table 7.8.

Table 7.8: Stepwise discriminant variable selection for the family data

Step	Variable Entered	Removed	Number In	Partial R^2	F Statistic	Prob > F
1	No_Child		1	0.446	45.69	0.00
2	Wife_Edu		2	0.191	13.27	0.00
3	Husb_Edu		3	0.158	10.48	0.00
4	Age_Baby		4	0.072	4.31	0.00
5	Income		5	0.037	2.13	0.09

As there were k = 4 classes in the data set, there were four group classification functions, $\hat{L}_i(\mathbf{x})$, and six group separation functions, $D_{ij}(\mathbf{x})$, created. The four group classification functions using SDA are shown in Table 7.9. Priors were set proportional to sample size (ppss).

Table 7.9: Group classification functions, $\hat{L}_i(x)$, using SDA for the family data

	$\hat{L}_1(x)$	$\hat{L}_2(x)$	$\hat{L}_3(x)$	$\hat{L}_4(x)$
Constant	-8.3951	-11.0312	-11.9301	-12.5852
Husb_Edu	-0.1483	0.4608	-0.2586	-0.3092
Wife_Edu	1.7542	0.4235	1.3594	2.0982
Income	0.0008	0.0013	0.0024	0.0001
No_Child	3.0541	4.7036	3.6580	3.5821
Age_Baby	0.0212	0.0531	0.0360	0.0553

Both the n-fold cross-validation and 0.632 error rates were calculated for the SDA classification rules above, with R(CV) = 0.195 and R(0.632) = 0.177. The corresponding values for LDA using all sixteen variables were R(CV) = 0.204 and R(0.632) = 0.194, showing that the stepwise model was more accurate. As numerous authors in the field of stepwise discrimination and the closely related topic of stepwise regression have pointed out, the best q variables found from the original sample may not be the best variables over the whole population of values. For instance, the five selected variables here may not necessarily be the most important variables for couples throughout all India seeking family planning advice.

Visually, the classification model is very hard to interpret. As the SDA group separation functions involve five variables, it is impossible to depict the full classification model. Figure 7.1 shows a three-dimensional graph of the three most important variables found during the stepwise selection process, that is, No_Child, Husb_Edu, and Wife_Edu. [Key to Figure 7.1: Club = IUD; Star = Tubectomy; Balloon = foam tablets; Diamond = oral pills.] From Figure 7.1 as well as the table of coefficients in Table 7.9, it is apparent that those who used IUD were characterised as having wives with a higher level of education and a small number of children. Those cases where the wife had little or no schooling led to the use of Tubectomy, while those cases where the wife had a higher level of education and a larger family also used Tubectomy. Distinguishing characteristics for the other two groups are not particularly relevant for these three variables.

As the class sample sizes were drastically different, it was decided to do another analysis using equal priors. The only difference that occurred from Table 7.9 is that the constants have change as evidenced from Section 4.5. Thus, each classification function changed by only one term after alteration of the priors.

7.4.3 CART

In contrast with LDA, CART used all twelve variables to perform the analysis. The CART tree, using PPSS, is as shown in Figure 7.2 (see Section 3.2 for a description of the CART tree analysis). To build this tree, the Gini splitting criterion was employed and the minimum node size was set at 5. The 1-SE rule was used to select the "right sized" tree. The twoing splitting criterion was also tried, but, in this instance, produced the same tree as that using Gini, though this is not always the case.

From Figure 7.2, it is clear that those cases where the wife had little or no schooling led to he use of Tubectomy. If the wife had a higher level of education and the husband was 32 or younger, then IUD was the predominant device used. Of those cases not already classified, husband's education was the final splitting variable used. Those cases where the husband had little or no education used Tubectomy in the main while those left over were more than likely to use IUD. For this tree, R(CV) = 0.213 and R(0.632) = 0.205.

Relationship of Wifes Education, Husbands Education and Number of Children in the Family to Contraceptive Device Used

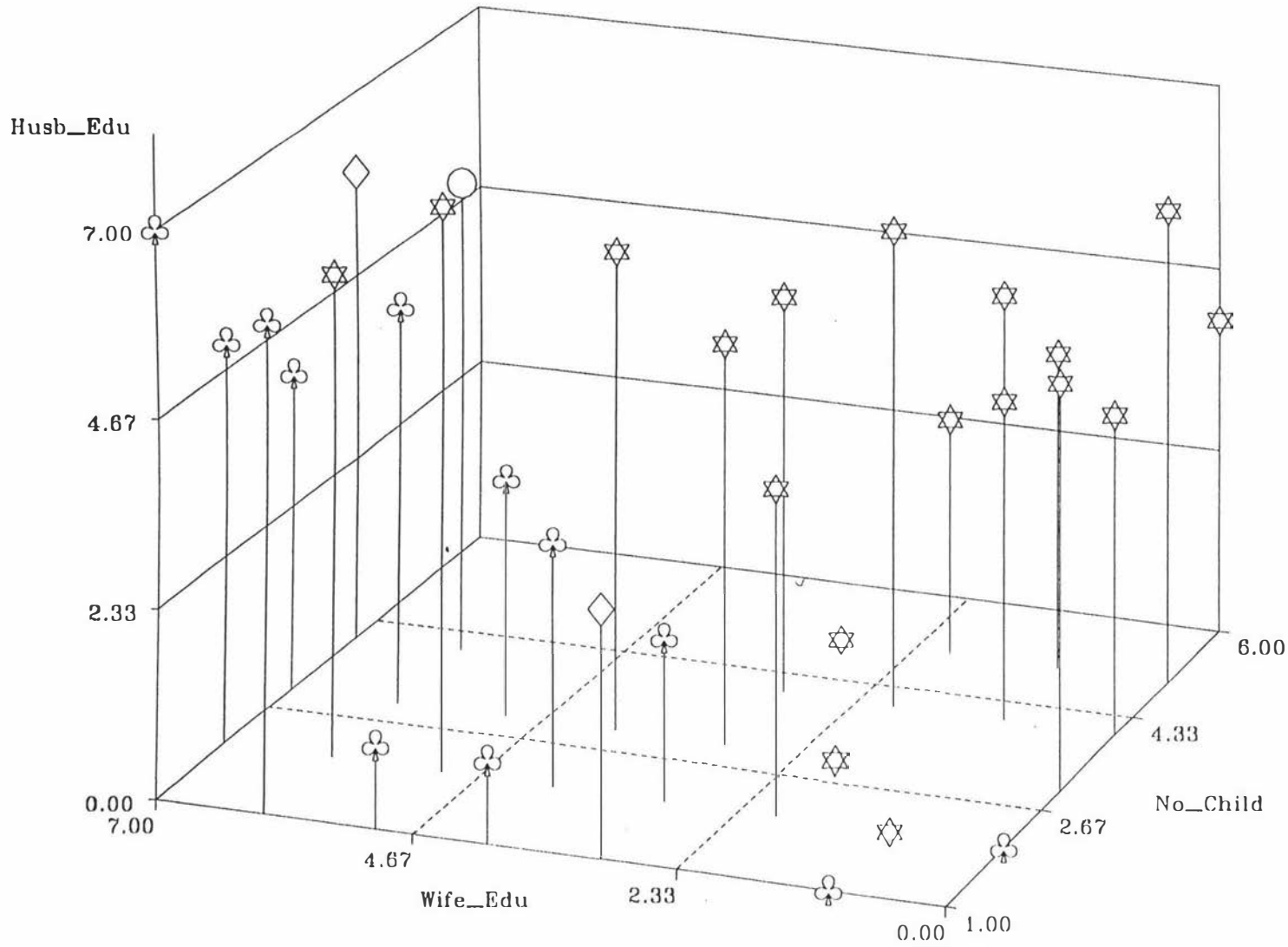
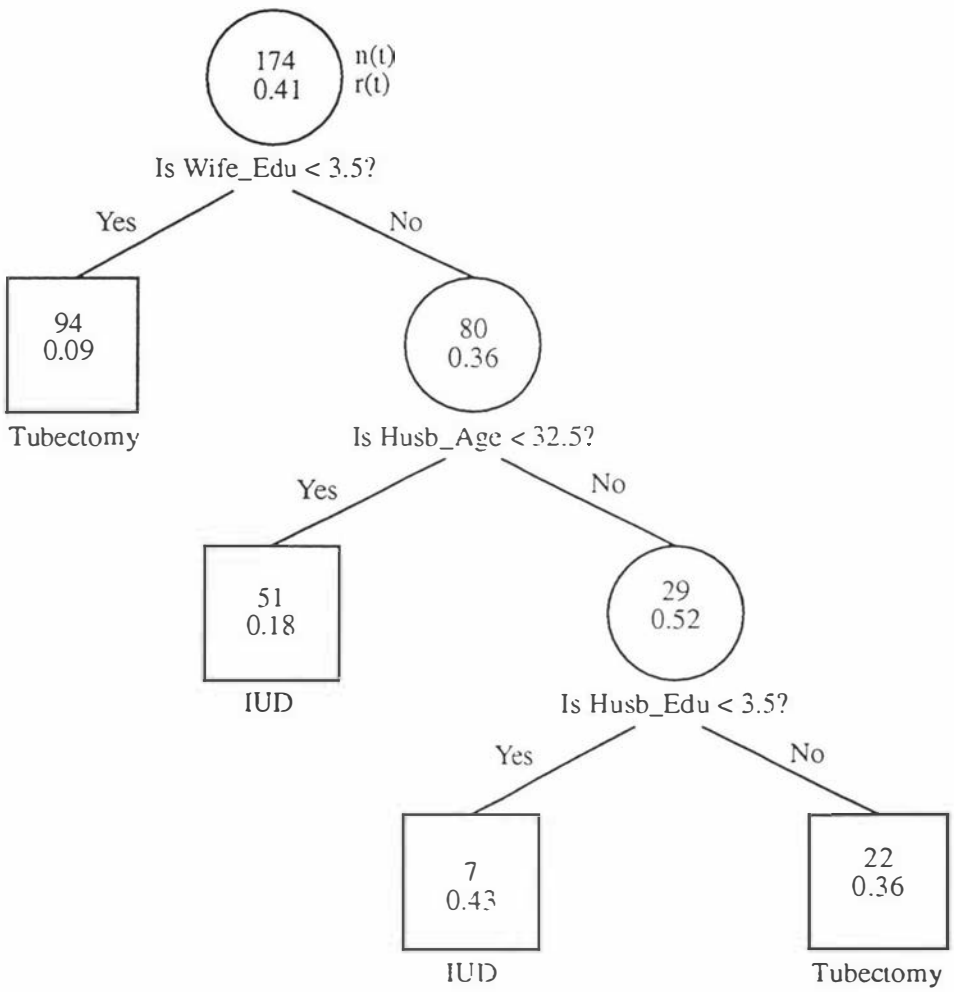


Figure 7.1

Club=IUD; Star=Tubectomy; Balloon=foam tablets; Diamond=oral pills



Variable	Relative Importance
No_Child	100
Wife_Edu	78
Husb_Age	63
No_Males	62
Wife_Age	60

n(t) = number of observations in the node.

r(t) = resubstitution error rate of the node.

Circles represent decision nodes which have to be split on while rectangles represent terminal nodes which are assigned to a particular class given below the node.

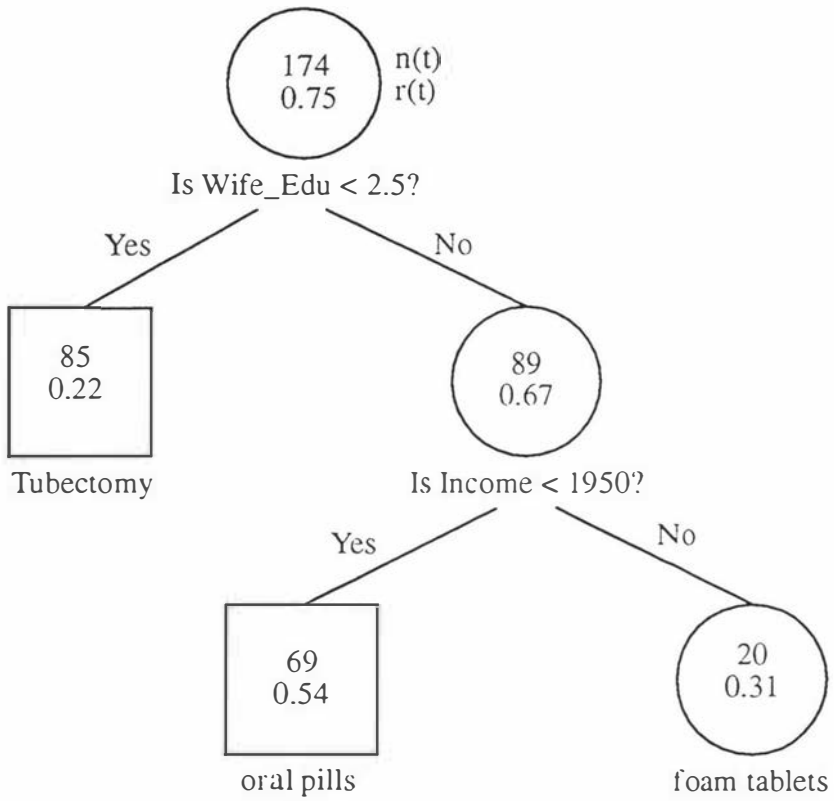
Figure 7.2: CART Tree with PPSS for the Family Data

It is clear from Figure 7.2 that there were a relatively small number of cases who used either foam tablets or oral pills, but were all classified as using either IUD or Tubectomy. This would not be a very good situation if these women were going to have adverse reactions when using either IUD or Tubectomy. In an attempt to counter this, another CART tree was grown, this time using equal priors and is given in Figure 7.3. The CART tree shown here has changed markedly from that of Figure 7.2. The first split is similar to that in Figure 7.2 with cases where the wife had less than three years education being classified as using Tubectomy. Cases where wives had three or more years education were next divided on the basis of income. Cases where wives had three or more years education were next divided on the basis of income, Figure 7.4 graphically depicts what happens in the CART tree. The solid line marks the first split so that all cases to the left of that line were classified as using Tubectomy. The interval line denotes the position of the second split. Those cases above the interval line were classified as using foam tablets while those below were predicted to be using oral pills. The disturbing feature about this tree, though, was the large number of IUD users in the sample, who were all misclassified, which as in the CART tree of Figure 7.2, could lead to very serious problems if this classification tree was put into practice.

As seen in Table 7.6, the number of misclassified observations from each class with CART was negatively related to sample size when ppss were used. When equal priors were used, only a small number of foam tablet and oral pill users were misclassified, but, as mentioned previously, all of those who used IUD were falsely classified.

7.4.4 FACT

In contrast with CART which is totally non-parametric, FACT uses F-ratios of between to within class variance to select the partitioning variable, then carries out LDA on the selected coordinate axis to partition the data.



Variable	Relative Importance
Wife_Edu	100
Husb_Age	90
Income	77
Wife_Age	75
No_Child	64

Figure 7.3: CART Tree with Equal Priors for the Family Data

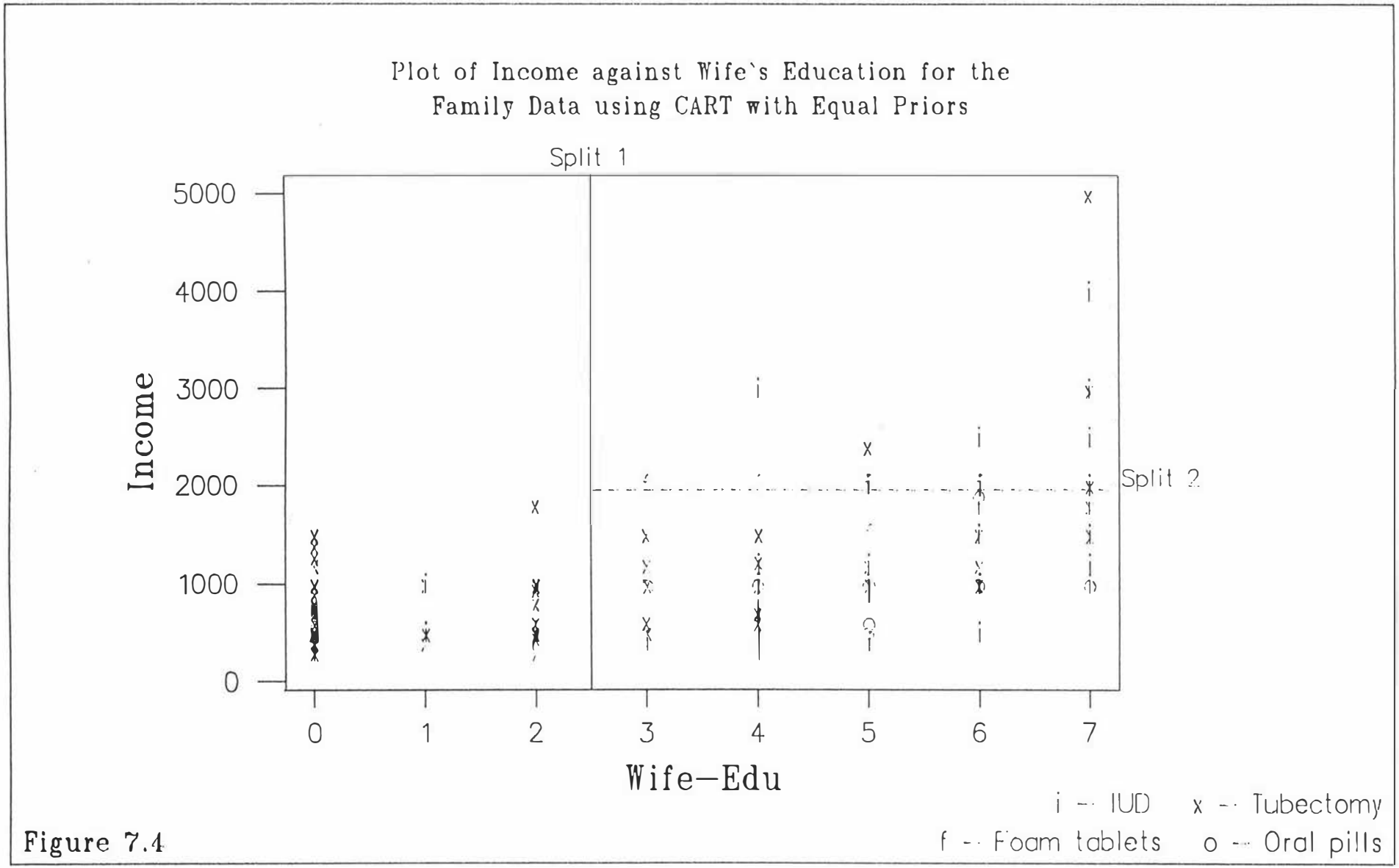

Plot of Income against Wife's Education for the
Family Data using CART with Equal Priors
Split 1
Split 2
Income
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
Wife—Edu
i -- IUD
x -- Tubectomy
f -- Foam tablets
o -- Oral pills

Figure 7.4

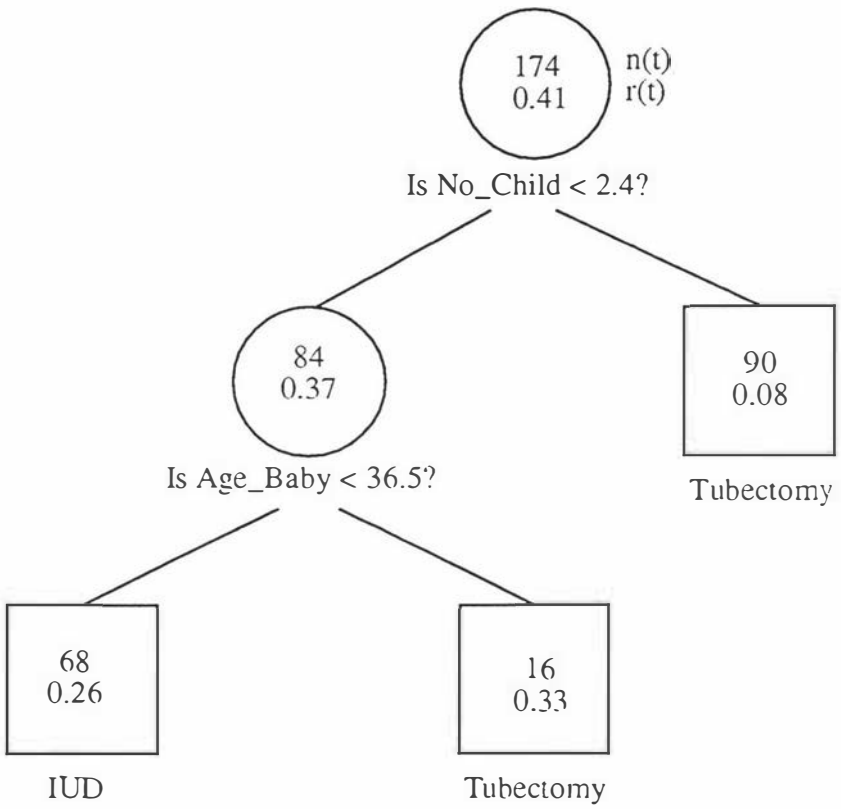
The FACT trees for the family data are given in Figures 7.5 and 7.6. Figure 7.5 illustrates the case of PPSS. The FACT tree is very different from the CART tree using ppss, shown in Figure 7.2. Interestingly, the two variables used to split on in this FACT tree were, however, two of the variables used in the SDA functions. This shows the link between LDA and FACT. Although FACT represents its output in a decision tree format, it is basically a parametric technique applied iteratively to each descendant subsample of observations. One might then expect the final classification structure to be similar to that produced by LDA, although this has not happened in this case.

As can be seen from Figure 7.5, FACT has failed to correctly classify any of those people who used either foam tablets or oral pills. Those with three or more children were more than likely to have a Tubectomy, while those with less than three children, and a youngest child who was three years old or more, were also more than likely to have a Tubectomy. Those with less than three children and whose youngest child was less than three years old were more than likely to use IUD. These rules seem to be straightforward and common sense compared with those found in Figure 7.2, which tended to be more sociological in nature. The FACT tree indicates that those with either larger families or who had not had any children for a while were more than likely to use a terminal contraceptive device. For this tree, R(0.632) = 0.197.

Figure 7.6 gives the FACT tree in the case of equal priors. Similar to CART and quite differently from LDA, the decision rules have changed quite dramatically after alteration of the priors. The tree contains only four splits but all are multiway rather than binary splits. The end result is a tree with ten terminal nodes, which are representative of all four classes (labelled 1 to 4 on the tree for the sake of space). Actually, the class misclassification error rates were not too dissimilar.

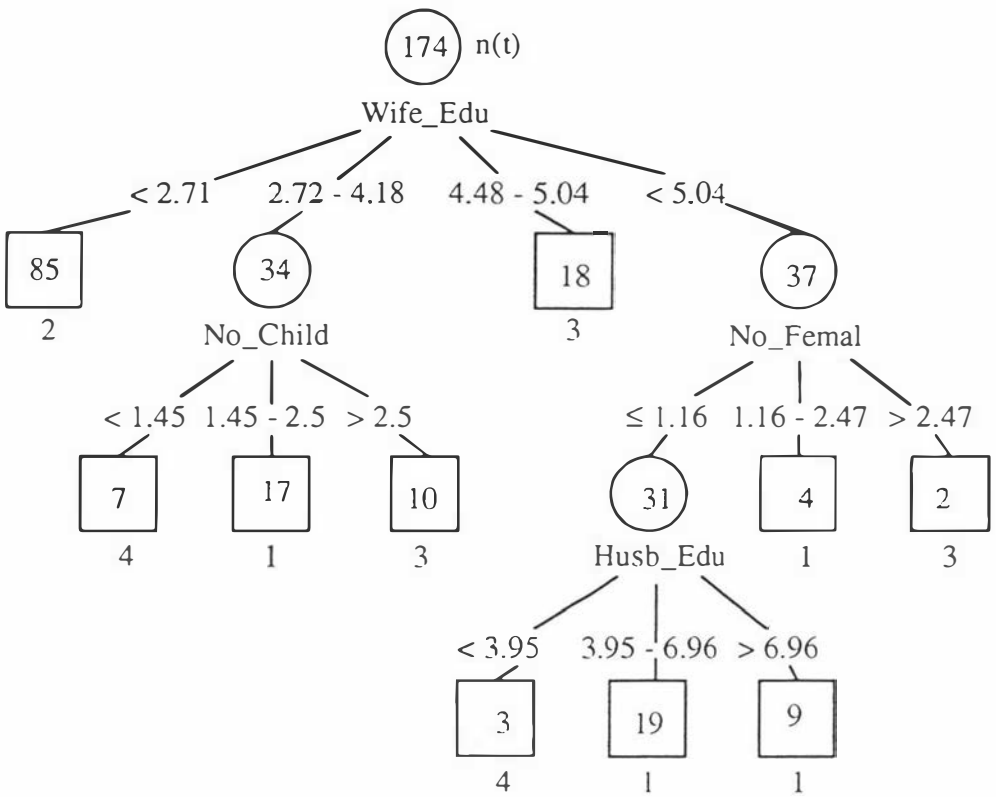
7.4.5 KnowledgeSeeker

KnowledgeSeeker is an example of a tree-based approach that uses, in contrast to CART, a statistical significance testing approach to splitting. Differently from FACT, however, χ^2 contingency table analysis is used to distinguish between the groups, rather than the use of means and covariances as employed by FACT. The method is based firmly on the refinements to AID carried out by Kass (1980), which resulted in the CHAID program.



Variable	Relative Importance
No_Child	100.0
Wife_Edu	98.7
No_Male	71.9
Wife_Age	49.1
Husb_Age	47.5

Figure 7.5: FACT Tree with PPSS for the Family Data



Variable	Relative Importance
Wife_Edu	100.0
No_Child	97.2
No_Male	81.6
Wife_Age	35.6
Husb_Age	34.8

1 = IUD

2 = Tubectomy

3 = foam tablets

4 = oral pills

Figure 7.6: FACT Tree with Equal Priors for the Family Data

The KnowledgeSeeker trees for the family planning data are given in Figures 7.7 to 7.12, using different partitioning methods, significance levels for splitting and splitting variables at the first node. Figures 7.7 to 7.10 show the trees with the first split being carried out on the most important variable. Version 2.1 of KnowledgeSeeker was used for all four trees. An obvious difference is the reduction in tree size when the significance level is decreased. Noticeable too is the slightly smaller trees produced using heuristic splitting as was suggested in the KnowledgeSeeker User's guide. Added to this is the increase in speed using heuristic partitioning. One trade off in using the heuristic partitioning algorithm is that a slightly different tree may be produced every time a new tree is created. This will not occur if exhaustive partitioning is used. One could argue that a tree may be "pruned" by decreasing the significance level used for splitting, but the tree that was originally created will only remain unchanged if exhaustive partitioning is used for tree construction.

Comparing the KnowledgeSeeker trees with those of CART and FACT, it is evident that this method has produced somewhat of a compromise between CART and FACT. No_Child was selected as the most important first splitting variable, as did FACT, and stepwise discriminant analysis also showed this to be the most important discriminating variable. Thus, all methods which use statistical significance to determine the classification rules, whether tree based or not, chose No_Child as the most important splitting variable. Using Figure 7.9 as the standard KnowledgeSeeker tree, it can be seen that those with either one or two children were more than likely to use IUD, with greater probability if there was only one child rather than two. Of those with three or more children, those with wives having zero to four years education were almost all users of Tubectomy. Those with wives having five or more years education were quite likely to use either foam tablets or oral pills.

No technique for handling priors exists in KnowledgeSeeker, though the tree of Figure 7.9 can be seen to have correctly classified at least some of the observations from every class. The larger trees produced here by KnowledgeSeeker should make the process more robust to large discrepancies in sample size. Reduction in the significance levels may affect group misclassifications to some degree, although this has not occurred for the examples presented here.

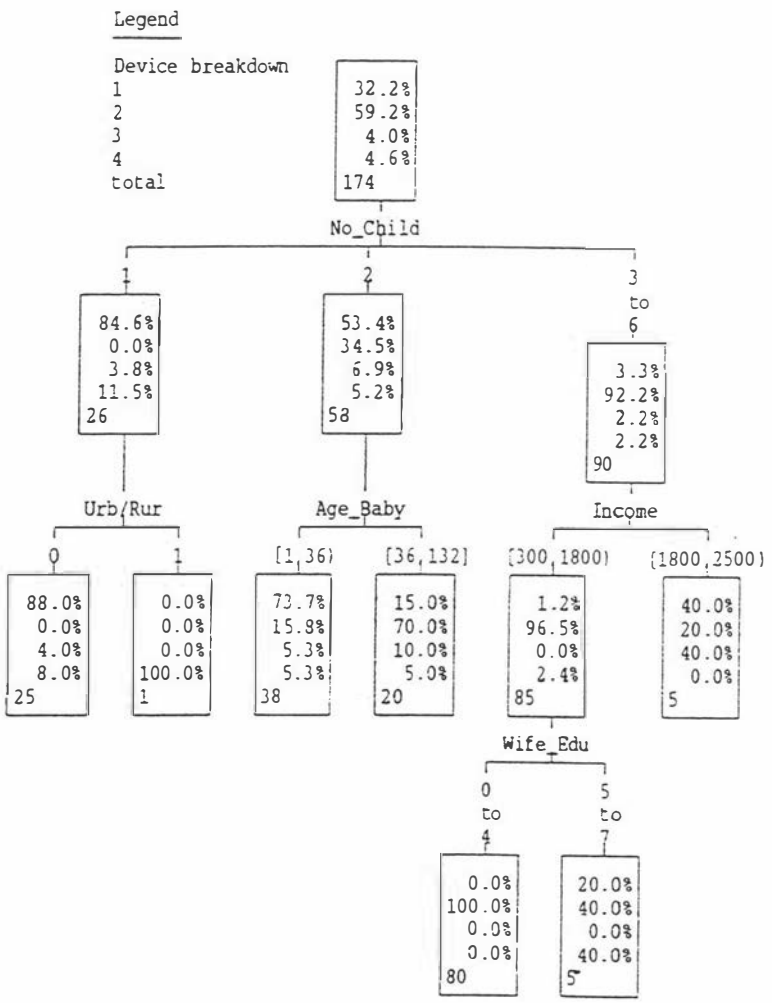


Figure 7.7: KnowledgeSeeker Tree Method = Heuristic, α = 0.05

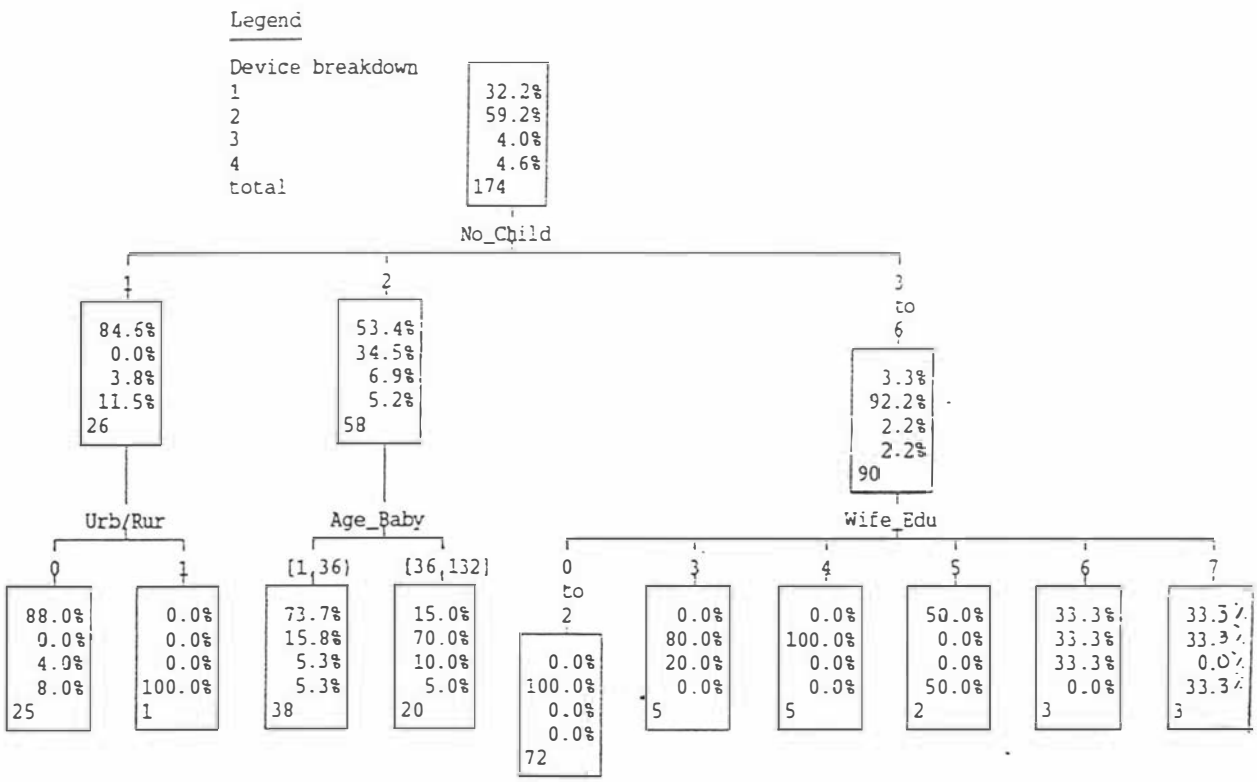


Figure 7.8: KnowledgeSeeker Tree Method = Exhaustive Partitioning, $\alpha = 0.05$

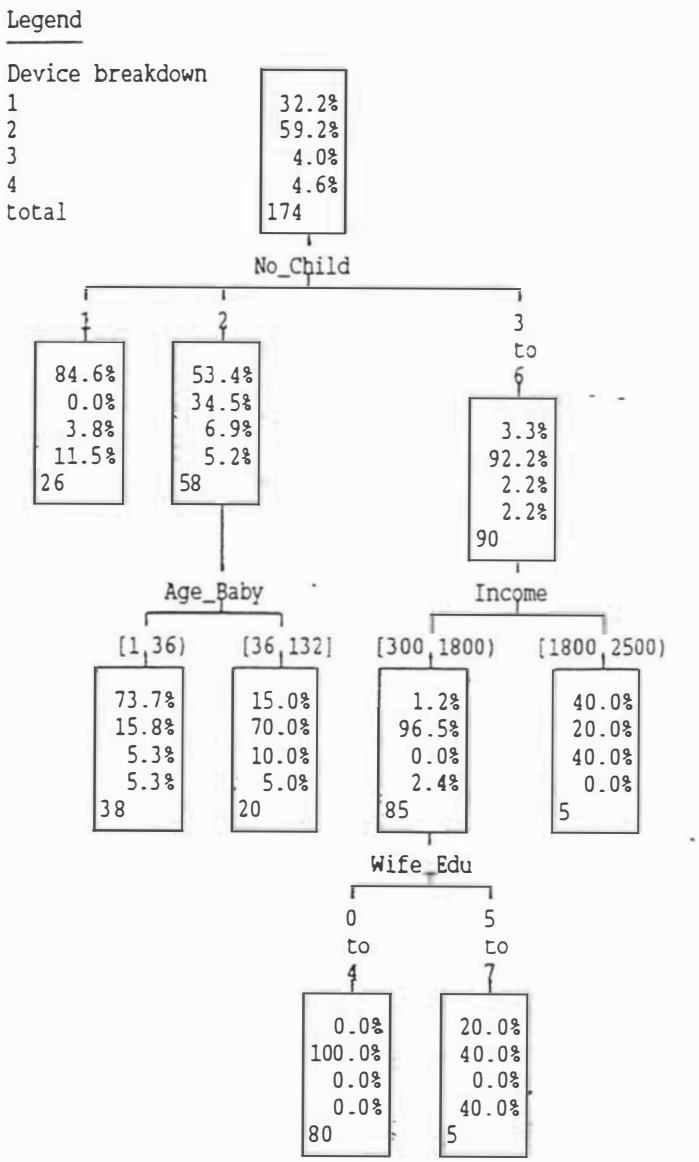


Figure 7.9: KnowledgeSeeker Tree Method = Heuristic, $\alpha = 0.01$

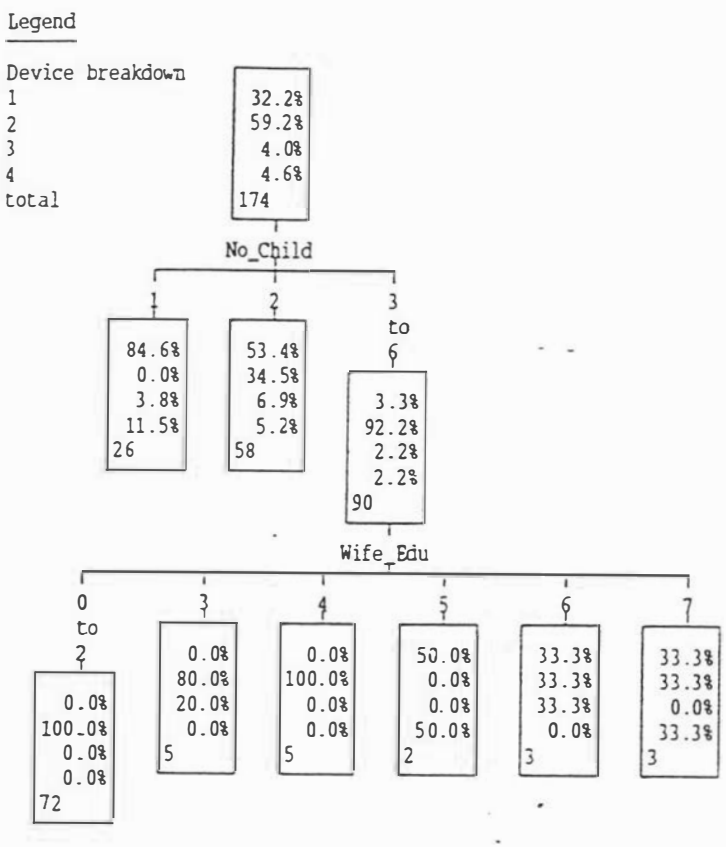


Figure 7.10: KnowledgeSeeker Tree Method = Exhaustive Partitioning, $\alpha = 0.01$

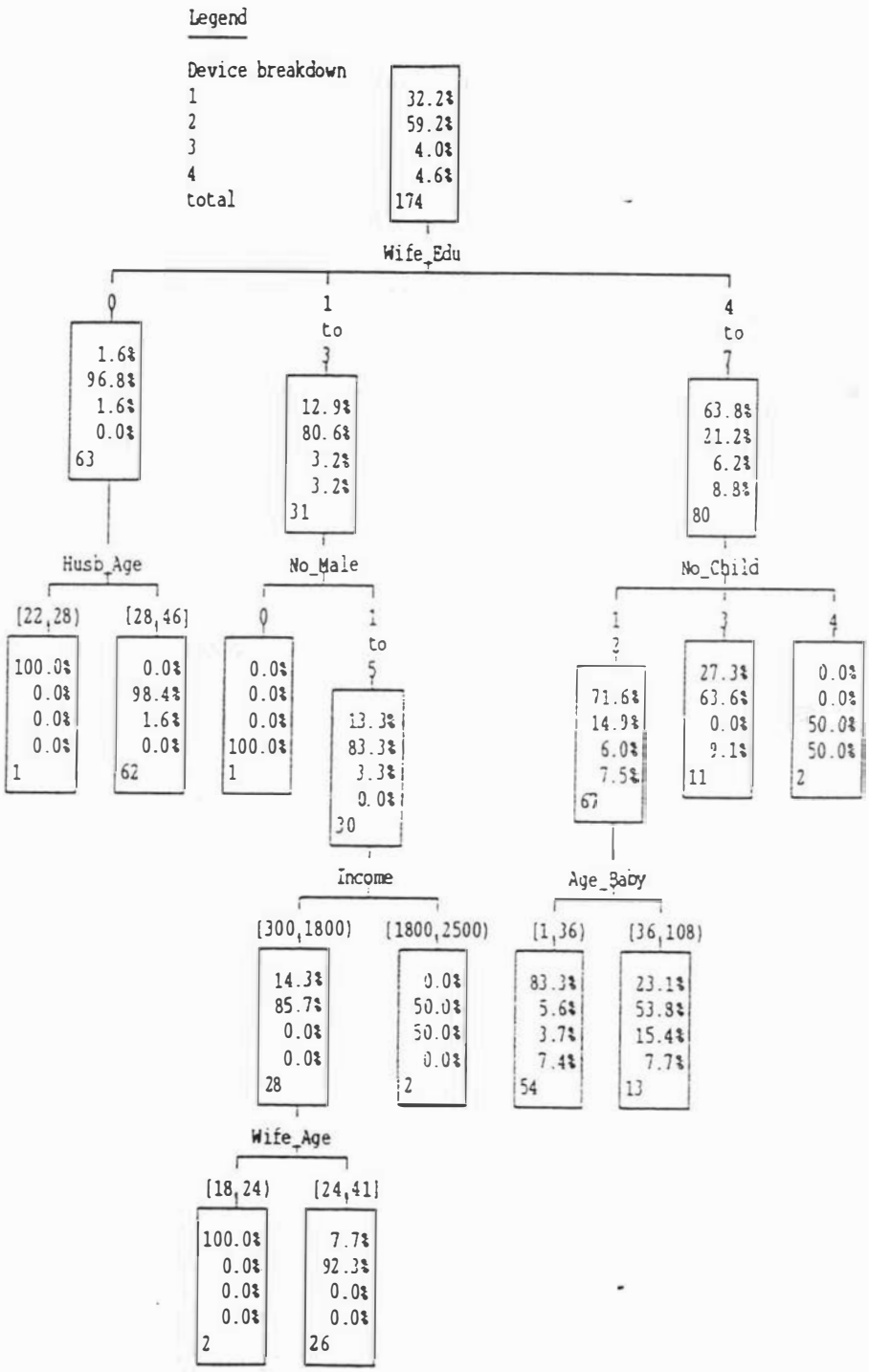
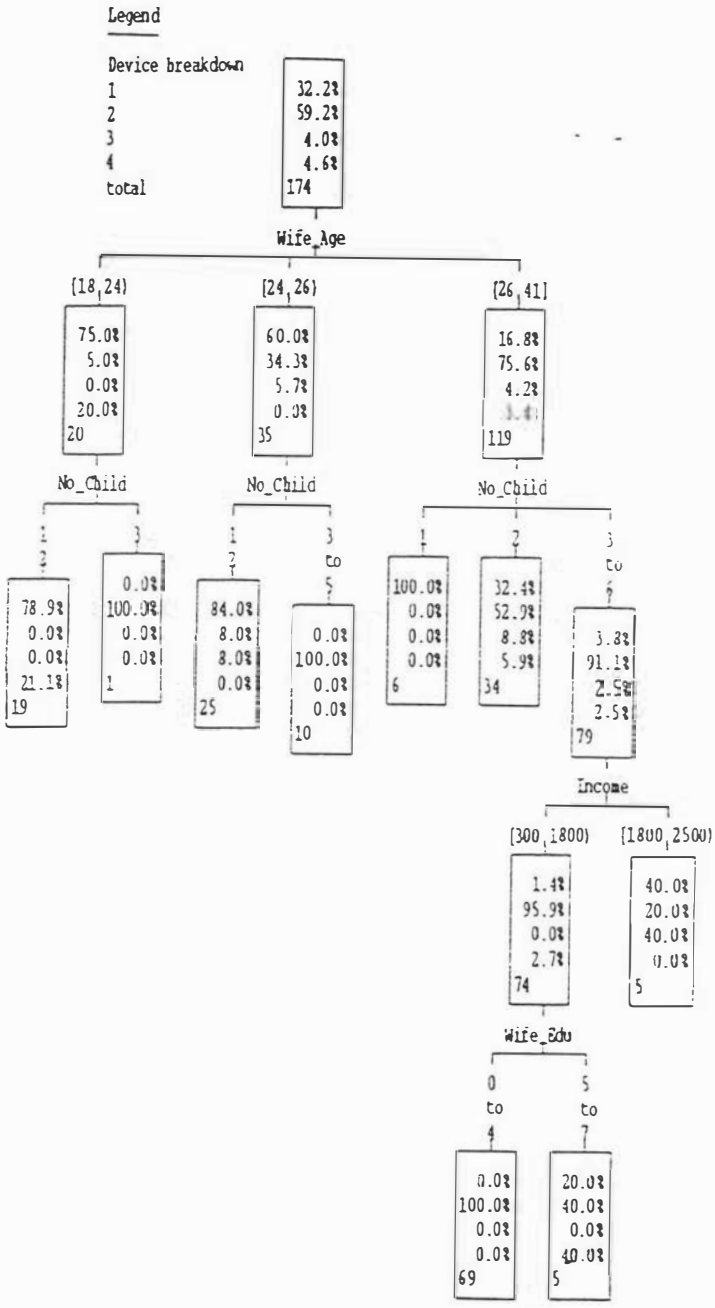


Figure 7.11: KnowledgeSeeker Tree: Method = Heuristic, $\alpha = 0.01$. Second Best Initial Split Used.



7.12: KnowledgeSeeker Tree: Method = Heuristic, $\alpha = 0.01$. Third Best Initial Split Used.

Figures 7.11 and 7.12 feature the KnowledgeSeeker trees with the first split being carried out on the second and third most important splits respectively. In Figure 7.11, the first split was done on Wife_Edu. This tree had an apparent error rate lower than the tree of Figure 5.9, though, as mentioned numerous times before, the apparent error rate is usually a biased estimate of the true performance of any classification rule. This is most particularly true of tree-based methods which are data intensive, using subsamples of the original data rather than all the data to construct a set of classification rules. Figure 7.12 has the first split carried out on Wife_Age, then on number of children in the family.

In order to truly test the validity of the KnowledgeSeeker trees, Version 2.0 was used so as to carry out some validation procedures. KnowledgeSeeker does not support the use of any form of cross validation so the rotation method was implemented as follows. The data set was first divided into two equal parts. Each half was in turn used as a learning sample and a test sample with the two test sample error rates being averaged to get the rotation estimate of the error rate. The technique was used on all six KnowledgeSeeker trees previously depicted (Figures 7.7 to 7.12). The rotation, resubstitution and 0.632 error rate estimates for the six trees are given in Table 7.10, with the minimum error rate for each estimate, over all six trees, given in bold.

Table 7.10: Rotation, resubstitution and 0.632 error rates for six KnowledgeSeeker trees for the family data

	Error Rate		
Tree	**Rotation**	**Resubstitution**	**0.632**
Figure 7.7	0.236	0.144	0.203
Figure 7.8	0.282	0.144	0.221
Figure 7.9	0.219	0.149	0.193
Figure 7.10	0.248	0.213	0.221
Figure 7.11	0.214	**0.138**	0.186
Figure 7.12	**0.184**	0-.172	**0.180**

From Table 7.10, it is apparent that the tree in Figure 7.12 had both the lowest rotation and 0.632 error rates, with the tree in Figure 7.11 having the second lowest error rates of the above type, showing that the best tree was not necessarily the one which produced the greatest separation of the classes at the first split. The ideal situation is to have the independent estimate of the error rate equal to the apparent error rate so that one can be confident in the classification rules generated from the learning sample. Using this as a criterion to choose the optimal KnowledgeSeeker tree, the tree in Figure 7.12 using heuristic partitioning with a 1% significance level, after initially splitting on Wife_Age was best. Now Wife_Age was noted as only the third most important variable for splitting at the first node. This perhaps shows that this tree was the most accurate classifier for the family data, backed up by the tree having the lowest overall independent error rates. One of the principal reasons for the initial creation of decision tree techniques was to identify complex interactions among the data, that could not be detected by parametric methods, such as LDA, without prespecifying the interaction terms directly. Tree-based methods, in general, appear to be very dependent on the initial split, that is, the largest main effect for the whole sample. If the first split is not particularly good, then it is unlikely that further partitionings of the data set will lead to a robust set of decision rules. If the first split is very good, but does not interact well with the other variables, the decision rules may also be rather weak. The interaction structure produced by the decision tree is going to be very dependent on the association with the first splitting variable. If splitting initially on a variable, x_i, does not lead to as purer descendant nodes as splitting on another variable, x_j, it may still produce a more robust tree than initially splitting on x_j, if the interaction structure is stronger between x_i and the other variables than between x_j and the other variables. In this example, Wife_Edu was rated the third most important variable or main effect, but in real terms, the interaction between Wife_Age and the other variables in the data set produced the more accurate set of decision rules using KnowledgeSeeker. Splitting initially on either No_Child or Wife_Edu did not produce as good a decision tree as that found through splitting first on Wife_Age.

7.4.6 Splus Trees( )

"The tree modelling interactive environment now available in Splus is to the batch mode program CART as graphical statistical packages such as Splus or JMP are to batch processing SAS for other statistical methods" (Morton, 1992, p 76). The method provides an example of the CART approach to decision tree growth, while incorporating all the advantages of an interactive environment for tree construction, pruning and graphics.

Figure 7.13 contains the fully grown (or overgrown Splus tree). Similarly to CART, and unlike FACT and KnowledgeSeeker, Wife_Edu was chosen as the initial splitting variable, thus having the largest main effect for the whole sample. The graph is rather messy as there are too many splits, hence terminal nodes in the tree. The residual deviance for this tree was 0.661 with R(A) = 0.149. Obviously, this error rate is optimistically biased. Figures 7.14 and 7.15 give the deviances for the sequences of subtrees produced by cost-complexity pruning and optimal shrinking respectively. As mentioned earlier (Section 3.11), the deviance always decreases as tree size increases, with the former being a step function because optimal subtrees remain constant between adjacent values of subtree sizes.

The rotation method was used to determine the optimal sized tree, that is, with minimum deviance. The average values for the deviance of given tree sizes, using rotational validation after optimal shrinking are given in Figure 7.16. This plot would tend to indicate that a tree with six terminal nodes would be optimal but perhaps the use of a standard error rule such as that used in CART would tend to suggest that a tree with four terminal nodes would suffice.

The resultant tree chosen by the rotation method is shown in Figure 7.17. The initial split is made on "Husb_Edu < 3.5" as occurred with the CART tree in Figure 7.2. The subgroup whose wives had less than four years education were next split on No_Child while the other subgroup was split on Husb_Age. For those cases where the husband was 33 or more years of age, a further split was made on "Husb_Edu < 3.5". A final split was made on the node whose husbands had four or more years education, splitting on the number of girls in the family. Those families with either no girls or only one girl were more than likely to use a terminal device (Tubectomy) while those with two or more girls opted for oral pills. The residual deviance for this tree was 0.933 with R(A) = 0.162, R(ROT) = 0.207 and R(0.632) = 0.190. In contrast with the CART tree of Figure 7.2, at least some of those cases who used oral pills were not misclassified. Notice too, that the 0.632 error rate is lower for this tree than that in Figure 7.2, due most probably to the much smaller apparent rate for the tree in Figure 7.17.

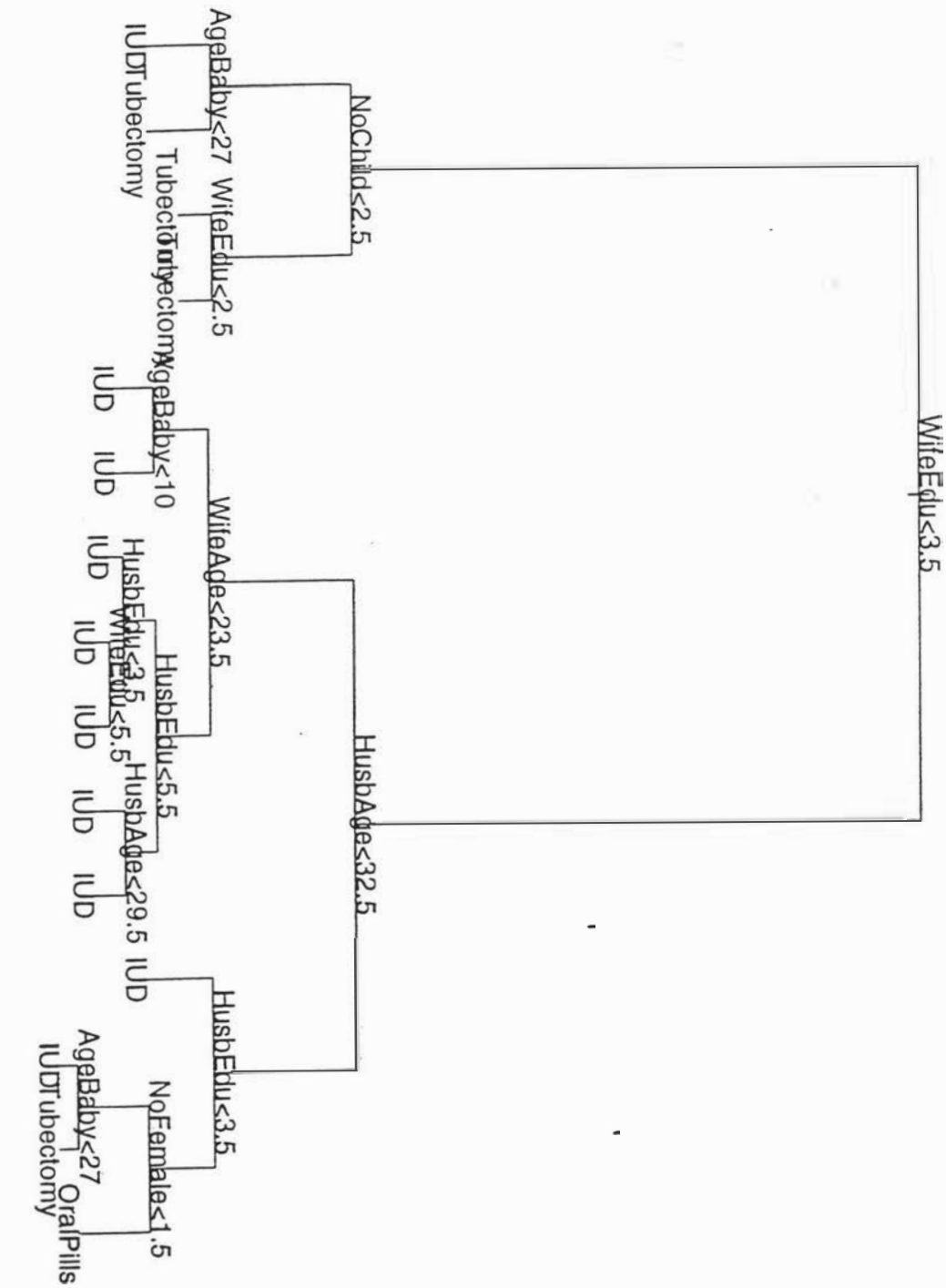


Figure 7.13: Splus Tree Before Pruning or Shrinking

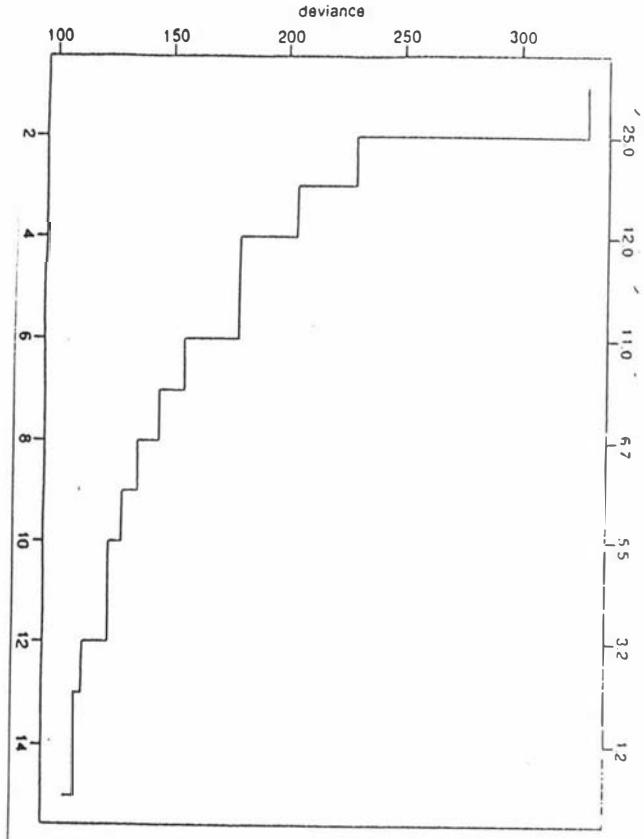


Figure 7.14: Deviance versus Tree Size for Subtrees Produced by Cost Complexity Pruning

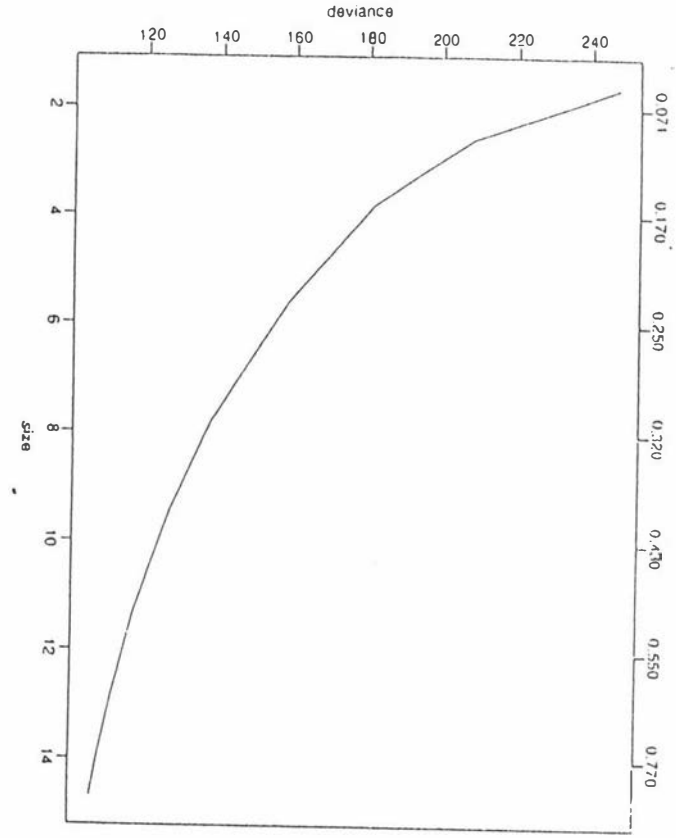


Figure 7.15: Deviance versus Tree Size for Subtrees Produced by Optimal Shrinking

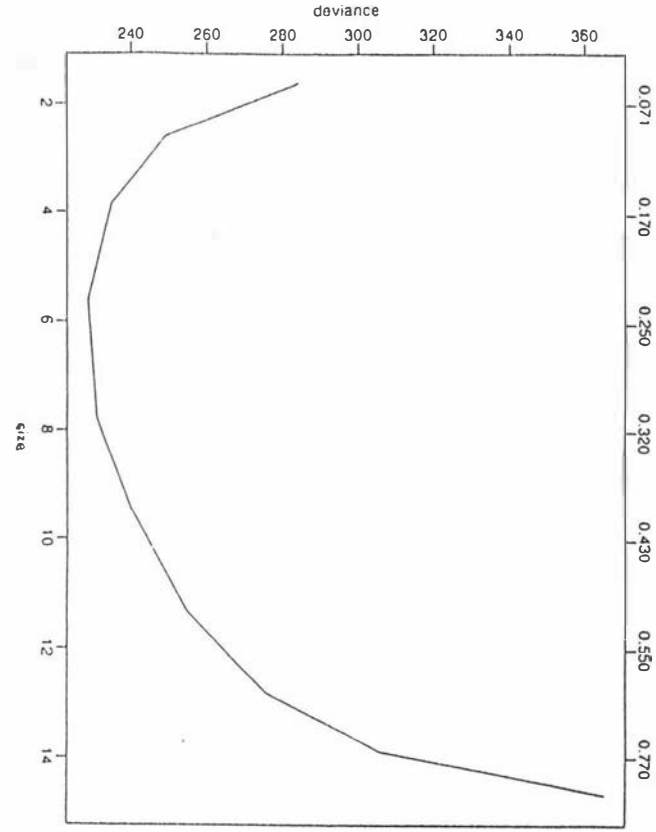


Figure 7.16: Rotational Deviance versus Tree Size for Subtrees Produced by Optimal Shrinking

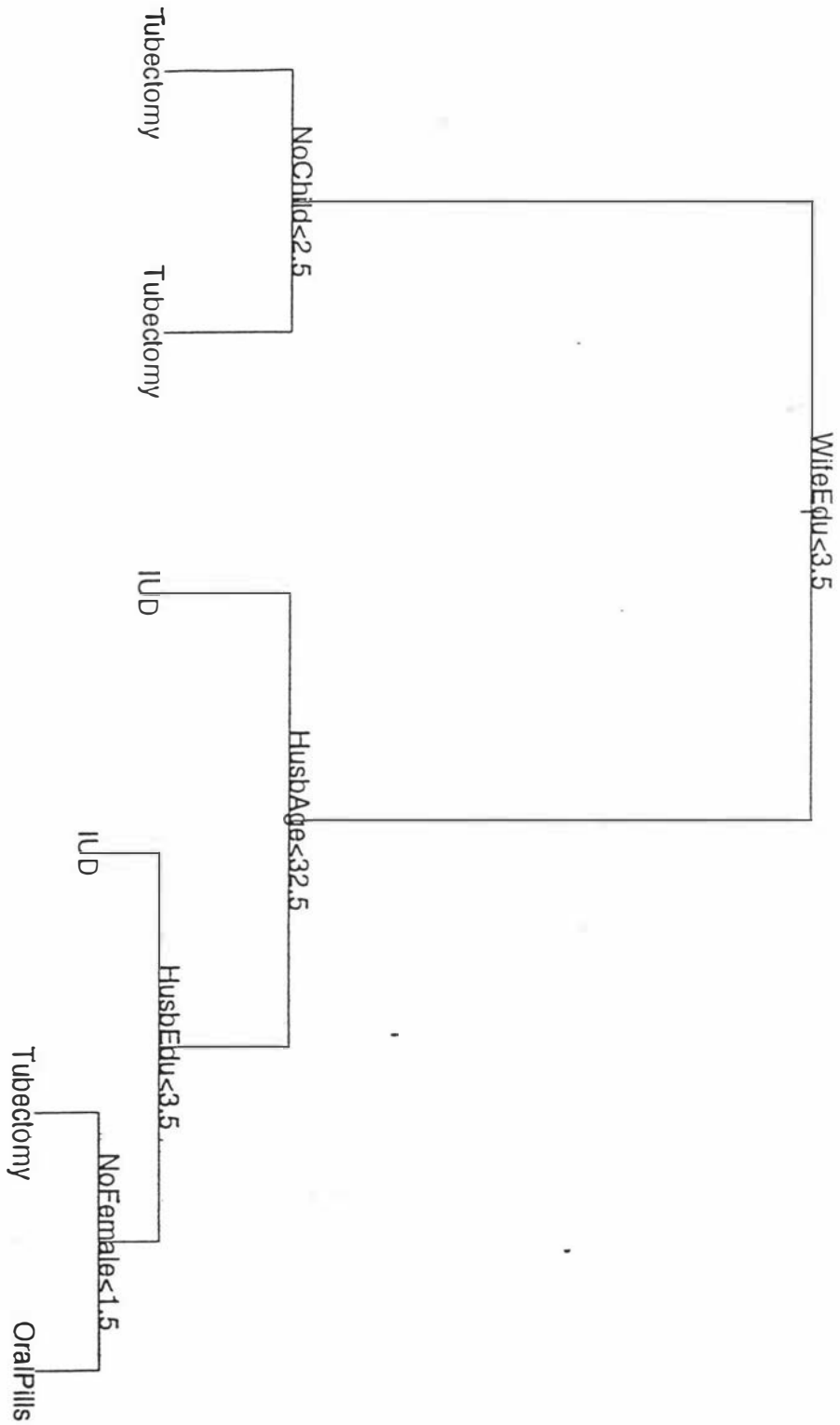


Figure 7.17: Splus Trees after Optimal Shrinking and Rotational Validation

7.4.7 Summary

This section has dealt with a set of data involving family planning information in a certain part of India. In terms of the accuracy of the rules produced and sensitivity to the prior probabilities of class membership, the parametric methods were marginally better, with stepwise LDA being the best overall. This case study has illustrated that CART is sensitive to the choice of priors used in the selection rule, and that care should be taken when interpreting the results from such a decision tree. Of the four tree based methods investigated, one of the trees constructed using KnowledgeSeeker had the lowest 0.632 estimate of error rate.

It has emerged from this case study that caution should be shown when interpreting the results from a decision tree. The choice of the first splitting variable can be very important to the future development of the tree. The main effect or first split chosen by a particular method may not necessarily lead to the most accurate set of decision rules.

In terms of accuracy of the models created, all methods were fairly similar. If, however, accuracy is not regarded as the sole criterion on which to judge the performance of particular methods, but other factors such as interpretability and comprehensibility of the models produced, ease of use etc, then different conclusions to those reached in this chapter as well as Chapters 4, 5 and 6 may be made. These performance criteria are investigated in Chapter 8.

8. WHICH CHARACTERISTICS OF TREE-BASED METHODS ARE PREFERRED

8.1 INTRODUCTION

In Chapter 3, ten tree-based methods were presented from a methodological point of view, examining a number of characteristics such as splitting criteria, stopping rules and tree pruning methods.

In this chapter, recommendations are made as to which options are preferred for each of the above characteristics. Thus the focus is not on comparing methods, but on comparing the approaches taken by each of the above methods to grow a classification tree. These recommendations are made both on what has been written in the literature, but also on the results of simulation and case studies undertaken in this thesis.

In Section 8.3, a review is carried out of what other authors from various fields of study have written about the various methods studied in this thesis. Recommendations are made as to what methods are preferred from the point of view of human comprehensibility and ease of use, based on the findings from the simulation studies in Chapters 4, 5 and 6, empirical studies in Chapter 7 and personal experience.

8.2 WHICH CHARACTERISTICS OF TREE-BASED METHODS ARE PREFERRED?

8.2.1 The Method of Splitting

Authors of early tree-based methods felt that their splitting rules were sufficient enough to grow an accurate decision tree, believing that the set of rules they developed for selecting the best variable at a node and value(s) of that variable to split on would produce an accurate and robust decision tree classifier. Most authors of more recent methods, starting with the CART algorithm of Breiman et al (1984), adopted the approach that regarded splitting rules more as a heuristic to form an overly large tree to be pruned rather than an end in itself, believing that the pruning process is perhaps the most important part of the tree growing procedures (see, for instance, Buntine and Caruana, 1993).

In the literature, there seems to be little clear-cut evidence as to which choice of splitting rules is best. Mingers (1989) shows that the accuracy of a decision tree is not affected by the choice of splitting rule, even when variables are selected randomly. Buntine and Niblett (1992) produce results indicating that random splitting leads to increased error, but other measures perform with similar acuracy to each other. Simulation and empirical studies in this thesis (see Sections 4.3, 4.4, 5.3 and 7.3.2) have shown that splitting using linear discriminant functions are not recommended for nonnormal data, examples where the covariance matrices are not equal, nor for categorical predictor variables. For normal data, with equal covariance matrices, linear discriminant splitting, as used by FACT, has been shown to perform satisfactorily. Comparisons between the information gain measure of C4.5 and the Gini/twoing goodness of split criteria of CART have proved inconclusive as have comparisons between the Gini and twoing criteria alone (Breiman et al, 1984). Buntine (1992) has found that the Bayesian quality measure approach employed in IND performs very similarly to the approaches used in CART and C4.5. Empirical studies undertaken in this thesis (see Section 7.3.6) have also shown that trees produced by the Splus deviance goodness of split criterion and CART's Gini criterion were often rather different even though the two splitting mechanisms produced partitions on the same variables and values of those variables at every node. This reinforces the point that the choice of splitting rule is not the most important step in the building of a reliable and accurate decision tree classifier.

Some authors of recent papers in the field of tree-based methods, however, have focused on the deficiencies of certain existing splitting methods and/or created a new type of splitting rule in the belief that this will lead to significantly "better" decision trees. For example, Todeschini and Marengo (1992) have used full p-variate LDA at each stage of the tree growing process in order to utilise the splitting power of LDA. Taylor and Silverman (1993) have emphasised the two main failings of the Gini splitting criterion, when used with CART, in the case of more than two classes, namely, the tendency to produce two offspring nodes that are as pure as possible and a bias towards splits which create descendant nodes of roughly the same size. They also noted that the twoing splitting criterion, which is also used in CART, failed to rectify these "weaknesses" of the Gini splitting criterion. This prompted Taylor and Silverman to develop an alternative splitting rule that placed less emphasis on creating pure offspring. This rule is known as the Mean Posterior Improvement (MPI) criterion. "[T]he MPI criterion is designed to be high when, for all k [classes], the individuals

of class k are all placed in the same offspring ... [but] does not directly strive for the offspring to be pure". (Taylor and Silverman, 1993, p 9.)

To correct against the bias towards equally sized samples, adaptive anti-end cut factors are introduced while still attempting to guard against splits which favour radically different sized descendant nodes. The main idea behind adaptive anti-end cut factors is to allow the differences between descendant node sizes to vary depending on the complexity of the problem.

Taylor and Silverman appear to place greater emphasis on interpretation of the tree and fitting a good model to the data at hand rather than constructing a robust tree that can be applied to other data sets of the same type. For the examples used in their paper, the independent misclassification rates for trees constructed using the MPI splitting criterion were no lower than those found using the Gini splitting criterion. As Einhorn (1972), Doyle (1973), Doyle and Fenwick (1975) and Breiman et al (1984), amongst others, have repeatedly stressed, a good classification tree is one which will work well on another data set of the same type.

8.2.2 Binary versus Multiway Splits

The choice of the type of splits, binary or multiway, is a question of debate amongst authors of decision tree methods. Quinlan (1979, 1986), Loh and Vanichsetakul (1988) and Biggs et al (1991) favour the use of multiway splits, whereas Breiman et al (1984) and Clark and Pregibon (1992) prefer the binary splitting approach. In the Bayesian approach adopted by IND, Buntine has used only binary partitioning. The advantage of the binary partitioning approach is simplicity. The cases in a node can be sent in only one of two ways. The direction a case is sent is dependent only on a yes/no question. Multiway splits may involve several conditional yes/no questions at the one node.

A major debate over whether binary or multiway splits are best occurred between Loh and Vanichsetakul (1988) and Breiman and Friedman (1988). Loh and Vanichsetakul argue that the use of binary splits has the following advantages: (i) categorical variables can be handled naturally as ordered variables, and (ii) the idea of surrogate splits is more straightforward to implement than if each node is split into varying pieces. They also see the disadvantage that they can produce a highly nested tree which leads to an increase in complexity and loss of interpretability . Multiway splitting, in their opinion, can reduce the level of nesting in the tree. If the number of partitions created at the root node is the same as the number of classes,

tree interpretation becomes much easier. However, Loh and Vanichsetakul warn that multiway splits may produce trees which are too short, stopping before any valuable information about the data set can be gained.

Breiman and Friedman (1988) criticise the belief that multiway splitting is superior to binary splitting. Using results from Friedman (1977), Breiman and Friedman argue that multiway splitting is not as effective in making use of the conditional information present in a tree as is binary splitting. Breiman and Friedman also argue that trees produced using multiway splits are no more interpretable than trees containing binary splits.

The results of simulation studies in Sections 4.3, 4.4 and 5.2, as well as for real-world data sets in Section 7.3, have shown that there are no major differences in the sizes of the final trees produced by the two methods. Using binary splits generally produces longer, narrower trees than using multiway splits which produce shorter, wider trees, where the size of a tree is determined by the number of terminal nodes contained in it. (Note: the number of terminal nodes in a tree equals the number of decision points plus one. For example, a tree with four terminal nodes has three decision points, while a m-way split contains (m-1) decision points.) For example, compare the CART tree for the Iris data, using binary splits (Figure 3.2) with that of the FACT tree, using multiway splits (Figure 3.4). Both trees have three terminal nodes thus contain two decision points. The CART tree, however, involves two splits while the FACT tree produces only one split, but partitions the data into three nodes.

8.2.3 Univariate versus Linear Combination Splits

On the question of whether linear combination splits are preferable over univariate splits, a major debate also took place between Loh and Vanichsetakul (1988) and Breiman and Friedman (1988). Loh and Vanichsetakul preferred the use of linear combination splits as their method involves carrying out LDA at each node. This approach has been taken further by Todeschini and Marengo (1992) with the use of full p-variate LDA at each stage of the tree growing process. Breiman and Friedman argued that linear combination splits are not better than univariate splits, stating that in most cases where recursive partitioning has performed better than traditional parametric methods, it has been through univariate splits. No complete comparisons of univariate and linear combination splits in either the CART or FACT program were made in this thesis. Comparing Figure 3.6 (CART tree with linear combination splits for the Iris data) with Figure 3.2 (CART tree with univariate splits for the Iris data), it is apparent that Figure 3.6 has the lower resubstitution error rate. However, the

tree in Figure 3.6 is no smaller than that of Figure 3.2, instead it is now more complex. In addition, the cross-validated error rate is 0.07 compared with 0.05 for the tree in Figure 3.2, implying that using linear combination splits in this example has led to a less accurate tree. Other instances where linear combination splits have been used have led to increases in accuracy, sometimes quite large, but these were for problems not suited to CART (normality, low dimension and sample size).

If the discriminatory variables are not correlated individually with the classification variable but are highly correlated in tandem with the classification variable then linear combination splits should perform better than univariate splits. An example of this scenario, is, in the case of a two class, two dimensional problem, where the scores of the first discriminatory variable are all higher than the scores of the second discriminatory variable for class one cases, while the opposite is true for class two cases. Generally, it appears that the choice of either univariate or linear combination splits involves a trade off between accuracy and simplicity, in the above situation. Linear combination splits may produce more accurate trees than univariate splits but the complexity of the rules produced is on a par with traditional discrimination methods. In the limited use of linear combination splits with more complex problems, there is no evidence of any increase in accuracy over univariate splits.

Linear combination splits provide the user with direct information about the splitting power of a number of variables at each stage of the tree growing process, via the coefficients associated with each variable in the linear combination. In contrast, univariate splits provide the user with direct information on only one variable at each node of a tree. Indirectly, however, most tree-based methods also provide information on the best competing splits. For instance, with CART, a list is provided of the best alternate and surrogate splits. In KnowledgeSeeker, the user can immediately investigate the effects on the tree of changing from the best possible partition to the best alternate/second best alternate partition etc. In Splus Trees(), graphical facilities are available to compare competing splits at a particular node. A procedure also exists for automatically changing the current split.

8.2.4 Costs and Priors

Feng et al (1993) compared a number of decision tree-based methods over a number of performance criteria. Their findings suggested that a method which incorporated a cost-handling mechanism should perform considerably better than methods without such a device. Methods such as CART and FACT which incorporate costs into the tree building process, therefore, have an advantage over C4.5, KnowledgeSeeker and other methods which have no capacity for handling costs. Buntine argues that IND has cost structures in that a vector of costs is combined with the class probability vector so that a minimum cost decision can be made.

On the question of class priors, both simulation and case studies in this thesis (see Sections 4.5 and 7.3.4) have shown that the choice of priors can dramatically alter the character of the final decision tree, thus leading to instability in the tree structure. Growing decision trees on data sets with grossly unequal class sample sizes, using priors proportional to sample size tends to lead to trees weighted very heavily in favour of the larger class(es). This results in all or nearly all of the observations from the smallest class(es) being misclassified. Using equal priors has been shown to sometimes have the opposite effect (see Section 7.4.3 of this thesis and Breiman et al, 1984, pp 112-113). The message, therefore, is that caution should be shown when viewing the classification trees generated from such data sets.

8.2.5 Stopping Rules and Tree Pruning

The use of stopping rules has been viewed by a number of authors of recent tree-based methods as unnecessary, if not inappropriate, due to the fact that "... a tree has to be grown out before any advantage is realised." (Buntine and Caruana, 1993, p 3-4). Some recent methods still use direct stopping rules, either stopping when node size falls below a certain value, or the number of terminal nodes is too large, or more commonly, use some measure of statistical significance to cease splitting. For instance, FACT stops splitting when the ratio of between to within group variance is less than a certain threshold value while KnowledgeSeeker ceases splitting if the optimal split on a predictor at a particular node does not exceed a specified significance level Breiman and Friedman (1988) criticised the top-down approach stating that it was one of the main reasons why early tree-based methods were not really recognised within the statistical community.

The illustrative case study of Section 7.4 showed that tree-based methods which used pruning algorithms were not always guaranteed to produce the most accurate tree. For that particular case study, a KnowledgeSeeker tree was more accurate than both the CART and Splus trees. Evidence from the literature, for instance, Breiman et al (1984), Quinlan (1987), Clark and Pregibon (1992) and Buntine (1992), amongst others, however, would tend to favour the view that pruning is preferred over direct stopping rules.

This naturally leads to the question of what pruning rule one should use? Quinlan (1987) conducted an empirical comparison of three pruning methods. The three methods tested were; cost-complexity pruning, as used in CART, pessimistic pruning, as used in C4.5, and reduced error pruning, a technique which reduces a subtree to the best terminal node and compares the test sample error rate of the new tree with that of the old. If the new tree has a test sample error rate less than or equal to that of the old, the subtree is replaced by the best terminal node. Quinlan's results showed that trees produced using cost-complexity pruning were usually the simplest, but also often the least accurate. He also stated that the method required an independent test sample of data although g-fold cross-validation can also be used. Buntine (1992) has incorporated Bayes pruning into the IND procedure. He mentions that comparisons with other pruning and smoothing techniques are difficult because the Bayesian methods are highly parametric. His belief, though, is that Bayesian pruning is the best approach as it allows the user to grow and evaluate more trees in less time.

From the literature, as well as simulation and empirical study results, it is difficult to determine which type of pruning algorithm is best as the pruning algorithm is dependent on the sequence of trees generated by the splitting rules, types of splits etc. When comparing cost-complexity pruning and optimal shrinking when used with Splus Trees(), it was found that which method worked best depended on the data sets, but often the final trees generated were exactly the same.

8.3 HUMAN COMPREHENSIBILITY AND USER-FRIENDLINESS OF CLASSIFICATION METHODS

In assessing the performance of a set of classification rules, one should not only be concerned with predictive accuracy but also how much information the classification methods provide the users, through the complexity and interpretability of the results. In other words, another major concern in the choice of classification method should be the explanatory power of the model. In making any such assessment, it is clear that all such recommendations are of a subjective nature, and that there are no right answers only opinions backed up by solid arguments.

As seen in Chapter 3, the ideas behind a decision tree, developed by Belson (1959), were an attempt to move away from the over complicated models of standard statistical techniques towards a much simpler approach. These ideas were incorporated into the AID algorithm of Morgan and Sonquist (1963). Einhorn (1972), Doyle (1973), Breiman et al (1984) and Quinlan (1986), amongst others, concluded that early AID (and THAID) constructed unnecessarily large trees, containing a number of redundant splits, resulting in a set of rules which were as incomprehensible, if not more so, than the standard statistical techniques they were designed to replace.

In contrast, Breiman et al (1984), the authors of CART, cite the example of a medical study where the objective was to identify high risk heart attack patients, those who will die within the next month, based on 19 measurements taken within their first 24 hours of being admitted to San Diego Medical Centre. The CART tree contained classification rules based on three yes-no questions. Standard statistical classification methods were far more complicated, and in this case, less accurate. It is stressed that "[t]he tree procedure output gives easily understood and interpreted information regarding the predictive structure of the data" (Breiman et al, 1984, p 58). They state that the method has been used in a wide variety of applications, with users finding "... that the classifier provides an illuminating and natural way of understanding the structure of the problem" (ibid, p 58). In contrast, they find that the standard statistical algorithms including stepwise discriminant analysis, kernel density estimation and Kth nearest neighbour methods are, except for relatively simple problems, difficult to interpret. In the case of the latter two techniques, very little useable information is gained regarding the structure of the data.

It must be remembered that, as the authors of CART, the opinions of Breiman et al must be treated with caution. The other authors of tree-based methods have also made claims as to why their method is best and why other methods fall down, though as Feng et al (1993) have pointed out, the studies were either biased in favour of the authors own method so demonstrating its effectiveness relative to other methods or were conducted over very similar data sets (similar in regards to dimension, sample size etc), so that only a subset of the parameter space is tested. Therefore, it is fairer to use those papers that were mentioned in the previous chapter, involving studies undertaken by authors with no deep seated inclinations towards one method or another.

Ildiko and Lanteri (1989) compared LDA, QDA, CART and SIMCA over various chemical data sets. From the point of view of complexity and interpretability of the model, CART was the clear winner with usually small, compact binary trees and classification rules that can be used to classify future unknowns from the same population.

Brown et al (1993) compared CART with a back propagation neural network algorithm, although neural networks are not covered in this thesis. They found that CART trees were simple and easy to read, providing a set of useable rules for the future.

Feng et al conducted a large scale comparative study across a variety of data structures from industrial settings. They found that the decision tree methods produced the most interpretable results, whereas the majority of traditional discrimination methods produced little or no explanation at all. They also found that tree-based methods were easy to use, though also noted that techniques such as LDA and QDA were user-friendly too. Of the tree-based methods used in their study, only two, CART and C4.5, have been mentioned in this thesis. It was suggested that CART produced the smallest, hence simplest, trees of all such methods, with evidence indicating that C4.5 trees were rather more complex than those of CART. No other direct comparisons of tree-based methods with those using traditional discrimination, have been discovered in the literature.

Based on the simulation studies undertaken in Chapters 4, 5 and 6 and empirical studies in Chapter 7, a subjective comparison of tree-based and traditional discrimination methods can be made. For bivariate problems, involving only two classes, the LDA rules are relatively simple. Only one group separation function is calculated and an observation is classified into

one of two classes based on whether the discriminant function is greater or less than zero. Graphically, the problem can also be depicted quite simply. (For example, see Figure 2.1.) In most cases where there exists a linear relationship between the variables within each class, decision trees may be more complicated, for example, the FACT trees for data sets R and S in Table 7.3, with each having four terminal nodes. In higher dimensional, two-class settings, the interpretation of the rules produced by LDA becomes more complex and difficult to understand, especially for non-statistically oriented users. One could use stepwise discrimination to obtain the best two variables, then do LDA on those two variables to calculate a linear discriminant function and graph the results. However, when p, the number of variables, is large, this is inadvisable as information on (p-2) variables in the data set is being thrown away. In situations where there are a large number of variables and/or a large number of classes, decision trees such as those produced by CART and FACT are recommended. As seen from the results for data sets B, C and D in Table 7.3, though, the decision tree approach is not guaranteed to produce the most easily understood classification rules. In those examples, one had to sort through twenty plus questions to classify a particular observation, leading to an unnecessary amount of complication.

The rules produced by QDA and kernel density estimation were completely unintelligible to anyone without a statistical background. One would have to be guaranteed a significant increase in predictive accuracy of the classification models produced by these methods to warrant their use. Empirical studies (and simulations) have suggested that this is not the case. From the point of view of complexity and interpretability of the model, decision trees are a clear winner. Compact decision trees, such as those created by CART and FACT, are clear and simple compared to the other complex, algebraic decision rules associated with traditional discrimination methods, though a decision tree model is not always going to be the simplest. One should always explore alternative approaches if possible.

Having decided that a tree-based approach is the most suitable for the data at hand, the question could be asked as to which method or program should be used? To help answer this question, four tree-based methods; CART, FACT, KnowledgeSeeker and Splus trees() were compared in terms of the complexity and interpretability of the models produced, as well as ease of use or user-friendliness of the computer package.

One of the prime motivations for the development of recursive partitioning, tree-based methods was to shy away from the often complex and unintelligible rules produced by traditional discrimination, at least to the statistically illiterate, as noted earlier. The decision tree output, however, should not be too simple. As seen in Section 7.4, both the CART tree (Figure 7.2) and the FACT tree (Figure 7.5) contained three terminal nodes. Both of these trees had rules which were too simple, as the number of terminal nodes was less than the number of classes in the data set. The Splus tree had six terminal nodes, two more than the number of classes, though all those cases who used foam tablets were misclassified. The KnowledgeSeeker tree on the other hand, through the use of multiway splits, had nine terminal nodes, but was still relatively easy to understand. In addition, at least some of the cases who used either foam tablets or oral pills were correctly classified. One could argue that the use of equal priors in both CART and FACT created rules which did correctly classify many of those who used either foam tablets or oral pills, but this was at the expense of the overall accuracy of the tree, and in the case of FACT, made the decision tree overly complicated.

Another criterion on which to judge the four methods is the ease of use or user-friendliness of the computer package. In ease of use, the menu driven approach of KnowledgeSeeker is difficult to beat. Both FACT and the Splus trees() procedures run on Splus, thus one must be familiar with the Splus language to be able to grow decision trees, then use all the add-on facilities that the program provides. With little or no knowledge of the Splus programming language, this would not be the preferred method of choice for the business manager, the medical researcher or the social scientist. CART, in its original form, requires data specifications and options files to be set up first, then is run in batch mode. When used with Systat`s menu driven approach the above problems disappear.

In terms of the ability to produce graphical displays, KnowledgeSeeker also seems to come out on top. The tree is displayed as it is grown and printed out with the click of a mouse button. CART, using Systat, can also display trees graphically, but the trees must be drawn separately after each analysis is done. Splus trees() and FACT both require the use of the Splus graphics facilities, a relatively easy task if one has mastered the intricacies of the Splus language! Both KnowledgeSeeker and Splus trees() have the ability to examine and change the variable to be split on at a particular node. Knowledge Seeker lists the most important splits at a node. If the user wants to see the effect on the tree of using the second or third

most important split at that node, a click of the mouse button allows the user to alter the split to be carried out. Splus trees() uses a function to change the split at a node. In contrast, CART and FACT split only on the value of a variable that is deemed optimal by each particular algorithm. The only way that one could see the effect of splitting on another variable at a node is to remove the most important varible from the analysis altogether. Naturally, this is an undesirable situation for the deleted variable may have had some impact in the latter stages of tree growth.

Although it has been stated that KnowledgeSeeker is the preferred method to use for this data in terms of comprehensibility of the models produced and ease of use, this does not mean that KnowledgeSeeker will always do the best for every problem encountered. The ease of use and interactive ability make KnowledgeSeeker an appealing method to use but the lack of a pruning algorithm may lead to trees that are overly large, hence complex, in some cases, and not applicable to other data sets of exactly the same type. The lack of a true validation procedure in the latest version of KnowledgeSeeker also provides some cause for concern. For those who have a good working knowledge of Splus, the Splus trees() routine, with its CART approach and functions for tree display, growth and modification, provides an excellent alternative.

9. CONCLUSIONS AND PROPOSALS FOR THE FUTURE

In this thesis, ten tree-based methods and the four most commonly used methods for estimating the conditional densities of observations, namely linear discriminant analysis, quadratic discriminant analysis, kernel density estimation and Kth nearest neighbour rules, were presented from a methodological point of view. Articles from the literature were used to identify and summarise where and when one should use each of the above methods.

A flow chart on a time scale is presented in Figure 9.1 showing the development of tree-based methods. Based on the ideas of Belson (1959), AID was developed as a technique using a sequential application of the one-way analysis of variance model, recursively partitioning the data into two subsets. The method was designed to predict the value of a continuous response variable. THAID was born out of AID in 1973 to handle categorical response variables. In the machine learning and artificial intelligence school of thought, the proposals of Hunt et al (1966) were developed further by Quinlan (1979) and put into the ID3 algorithm. All three of these early methods were criticised for, amongst other things, producing overly large and unreliable decision trees. Kass (1980) incorporated a significance testing approach into THAID to produce the CHAID method, in an attempt to solve the above mentioned problems. In contrast, Breiman et al (1984) developed tree pruning and validation procedures to build on the THAID algorithm. The end result of their work was the CART program. In 1986, in response to criticisms of ID3, Quinlan introduced its direct predecessor, C4.5, which also incorporated the idea of a pruning algorithm. FACT incorporated some of the ideas behind CART, but used statistical theory to carry out the splitting process. Both CART and FACT were criticised by various authors for, most particularly, the pruning algorithms and tree sizes. KnowledgeSeeker was developed directly out of CHAID but also included some of the approaches used by CART and C4.5. Splus Trees() was the incorporation of the CART method into the Splus programming environment, providing the user with many more options and flexibility than were available in the old CART program. The final method mentioned, IND, is a combination of the CART and C4.5 approaches to tree growth, tied together with Bayesian statistics. This last algorithm seems to be the complete package, allowing the user to implement either the CART, C4.5, minimum message length or Bayesian tree growing routines.

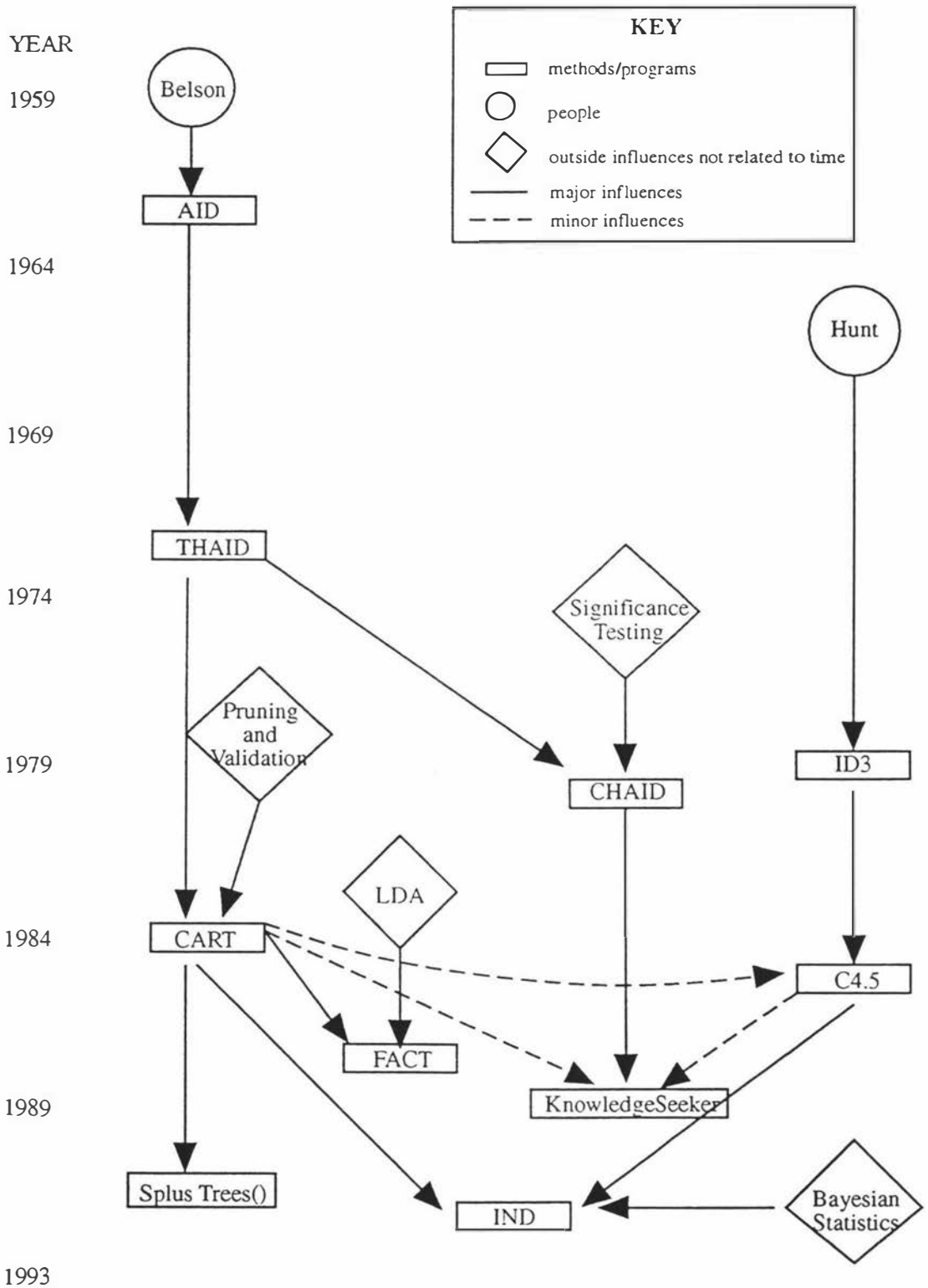


Figure 9.1: The Development of Tree-Based Methods Over Time

A range of simulation studies were undertaken in Chapters 4, 5 and 6 for both continuous and categorical data, involving a wide range of factors, while a number of empirical comparisons were carried out in Chapter 7. In Chapter 8, some recommendations were made, both from the literature and personal findings as to which characteristics of tree-based methods are important, as well as which method is preferred from the point of view of complexity and interpretability of the models produced.

In this, the concluding chapter of the thesis, the recommendations from the penultimate chapter are tied together with the findings of Chapters 4 to 7, to classify which methods should be used in particular situations. Some proposals for the development of future tree-based methods are also provided. The focus in this thesis has been on when to use a tree-based discrimination method in preference to either a parametric method, such as linear discriminant analysis, or a non-parametric technique, such as kernel density estimation. It has been established (see Chapter 8) that the tree-based approach, in the main, provides a more user-friendly approach to examining a set of data. Tree-based methods have also been used in conjunction with other methods, providing an alternative way of looking at a data set and suggesting possible interactions of variables and uncovering various subgroups. Though tree-based methods have been in existence for thirty years, there still appears to be a reluctance to use a tree-based method on its own to analyse a set of data. A primary objective of this thesis has been to compare tree-based methods with other discrimination techniques, through both simulation and empirical studies, to determine in which situations a tree-based method is most appropriate. The misclassification error rate of a prediction rule has been used as the performance criterion, providing a measure of the statistical power of each method. The results of these studies have led to the following set of recommendations.

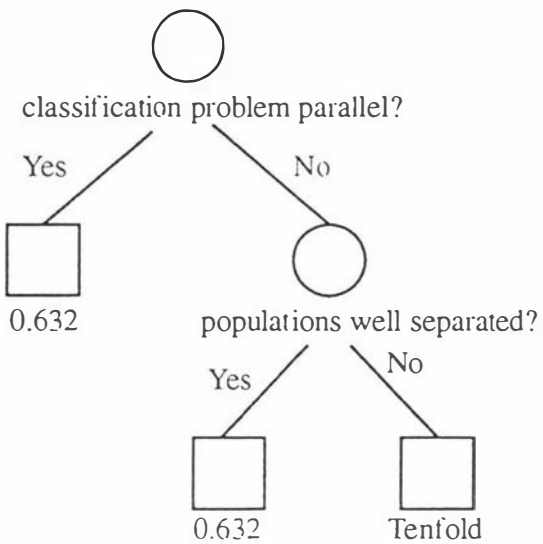
For continuous explanatory variables (see Sections 4.3, 4.4 and 7.3.2), the distribution of the data is the most important factor in deciding which method to use. It is well known that for normally distributed data a parametric technique such as linear or quadratic discriminant analysis is likely to be best. For lognormal data, a non-parametric technique such as CART is recommended, or indeed Splus trees() which utilises the basic CART approach. Tree-based methods, such as FACT, which use traditional statistical methods have unfortunately been shown to perform poorly, especially for unequal sample sizes, differing covariances and categorical data (see Sections 4.3, 4.4, 4.5 and 5.3). The second most important factor in choosing a classification method would appear to be the type of classification problem.

Sequential classification problems, where only a few out of many variables are important, are suited to CART-like methods. Parallel problems, where all the variables have approximately equal weighting, should be handled by traditional discrimination methods. Other factors which are linked to the complexity of the problem, that is, sample size, dimension and covariance structure are less important than the type of distribution. Results indicate that the more complex a data set is (larger sample size, higher dimension and unequal covariance matrices) the better the performance of CART-like methods over other techniques. A highly complex problem as defined here, if the variables are normally distributed, may be best analysed by CART-like methods while for less complex problems where the variables are normally distributed a traditional discrimination approach will perhaps be preferred. Another consideration in the choice of classification method is the distance between populations as measured by the difference in class means or the Mahalanobis distance. For well separated populations, traditional discrimination methods are preferred while the CART-like methods are better for poorly separated populations. As for the complexity of the data, however, these recommendations are very dependent on the type of distribution and classification problem.

For categorical explanatory variables (see Sections 5.3 and 7.3.2), in lower dimensional settings, a first consideration is the type of classification problem. CART-like methods do best for sequential problems while traditional discrimination methods work better for parallel problems, as occurred for continuous data. When there are a larger number of categorical variables, the same rules given above also apply. Other considerations are sample size where CART does best for smaller samples, while traditional discrimination methods are preferred for larger samples, in contrast to the recommendations for continuous data. Slightly less important is the question of distance between populations. CART-like approaches are suited for less well separated populations while linear discriminant analysis etc are preferred for highly separated populations in accordance with results for continuous data.

If CART was chosen as the tree-based method to use, which error rate estimator should be used to choose the optimum-sized tree from the pruned sequence of subtrees produced by CART's pruning algorithm? Using the twin criteria of accuracy, that is, how close the error rate of the tree is to the error rate found from running a very large test sample down the tree and simplicity of the rules produced, as well as the size of the final decision tree, the following recommendations can be made. If the explanatory variables are continuous (see Sections 4.6, 6.2 and 7.3.3), with small samples, Figure 9.2 shows that the 0.632 estimator

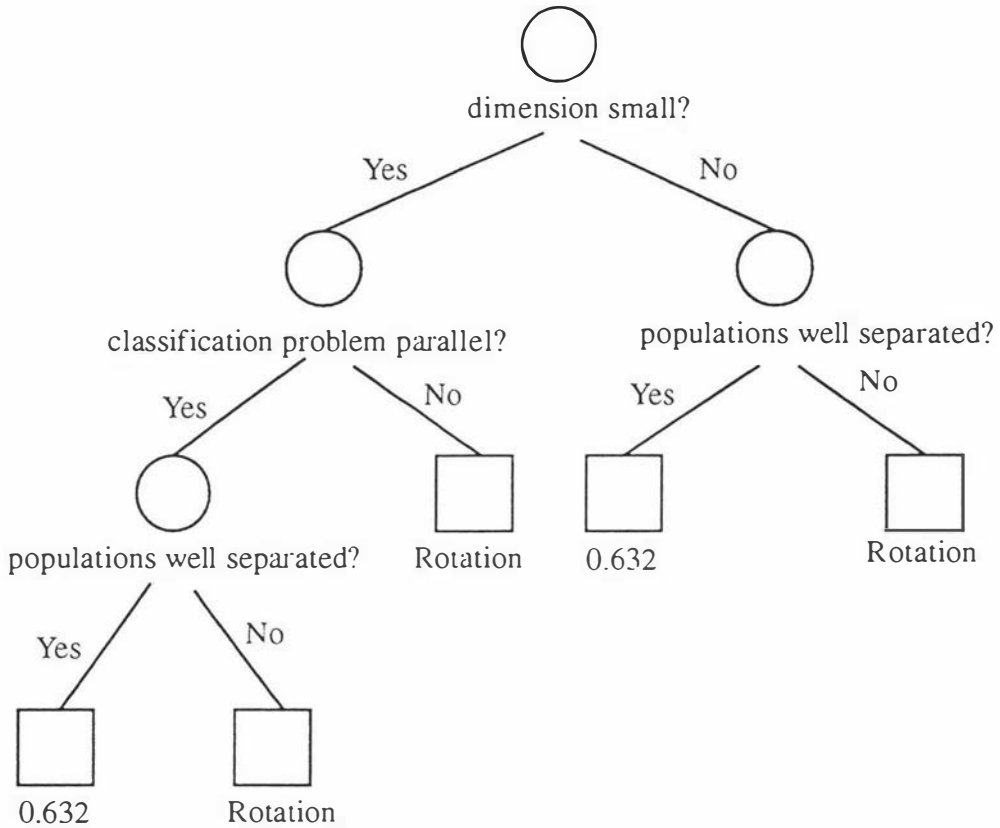
should be used unless only a few variables are important (sequential classification problem) and the classes are poorly separated. For continuous explanatory variables, with large samples, as above, the 0.632 estimator should be used unless the classification problem is sequential and the classes are not well separated. Simulation study results suggest that in such situations, it would be advisable to use the tenfold error rate estimate.



Circles represent decision nodes which have to be split while rectangles represent terminal nodes which are assigned to a particular class given below the node.

Figure 9.2: Decision Tree for deciding which Error Rate Estimator to use in CART: Continuous Explanatory Data

For categorical explanatory variables (see Sections 6.3 and 7.3.3), Figure 9.3 shows that the 0.632 estimator should be used for small samples, parallel classification problems and well separated populations, when the number of variables is not large. For higher dimensional problems, the 0.632 error rate estimate is recommended for either small or moderate sample sizes (and suitable for large sample sizes) or for moderately to well separated populations. For other situations, either the rotation or tenfold cross-validation error rate estimates should be used, with the former preferred for smaller samples. The n-fold cross-validation estimate should be used with a deal of caution, especially for small samples and poorly separated populations. In the latter situation, n-fold cross-validation was discovered to produce excessively optimistic error rates, hence overly large trees.

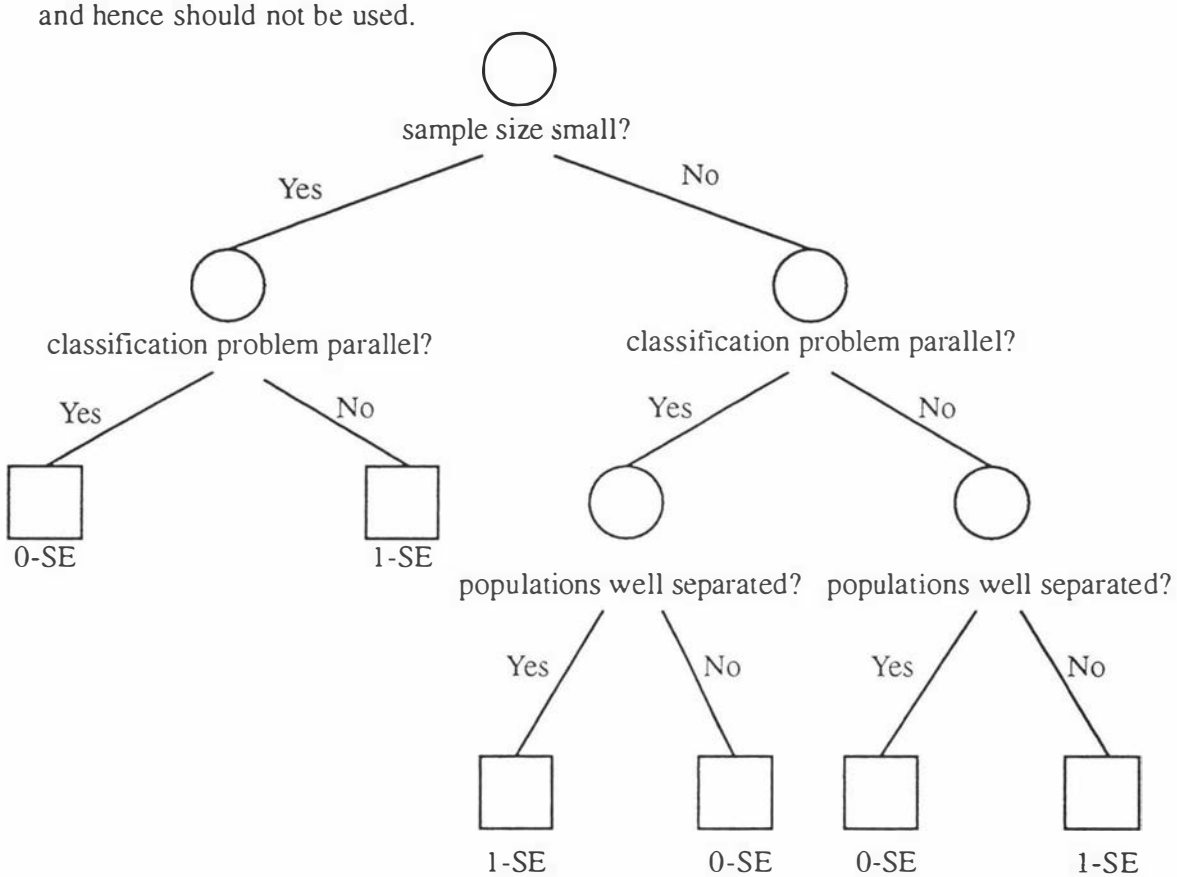


Circles represent decision nodes which have to be split while rectangles represent terminal nodes which are assigned to a particular class given below the node.

Figure 9.3: Decision Tree for deciding which Error Rate Estimator to use in CART: Categorical Explanatory Data

Having decided on which error rate estimator to use in CART, the next question to be asked is whether one should use the one standard error rule or not to select the right sized tree (see Sections 6.4 and 7.3.5). A set of recommendations on the use of the one standard error rule is displayed in Figure 9.4, in the form of a decision tree. In summary, the one standard error rule should be employed for small samples and sequential classification problems, and for large samples, when the classification problem is parallel and the populations are well separated or for sequential classification problems where the populations are poorly separated. Otherwise, the zero standard error rule should be used. The one standard error rule is designed to both correct the optimistic bias of the cross-validation estimate of error and produce as simpler tree as possible. When there is a large amount of noise in the data, the one

standard error rule should be used to remove unwanted splits. If, however, there is very little noise in the data, the one standard rule could lead to some important splits being removed, and hence should not be used.



0-SE = use the zero standard error rule.
1-SE = use the one standard error rule.

Circles represent decision nodes which can be split while rectangles represent terminal nodes which are assigned to a particular class given below the node.

Figure 9.4: Decision Tree for deciding when to use the One Standard Error Rule in CART

The rest of the chapter is devoted to future trends and developments in tree-based methods. With the tremendous advance of technology and computing power in the last few decades, there has been a corresponding increase in the number of tree-based methods appearing, usually with greater sophistication than their immediate predecessors. In ten, or even five years, most if not all of the tree-based methods that have been studied in this thesis may be

regarded as obsolete or discarded in favour of new, innovative and faster techniques. A number of ideas are presented here for the further development and refinement of tree-based methods.

The IND procedure also incorporates the decision graph algorithm of Oliver (1992), which Buntine (1993) suggests may be a hint of what is to come. The method involves a three stage process for growing decision graphs:

1. For each node, t, determine the variable, x_j, to be split on. Do not carry out the split but note the saving in message length.

2. For each pair of nodes, t_1 and t_2, calculate the saving in message length from amalgamating the two nodes into one. Do not perform the amalgamation.

3. Choose the alteration from (1) and (2) which had the greatest saving. Carry out this alteration.

This approach looks very appealing though empirical results have shown that there is no real increase in accuracy over C4.5 from using this method. If, instead, one wishes to take the simple CART-like binary-tree approach, how should one proceed?

From simulation studies undertaken in this thesis, it is apparent that the performance of tree-based methods is determined to a large extent by the characteristics of the data set. A first step in any decision tree program should be the printing of summary statistics of a data set. These statistics would include dimension, sample size, Mahalanobis distance between classes, some measure of skewness, a measure of equality of covariance matrices, a variable ranking procedure and a correlation matrix for all variables in the data set. From these summary statistics, the user should be able to know what sort of problem he/she is dealing with. For example, the information may indicate that the data is positively skewed, with poorly separated classes and only a few of the many variables being important. It would be very helpful for the program, on the basis of these summary statistics, to make recommendations as to which parameters should be used in the tree building process.

The purpose of the above procedure is to provide an option of almost complete automation in the tree building process, if the user so desires. Thus, the simplicity of the procedure would be increased. Simplicity in both running the program and interpreting the results should remain a key feature in any tree-based program. Recently, a paper by Todeschini and Marengo (1992) appeared detailing the linear discriminant classification tree (LDCT) method. As with FACT, the method is designed to combine the best features of LDA and classification tree methods, but unlike FACT, the algorithm uses full p-variate LDA at each stage of the tree growing process. Although Todeschini and Marengo claim an increase in accuracy, the method is not, as they also claim it to be, characterised by low complexity and ready interpretability. Such a method is indeed outside the aims of a tree-based method. Future tree-based methods should consider the simplicity of the interpretation of results as a primary objective.

This approach has been taken up by Taylor and Silverman (1993). They have produced a new form of displaying a classification tree, known as a block diagram, using a reimplementation of the CART algorithm. They focus on the use of tree-based methods as a means of better exploring and interpreting the data rather than providing a predictive classification model.

Without going into too much detail, block diagrams provide the user, through the colour coding of the nodes, with an indication of the splitting power of the discriminatory variables. Terminal nodes which are predominantly one colour indicate that CART has been relatively successful for a particular problem, while multicoloured nodes show examples of unreliable predictions with a strong overlapping of the classes. "[Taylor and Silverman] have found that block diagrams make it possible to identify and rectify failures in the classification method itself, rather than just to identify features of the classification of the particular data set under consideration." (Taylor and Silverman, 1993, p 6.) With the incorporation of text to help in the understanding of each split and makeup of each node, an inexperienced user would have comparatively little difficulty in interpreting the decision tree.

It has become apparent in this thesis that tree size is the major factor in determining the comprehensibility of a particular decision tree. Therefore, it would be appropriate to introduce a criterion that restricts the size of a decision tree to between preset lower and upper limits. This would be incorporated into the pruning algorithm with the tree having the

minimum cross-validation (or 0.632) error rate within those limits being chosen as the final tree. Naturally, the lower bound should not be set too low so as to make the tree too small, nor should the upper bound be set too high, thus leading to lack of interpretability.

A sensible choice of a lower limit could be k, the number of classes in the data set, with an upper limit of say 2k. Alternatively, a better measure may be to take into account how well each class is represented in the data set. If certain classes are not well represented in the data set, it may be wasteful to attempt to produce extra terminal nodes in order to incorporate these classes. One way of determining how many well represented classes there are in a particular data set is to use $m^*(t)$, the reciprocal entropy index as used by Taylor and Silverman, where

$$m^*(t) = \frac{1}{\boldsymbol{\pi}' \boldsymbol{\pi}}$$

and $\boldsymbol{\pi}' = (\pi_1, \pi_2, ..., \pi_k)$ is the vector of class probabilities. For data sets where all class sample sizes are equal, $m^*(t) = k$. Otherwise, $m^*(t)$ decreases as class sample sizes become more disparate. The lower and upper bounds for tree size could be set at $m^*(t)$ and $2m^*(t)$ respectively.

The question of correlation between variables raises another point in the formulation of a new tree-based method. It may happen, at a node, that a variable which, while not giving the best split, provides the second or third best split. As well, most of the cases sent left or right by the best split may be sent the same way by the alternate split. This implies that the two variables are highly correlated. At each node of the splitting process, some notification of the correlation between variables should be given so that the user knows what would happen if the split changed from say $x_j < c$ to $x_m < d$. Such facilities as those provided by KnowledgeSeeker for changing splits automatically should be a requirement for any future tree-based program.

Allied with the idea of correlation between variables at a single node is the question of correlation between a split at a current node and future partitions. As seen in Section 7.4, and mentioned elsewhere in the thesis, most decision tree methods are one-stage optimal in that they are only guaranteed to find the maximal separation of the k classes at each stage of the tree-growing process, that is, at the current node. No account is made of what will happen to

future tree growth if this so-called "optimal" split is carried out. The "optimal" split at each stage of the tree-growing process may not correlate well with future partitions and hence not lead to the most accurate tree possible. A lesser split at the current node, may, in contrast, correlate well with future partitions and so be best for future tree growth. As mentioned previously, such a r-stage lookahead option becomes infeasible for large r as p^r possible splits have to be examined where p is the number of variables in the data set. Morgan (1993) stated that such an option did not lead to any real improvements with AID (see Section 3.3), though Buntine (1992) has incorporated such a facility into the IND program. The case study of Section 7.4 showed that such instances of improvement can and do happen.

Future research into this area with more detailed simulation and/or empirical studies are required to decide whether the lookahead option does provide any significant improvement in accuracy over the one-stage optimality procedure. With the tremendous advances in computing power occurring today, the computing and information storage required is not the major drawback it was for the developers of earlier tree-based methods.

Other unanswered questions requiring further research and tests are whether univariate splits are preferred to linear combination splits as well as how sequential/parallel a classification problem appears. As discussed in Section 8.2.3, various views abound as to whether univariate splits are better than linear combination splits. Further simulation and/or empirical studies should be undertaken to determine which method is preferable both in terms of accuracy and overall tree size. In terms of the amount of usable information provided by the decision tree, the question of which approach is best remains rather subjective in nature.

A question mark also hangs over the issue of the type of classification problem, that is sequential, where relatively few of the variables are important, or parallel, where most of the variables are important in forming the classifier.. In Section 7.3, a "mixed" classification problem was defined as one which did not fit neatly into being either sequential or parallel. Other criteria should be set up so as to define where a problem is best suited to tree-based methods (sequential problems), or where the problem is best suited to traditional discrimination methods (parallel problems) though such criteria are unlikely to completely eliminate the fuzzy area between the two.

A final proposal for future methods is the further development of the FACT splitting criterion. Simulation studies have shown that splitting using F-ratios and univariate LDA was inappropriate for lognormally distributed data. One possible way of getting around this problem is to determine the best split by using non-parametric tests. In the case of two classes, the Mann-Whitney, or two sample Wilcoxon rank, test (see Lehmann, 1975), could be used for calculating the difference between two class medians. The assumptions for the test are that the two samples have the same shape and variances. For lognormal data, the assumption of equal variances may be violated. An alternative is the Kruskal-Wallis (or k-sample Mann-Whitney) test which has the sole assumption that the k classes all have the same shape. The observations from all classes are pooled together and ranked from 1 to n. A test statistic involving the average rankings for each of the k classes is then calculated and the variable with the largest of these is used to split on.

Another alternative is to use the MOOD median test (see Lehmann), which carries out a form of contingency table analysis. Firstly, the overall median is calculated for all the k classes pooled together. Then, for each class, MOOD determines the number of observations less than or equal to the overall median, and the number of observations greater than the overall median. This gives a 2 * k table of counts. A χ^2 test of independence or association is carried out on the table and the significance of the result calculated. The MOOD test is more robust to outliers than the Kruskal-Willis test, but is less efficient for normally distributed data. In such cases, the use of parametric tests would be preferable. After determining the best variable to split on, splitting can then be carried out on that variable by means of Raveh's non-metric discriminant analysis method (see Section 2.4.1), so that as many observations from the first class are greater than or less than those in the second class (assuming two classes).

This thesis has been designed to serve a number of purposes. Firstly, it provides a critical reference guide for current users of tree-based methods. Secondly, it gives guidelines as to when and where tree-based methods are best used. Thirdly, it offers recommendations as to which options should be employed when using the CART method. Finally, and no less importantly, some suggestions are made as to what a future tree-based method should look like. With increasing memory capabilities and processing speed, tomorrow's computers will provide a mechanism, ready and able to handle the development of more sophisticated and accurate, yet also more user-friendly, decision tree packages.

NOTATION INDEX

Below is given a list of the notation used in this thesis. The list is ordered by the number of the page that the term first appears. The general rule adopted here is that vectors and matrices appear in bold.

CART	2	Classification and Regression Trees
FACT	2	Fast Algorithm for Classification Trees
$\mathbf{x}$	5	vector of measurements for an observation
$D_B(\mathbf{x})$	5	Bayes classification rule, optimal rule of allocation
$f_i(\mathbf{x})$	5	conditional density function of $\mathbf{x}$
π_i	5	prior probability that $\mathbf{x}$ belongs to class i
R(B)	5	Bayes misclassification error rate, optimal error rate of any classifier
k	5	number of classes/populations/groups in the data set
Π_i	6	class/population/group i
$\boldsymbol{\mu}_i$	6	mean vector for class i
$\boldsymbol{\Sigma}_i$	6	covariance matrix for class i
$\boldsymbol{\Sigma}$	6	pooled covariance matrix for k classes
T(A, f)	6	total probability of misclassification
$D(\mathbf{x})$	7	the true discriminant function
δ^2	7	square of the true Mahalanobis distance between two classes
$R_1(T)$	7	true probability of misclassifying an observation belonging to class 1
$\Phi(.)$	8	cumulative normal distribution function
$E(D(\mathbf{x}))$	8	expected value of the true discriminant function
$se[D(\mathbf{x})]$	8	standard error of the true discriminant function
n_i	8	number of sample observations from class i
$\bar{\mathbf{x}}_i$	8	sample mean vector of the observations from class i
$\mathbf{S}_i$	8	sample covariance matrix of the observations from class i
$\mathbf{S}_p$	8	pooled estimate of the sample covariance matrix
$\hat{D}(\mathbf{x})$	8	linear discriminant function in the case of two populations
C(i/j)	8	cost of misclassifying an observation from class j to class i
$\hat{L}_i(\mathbf{x})$	9	group classification function for class i
$\hat{D}_{ij}(\mathbf{x})$	9	group separation function, linear discrimination function in the case of more than two groups
$Q(\mathbf{x})$	12	optimal rule of allocation in the case of unequal class covariance matrices
$\hat{Q}(\mathbf{x})$	12	quadratic discriminant function
LDA	12	linear discriminant analysis
QDA	12	quadratic discriminant analysis
p	12	number of variables/dimension of the data set
RDA	14	regularised discriminant analysis

λ, γ	14	regularisation parameters used in regularised discriminant analysis
NDA	14	non-metric discriminant analysis
$v(x/\Pi_m)$	15	number of class m sample points with values less than or equal to x
$\hat{F}(x/\Pi_m)$	15	cumulative distribution function estimate
$\hat{f}(x/\Pi_m)$	15	density function estimate
$k_0(z)$	15	smoothing weighting function used in kernel density estimation
$\hat{K}(x)$	16	kernel density discriminant function
K-NN	17	Kth nearest neighbour method
$d(x_j, x)$	17	distance function between x_j and x
AID	21	Automatic Interaction Detector
THAID	21	THeta AID
CHAID	21	CHi-squared AID
n	26	sample size, sample size at a node
r(t)	26	proportion of observations not from the class with the largest number of observations at each node
t	31	subset of the data, node
TSS_t	31	total sums of squares for node t
y_t	31	response in node t
BSS_j	31	between group sum of squares found after splitting on variable x_j
$\bar{y}_1$	31	mean of responses in the first subgroup of node t
BSS_m	31	maximum between group sums of squares over all variables
TSS_1	31	total sum of squares for the whole data set
R, Q, P, L	31	parameters used in AID stopping rules
$\theta_{y/x}$	34	Theta splitting criterion
n_t	34	total number of observations at node t
m_i	34	total number of misclassified observations in the ith split group
$\delta_{y/x}$	34	Delta splitting criterion
p_j	34	proportion of observations from class j in node t
p_{1j}	34	proportion of observations from class j in split group 1
$n_{min}, \theta_{min}, \delta_{min}$	35	parameters used in THAID stopping rules
m, n	36	the number of observations from classes 1 and 2 respectively (ID3 only)
I(m, n)	37	expected information needed to classify an object using an ID3 tree.
$A_1, ..., A_v$	37	v distinct categories of a variable
$t_1, ..., t_v$	37	descendant nodes of t
$I(m_i, n_i)$	37	information required to classify an object using a subtree from t_i
$E(x_j)$	37	expected information required for trees partitioned on variable x_j at the root node
$gain(x_j)$	37	information gained through branching on x_j
p(i/t)	41	probability an object can be assigned to class i at node t
p(j/t)	41	probability an object can be assigned to class j at node t

$i(t)$	42	estimated probability of misclassification under the Gini index
s	42	a split
$\Delta i(s, t)$	42	Gini splitting criterion
p_L, p_R	42	proportion of observations at node t sent left or right respectively by the split
$i(t_L)$	42	estimated probability of misclassification for the observations sent left by the split
C_1, C_2	42	amalgamation of classes, superclasses
$\Phi(s/t)$	42	twoing splitting criterion
T_{max}	43	a fully grown tree
T	43	a subtree of T_{max}
$R_\alpha(T)$	43	cost-complexity measure for T
$R(A_T)$	43	resubstitution error rate for T
α	43	cost-complexity parameter
$L(T)$	43	number of terminal nodes in T
T_α	43	subtree that minimises $R_\alpha(T)$
$R(\hat{T})$	44	independent estimate of the error rate, test sample or cross-validation error rate estimate
$se(R(\hat{T}))$	44	standard error of $R(\hat{T})$
s^*	44	optimal partition of a node t into t_L and t_R
s_j	44	split carried out on variable x_j
$p(s^*, s_j)$	45	probability that s_j sends the cases in t the same way as s^*
$p_{LL}(s^*, s_j)$	45	probability that both s^* and s_j send the cases in t, to the left
t'_L	45	set of observations sent left by s
$p(t)$	45	probability that an observation is in node t
$\tilde{s}_j$	45	surrogate split on variable x_j
$IV(x_j)$	47	correctness of the answer from splitting on x_j
$\sum K$	48	total number of observations in subtree T
$\sum J$	48	total number of misclassified observations in subtree T
L	48	pessimistic view of the number of misclassified observations in subtree T
$se(L)$	48	standard error of L
E	48	number of observations misclassified by the best terminal node within T
γ_{ij}	56	probability that the ith response falls in the jth class
$D(\mu_i, y_i)$	56	deviance function for an observation y_i
$D(\hat{\mu}; y)$	56	deviance of a node, sum of the deviances of all observations in the node
$D(\hat{\mu}_L, \hat{\mu}_R; y)$	56	combined deviance of the two descendant nodes
ΔD	56	difference between deviance of a node and the combined deviance of the two descendant nodes
$D_\alpha(T)$	57	cost-complexity measure for T using deviances
$\bar{y}(node)$	57	fitted value for each node

$\hat{y}$(parent)	57	shrunken fitted value for the node's parent
$\hat{y}$(node)	57	shrunken fitted value for the node's parent
$\bar{y}$(parent)	57	fitted value for the node's parent
R(T)	66	actual or true error rate, expected probability of misclassification when class conditional density functions are known
R(E)	67	the expected error rate for a learning sample of a given size
R(TS)	68	test sample error rate estimate
R(A)	68	apparent/resubstitution error rate estimate
R(H)	69	holdout error rate estimate
R(ROT)	69	rotation error rate estimate, twofold cross-validation error rate estimate
R(CV)	69	n-fold cross-validation error rate estimate, leave-one-out estimate, U estimate
$R_j^*(A)$	69	apparent error rate for the learning sample with the jth observation omitted
$R_J^*(A)$	70	average jackknife error rate estimate
$\hat{w}_J$	70	jackknife estimate of the bias of the apparent error rate
R(J)	70	jackknife error rate estimate
$\theta_i(\mathbf{x})$	70	posterior probability that $\mathbf{x}$ belongs to class i
R(PP)	70	posterior probability error rate estimate
C_j	71	class of observation $\mathbf{x}_j$
$\hat{C}_j$	71	predicted class of observation $\mathbf{x}_j$
$Q(C_j, \hat{C}_j)$	71	(0, 1) loss function
w	71	true bias involved in using the apparent error rate as an estimate of the actual error rate
$\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$	71	random sample of observations drawn with replacement from the learning sample, bootstrap sample
$R^*(A)$	71	apparent error rate of the bootstrap sample
$R^*(T)$	72	actual error rate of the bootstrap sample
p_{jb}^*	72	resampled proportion of observations in the bootstrap sample
$\hat{w}_b$	72	bias involved in using the apparent error rate of the bootstrap sample
$\hat{w}_B$	72	bootstrap estimate of the bias of the apparent error rate
R(BOOT)	72	bootstrap error rate estimate
R(0.632)	72	0.632 error rate estimate
$R(\varepsilon)$	72	average error rate for all observations not in the boostrap sample
$\hat{w}_{0.632}$	73	0.632 estimate of the bias of the apparent error rate
R(GCV)	73	g-fold cross-validation error rate estimate
$R(\tau)$	73	weighted estimate of the g-fold cross-validation and apparent error rates
$\mathbf{z}_{ij}$	74	standardised distribution with mean zero and standard deviation one
e	74	combination or prior probabilities and covariance matrices factor
R	76	classification method factor

ANOVA	76	analysis of variance
$se_{R(CV)}$	82	standard error of the misclassification cost
PPSS	93	priors proportional to sample size
R(i/j)	94	group/class misclassification error rates, proportion of observations from class i classified as class j
R(TEN)	108	tenfold cross-validation error rate estimate
R(ACV)	108	apparent error rate for CART trees chosen by n-fold cross-validation
R(AR)	108	apparent error rate for CART trees chosen by rotation
R(AT)	108	apparent error rate for CART trees chosen by tenfold cross-validation
$R(\hat{T})$	108	Any error rate estimating the actual error rate
MSE	108	expected value of the squared distance between an error rate estimate and the actual error rate
c_i	115	category i in a categorical variable
p_{ij}	116	probability of getting a response $x_j = 1$ for class i, probability pattern factor
r_{ijk}	116	correlation coefficient between x_j and x_k for class i, correlation factor
q	131	factor combination of means, dimension and correlation
ρ	131	population correlation coefficient
$\mathbf{P}_i$	131	population correlation matrix for class i
bias	132	expected value of the difference between the actual error rate and the error rate estimate
COUNT	132	proportion of samples for each factor combination in which the estimated error rate was less than the actual error rate
LR(T)	158	actual error rate after a logit transformation
$LR(\hat{T})$	158	error rate estimate after a logit transformation
PR(T)	158	actual error rate after a proportion transformation
$PR(\hat{T})$	158	error rate estimate after a proportion transformation
$m^*(t)$	230	reciprocal entropy index

BIBLIOGRAPHY

Aitchison, J. and Brown, J.A.C. (1957) The lognormal distribution with special reference to its uses in economics, Cambridge University Press.

Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977) A critical comparison of two methods of statistical discrimination, Applied Statistics, **26**, pp 15-25.

Anderson, J.A. (1966) Some nonparametric multivariate procedures based on statistically equivalent blocks, in "Multivariate Analysis", P.R. Krishnaiah (ed), New York: Academic Press, pp 5-27.

Ashikaga, T. and Chang, P.C. (1981) Robustness of Fisher's linear discriminant function under two-component mixed normal models, Journal of the American Statistical Association, **76**, pp 676-680.

Assael, H. (1970) Segmenting markets by group purchasing behaviour: an application of the A.I.D. technique, Journal of Marketing Research, **7**, pp 153-158.

Bellman, R.E. (1961) Adaptive Control Processes, Princeton, New Jersey: Princeton University Press.

Belson, W.A. (1959) Matching and prediction on the principle of biological classification, Applied Statistics, **8**, pp 65-75.

Biggs, D., de Ville, B. and Suen, E. (1991) A method of choosing multiway partitions for classification and decision trees, Journal of Appied Statistics, **18**, pp 49-62.

Bradford, E. (1993) Tree-based models in S, New Zealand Statistician, **28**, pp 36-51.

Breiman, L. (1968) Probability, Reading, Massachussets: Addison-Wesley.

Breiman, L. (1978) Description of chlorine tree development and use, Technical Report, Santa Monica, California: Technology Service Corporation.

Breiman, L. and Friedman, J.H. (1988) Commentary on "Tree-structured classification via generalised discriminant analysis", Journal of the American Statistical Association, **83**, pp 725-727.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees, Belmont, California: Wadsworth.

Bumpus, H.C. (1898) The elimination of the unfit as illustrated by the introduced species, Passer Domesticus, Biological Lectures, Marine Biology Laboratory, Woods Hole, 11th Lecture.

Buntine, W.L. (1992a) Learning classification trees, Statistics and Computing, **2**, pp 63-73.

Buntine, W.L. (1992b) Tree classification software, Presented at the Third National Technology Transfer Conference and Exposition, Baltimore.

Buntine, W.L. (1993) Personal Communication.

Buntine, W.L. and Caruana, R. (1993) Introduction to IND version 2.1 and recursive partitioning, Technical Report, Moffet Field, California: NASA Ames Research Centre.

Buntine, W.L. and Niblett, T. (1992) A further comparison of splitting rules for decision-tree induction, Machine Learning, **8**, pp 75-85.

Chernick, M.R., Murthy, V.K. and Nealy, C.D. (1985) Application of bootstrap and other resampling techniques: evaluation of classifier performance, Pattern Recognition Letters, **3**, pp 167-178.

Chernick, M.R., Murthy, V.K. and Nealy, C.D. (1986) Correction note to "Application of bootstrap and other resampling techniques: evaluation of classifier performance", Pattern Recognition Letters, **4**, pp 133-142.

Chittineni, C.B. (1977) On the estimation of probability of error, Pattern Recognition, **9**, pp 191-196.

Chou, P.A. (1991) Optimal partitioning for classification and regression trees, IEEE Transactions on Pattern Analysis and Machine Intelligence, **13**, pp 340-354.

Ciampi, A., Hogg, S.A., McKinney, S. and Thiffault, J. (1988) RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occuring in biostatistics, Computer Methods and Programs in Biomedicine, **26**, pp 239-256.

Ciampi, A., Lawless, J.F., McKinney, S.M. and Singhal, K. (1988) Regression and recursive partition strategies in the analysis of medical survival data, Journal of Clinical Epidermiology, **41**, pp 737-748.

Ciampi, A., Thiffault, T., Nakache, J-P. and Asselain, B. (1986) Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates, Computational Statistics and Data Analysis, **4**, pp 185-204.

Ciampi, A., Schiffrin. A., Thiffault, J., Quintal, H., Weitzner, G., Poussier, P. and Lalla, D. (1990) Cluster analysis of an insulin-dependent diabetic cohort towards the definition of clinical subtypes, Journal of Clinical Epidermiology, **43**, pp 701-715.

Clark, L.A. and Pregibon, D. (1992) Tree-based models, in "Statistical Models in S", J.M. Chambers and T.J. Hastie (eds), Pacific Grove, California: Wadsworth and Brooks/Cole, pp 377-419.

Cochran, W.G. (1968) Commentary on "Estimation of error rates in discriminant analysis", Technometrics, **10**, pp 204-205.

Cook, E.F. and Goldman, L. (1984) Empiric comparison of multivariate analysis techniques: advantages and disadvantages of recursive partitioning analysis, Journal of Chronic Diseases, **37**, pp 721-731.

Cover, T.M. and Hart, P.E. (1967) Nearest neighbor pattern classification, IEEE Transactions on Information Theory, **IT-13**, pp 697-722.

Crawford, S.L. (1989) Extensions to the CART algorithm, Journal of Man-Machine Studies, **31**, pp 197-217.

Crawford, S.L. and Souders, S.K. (1990) A comparison of two conceptual clustering algorithms, International Journal of Pattern Recognition and Artificial Intelligence, **4**, pp 409-420.

Crawford, S.L., Fung, R.M. and Tse, E. (1989) Data-driven assessment and decision making, in "Expert Systems in Economics, Banking and Management", L.F. Pau et al (eds), Amsterdam: North-Holland, pp 399-408.

de Ville, B. (1990) Applying statistical knowledge to database analysis and knowledge base construction, Proceedings of the Sixth Conference on Artificial Intelligence Applications, pp 30-36.

de Ville, B. (1994) Personal Communication.

Doyle, P. (1973) The use of Automatic Interaction Detector and similar search procedures, Operational Research Quarterly, **24**, pp 465-467.

Doyle, P. and Fenwick, I. (1975) The pitfalls of AID analysis, Journal of Marketing Research, **12**, pp 408-413.

Efron, B. (1979) Bootstrap methods: another look at the jackknife, Annals of Statistics, **7**, pp 1-26.

Efron, B. (1982) The Jackknife,the Bootstrap and Other Resampling Methods, Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., (1983) Estimating the error rate of a prediction rule: Improvement on cross-validation, Journal of the American Statistical Association, **78**, pp 316-331.

Einhorn, H.J. (1972) ~Alchemy in the behavioural sciences, Public Opinion Quarterly, **36**, pp 367-378.

Euromonitor (1979) European Marketing Data and Statistics, London: Euromonitor Publications.

Feng, C., Sutherland, A., King, R., Muggleton, S. and Henery, R. (1993) Comparison of machine learning classifiers to statistics and neural networks, Personal Communication.

Fielding, A. (1977) Binary segmentation: the Automatic Interaction Detector and related techniques for exploring data structure, in "The Analysis of Survey Data, Volume 1", C.A. O'Muircheartaigh and C.Payne (eds), London: John Wiley, pp 221-257.

Fisher, R.A. (1936) The use of multiple measurement in taxonomic problems, Annals of Eugenics, **7**, pp 179-188.

Fitzmaurice, G.M., Krzanowski, W.J. and Hand, D.J. (1991) A Monte Carlo study of the 632 estimator of error rate, Journal of Classification, **8**, pp 239-250.

Fix, E. and Hodges, J.L. (1951) Discriminatory analysis, nonparametric discrimination: consistency properties, Report No. 4, Project No. 21-49-004, USAF School of Medicine, Brooks Air Force Base, Randolph Field, Texas.

Friedman, J.H. (1977) A recursive partitioning decision rule for nonparametric classification, IEEE Transactions on Computers, **E-26**, pp 404-408.

Friedman, J.H. (1989) Regularised discriminant analysis, Journal of the American Statistical Association, **84**, pp 165-175.

Fukunaga, K. (1990) Introduction to Statistical Pattern Recognition, Second Edition, Boston: Harcourt Brace Jovanovich.

Ganesalingam, S. and Lynn, R.D. (1991) Posterior probability based estimator for the overall error rate associated with a linear discriminant function, Occasional Publications in Mathematics and Statistics, **23**, Massey University.

Ganeshanandam, S. and Krzanowski, W.J. (1990) Error rate estimation in two-group discriminant analysis using the linear discriminant function, Journal of Statistical Computation and Simulation, **36**, pp 157-175.

Glick, N. (1978) Additive estimators for probabilities of correct classification, Pattern Recognition, **10**, pp 211-222.

Gnanadesikan, R. (1977) Methods for Statistical Data Analysis of Multivariate Observations, New York: John Wiley.

Gong, G. (1986) Cross-validation, the jackknife and the bootstrap: excess error estimation in forward logistic regression, Journal of the American Statistical Association, **81**, pp 108-113.

Gordon, L. and Olshen, R.A. (1980) Consistent nonparametric regression from recursive partitioning schemes, Journal of Multivariate Analysis, **10**, pp 611-627.

Grajski, K.A., Breiman, L., Viana Di Prisco, G. and Freeman, W.J. (1986) Classification of EEG spatial patterns with tree-structured methodology: CART, IEEE Transactions on Biomedical Engineering, **33**, pp 1076-1086.

Habbema, J.D.F., Hermans, J. and van den Broek, K. (1974) A stepwise discriminant analysis program using density estimation, in "Compstat 1974", G.Bruckmann, F.Ferschl and L.Schmetterer (eds), Vienna: Physica-Verlag, pp 101-110.

Hadjimichael, M. and Wasilewska, A. (1993) Interactive inductive learning, International Journal of Man-Machine Studies, **38**, pp 147-167.

Hand, D.J. (1981) Discrimination and Classification, New York: John Wiley.

Hand, D.J. (1986) Recent advances in error rate estimation, Pattern Recognition Letters, **4**, pp 335-346.

Hanson, S.J. (1990) What connectionist models learn: learning and repesentation in connectionist neural networks, Brain and Behavioral Sciences, **13**, pp 471-511.

Heald, G.I. (1972) The application of the Automatic Interaction Detector (A.I.D.) programme and multiple regression techniques to the assessment of store selection and site performance, Operational Research Quarterly, **23**, pp 445-457.

Hosmer, D.W. and Lemeschow, S. (1989) Applied Logistic Regression, New York: Wiley.

Ildiko, E.F. and Lanteri, S. (1989) Classification models: discriminant analysis, SIMCA, CART, Chemometrics and Intelligent Laboratory Systems, **5**, pp 247-256.

Jennrich, R.I. (1977) Stepwise discriminant analysis, in "Statistical Methods for Digital Computers, Volume 3", K.Enslein, A.Ralston and H.S.Wilf (eds), New York: John Wiley, pp 76-95.

Kanal, L. (1974) Patterns in pattern recognition: 1968-1974, IEEE Transactions on Information Theory, **IT-20**, pp 697-722.

Kass, G.V. (1975) Significance testing in automatic interaction detection (A.I.D.), Applied Statistics, **24**, pp 178-189.

Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data, Applied Statistics, **29**, pp 119-127.

Kendall, M.G. (1966) Discrimination and classification, in "Multivariate Analysis", P.R. Krishnaiah (ed), New York: Academic Press, pp 165-185.

Konishi, S. and Honda, M. (1990) Comparison of procedures for estimation of error rates in discriminant analysis under non-normal populations, Journal of Statistical Computation and Simulation, **36**, pp 105-115.

Krzanowski, W.J. (1977) The performance of Fisher's linear discriminant function under non-optimal conditions, Technometrics, **19**, pp 191-200.

Kumar, K. (1993) Application of discriminant analysis and Mahalanobis distance to the family planning data, Personal Communication.

Kumar, K. and Srivastava, S. (1989) An analysis of the profile of accepters of family welfare programme in India, Paper presented in the International Union of the Scientific Study of the Population Conference, New Delhi.

Lachenbruch, P.A. (1967) An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis, Biometrics, **23**, pp 639-645.

Lachenbruch, P.A. (1975) Discriminant Analysis, New York: Hafner Press.

Lachenbruch, P.A. and Mickey, M.R. (1968) Estimation of error rates in discriminant analysis, Technometrics, **10**, pp 1-11.

Lachenbruch, P.A., Sneeringer, C. and Revo, L.T. (1973) Robustness of the linear and quadratic discriminant function to certain types of non-normality, Communications in Statistics - Simulation and Computation, **18**, pp 39-56.

Laurie, P. (1979) Beneath the City Streets, London: Granada Publishing.

LeBlanc, M. and Crowley, J. (1993) Survival trees by goodness of split, Journal of the American Statistical Association, **88**, pp 457-467.

Lehmann, E.L. (1975) Nonparametrics: Statistical Methods Based on Ranks, San Francisco: Holden-Day, Inc.

Lesaffre, E., Williams, J.L. and Albert, A. (1989) Estimation of error rates in multiple group logistic discrimination analysis. The approximate leaving-one-out method, Communications in Statistics - Simulation and Computation, **18**, pp 2989-3007.

Loh, W.Y. and Vanichsetakul, N. (1988) Tree-structured classification via generalised discriminant analysis, Journal of the American Statistical Association, **83**, pp 715-725.

Lubischew, A.A. (1962) On the use of discriminant functions in taxonomy, Biometrics, **18**, pp 455-477.

Lynn, R.D. and Brook, R.J. (1991) Classification by decision trees and discriminant analysis, New Zealand Statistician, **26**, pp 18-26.

Lynn, R.D., Brook, R.J. and Arnold, G.C. (1993) A comparison of four classification methods: linear and quadratic discriminant analysis, CART and FACT, Mathematical and Information Sciences Report, Series B: **1**, Massey University.

McLachlan, G.J. (1974) Estimation of the errors of misclassification on the criterion of asymptotic mean square error, Technometrics, **16**, pp 255-260.

McLachlan, G.J. (1977) A note on the choice of a weighting function to give an efficient method for estimating the probability of misclassification, Pattern Recognition, **9**, pp 147-149.

McLachlan, G.J. (1986) Assessing the performance of an allocation rule, Computing and Mathematics with Applications, **12A**, pp 261-272.

McLachlan, G.J. (1987) Error rate estimation in discriminant analysis: recent advances, in "Advances in Multivariate Statistical Analysis", A.K. Gupta (ed), Amsterdam: Dordrecht, pp 235-252.

McLachlan, G.J. (1992) Discriminant Analysis and Statistical Pattern Recognition, New York: John Wiley.

Mabbett, A., Stone, M. and Washbrook, J. (1980) Cross-validatory selection of binary variables in differential diagnosis, Applied Satistics, **29**, pp 198-204.

Margolese, M.M. (1970) Homosexuality: A new endocrine correlate, Hormones and Behaviour, **1**, pp 151-155.

Marks, S. and Dunn, O.J. (1974) Discriminant functions when covariance matrices are unequal, Journal of the American Statistical Association, **69**, pp 555-559.

Marshall, R.J. (1986) Partitioning methods for classification and decision making in medicine, Statistics in Medicine, **5**, pp 517-526.

Marshall, R.J. (1990) Non-hierarchical partition analysis for classification and for identification of subgroups in medical research, Personal Communication.

Messenger, R.C. and Mandell, M.L. (1972) A modal search technique for predictive nominal scale multivariate analysis, Journal of the American Statistical Association, **67**, pp 768-772.

Michie, D. (1986) Current developments in expert systems, in "Applications of Expert Systems, Volume 1", J.R. Quinlan (ed), Wokingham: Addison-Wesley, pp 137-156.

Michie, D. (1989) Problems of computer-aided concept formation, in "Applications of Expert Systems, Volume 2", J.R. Quinlan (ed), Wokingham: Addison-Wesley, pp 310-333.

Mingers, J. (1989) An empirical comparison of selection measures for decision-tree induction, Machine Learning, **3**, pp 319-342.

Minitab Inc., Version 9, State College, Pennsylvania.

Moore, D.H. II. (1973) Evaluation of five discrimination procedures for binary variables, Journal of the American Statistical Association, **68**, pp 399-404.

Moore, D.S. and McCabe, C.P. (1989) Introduction to the Practice of Statistics, San Francisco: Freeman.

Morgan, J.N. (1990) A conditional analysis of movers' housing responses, Journal of Economic Behaviour and Organisation.

Morgan J.N. (1993) Personal Communication.

Morgan, J.N. and Messenger, R.C. (1973) THAID: a Sequential Search Program for the Analysis of Nominal Scale Dependent Variables, Ann Arbor: Institute for Social Research, University of Michigan.

Morgan, J.N. and Sonquist, J.A. (1963) Problems in the analysis of survey data, and a proposal, Journal of the American Statistical Association, **58**, pp 415-434.

Morton, S.C. (1992) Personal crunching: new advances in statistical dendrology, Chance, **5**, pp 76-79.

Muxworthy, D.T. (1972) Review of AID III, British Sociological Association Maths, Statistics and Computing Applications Group Newsletter, **9**.

Oliver, J.L. (1993) Decision graphs - an extension of decision trees, to appear in Artificial Intelligence and Statistics, **14**.

O'Neill, J.L. (1986) Knowledge acquisition for radar classification, in "Applications of Expert Systems, Volume 1", J.R. Quinlan (ed), Wokingham: Addison-Wesley, pp 184-199.

Pawlak, Z., Wong, S.K.M. and Ziarko, W. (1988) Rough sets: probabilistic versus deterministic approach, International Journal of Man-Machine Studies, **29**, pp 81-95.

Picard, R.R. and Berk, K.N. (1990) Data splitting, American Statistician, **44**, pp 140-147.

Prada Sanchez, J.M. and Otero Cepeda, X.L. (1989) The use of smooth bootstrap techniques for estimating the error rate of a prediction rule, Communications in Statistics - Simulation and Computation, **18**, pp 1169-1186.

Quenouille, M. (1949) Approximate tests of correlation in time series, Journal of the Royal Statistical Society Series B, **11**, pp 18-84.

Quenouille, M. (1956) Notes on bias estimation, Biometrika, **43**, pp 356-360.

Quinlan, J.R. (1979) Discovering rules from large collections of examples: a case study, in "Expert Systems in the Micro Electronic Age", D. Michie (ed), Edinburgh: Edinburgh University Press.

Quinlan, J.R. (1983) Learning efficient classification procedures and their application to chess end games, in "Machine Learning, an Artificial Intelligence Approach, Volume 1", R.S. Michalski, J.G. Carbonell and T.M. Mitchell (eds), Palo Alto, California: Tioga, pp 463-482.

Quinlan, J.R. (1986) Induction of decision trees, Machine Learning, **1**, pp 81-106.

Quinlan, J.R. (1987) Simplifying decision trees, International Journal of Man-Machine Studies, **27**, pp 221-234.

Quinlan, (1993) Comparing connectionist and symbolic learning methods, in "Computational Learning Theory and Natural Learning Systems: Constraints and Prospects", S. Hanson, G. Drastal and R. Rivest (eds), Cambridge, Massachussets: MIT Press.

Quinlan, J.R. and Rivest, R.L. (1989) Inferring decision trees using the minimum description length principle, Information and Computation, **80**, pp 227-248.

Quinlan, J.R., Compton, P.J., Horn, K.A. and Lazarus, L. (1986) Inductive knowledge acquisition: a case study, in "Applications of Expert Systems, Volume 1", J.R. Quinlan (ed), Wokingham: Addison-Wesley, pp 157-173.

Raveh, A.A. (1989) A nonmetric approach to linear discriminant analysis, Journal of the American Statistical Association, **84**, pp 176-183.

Rawlings, J.O. (1988) Applied Regression: A Research Tool, Belmont, California: Wadsworth.

Rutter, C., Flack, V. and Lachenbruch, P.A. (1991) Bias in error rate estimates in discriminant analysis when stepwise variable selection is employed, Communications in Statistics, **20**, pp 1-22.

Safavian, S.R. and Landgrebe, D. (1991) A survey of decision tree classification methodology, IEEE Transactions on Systems, Man and Cybernetics, **21**, pp 660-674.

Salahuddin and Hawkes, A.G. (1991) Cross-validation in stepwise regression, Communications in Statistics - Theory and Methods, **20**, pp 1163-1182.

Sankar, A. and Mammone, R.J. (1991) Neural tree networks, in "Neural Networks: Theory and Applications", R.J. Mammone and Y.Y. Zeeve (eds), Boston: Harcourt Brace Jovanovich, pp 281-302.

SAS Institute Inc., Version 6.04, Cary, North Carolina.

Schaffer, C. (1993) Selecting a classification method by cross-validation, Personal Communication.

Schwartz, S., Wiles, J., Gough, I. and Phillips, S. (1993) Connectionist, rule-based and Bayesian decision aids: an empirical comparison, in "Artificial Intelligence Frontiers in Statistics", D.J. Hand (ed), London: Chapman & Hall, pp 264-278.

Seber, G.A.F. (1984) Multivariate Observations, New York: John Wiley.

Segal, M.R. (1988) Regression trees for censored data, Biometrics, **48**, pp 35-47.

Shlien, S. (1990) Multiple binary decision tree classifiers, Pattern Recognition, **23**, pp 757-763

Snapinn, S.M. and Knoke, J.D. (1985) An evaluation of smoothed classification error-rate estimators, Technometrics, **27**, pp 199-206.

Snapinn, S.M. and Knoke, J.D. (1989) Estimation of error rates in discriminant analysis with selection of variables, Biometrics, **45**, pp 289-299.

Sonquist, J.A. (1970) Multivariate Model Building: The Validation of a Search Strategy, Ann Arbor, Institute for Social Research, University of Michigan.

Taylor, P.C. and Silverman, B.W. (1993) Block diagrams and splitting criteria for classification trees, Statistics and Computing, **3**, pp 147-161.

Todeschini, R. and Marengo, E. (1992) Linear discriminant classification tree: A user-driven multi-criteria classification method, Chemometrics and Intelligent Laboratory Systems, **16**, pp 25-35.

Toussaint, G.T. (1974) Bibliography on estimation of misclassification, IEEE Transactions on Information Theory, **20**, pp 472-479.

Wahl, P.W. and Kronmal, R.A. (1977) Discriminant functions when covariances are unequal and sample sizes are moderate, Biometrics, **33**, pp 479-484.

Wernecke, K-D. and Kalb, G. (1983) Further results in estimating the classification error in discriminance analysis, Biometrical Journal, **25**, pp 247-258.

Wernecke, K-D., Kalb, G. and Sturzebecher, E. (1980) Comparison of various procedures for estimation of the classification error in discriminance analysis, Biometrical Journal, **22**, pp 639-649.

Wolberg, W.H., Tanner, M.A., Loh, W.Y. and Vanichsetakul, N. (1987) Statistical approach to fine needle aspiration diagnosis of breast masses, Acta Cytologica, **31**, pp 731-741.

ADDENDA

*p*5, *l* 2: change "is defined to be" to "assigns a random observation **x** to population Π_i if".

*p*5, *l* -5: change "Let us" to "In the case of two p-dimensional multivariate normal populations (the general multivariate ellipsoidal case is not considered here), let us".

*p*5, *l* -1: in formula (2.2.1) change "Σ^{-1}" to "Σ_i^{-1}".

*p*6, *l* 11: omit "much".

*p*7: in the formula after (2.2.9) change "$R_1(T) = \Phi[...]$" to "$R_1(T) = Pr[...]$" and change "$\ln(\pi_1/\pi_2)$.

*p*7, *l* -1: after "cumulative" add "standardised".

*p*9, *l* 11: change "$\hat{D}_{ji} < 0, \ \forall\ i > j$".

*p*9, *l* 16: change "$\hat{D}_{12}(\mathbf{x})$" to "$\hat{D}_{13}(\mathbf{x})$".

*p*42, *l* 7: change "C_1 and C_2" to "SC_1 and SC_2".

*p*48, *l* -10: change "$L = \Sigma J + L(T)/2$" to "$E + 1/2$".

*p*65, *l* 5: after "error rate" add "estimate".

*p*66, Note: The discussion from (4.2.2) on is restricted to multivariate normal data using LDA.

*p*69: on a new line after (4.2.7) add "where n_{ij} is the number of observations from Π_j falsely classified to Π_i, $i \neq j$".

*p*74, *l* 12: after "(e)." add "The above model involves a full factorial design. It may have been better to adopt some form of fractional design, so allowing a wider range of factors to be explored. However, as is almost always the case in simulation studies, the possible range of factors that can be explored is vast, so that a line has to be drawn somewhere."

*p*74, *l* -4: change "$\mathbf{x}_{ij}$ is lognormal" to "log ($\mathbf{x}_{ij}$) is normal (0,1)".

*p*74, *l* -3: delete "which is lognormal (0, 1)".

*p*74: *l* -1 On a new line add "In summary, univariate normal data, $\mathbf{x}_j$, was generated for each dimension j, then transformed to $\mathbf{y}_j = \exp(\mathbf{x}_j)$. Finally, the data was standardised giving marginals with mean 0 and standard deviation 1, that is,

$$\mathbf{z}_j = (\mathbf{y}_j - \bar{\mathbf{y}}_j) \,/\, \text{estimated s.d.}(\mathbf{y}_j).$$ "

The ranking of observations has changed after standardisation though the $\mathbf{z}_j$ are independent (uncorrelated) as the $\mathbf{x}_j$ are independent."

*p*75, *l* -3: after "(1990)." add "The two tree-based methods were chosen because of their ready availability and representing two different approaches to tree-based classification. LDA and QDA were selected because they are the two most commonly used classification methods."

*p*76, *l* -14: After "here." add "It may have been preferable to have used a separate test set instead of the cross-validation method which does introduce possible error. It was decided to use cross-validation instead of an independent test set as the former is used more often in the real world as large test sets are usually unavailable."

*p*82, *l* 8: after "was used." add "This decrease in LDA error rate between normal and lognormal is most probably due to the effects of standardisation which maintains the theoretical covariance differences, hence distances between populations. However, as the distributions of the two populations are skewed, the lower 75% of the distribution will be bunched together around a high peak, thus closer to the respective class mean than in the case of normally distributed data. The net effect is that fewer observations are misclassified for standardised lognormal data."

*p*83, *l* -9: after "transformation." add "In this case, for pure lognormally distributed data, the actual values of δ will be different from those given in Section 4.3.1."

*p*83, *l* -7: after "true" add "(pure)".

*p*83, *l* -6: change "log[f(.)]" to 'ln[f($\mathbf{x}$)]".

*p*91, *l* -2: after "data" add ", though this alters the correlations between variables so that the covariance matrices are different from those in (4.3.2)."

*p*92, *l* 3: change "covariance" to "variance".

*p*93, *l* -5: insert a transpose symbol (') between ")" and "$\mathbf{S}^{-1}$".

*p*94, *l* -1: on a new line add "though absolute differences are used in the graphs to make for easier comparison between methods."

*p*96, *l* 3: after "rates" add "(to make the graph easier to read)".

*p*97*ff*: "e" classifications as on *p* 75.

*p*116: after *l* -7 add "with $\beta_j > 0$, j = 1, 2."

*p*117, *l* -13: after "theirs." add "Binary data rather than general categorical data was also used for simplicity. Using categorical variables containing more than two categories would involve creating a large number of binary variables to use in LDA and QDA."

*p*117, *l* -11: after "n" add "(total sample size)".

*p*117, *l* -8: after "$r_{ijk} = 0.25$" add ", $j \neq k$"

*p*119: omit "Level" from Table 5.2.

*p*121: omit "p_{ij} =" from Table 5.4.

*p*122, *l* -10: after "of R(T)." add "The mean square error (MSE) criterion was used to compare different error rate estimators for each method (see p 108)."

*p*126: omit "Level" from Table 5.7.

*p*130, *l* -7: change "of the sample" to "associated with the classification tree".

*p*132, *l* -2: change "the F-ratio should not be regarded as a true measure of the statistical significance of each result." to "a statistically significant F-ratio may not be of substantive significance."

*p*165, *l* -9: change "variances" to "covariance matrices".

*p*171, *l* -12: change "varible" to "variable".

*p*178: in Table 7.6 (and at the bottom of *p*178), change "**R(i/j)**" to " Σ_j**R(i/j)**".

*p*184, *l* 16: after "significant" add "(with univariate tests)".

*p*186, *l* 14: convert "change" to "changed".

*p*187: Note: Husb_Edu and Wife_Edu are ordinal categorical variables with values ranging from zero to seven and refer to the education level of the husband and wife respectively. No_Child refers to the number of children in the family.

*p*209, *l* -8: omit "enough".

*p*231, *l* 11: change "significant" to "major".

*p*236: on a new line after *l* 12 add "US 23 univariate split".

*p*237, *l* 8: change "C_1, C_2" to "SC_1, SC_2".

*p*238: on a new line after *l* 12 add "n_{ij} 69 the number of observations from Π_j falsely classified to Π_i".

*p*239: after "Breiman, ... (1984)..." add the reference "Brown, D.E., Corruble, V. and Pittard, C.L. (1993)" A comparison of decision tree classifiers with back-propagation neural networks for multimodal classification problems, Pattern Recognition, **26**, pp 953-961."

*p*239 (2): Note: The page numbers for the Bibliography are wrong. This section should start on *p*241 and all other pages should be put back two pages.