Variable selection for classification in complex ophthalmic data: a multivariate statistical framework

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by:

Padraic Eoin Walsh

September 2017

Abstract

Author: Padraic Eoin Walsh
Thesis title: Title: Variable selection for classification in complex ophthalmic data: a multivariate statistical framework

Variable selection is an essential part of the process of model-building for classification or prediction. Some of the challenges of variable selection are heterogeneous variance-covariance matrices, differing scales of variables, non-normally distributed data and missing data. Statistical methods exist for variable selection however these are often univariate, make restrictive assumptions about the distribution of data or are expensive in terms of the computational power required.

In this thesis I focus on filter methods of variable selection that are computationally fast and propose a metric of discrimination. The main objectives of this thesis are (1) to propose a novel Signal-to-Noise Ratio (SNR) discrimination metric accommodating heterogeneous variance-covariance matrices, (2) to develop a multiple forward selection (MFS) algorithm employing the novel SNR metric, (3) to assess the performance of the MFS-SNR algorithm compared to alternative methods of variable selection, (4) to investigate the ability of the MFS-SNR algorithm to carry out variable selection when data are not normally distributed and (5) to apply the MFS-SNR algorithm to the task of variable selection from real datasets.

The MFS-SNR algorithm was implemented in the R programming environment. It calculates the SNR for subsets of variables, identifying the optimal variable during each round of selection as whichever causes the largest increase in SNR. A dataset was simulated comprising 10 variables: 2 discriminating variables, 7 non-discriminating variables and one non-discriminating variable which enhanced the discriminatory performance of other variables. In simulations the frequency of each variable's selection was recorded. The probability of correct classification (PCC) and area under the curve (AUC) were calculated for sets of selected variables. I assessed the ability of the MFS-SNR algorithm to select variables when data are not normally distributed using simulated data.

I compared the MFS-SNR algorithm to filter methods utilising information gain, chi-square statistics and the Relief-F algorithm as well as a support vector machines and an embedded method using random forests. A version of the MFS algorithm utilising Hotelling's T^2 statistic (MFS-T2) was included in this comparison. The MFS-SNR algorithm selected all 3 variables relevant to discrimination with higher or equivalent frequencies to competing methods in all scenarios. Following non-normal variable transformation the MFS-SNR algorithm still selected the variables known to be relevant to discrimination in the simulated scenarios.

Finally, I studied both the MFS-SNR and MFS-T2 algorithm's ability to carry out variable selection for disease classification using several clinical datasets from ophthalmology. These datasets represented a spectrum of quality issues such as missingness, imbalanced group sizes, heterogeneous variance-covariance matrices and differing variable scales. In 3 out of 4 datasets the MFS-SNR algorithm out-performed the MFS-T2 algorithm. In the fourth study both MFS-T2 and MFS-SNR produced the same variable selection results.

In conclusion I have demonstrated that the novel SNR is an extension of Hotelling's T^2 statistic accommodating heterogeneity of variance-covariance matrices. The MFS-SNR algorithm is capable of selecting the relevant variables whether data are normally distributed or not. In the simulated scenarios the MFS-SNR algorithm performs at least as well as competing methods and outperforms the MFS-T2 algorithm when selecting variables from real clinical datasets.

Acknowledgements

I wish to extend my most sincere thanks and appreciation to the many people who have assisted me over the last 4 years and provided their support, both material and moral. I wish to thank my supervisors, Dr. Gabriela Czanner, Dr. Marta (Garcia-Finana) Van der Hoek and Prof. Simon Harding. It would not have been possible for me to complete this thesis without the vast breadth of their expertise and patience.

I'm very grateful to the staff of St. Paul's Eye Unit of the Royal Liverpool University Hospital for all of their efforts collecting the real clinical data which I used in my work. Without the diligence and hard work of everyone involved I would have had no real data to use.

I would also like to thank my colleagues within the Biostatistics and the Eye and Vision science departments. There was a great willingness to assist and render constructive feedback within both departments during my time in Liverpool.

Finally I wish to express my deepest gratitude to my parents both of whom have gone above and beyond with their unwavering support and patience over the last 4 years. I only hope I can muster as much when my turn comes around!

Contents

Abstract II

Acknowledgements III

Abbreviations IX

List of figures XVII

Symbols and statistical notation XIX

Chapter 1. Introduction 1

1.1 The importance of biomarkers 1

1.2 The search for biomarkers is a statistical problem of variable selection for discrimination 2

1.3 Aim of this thesis 3

1.4 The structure of this thesis 3

Chapter 2. Literature review 5

2.1 Introduction 5

2.2 PCA as a variable selection method for classification 5

2.3 Review of filter methods of variable selection 8

2.3.1 Filter methods based on Kullback-Leibler divergence 8

2.3.2 Filter methods based on Biomarker Identifier Measure 11

2.3.3 Filter methods for variable selection based on correlation between predictors and outcome group variable 14

2.3.4 Filter variable selection methods based on correlations between predictor variables 17

2.3.5 Filter variable selection methods based on conditional mutual information 20

2.3.6 Filter variable selection methods based on t-scores 25

2.3.7 Probabilistic approach to variable selection 27

2.4 Wrapper methods for variable selection 29

2.5 Embedded methods for variable selection 30

2.6 Information evaluation in signal-processing theory and SNR measures 38

2.7 Discussion 41

2.7.1 DREFUS 42

2.7.2 ISDR 43

2.7.3 MRet 44

2.7.4 Keratoconus 45

Chapter 3. Methodology: a multivariate SNR metric for variable selection in classification 46

3.1 Introduction 46

3.2 Discrimination methods 47

3.2.1 Linear discriminant analysis 47

3.2.2 Quadratic Discriminant Analysis 51

3.3 Discriminatory performance of a classifier 53

3.4 Hotelling's T^2 statistic 56

3.5 Proposed SNR statistic 58

3.6 Proof of equivalency of Hotelling's T^2 statistic and SNR when variance-covariance matrices are homogeneous 60

3.7 A variable selection algorithm that is based on Hotelling's T^2 64

3.8 Proposed variable selection algorithm that is based on SNR 66

3.9 Computational challenges 67

3.9.1 Stopping criteria 67

3.9.2 Estimation of the probability of correct classification 68

3.9.3 Deciding on the size of simulations 70

3.10 Discussion 73

CHAPTER 4. A comparison of the variable selection performance of the MFS-SNR algorithm with existing alternative methods using simulated normal data 75

4.1 Introduction 75

4.2 Data-generating mechanisms 75

4.3 Comparators 80

4.4 Performance measures 82

4.5 Results of comparison of the MFS-SNR algorithm to existing variable selection methods in simulated data 83

4.5.1 Variable selection in the presence of a small difference between the correlation matrices of the groups. 84

4.5.2 Variable selection in the presence of a large difference between the correlation matrices of the groups. A case of high correlation between the variables $X1$ and $X3$. 87

4.5.3 Variable selection in the presence of a large difference between the correlation matrices of the groups. A case of high correlation between the variables $\boldsymbol{X2}$ and $\boldsymbol{X3}$. 91

4.6 Discussion 94

4.6.1 Effect of sample size 94

4.6.2 Computing time considerations 95

4.6.3 Effect of correlation differences across groups 96

4.7 Conclusion 97

Chapter 5. Performance of the MFS-SNR algorithm for non-normal data 98

5.1 Aim of simulation 98

5.2 Data generating mechanism 99

5.2.1 Simulated data 101

5.2.2 The simulation of dichotomised, trichotomised and log-normal transformed data 103

5.3 Evaluation of MFS-SNR algorithm and of variable selection in simulated non-Normal data ... 104

5.4 Discussion 106

Chapter 6. Application to real data 108

6.1 Introduction 108

6.2 Application of methods to discriminate between disease stages in diabetic retinopathy [DREFUS dataset] 109

6.2.1 Introduction 109

6.2.2 Methods 109

6.2.3 Results 110

6.2.4 Discussion 115

6.3 Application of methods to predict conversion to sight threatening diabetic retinopathy [ISDR] 116

6.3.1 Introduction 116

6.3.2 Methods 116

6.3.3 Results 119

6.3.4 Discussion 126

6.4 Application of methods to discriminate between survival outcome groups in malarial retinopathy 127

6.4.1 Introduction 127

6.4.2 Methods 128

6.4.3 Results 129

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows. 135

6.4.4 Discussion 135

6.5 Application of methods to discriminate between disease stages in keratoconus 136

6.5.1 Introduction 136

6.5.2 Methods 136

6.5.3 Results 136

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows. 142

6.5.4 Discussion 142

6.6 Conclusions 143

Chapter 7. Conclusions and further work 146

7.1 Introduction 146

7.2 Summary of findings 146

7.2.1 Summary of findings from literature review 146

7.2.2 Summary of findings on the SNR metric 147

7.2.3 Summary of findings on implementation of the MFS-SNR algorithm 148

7.2.4 Summary of findings from simulated data 148

7.2.5 Summary of findings from real data 151

7.4 Recommendations for practice 153

7.4.1 Discriminatory metric 153

7.4.2 Variable selection for classification 153

7.5 Recommendations for future research 154

7.5.1 Future research on the discriminatory SNR metric 154

7.5.2 Future research on MFS-algorithms for variable selection in classification 154

7.6 Conclusions 155

Bibliography 157

Appendix – R code 169

MFS-T2 algorithm 170

mfs function: 170

stat function: 173

MFS-SNR algorithm 175

mfs function: 175

stat function: 178

Abbreviations

ALL:	Acute lymphocytic leukemia
AMD:	Age-related macular degeneration
AML:	Acute myeloid leukemia
ARI	Adjusted Rand index
AUC:	Area under the curve
AUROC:	Area under the ROC curve
BIC:	Bayesian information criterion
BMI:	Body mass index
BMI:	Biomarker identifier
BP:	Blood pressure
BRCA:	Breast Cancer susceptibility gene
CDF:	Cumulative distribution function
CFS:	Correlation-based feature selection
Chol:	Cholesterol
CMI:	Conditional mutual information
CS:	Contrast sensitivity
CSMO:	Clinically significant macular oedema
DC-SIS:	Distance correlation-based method with the sure independence screening property
DDA:	Diagonal discriminant analysis
DEG:	Differentially expressed gene
DP:	Diastolic pressure

DR:	Diabetic retinopathy
DREFUS:	Diabetic retinopathy: Functional and structural study
DVA:	Distance visual acuity
ECM:	Expected cost of misclassification
ERCC1:	Excision Repair Cross-Complementation Group 1
ERG:	Electroretinogram
ETDRS:	Early treatment of diabetic retinopathy study
FCBF:	Fast correlation based filter
fMRI:	Functional magnetic resonance imaging
FNDR:	False non-discovery rate
GFR:	Glomerular filtration rate
GP:	General practitioner
GLU:	Glutamic acid
HbA1c:	Haemoglobin A1c
HDL:	High density lipoprotein
HVDM:	Heterogeneous value difference metric
IG:	Information gain
ISDR:	Individual risk-based screening for diabetic retinopathy
KL:	Kullback-Liebler
LASSO:	Least absolute shrinkage and selection operator
LDA:	Linear discriminant analysis
LDL:	Low density lipoprotein
LOOCV:	Leave-one-out cross-validation

LR:	Logistic regression
LTG:	Low tension glaucoma
MAR:	Missing at random
MAR:	Minimum angle of resolution
MCAR:	Missing completely at random
mfERG:	Multifocal electroretinography
MFS:	Multiple forward selection
MI:	Mutual information
MNAR:	Missing not at random
MP:	Microperimetry
MPOD:	Macular pigment optical density
MrMr:	Minimal-redundancy maximal-relevance
MR:	Malarial retinopathy
MRet:	Research programme, “The retinal microvasculature in cerebral malaria in African children” (Wellcome Trust, 092668/Z/10/Z)
nNOS:	Neuronal nitric oxide synthase
NO:	Nitric oxide
NOSs:	Nitric oxide synthases
NP:	Non-deterministic polynomial time
NPV:	Negative predictive value
OCT:	Optical coherence tomography
OP:	Oscillatory potential
PC:	Principal component
PCC:	Probability of correct classification

PCA:	Principal components analysis
PCR:	Polymerase chain reaction
PPV:	Positive predictive value
QDA:	Quadratic discriminant analysis
RF:	Random forest
RFE:	Recursive feature elimination
ROC:	Receiver operating characteristic
SIS:	Sure independence screening
SIRS:	Sure independent ranking and screening
SNP:	Single nucleotide polymorphism
SNR:	Signal-to-noise ratio
SP:	Systolic pressure
STDR:	Sight-threatening diabetic retinopathy
SU:	Symmetrical uncertainty
SVM:	Support vector machine
tSNR:	Temporal SNR
TC:	Total count
TNF-α:	Tumour necrosis factor-alpha
TPM:	Truncated product method
VDM:	Value difference metric
VEGF:	Vascular endothelial growth factor
VSCC:	Variable selection for clustering and classification

List of tables

Table 3.1.1: Summary of the known and developed novel concepts explained in Chapter 3. 46

Legend: The text in italic represents the two of the three contributions of this thesis. 46

Table 3.3.1 Outline of a contingency table 54

Table 3.6.1 SNR values, Hotelling's T^2 statistics, means, variances and correlations for the variable set containing X1 and X2 for each of the variance and correlation scenarios. 63

Table 3.9.1.1 Hotelling T^2 statistics, PCC estimates and p-values for subsets of normally distributed, uncorrelated variables 68

Table 3.9.2.1 Contingency table showing hypothetical classification results....................................69

Table 3.9.3.1 Selection frequencies and PCC ranges (maximum-minimum) for all variables simulated 10,000 times for group sizes of 20, 100 and 1,000. 72

Table 3.9.3.2 Selection frequencies (in %) and PCC ranges (maximum-minimum) for all variables simulated 1,000 times for group sizes of 20, 100 and 1,000. 72

Table 4.5.1.1 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n0 = n1 = 40$ (Scenario 1). 85

Table 4.5.1.2 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n0 = n1 = 400$ (Scenario 4). 86

Table 4.5.1.3 Selection frequencies (in %) and AUC/PCC estimates (in %) for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n0 = n1 = 1{,}000$ (Scenario 7). 86

Table 4.5.1.4 Selection frequencies (in %) and AUC/PCC estimates (in %) for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n0 = 50$, $n1 = 150$ (Scenario 10). 86

Table 4.5.2.1 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X1 and X2 in group 1 and $n0 = n1 = 40$ (Scenario 2). 88

Table 4.5.2.2 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X1 and X2 in group 1 and $n0 = n1 = 400$ (Scenario 5). 89

Table 4.5.2.3 Selection frequencies (in %) and AUC and PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X1 and X2 in group 1 and $n0 = n1 = 1{,}000$ (Scenario 8). 89

Table 4.5.2.4 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $\mathrm{n0} = 50$, $\mathrm{n1} = 150$ (Scenario 11). 90

Table 4.5.3.1 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between $\mathrm{X1}$ and $\mathrm{X3}$ in group 1 and $\mathrm{n0} = \mathrm{n1} = 40$ (Scenario 3). 92

Table 4.5.3.2 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between 92

Table 4.5.3.3 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between $\mathrm{X1}$ and $\mathrm{X3}$ in group 1 and $\mathrm{n0} = \mathrm{n1} = 1{,}000$ (Scenario 9). 93

Table 4.5.3.4 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $\mathrm{n0} = 50$, $\mathrm{n1} = 150$ (Scenario 12). 93

Table 4.6.2.1 Run-time (mins:secs) for each variable selection method applied to 1,000 simulated datasets under the conditions of scenario 2 96

Table 5.2.1 Dummy variable coding for a factor with 3 levels; X, Y and Z. Level Z used as the reference level. 100

Table 5.2.2 Effect variable coding for a factor with 3 levels; X, Y and Z. Level Z used as the reference level. 101

Table 5.2.3 Contrast variable coding for a factor with 3 levels; X, Y and Z. Level Z used as the reference level. 101

Table 5.3.1 Variable selection frequencies and performance estimates for dichotomised and trichotomised variables for groups sizes of $\mathrm{n0} = \mathrm{n1} = 100$ and $\mathrm{n0} = \mathrm{n1} = 500$ from scenarios 1, 2, 3, 4, 5, 6, 7, and 8 104

Table 6.2.3.1. Univariate analysis of all 27 variables in the early DR and no DR groups 112

Table 6.2.3.2 Multivariate selections using the MFS-T2 algorithm for discrimination between the early DR and no DR groups 113

Table 6.2.3.3 Multivariate selections using the MFS-SNR algorithm for discrimination between the early DR and no DR groups 113

Table 6.3.2.1 Definition of risk scores assigned to each patient and definition of the two groups used in our analyses 117

Table 6.3.2.2 Definition of variables in dataset for discrimination between referable and non-referable STDR 118

Table 6.3.2.3 Coding scheme for dummy variables X1, X2, X3 and X4 119

Table 6.3.3.1 Univariate analysis of all variables in the background DR and proliferative DR groups 120

Table 6.3.3.2 Selection frequencies for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, frequencies are only shown for selected variables 122

Figure 6.3.3.1 Average diagnostic estimate values for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset 122

Table 6.3.3.3 Selection frequencies for variable selections using MFS-SNR made from each of 1,000 samples taken without replacement, frequencies are only shown for selected variables 123

Figure 6.3.3.2 Average diagnostic estimate values for variable selections using MFS-SNR made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset 123

Table 6.3.3.4 Selection frequencies for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, frequencies are only shown for selected variables 123

Figure 6.3.3.3 Average diagnostic estimate values for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset 124

Table 6.3.3.5 Selection frequencies for variable selections using MFS-SNR made from each of 1.000 samples taken without replacement, frequencies are only shown for selected variables 124

Figure 6.3.3.4 Average diagnostic estimate values for variable selections using MFS-SNR made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset 125

Table 6.4.3.1 Univariate analysis of all variables in the survival and death groups. 130

Table 6.4.3.2 Multivariate selections by the MFS-T2 algorithm for discriminating between the death and survival groups. 132

Table 6.4.3.3 Multivariate selections by the MFS-SNR algorithm for discriminating between the death and survival groups 132

Table 6.4.3.4 Multivariate selections by the MFS-T2 algorithm for discriminating between the dead and survivor groups following imputation of the malarial retinopathy dataset 132

Table 6.4.3.5 Multivariate selections by the MFS-SNR algorithm for discriminating between the death and survival groups following imputation of the malarial retinopathy dataset 133

Table 6.4.3.6 Multivariate selections by the MFS-T2 algorithm for discriminating between the death and survival groups using optimal priors identified for the dataset produced following imputation of the malarial retinopathy dataset 133

Table 6.4.3.7 Multivariate selections by the MFS-T2 algorithm for discriminating between the death and survival groups using optimal priors identified for the dataset produced following imputation of the malarial retinopathy dataset. 133

Table 6.4.3.8 Variable selected by MFS-T2, T^2 statistics and PCC estimates for selections made from combinations of image sectors 1-24, statistics extracted from sector variables and clinical variables. 134

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows........ 134

Table 6.4.3.9 Variable selected by MFS-SNR, SNR values and PCC estimates for selections made from combinations of image sectors 1-24, statistics extracted from sector variables and clinical variables. 134

Table 6.5.3.1 Univariate analysis of all variables in the healthy and keratoconus groups for a dataset using the averages of each variable........ 137

Table 6.5.3.2 Univariate analysis of all variables in the healthy and keratoconus eyes groups for a dataset using the first measurement of each variable 139

Table 6.5.3.3 Variable selection by the MFS-T2 algorithm from a dataset composed of the variable averages from the Keratoconus data........ 140

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows......... 140

Table 6.5.3.4 Variable selection by the MFS-SNR algorithm from a dataset composed of the variable averages from the Keratoconus data........ 141

Table 6.5.3.5 Variable selection by the MFS-T2 algorithm from a dataset composed of the first measurement for each variable from the Keratoconus data. PCC estimates in each row are calculated for the variable in that row and all previous rows........ 141

Table 6.5.3.6 Variable selection by the MFS-SNR algorithm from a dataset composed of the first measurement for each variable from the Keratoconus data 142

List of figures

Figure 3.2.1.1 Regions R1 and R2 for two populations defined by variables X1 and X2 50

Source: Johnson & Wichern, Applied multivariate statistical analysis, 2008 50

Figure 3.2.2.1 Regions R_1 and R_2 for tow normally distributed groups with heterogeneous variance-covariance matrices 52

Source: Johnson & Wichern, Applied multivariate statistical analysis, 2008 52

Figure 3.3.1 An example of an ROC curve.. 55

Source: Hosmer & Lemeshow, Applied logistic regression, 2001 55

Figure 3.6.1 Bivariate plots of X1 vs. X2 across simulated groups under three different scenarios 63

Legend: (left) same variances and different correlations, (middle) different variances and same correlation; and (right) same variances and same correlations. 63

Table 4.2.1 Group sizes and variance-covariance matrices for each of 12 simulation scenarios......... 79

Figure 4.2.1 Bivariate plots of X1, X2, X3, for $\mathrm{n0} = \mathrm{n1} = 400$ using Cor(X1/X3)=0.42 in Group 1 and Cor(X1/X3)=0.11 in Group 1 (Scenario 4) 80

Figure 6.2.3.1 Pairwise bivariate plots of Cholesterol, mfERG Central density and HbA1c across the early DR and no DR groups 114

Legend: Correlation of HbA1c with Cholesterol in the early and no DR groups is 0.5 and -0.4, respectively, and these correlations are different across the groups (p=0.02, Bartlett test). mfERG CD is mfERG Central Density.. 114

Figure 6.3.3.1 Pairwise plots of HbA1c against Cholesterol, Systolic BP and diastolic BP for the background and proliferative groups at t0. 125

Legend: t0 is the time of each patient's last visit. Refer is the group with referable STDR, non-refer is the group with non-referable STDR. 125

Figure 6.3.3.2 Pairwise plots of HbA1c against Cholesterol, Systolic BP and diastolic BP for the background and proliferative groups at t1 126

Legend: t1 is 1 year prior to t0. Refer is the group with referable STDR, non-refer is the group with non-referable STDR. 126

Figure 6.3.3.3 Pairwise plots of HbA1c against Cholesterol, Systolic BP and diastolic BP for the background and proliferative groups at t2 126

Legend: t2 is 2 years prior to t0. Refer is the group with referable STDR, non-refer is the group with non-referable STDR. 126

Symbols and statistical notation

n_1 = sample size in group 1

n_2 = sample size in group 2

n_T = total sample size in group 1 and 2

X = random variable

X_1 = random variable in group 1

X_2= random variable in group 2

x= realisation of random variable (i.e. the measurement)

x_1 = realisation of random variable in group 1 (i.e. the measurement)

x_2= realisation of random variable in group 2 (i.e. the measurement)

$\bar{\bar{x}}$= sample mean of random variable X i.e. the total mean over groups 1 and 2

$\bar{x}_1$ = sample mean of random variable X in group 1

$\bar{x}_2$ = sample mean of random variable X in group 2

s_p^2 = overall sample variance, i.e. pooled sample variance over groups 1 and 2

s_1^2 = sample variance in group 1

s_2^2 = sample variance in group 2

$\boldsymbol{X}$ = random vector (several variables)

$\boldsymbol{X}_1$ = random vector in group 1

$\boldsymbol{X}_2$= random vector in group 2

$\boldsymbol{x}$ = realisation of random vector (i.e. the vector of measurements)

$\boldsymbol{x}_1$ = realisation of random vector in group 1 (i.e. the vector of measurements)

$\boldsymbol{x}_2$= realisation of random vector in group 2 (i.e. the vector of measurements)

$\overline{\overline{\boldsymbol{x}}}$ = vector of sample means of random vector $\boldsymbol{X}$ i.e. the total mean over groups 1 and 2

$\overline{\boldsymbol{x}}_1$ = vector of sample means of random vector $\boldsymbol{X}$ in group 1

$\overline{\boldsymbol{x}}_2$ = vector of sample means of random vector $\boldsymbol{X}$ in group 2

$\boldsymbol{S}_p^2$ = overall sample variance-covariance matrix, i.e. pooled over groups 1 and 2

$\boldsymbol{S}_1^2$ = sample variance-covariance matrix in group 1

$\boldsymbol{S}_2^2$ = sample variance-covariance matrix in group 2

T^2 = Hotelling T^2 statistic

Chapter 1. Introduction

1.1 The importance of biomarkers

In any effort to accurately classify observations or cases to the appropriate groups it is necessary to identify those variables which have the greatest discriminatory potential. The process of identifying these variables is termed variable selection. In the arena of clinical science the variables we are interested in are referred to as “biomarkers”. The World Health Organisation (WHO) has defined a biomarker as “any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease” (WHO, 2001).

The definition of biomarkers is quite broad and accordingly there are several biomarker sub-types. Surrogate biomarkers are biomarkers which are used as substitutes for clinically relevant end-points (Strimbo and Tavel, 2010). A surrogate biomarker may not be directly associated with the underlying condition being monitored but it will have been demonstrated to reliably and accurately predict a relevant clinical outcome. An example is measuring a patient’s blood pressure (BP) as an indirect assessment of left ventricular function instead of using echocardiography (Aronson, 2005).

A prognostic biomarker may be used to determine the likelihood of a patient experiencing a clinical event, disease recurrence or progression given that they are already afflicted with a particular medical condition or disease. An example is the breast cancer susceptibility gene (BRCA) 1. A high level of expression of BRCA1 in untreated breast cancer patients is associated with a worse prognosis (James et al., 2007).

A predictive biomarker may be used to determine whether one of a pair of individuals who are physiologically similar (except in respect of the presence or absence of the predictive biomarker) may experience a reaction to some medical, chemical or environmental agent. An example is the Excision Repair Cross-Complementation Group 1 (ERCC1) in non-small cell lung cancer (NSCLC). High expression of the ERCC1 gene predicts resistance to cisplatin-based chemotherapy in NSCLC patients (Olaussen et al., 2006).

An example of a disease in which biomarkers are important is diabetes mellitus, a chronic long-term disease. Patients with diabetes mellitus are at risk of developing diabetic retinopathy. There is significant literature on identifying biomarkers which may be used to predict the progression of diabetic retinopathy and associated visual loss. Well established examples that are used in clinical practice include haemoglobin A1c (HbA1c), BP and lipid profile. The link between a lower blood sugar concentration and reduced levels of HbA1c was first demonstrated in 1976 (Keonig et al., 1976) and has since become the standard for measuring blood glucose levels. HbA1c is an example of a metabolite which is also a biomarker for progression (Caveney and Cohen, 2011). Lactate is another metabolite with potential as a biomarker in diabetes mellitus as a strong association with lactate levels and type 2 diabetes has been established (Crawford et al., 2010). It is commonly used as a measure of the difference between energy expenditure and oxidative capacity of muscle tissue. For BP it has been demonstrated that hypertension and diabetes are risk factors for atherosclerosis and that both hypertension and diabetes may occur sequentially in individuals as they share several metabolic pathways (Cheung and Li, 2012).

A number of less well established biomarkers have been suggested as potential candidates including several cytokines which have been found to play a role in the pathology of diabetic retinopathy. Tumour necrosis factor-α (TNF-α) is a cytokine which is active in the acute phase of the immune response. It has been demonstrated that TNF-α plays an important role in the development of diabetic retinopathy (Joussen et al., 2009). Vascular endothelial growth factor (VEGF) is active in stimulating angiogenesis. In healthy individuals this is an important aspect of normal somatic cell proliferation and development. However it has been shown that VEGF levels are upregulated in hypoxic conditions and that the upregulated expression of VEGF is active in mediating active intraocular neovascularisation in patients with ischaemic retinal diseases (Aiello et al, 1994). Similar conclusions about the role of VEGF and cytokines in general have been reached by other researchers. Monocyte-chemoattractant protein 1 (MCP-1) as well as VEGF have been identified as regulators of diabetic retinopathy which might be suitable as biomarkers for risk assessment in diabetic patients (Ozturn et al, 2009).

1.2 The search for biomarkers is a statistical problem of variable selection for discrimination

The main question is how to find the best marker or set of markers for disease stage diagnosis. This question can be seen as a statistical problem of finding the variable or set of variables that can best discriminate between disease groups. The question is how to select a subset of variables which will capture enough information about the groups of interest to be capable of discriminating between them and thus make appropriate assignments for new patients. Suppose that we have a set of potential explanatory variables $X_1 \ldots X_p$ measured on n patients which may be used to predict membership of the appropriate group. How do we select from these potential explanatory variables the subset that will best discriminate between each of our groups of interest? Furthermore how do we select from among those variables which can distinguish between the groups of interest where there may be multiple variables with the same discriminatory potential?

One possible solution to the variable selection problem is simply to evaluate every possible subset of variables. This is not a practical solution as it requires increasing computing power as the number of variables increases. Alternatively we might consider selecting variables based on some prior knowledge of the condition being studied. For example in studying patients with diabetes it is intuitively obvious that the insulin level of a patient is an important variable. However, diabetes is a very well researched condition on which a significant amount of research has been conducted. For conditions which are less well-known and less well-researched there may not be as much information at the disposal of researchers and clinical workers.

The difficulties of the variable selection problem are not limited to logistic considerations such as the number of variable subsets which must be analysed or the size of the dataset. The quality of the collected data is also a complex issue for variable selection. Problems can occur in how well data are balanced across groups as a large disparity in group sizes risks biasing any conclusions or inferences made using a given dataset. We must also consider the composition of the datasets. In addition to continuous variables we can expect to have to deal with ordinal and nominal variables possessing multiple levels which may require specialised techniques for their analysis. Missingness further complicates the task of variable selection as does the presence of noise in the data. While some issues may be addressed through effective study design there is a limited capacity to anticipate potential issues which means that researchers will always be faced with some of these problems. This makes the task of variable selection particularly complex.

1.3 Aim of this thesis

The aim of this thesis is to identify and improve the statistical methods of variable selection for discrimination that are suitable for complex data such as clinical ophthalmic data that is associated with studies of diabetic retinopathy.

1.4 The structure of this thesis

The structure of this thesis is as follows. In Chapter 2 I present a literature review outlining the major categories of current variable selection methods the filter, wrapper and embedded methods. In Chapter 3 I outline the theoretical background of the linear discriminant analysis (LDA) and quadratic discrimination analysis (QDA) classifiers and Hotelling's T^2 statistic. I also propose an extension to Hotelling's T^2 statistic to produce the novel SNR and describe the multiple forward selection algorithm I have created which uses either the SNR (MFS-SNR) or Hotelling's T^2 statistic (MFS-T2) to carry out variable selection. In Chapter 4 I present the results of a study using simulated datasets to compare the novel MFS-SNR algorithm with several existing variable selection methods including univariate and multivariate filter methods and an embedded method utilising random forests. In Chapter 5 I present the results of an assessment of the MFS-SNR algorithm's ability to carry out variable selection when data are not normally distributed. In Chapter 6 I present the results of variable selection from 4 ophthalmological datasets using the MFS-SNR and MFS-T2 algorithms. Finally in Chapter 7 I present a summary and discussion of the work I have completed over the course of my studies and outline possible future work.

Chapter 2. Literature review

2.1 Introduction

The problem of variable selection for classification is complex (see Section 1.2) and has been the subject of considerable research. Consequently there are numerous methods of variable selection for classification available (Bolòn-Canedo et al., 2011; Chandrashekar & Sahin, 2004; Lazar et al., 2012; Pacheco et al., 2006; Saeys et al., 2007). However, these are often univariate in nature, failing to take account of relationships between variables. Methods which are designed to take account of such relationships often do so by analysing large numbers of variable subsets which makes the application of these methods to high dimensional datasets impractical. Existing methods may not be robust to quality issues that are common with datasets. These quality issues can include large proportions of missingness in datasets, imbalances in group sizes and mixtures of variable types.

This chapter is a literature review. It is focused on two methodological areas:

- current statistical methods of variable selection for classification focusing on clinical datasets where data can be missing, where correlation structure can be different across disease groups and where variables can be measured in different units.
- measurement of information via signal-to-noise ratio.

As the goal of my work was to produce a novel multivariate filter method the focus of my literature review is on filter methods with a view to investigating if existing methods addressed the shortcomings of filter methods. The structure of this chapter is as follows. I first review the method of principal components analysis (PCA) (in Section 2.2) which is an alternative to variable selection that achieves dimensionality reduction. Then I provide an overview of the general classes of variable selection methods: filter methods (Section 2.3), wrapper methods (Section 2.4) and embedded methods (Section 2.5). I then discuss the current relevant methods of information evaluation via signal-to-noise ratio (Section 2.6). Finally, the chapter concludes with a discussion (Section 2.7) on existing gaps in the literature and a description of which gaps this thesis is addressing.

2.2 PCA as a variable selection method for classification

The purpose of variable selection is to identify those variables which are most effective at discriminating between the groups of interest. Variable selection tacitly assumes that some variables contain the same information that is useful for discriminating between groups, hence the number of variables can be reduced before classification. By identifying these variables the optimal discriminatory performance can be achieved while also reducing dimensionality. PCA is an alternative to variable selection which generates new variables using the starting set of variables and it ignores information about group membership.

PCA is a means of achieving dimensionality reduction by generating linear combinations of the initial variables which serve to account for the largest amount of information (measured via variance) possible in the new orthogonal variables produced (Cox, 2005). The application of PCA will be most constructive when variables are not independent, i.e. some correlation exists between variables. This can be assessed using tests such as Bartlett's chi-squared test or the Kaiser-Meyer-Olkin statistic (Lattin, 2003). Once correlation between variables has been established eigenvalues may be calculated using the correlation matrix (or variance-covariance matrix) for a given set of variables with each matrix having a number of eigenvalues (ev) less than or equal to the number of variables/parameters ($ev \leq p$). The eigenvalues λ_j ($j = 1..p$, where p is number of variables) can be calculated by solving

$$det(R - \lambda I) = 0 \qquad (2.2.1)$$

here R is the correlation matrix. Similarly the p eigenvectors, represented by v_j, can be found by solving

$$\left(R - \lambda_j I\right) v_j = 0 \qquad (2.2.2)$$

here R is the correlation matrix. Principal components (PCs) are then obtained as a linear combination following multiplication of the variables by the eigenvectors corresponding to the calculated eigenvalues.

$$y = A'x \qquad (2.2.3)$$

In this equation y is the vector of principal components derived from the original variables. The matrix A' is the matrix of eigenvectors and x Is a variable which has mean vector μ and variance-covariance matrix Σ. The principal components are orthogonal by design with each one accounting for decreasing amounts of variance in a given system so the first PC accounts for the largest proportion of variance with each subsequent component accounting for a smaller proportion of total variance. The decreasing variance associated with each PC is ensured by choosing the largest eigenvalue and associated eigenvector as it is true by construction that the variance is equal to the eigenvalue (Cox, 2005; Lattin et al., 2003). Therefore choosing the largest eigenvalue ensures that the first PC accounts for the largest proportion of variance. Of the remaining eigenvalues the largest will be selected for the second PC and so on until all of the PCs have been calculated.

Once the PCs have been calculated there are numerous ad hoc methods available on which to base the decision of how many PCs to use. The number of PCs can be dictated by the minimum amount of variance we wish to account for using the new PCs, for example we may wish to account for at least 70 % of total variance within the system and therefore we will select the lowest number of PCs which has a cumulative variance equal to or greater than 70 %. In addition to reducing dimensionality, analysis of PCs can also tell us a great deal about the underlying relationships and structures of the correlations between variables being studied. PCs can reveal relationships between different variables by studying the coefficients applied to calculate them. Principal component loadings are also useful in this regard as they represent the correlations between the original variables and the principal components produced. In terms of analysing PCA output commonly used graphical devices include scree plots which represent the relative contribution of PCs to the cumulative variance accounted for and biplots which plot PCs against each other and allow visualisation of variables which are shared by different PCs.

While PCA is not strictly speaking a method for selecting variables the logic behind the method of PCA is focused on identifying those variables which account for the largest proportions of variance in a particular dataset. The motivation for this approach is the idea that the variance is related to the amount of information. This aspect of PCA has been exploited for the purpose of variable selection. Paul et al. (2008) describe a method for carrying out variable selection which uses supervised principal components analysis to pre-condition the variable selection process. Those variables with the highest correlation with the outcome of interest (or highest discriminatory strength) are subjected to supervised principal components analysis. A least squares regression is then carried out using the principal components produced. This regression model is then used to produce anan estimate of the outcome of interest. Variable selection is then carried out using this estimate of the outcome of interest and the initial set of variables. By utilising the estimate of the outcome of interest the resulting data are less noisy and so the selection of variables using standard procedures such as LASSO or forward selection is more effective.

A disadvantage of PCA is that its effectiveness is proportional to the level of correlation between variables. For a dataset where there is little or no correlation between variables PCA will be of very limited benefit. Additionally because PCA focuses on correlations and covariances between variables while not taking into account group membership any variable selection method utilising it may fail to select variables which are important to discriminating between groups.

2.3 Review of filter methods of variable selection

Filter methods of variable selection operate by using a suitable data summary metric. The data summary metric should be easy to use and should require much less computational time than classification.

One way to utilise the filter metric is to rank each of the candidate variables i.e. as a univariate approach. The ranked list of variables produced can then be used to select those variables which have the highest ranks and are therefore assumed to be the most relevant to classification and discrimination. There are many different kinds of data summary metrics which may be used by filter methods to rank variables.

Another way to utilise the filter metrics is taking a multivariate approach to selecting a set of variables. Multivariate filter methods can take correlations and covariances between variables into account when calculating values for the data summary metric. Forward and backward selection protocols may also be used with filter methods. Once a set of variables has been chosen, this set is then be evaluated and estimates of performance calculated. In the following subsections I list some of the most frequently used filter variable selection methods.

2.3.1 Filter methods based on Kullback-Leibler divergence

The first filter method for variable selection among $X_1, \ldots, X_p$ that I will mention is a method based on the Kullback-Leibler (KL) divergence. There are several concepts for utilizing the KL divergence metric. In this section I will mention two filter methods based on KL divergence: the method of Mahat et al. (2007) and the method of Dasgupta (2015).

The method of Mahat et al. (2007) assesses differences between groups by calculating the smoothed Kullback-Leibler divergence and uses this as a metric to carry out variable selection. Mahat calls this concept the location model-based variable selection method. It utilises forward, backward and stepwise selection. Each variable set is assessed by calculating a test statistic using the sample-based divergence for the addition/removal of a variable. Where no change in the divergence occurs following the addition or removal of a given variable the difference in the divergence values calculated with or without that variable should approximately follow a chi-squared distribution with altered degrees of freedom to accommodate whether that variable is categorical or continuous. Analytically, this test statistic can be written as:

$$X_0^f = \frac{n_1 n_2}{n_1+n_2}\left(D_{J_j} - D_{J_{j-1}}\right) \sim \chi^2_{(v=v_j-v_{j-1})} \qquad (2.3.1.1)$$

where X_0^f is the rescaled difference in KL distances and its asymptotic distribution is the chi-squared distribution χ^2 with v degrees of freedom. The values $n_1$and n_2 are the group sizes, D_j and D_{j-1} are the estimated Kullback-Leibler divergences calculated for (forward or backward) steps j and $j-1$ while v_j and v_{j-1} are the degrees of freedom associated with the chi-squared distribution of D_j and D_{j-1}. When a continuous variable is being assessed using forward or backward selection in step j the test statistic X_0^f is compared with

$$\chi^2_{(v=2^q-1,1-\alpha)} \qquad (2.3.1.2)$$

When a binary variable is being assessed using forward selection in step j the test statistic X_0^f is compared with

$$\chi^2_{(v=2^q(2+p),1-\alpha)} \qquad (2.3.1.3)$$

Here p and q are the numbers of continuous and binary variables, respectively, at step $j-1$ and α is the type 1 error.

Using backward selection where a binary variable is being assessed the test statistic X_0^f is compared with

$$\chi^2_{(v=2^q+p2^{q-1},1-\alpha)}. \qquad (2.3.1.4)$$

When forward selection is used a variable is added to the selection if the test statistic is larger than the chi-squared critical value otherwise variable selection is terminated. When backward selection is used a variable is selected if the test statistic is smaller than the chi-squared critical value otherwise selection is terminated.

Where stepwise selection is used every time a new variable is added to the set of discriminatory variables, the significance of the other variables in that set is reassessed, to ensure that they all remain significant for discrimination. The criterion used for adding a variable is the same as when forward selection is used alone. The criterion for removing a variable becomes

$$X_0^s = \frac{n_1 n_2}{n_1+n_2}\left(D_{j_i}^f - D_{j_i}^b\right) \qquad (2.3.1.5)$$

Here $D_{j_i}^f$is the estimated distance at step j calculated using forward selection and $D_{j_i}^b$ is the estimated distance at step j obtained using backward selection. This test statistic is compared to a chi-squared distribution with the degrees of freedom described above for backward selection. Stepwise selection terminates when both the forward and backward selection sequences satisfy the stopping criteria.

Variable selections made using this method are validated by splitting the data into training and validation sets. The training set is used to train a classifier which is then used to classify cases from the validation set. The results are evaluated by counting the proportion of misclassified cases. Mahat et al. (2007) simulated data sets to evaluate their method. The data sets contained a mixture of 3, 4 or 8 binary variables and 4 continuous variables with group sizes 25 and 100. A mixture of both discriminating and non-discriminating variables was present in each of the data sets. Results were evaluated in terms of the proportion of discriminating variables selected by each method out of the total number of discriminating variables and the error rate calculated for a model using the selected variables. Overall backward selection performed better than either forward or stepwise selection in terms of the selection of discriminating variables and the error rates calculated. However, Mahat et al. (2007) note that given the lack of prior knowledge regarding which variables are discriminating in a given real data set the optimal method of variable selection must be determined on a case-by-case basis.

The main limitation of the KL filter method of variable selection of Mahat et al. (2007) is that it is designed to work only with continuous variables and binary variables. While it is possible to apply this method to datasets that contain mixtures of continuous and categorical variables it is necessary to transform ordinal and nominal variables into binary variables to facilitate their use with this method. A problem with this approach is that any transformation can result in a loss of information. This can lead to misrepresentation of the importance of a given variable following transformation.

A second filter variable selection method based on KL divergence was proposed by Dasgupta (2015). Their work presents a method which utilises the KL divergence in place of the squared error loss when calculating regression coefficients for each of the variables in a dataset. Using the formula

$$\bar{\beta}^{KL} = \arg min \left\{ \sum_{i=1}^{n} \left[log \frac{f(y_i|x_i,\beta^*)}{f(y_i|x_i,\beta)} \right] + \lambda_n \sum_j \widehat{w}_j |\beta_j| \right\} \quad (2.3.1.6)$$

the “Adaptive Penalized KL Divergence” estimator ($\bar{\beta}^{KL}$) is calculated. In this formula $\widehat{w}_j$ is a weight vector, $f(y_i|x_i,\beta^*)$ and $f(y_i|x_i,\beta)$ are the normal probability densities associated with the true regression coefficients β^* and β. λ_n Is one of two tuning parameters used with this method.

The method of Dasgupta (2015) works by first finding the ordinary least squares estimate of β^*. This is then used to calculate $\widehat{w}$ for some value of $\gamma > 0$, γ is the second tuning parameter used with this method. Coefficient estimates are calculated by solving

$$\hat{\beta}^* = \arg min \left\| \hat{y} - \sum_{j-1}^{p} x_j^* \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \qquad (2.3.1.7)$$

For all possible values of λ_n. The "Adaptive Penalized KL Divergence" is then calculated by dividing these estimates by the weight vector $\widehat{w}_j$. A list of regression coefficients associated with each of the predictor variables is produced in this way and can be used to select which variables are the most important to predicting the outcome of interest.

Dasgupta (2015) successfully demonstrated his method on a dataset of measurements taken from n = 442 patients with diabetes. The variables age, sex, body mass index (bmi), average blood pressure and six blood serum measurements (total count (tc), low-density lipoprotein (ldl), high density lipoprotein (hdl), total cholesterol (tch), low tension glaucoma (ltg), glutamic acid (glu)) were considered for each of the patients. Age and the blood serum measurements hdl and glu were determined to have no significant influence on the progression of the disease one year after baseline.

The use of regression coefficients in carrying out variable selections is a logical approach as the regression coefficients are designed to reflect the influence of a particular variable on the outcome of interest. However, Dasgupta's method (2015) involves the use of two tuning parameters λ_n and γ, which is a disadvantage. The quality of the final subset of variables will be directly affected by the values of these tuning parameters. The researcher must specify the optimal value of γ. As part of the procedure the estimated $\hat{\beta}^*$ is calculated for all possible values of λ_n. For sufficiently high dimensional datasets this may not be a practical approach as it can be time consuming. It should also be noted that while Dasgupta (2015) provides some assessment of his method using a diabetes dataset there is no effort to validate the selections made from the dataset. There are also no other datasets used for comparison in his paper and there is no attempt to compare this method to any alternative method of variable selection.

2.3.2 Filter methods based on Biomarker Identifier Measure

Another example of filter methods is based on the Biomarker Identifier (BMI) developed by Lee et al. (2011). BMI assesses differences in distributions of the variables across two outcome groups. The BMI measure involves calculation of several terms: the ratio of overall control group variance to the overall variance, the ratio of the two groups' means and the product of the true positive rates obtained using a logistic regression to assign subjects to each of the two groups. The formula for the BMI is

$$BMI(X) = \lambda\ TP^2 \sqrt{|\Delta_{diff}| \frac{CV_{ctr}}{CV}} \qquad (2.3.2.1)$$

Here λ is a scaling factor, TP^2 is the product of the true positive rates across each group calculated using logistic regression with the format 'outcome ~ variable'. The terms CV and CV_{ctr} denote the coefficients of total variance in both groups and in the control group, respectively. The factor Δ_{diff} is based on Δ which is calculated according to

$$\Delta = \frac{\bar{x}}{\bar{x}_{ctr}} \qquad (2.3.2.2)$$

$\bar{x}$ and $\bar{x}_{\mathrm{ctr}}$ are the overall mean value of the variable X in both groups and in the control group, respectively. If the value of Δ is greater than or equal to 1 then

$$\Delta_{diff} = \Delta \qquad (2.3.2.3)$$

otherwise

$$\Delta_{diff} = -\frac{1}{\Delta} \qquad (2.3.2.4)$$

The quantity Δ_{diff} is designed to take account of the differences in the distribution of a given variable across the disease and control groups as measured by the differences in the mean values. The logic is that if a variable is capable of discriminating between two groups there will be a difference in the means across the two groups and hence Δ_{diff} will have a larger deviation from a value of 1. The ratio of control group variation to overall variation is designed to act as a measure of how well defined or how noisy the control group is. A smaller ratio means that the overall variance is different from the variance of the control group indicating that the disease group is noisier relative to the control group. Conversely a larger ratio means that the difference between the overall variation and the variation of the control group is small indicating that the disease group is less noisy relative to the control group. The true positive rate for a given variable is calculated using logistic regression. The incorporation of the true positive rate in the BMI formula takes the discriminative performance of the variable into account when calculating the BMI.

BMI was compared to 6 alternative variable selection methods by Lee et al. (2011): Information gain, Relief-F, two versions of the t-test (moderated t-test and window t-test) and a chi-squared test.

These methods and the BMI were evaluated using two real datasets. The first comprised 129 sets of microarray data from 60 smokers with lung cancer and 69 smokers without lung cancer. Seven genes had previously been identified from this data set as being able to differentiate between cancerous and non-cancerous samples evaluated using quantitative polymerase chain reaction (PCR). The second dataset contained breast cancer data where subjects were grouped based on receptor status (i.e. whether estrogen receptors were present or not). This data set contained 130 samples used as training data and 100 samples used as validation data.

Information gain assesses whether or not a given variable adds information regarding the group variable. This is done by looking at the differences in the marginal and conditional distributions of the grouping variable. The marginal distribution gives the probability of the group variable having certain values independent of the predictor variable. Conversely the conditional distribution gives the probability of the group variable having certain values when it is dependent on the predictor variable. RELIEF-F assesses the discriminatory ability of a variable by selecting k nearest neighbours for a random sample evaluating how well the variable differentiates between neighbours in the same and different groups (Kononenko, 1994).The moderated t-test is an adaptation of the student's t-test using a hierarchical Bayesian approach and knowledge of posterior residual standard deviations while the window t-test uses multiple genes with similar expression levels to calculate the variance to be incorporated into the t-test. The Chi-squared test is used to select variables by testing if the distribution of a given variable is different across the groups.

In Lee et al. (2011) the performance of each of the variable selection methods on the lung cancer dataset was evaluated using classifier performance as well as the ranking of features known to discriminate between groups. The breast cancer dataset was evaluated using classifier performance only as the variables relevant to discriminating between the groups were not known. For both datasets 10-fold cross-validation was used to validate the classifier developed with the training dataset, and the resulting classification data was used to calculate area under the curve (AUC) values. For the lung cancer dataset performance was also evaluated by observing how the 7 genes known to differentiate between cancerous and non-cancerous samples were ranked. Multiple different classifiers were used with both datasets.

In terms of classifier performance it was found that the performance of BMI in terms of the estimated AUC values was comparable to that of other methods. The set of variables selected using the BMI was stable regardless of which classifier was used for both datasets. For the lung cancer dataset the performance of BMI and the alternative methods was also evaluated in relation to the ranking of genes known to successfully differentiate between cancerous and non-cancerous samples. In this respect BMI exhibited superior performance to the alternative methods ranking these genes within the top 4,000 ranked genes.

One limitation is that the BMI is univariate in nature. Each of the variables is assessed independently and relationships between variables are not considered. The authors also do not explain how the scaling factor λ is calculated for different datasets. It is also reasonable to question how appropriate it is to combine multiple statistical measures into a single composite measure. The conclusions that may be made based on individual statistical measures may not extend to the use of a composite measure. It should be noted that while this is not addressed directly in the article, the results of the comparison study involving other methods reflect favourably on the BMI index.

2.3.3 Filter methods for variable selection based on correlation between predictors and outcome group variable

Here I present a filter method of variable selection based on correlation between predictor variables and the outcome group variables proposed by Li et al. (2012). They describe a filter method that is a sure independence screening procedure based on a distance correlation metric (DC-SIS). The sure independence screening property states that as the sample size approaches infinity all variables relevant to predicting the outcome of interest can be selected. Distance correlation measures the dependence between two random variables. Two properties of the distance correlation that make it suited to DC-SIS are that it is equal to zero for random vectors when they are independent and that it is a strictly increasing function of the absolute value of the Pearson correlation of the two normal random variables. The DC-SIS method considers correlations between predictor variables and the outcome variable. Using the DC-SIS metric (Li et al., 2012) it is possible to split variables into active and inactive predictors. The sample distance correlation is calculated according to

$$\widehat{dcorr}(X,Y) = \frac{\widehat{dcov}(X,Y)}{\sqrt{dcov(X,X)dcov(Y,Y)}} \qquad (2.3.3.1)$$

where X and Y are two random vectors from a joint probability distribution. The term $dcov$ is the sample distance covariance defined as

$$\widehat{dcov}^2(X,Y) = \hat{S}_1 + \hat{S}_2 + \hat{S}_3, \qquad (2.3.3.2)$$

where $\hat{S}_1$, $\hat{S}_2$ and $\hat{S}_3$ are calculated according to

$$\widehat{S_1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|x_i - x_j\right\|_{d_X}\left\|y_i - y_j\right\|_{d_y} \qquad (2.3.3.3)$$

$$\widehat{S_2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|x_i - x_j\right\|_{d_X}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|y_i - y_j\right\|_{d_y} \qquad (2.3.3.4)$$

$$\widehat{S_3} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|x_i - x_l\|_{d_X}\|y_j - y_l\|_{d_Y} \qquad (2.3.3.5)$$

where (x_i, y_i) is a random sample taken from the joint probability distribution of (X, Y). Once the sample distance correlation is calculated the metric ω_k is calculated

$$\widehat{\omega}_k = \widehat{dcorr}^2(X_k, Y) \qquad (2.3.3.6)$$

ω_k is then used to rank the importance of each variable X_k. The random vector X_k represents the k^{th} potential predictor variable and Y is a vector representing the response variable. Once the variables X_k have been ranked in this way the final set of active predictors is chosen using the following rule;

$$\widehat{D}^* = \{k: \widehat{\omega}_k \geq cn^{-\kappa}, for\ 1 \leq k \leq p\} \qquad c > 0,\ \ 0 \leq \kappa < \frac{1}{2} \qquad (2.3.3.7)$$

Here $\widehat{D}^*$ is a set which will contain the index values for the variables which are determined to be active in predicting the outcome of interest. c and κ in this equation are parameters with values set in accordance with the rules $c > 0$ and $0 \leq \kappa < \frac{1}{2}$.

The performance of the DC-SIS based filter method of variable selection was assessed using both real and simulated data (Li et al., 2012). The simulated data were normally distributed with varying correlations and dimensionality. The real data was a cardiomyopathy microarray dataset containing expression data for 6,319 genes. Three criteria for evaluating performance were used: S, the minimum size of the models including all active predictors; P$_S$, the probability that an individual active predictor is selected for a particular model over 500 replications and P$_a$, the proportion that all active predictors are selected for a particular model over 500 replications. The DC-SIS method was compared to the sure independence screening (SIS) methods of Fan & Lv (2008) and the sure independent ranking and screening method of Zhu et al (2011).

For each of the models used with the simulated data the DC-SIS method out-performed the SIS and sure independent screening and ranking (SIRS) methods in terms of the minimum size of the models S; DC-SIS models contained fewer variables out of 500 simulations. Similarly the DC-SIS method had P$_a$ values larger than those of the competing SIS and SIRS models in the majority of the simulations (Li et al., 2012). The results for P$_s$ exhibit similar patterns with DC-SIS values being larger than those for SIS and SIRS in a majority of scenarios. Data were also simulated to investigate the performance of the DC-SIS method when selecting grouped predictors. The results of this indicate that the minimum model size to ensure the inclusion of all active predictors is small. The last example using simulated data assessed the ability of the DC-SIS method to handle multivariate responses. The results indicate that the DC-SIS method retains the sure screening property even when dealing with multivariate responses.

Lastly the DC-SIS method was used to select variables from a microarray dataset with the goal of identifying the most influential genes for expression of the G protein-coupled receptor in mice. Previous work with this dataset identified Msa.2877.0 and Msa.1166.0 (Hall & Miller, 2009) however the DC-SIS method identified Msa.2134.0 and Msa.2877.0. When the two selections were compared in terms of the performance of a model using these predictors it was found that the DC-SIS selections achieved better performance with a larger adjusted R^2 value of 96.8 % and a larger explained deviance value of 98.3 % compared to 84.5 % and 86.6 % respectively for the previously determined selections.

While the method of Li et al. (2012) does take correlations into account it is only concerned with correlations that exist between predictor variables and the outcome variable. It does not consider correlations that may exist between predictor variables. Therefore the DC-SIS method is a univariate method in the sense that the correlations are only considered between each predictor variable and the outcome variable. Any interactions between predictor variables that may exist are not considered when carrying out variable selection.

Bouhamed et al. (2012) present a method for selecting categorical predictor variables using the correlations between the predictor variables and the response variable. Their method analyses variables in a univariate and a multivariate context over a 5-stage procedure. In the first stage redundant variables are eliminated from the dataset. In the second stage the chi-square statistics and the associated p-values are calculated for each of the predictor variables and the response variable. Those variables with a p-value above the designated significance level are considered redundant and eliminated from the process at this stage. In the third stage bootstrapped k-means clustering based on proximity to a principal component generated using the `kmeansvar` function of the `ClustOfVar` package is used to produce clusters of predictor variables. In the fourth stage the clusters are then analysed using the Truncated Product Method (TPM). TPM scores are subjected to the logarithmic transformation. The logarithmic transformation makes it easier to identify which clusters of variables are most important as it is easier to identify high scores (which are associated with a high degree of significance). In the fifth and final stage clusters of variables are selected based on a composite scoring of each cluster using the results of analyses from each of the previous steps.

Bouhamed et al. (2012) applied their method to 2 datasets. The first is a cardiac database. It comprises 23 variables measured on 267 patients. The second is an automotive database comprising

18 variables. Bouhamed et al. use their method to identify the optimal subset of variables from each of these datasets. Several other methods are also used to identify the optimal subset from each of these datasets and the results are compared to those obtained using the method of Bouhamed et al. The competing methods include wrapper methods employing forward and backward selection. The conclusion presented in their manuscript is that the method of Bouhamed et al. (2012) is more effective than the competing methods.

The most obvious shortcoming of the work of Bouhamed et al. (2012) is that their method is only effective when selecting categorical variables. It cannot be applied to continuous data or any mixture of categorical and continuous predictor variables. A further disadvantage with the work relates to the comparison of their method with alternative methods. Results are presented demonstrating the functioning of this novel method and defending the variable selections it makes from the automotive and cardiac datasets. However, there is no data on the performance of the final selections made using each method (including the method of Bouhamed et al., 2012). It appears that no work was done to validate the selections of any of the methods.

2.3.4 Filter variable selection methods based on correlations between predictor variables

Another category of filter variable selection methods is based on correlations between predictor variables. Andrews & McNicholas (2014) describe a variable selection method (VSCC – Variable Selection for Clustering and Classification) which utilises the within-group variances as well as the correlations between variables. The within-group variances are calculated according to

$$W_j = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n} y_{ig}\left(x_{ij}-\mu_{gj}\right)^2}{n} \qquad (2.3.4.1)$$

Here x_{ij} is the observation for subject i on variable x_j, μ_{gj} is the mean of variable x_j in group g, n is the number of subjects and G is the total number of groups. The term y_{ig} is set to 1 if a given subject belongs to group g and 0 otherwise.

Data are standardised so that all variables have mean and variance equal to 1. Within-group variances of the variables are then calculated and listed in ascending order so that the first variable has the lowest within-group variance and is thus automatically selected. Each of the unselected variables remaining is then considered for addition to the selected variable subset in order of their within-group variances and whether their correlation with each of the selected variables is less than some threshold. The stopping criterion for the algorithm is that $k \geq p$. k is initially set equal to 1 and is incremented by 1 each time a variable is selected. The term p is the total number of considered predictor variables. The stopping criterion is designed to ensure that the correlation between the selected variables and each of the candidate variables is assessed (i.e. all variables are analysed but only those with a sufficiently low correlation when paired with each of the previously selected variables are added to the set of selected variables). This is in contrast to other methods where the stopping criterion is designed to terminate variable selection when a sufficient number of variables have been picked (as determined by the stopping criterion).

The simplest relationship illustrating the acceptable threshold of correlation between variables is

$$|\rho_{kr}| < 1 - W_k \tag{2.3.4.2}$$

where k is the index of the variable under consideration for selection, r is the index of each of the variables which has already been selected. Thus the correlation between variable k and each of the already selected variables must be less than the difference between 1 and the within-group variance of variable k or this variable will not be added to the selection. However this is just one possible threshold and not necessarily the optimal threshold. Higher order relationships exist and those considered are quadratic, cubic, quartic and quantic in order of the increasing exponent of the within-group variance.

The algorithm is run using each of these thresholds which results in a maximum of 5 different variable subsets. The authors (Andrews & McNicholas, 2014) propose identifying the subset that is the optimal for classification by calculating the fuzzy classification matrix for each of the possible variable selections. This is an $n\,x\,G$ matrix containing the $\hat{y}_{ig}$ values which are estimates of the quantity y_{ig}. The quantity y_{ig} is a measure of the strength of the evidence that observation i belongs to group g. Where clusters are well-defined each row of the fuzzy classification matrix will contain one entry approximately equal to 1 and all of the other entries approximately equal to zero. The fuzzy classification matrix is used to calculate the uncertainty associated with each variable selection according to the formula;

$$n - \sum_{i=1}^{n} max_g\{\hat{y}_{ig}\} \tag{2.3.4.3}$$

Unfortunately it is not possible to use this novel uncertainty calculation. This is because the VSCC algorithm cannot be computed as described when there is only 1 group (i.e. $G = 1$), it implicitly assumes that $G > 1$.

Therefore the optimal variable subset for classification is identified in one of two ways. The known y_{ig} values may be used to calculate W$_j$ initialising the algorithm or the $\hat{y}_{ig}$ values are calculated using model-based classification and these are used with the known y_{ig} values to initialise the algorithm. The optimal subset is then determined by selecting the associated model using the Bayesian Information Criterion (BIC).

Andrews & McNicholas (2014) compared the model-based classification of their selection algorithm to model-based classification using the MCLUST family of models on a simulated data set containing 15 groups. Classification performance was assessed using the adjusted Rand index (ARI) calculated as the number of pairwise agreements between the estimated classifications and the known groups. The ARI is adjusted to allow for the occurrence of pairwise agreements between random groups. The VSCC algorithm achieved mean ARI of 0.85 with a standard deviation of 0.02 compared to 0.81 and 0.05 respectively for the MCLUST family of models. The VSCC algorithm also selected the relevant variables for discriminating between the groups in over 90 % of the data sets analysed.

The method of Andrews & McNicholas (2014) has the advantages of being multivariate and considering the correlations that may exist between variables. However in order to effectively use this method it is necessary to determine the optimal threshold for the correlation between one of the variables which has not yet been selected and each of the variables which has already been selected. To address this the variable selection procedure needs to be carried out multiple times to determine what the optimal variable subset is. This greatly increases the work and time required to use this method.

In the paper entitled "Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays", Lu et al. (2005) outline a method for identifying groups of differentially expressed genes (DEGs) from microarray data. The method of Lu et al. works by using the Hotelling's T^2 statistic associated with a variable as a measure of the discriminatory potential of that variable. During each round of variable selection the variable with the largest Hotelling's T^2 statistic is identified and selected. The p-value associated with a particular variable (or set of variables) is calculated. Provided the p-value is below the specified significance level (and lower than the p-value calculated during the previous round of selection) another round of variable selection will take place. Variable selection continues until the p-value associated with a particular round of variable selection is not lower than the p-value for the previous round, or until all the number of selected variables is larger than $n_1 + n_2 - 2$ (n_1 and n_2 being the sizes of groups 1 and 2).

The main limitation of the method of Lu et al. is in the use of Hotelling's T^2 statistic as the metric of discriminatory potential. Hotelling's T^2 statistic assumes that the variance-covariance matrices are homogeneous across groups. This is often an invalid assumption when dealing with real data and is therefore a limiting factor in its application.

2.3.5 Filter variable selection methods based on conditional mutual information

There are filter methods for variable selection which utilise mutual information (MI) for variable selection. Here I describe three methods and give their advantages and disadvantages.

Todorov and Setchi (2014), describe the use of the conditional mutual information as an index for the selection of variables. The standard measure of information exchange is entropy, defined as;

$$H(X) = -\sum_u p(X = u)\log\big(p(X = u)\big) \qquad (2.3.5.1)$$

where X is a random variable, u is a value of X and $p(X = u)$ is the probability that X will take the value u. The entropy calculated in this way is always greater than zero and is bounded by the logarithm of the set of values which X may take. The larger the value of $H(X)$ the more information we obtain by observing the values of X. If on the other hand the value is zero, when there is only one value u with $p(X = u) = 1$, no information is obtained by observing X.

This definition can be extended to produce the conditional entropy which is a measure of the amount of information that a given variable X contributes to the outcome of interest Y. The conditional entropy is calculated according to

$$H(Y|X) = -\sum_{u,v} p(Y = u, X = v)\log\big(p(Y = u|X = v)\big) \qquad (2.3.5.2)$$

where u and v are the values that Y and X can take respectively. The term $p(Y = u, X = v)$ is the probability of Y equal to the value u given that X is equal to the value v.

Conditional entropy can be further extended to produce the mutual information which is a measure of the distance between the probability distributions of X and Y as the following

$$I(Y; X) = H(Y) - H(Y|X) \qquad (2.3.5.3)$$

The mutual information calculated in this way is always greater than zero with an upper bound that is the minimum value of $\big(H(Y), H(X)\big)$. The correct interpretation of the mutual information is that it represents the reduction in uncertainty of Y when we have knowledge of X.

The naïve approach to variable selection using mutual information involves calculating the mutual information $I(Y; X_{1\ldots k})$ for all subsets of variables from a p-dimensional multi-variate random vector X, where $k < p$. However, as the number of variables increases the computational requirements become prohibitive. The classical approach to variable selection using mutual information addresses this problem by calculating $I(Y; X)$ for all X_i and Y pairs of k variables then selecting those variables with the highest mutual information with Y. However, this method is univariate and hence does not consider redundancy between variables.

One way to include more than one X_i variable is via the conditional mutual information (CMI). CMI is the change in the entropy of a variable Y conditional on variable X_i compared to the entropy of the variable Yconditional on variables X_i and X_j. The conditional mutual information is calculated as;

$$I\left(Y; X_i|X_j\right) = H\left(Y|X_j\right) - H\left(Y|X_i, X_j\right) \qquad (2.3.5.4)$$

where $H\left(Y|X_j\right)$ is the entropy of the variable Y conditional on X_j and $H\left(Y|X_i, X_j\right)$ is the entropy of the variable Y conditional on X_i and X_j. A further reduction in the computational requirements is achieved by utilising the mutual information instead of the conditional mutual information. Under the assumption that the ratio of information between variables X_i and X_j is not altered where conditioning is based on the class variable Y. On that basis the CMI may be approximated as;

$$I\left(Y;\ X_i|X_j\right) = I(Y;\ X_i) - \frac{\left(I(Y; X_i)*I\left(X_i; X_j\right)\right)}{H\left(X_j\right)} \qquad (2.3.5.5)$$

Another filter method that utilises the mutual information is proposed by Yu and Liu (2003). They developed a method for variable selection which exploits knowledge of correlation between variables and thus their method is multivariate. The principle of this method is that a variable which is useful for discriminating between two or more groups will have relatively high correlation with the groups of interest but will have low correlation to other variables. By selecting variables with low correlation to other variables redundancy in the final variable selection is minimised. The linear correlation r between two variables (X, Y) is calculated using the Pearson correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x_i})(y_i - \bar{y_i})}{\sqrt{\sum_i (x_i - \bar{x_i})^2}\sqrt{\sum_i (y_i - \bar{y_i})^2}} \qquad (2.3.5.6)$$

Here $\bar{x}_i$ and $\bar{y}_i$ are the mean values of the variables X and Y respectively. It is possible to use this value to determine the level of correlation and redundancy between variables. However the Pearson correlation assumes associations are linear which may not be the case. The Pearson correlation also assumes equal correlations across the groups which may not be the case. For this reason Yu and Liu (2003) do not use the Pearson correlation coefficient and rather they calculate the information gain using the entropy

$$IG(X|Y) = H(X) - H(X|Y) \qquad (2.3.5.7)$$

According to this measure, a feature Y is regarded as being more correlated to variable X than to variable Z, if $IG(X|Y) > IG(Z|Y)$. There is still a limitation to using the information gain in that its calculation is biased toward variables which have more values. To address this problem and also normalize the calculated values Yu and Liu use the symmetrical uncertainty (SU) calculated according to the formula

$$SU(X, Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \quad (2.3.5.8)$$

A value of 1 for the SU indicates that knowledge of the value of one variable will allow us to predict the value of the other while a value of 0 indicates that the two variables are independent.

To determine which variables are most relevant to classification and least redundant to other variables two parameters are calculated: the predominant correlation and the predominant feature. The predominant correlation is the correlation between some variable $X_i$and the group g which is larger than the user-specified threshold δ and has the largest SU with the group g. The predominant feature is a feature X_i that has the predominant correlation to the group g of interest. These calculations are implemented in the fast correlation based filter (FCBF) algorithm. The FCBF first calculates the SU values for all variables and uses these to identify the predominant variables as any variable with a symmetrical uncertainty greater than δ. This selection is then refined to identify predominant variables and remove redundant variables.

The FCBF algorithm was compared to Relief-F, CorrSF and ConsSF by Yu and Liu (2003). CorrSF and ConsSF utilize correlation and consistency measures to select variables. The Relief-F algorithm selects nearest neighbours to probe the classification accuracy of variables. 10 datasets were subjected to each of the variable selection methods and the results recorded. Performance was analysed in terms of the number of variables selected and the running time for each method. In this regard the FCBF algorithm was found to run faster than the competing methods as well as selecting the smallest number of variables. When the data were analysed for accuracy the FCBF selections were shown to be at least equivalent to those made by competing methods.

By analysing the correlations between variables and groups of interest the method of Yu & Liu is designed to eliminate redundancy between selected variables while favouring those variables which are highly correlated with the group(s) of interest. However the process of removing variables whose correlations are above a certain threshold implicitly assumes that the inclusion of such variables would not have any positive influence on the discriminatory performance of the final subset of variables. The method may eliminate variables which may have limited discriminatory potential by themselves but may be capable of enhancing the discriminatory performance of other variables.

Zhang and Zhang (2011) propose a 2-stage method for selecting variables based on the concept of cumulate conditional mutual information minimization. In the first stage of their method redundant variables are removed from the full set of variables. This is achieved using a dynamic sequential forward strategy to obtain a set of variables for which the mutual information with the class C is equal to the mutual information for the full set of variables. This is represented as

$$\mathrm{I}(\mathrm{S},\mathrm{C}) = \mathrm{I}(\mathrm{F},\mathrm{C}) \tag{2.3.5.9}$$

where F is the starting set of variables and S is a subset of variables obtained following the removal of redundant variables in stage 1. Zhang & Zhang demonstrate that the first step of their method is theoretically guaranteed to produce a subset of variables containing all of the relevant variables. However as dimensionality increases so does the likelihood that S will still contain redundant variables. At higher dimensionalities the estimation of the conditional mutual information also becomes less reliable. In order to address these issues in the second step the conditional mutual information for each variable found in S is compared to a predefined threshold ε.

This method is compared to Relief-F, FCBF and correlation based feature selection (CFS). Nine benchmark datasets were used in this comparison ranging in sample size from 2,000 to 72,626 and in number of variables from 22 to 5,000. The methods were all compared in terms of the number of selected variables, the time taken to carry out variable selection and the accuracy of a naïve Bayesian classifier, k-nearest neighbour classifier and c4.5 classifier. The method of Zhang and Zhang exhibited similar performance to each of the methods used in the study.

The quality of the variable subset returned by the algorithm of Zhang & Zhang (2011) is heavily dependent on the value assigned to the threshold ε. Zhang & Zhang (2011) do not explain how an appropriate value of ε can be identified. Presumably researchers are required to find the optimal threshold value by trial and error. Alternatively it may be possible to determine the optimal value analytically using prior knowledge of the dataset under study but this may not always be possible. Either way the performance of the method is heavily dependent on the threshold value but there is no means of identifying the optimal threshold presented in the article. A question remains about the reproducibility of the performance reported by Zhang and Zhang (2011).

The RELIEF algorithm is another example of a univariate filter method. It was originally proposed by Kira and Rendell (1992). The RELIEF algorithm operates by calculating the quantity $W[X]$ for each variable. $W[X]$ is first set to 0.0.

$$W[X] := W[X] - \frac{diff(X,R,H)}{m} + \frac{diff(X,R,M)}{m} \qquad (2.3.5.10)$$

R is an observation of the variable X. H is an observation of variable X from the same group (called a nearest hit). M is an observation of variable X from a different group (called a nearest miss). The term m specifies the number of observations of variable X which are used to estimate the weight $W[X]$. Dividing the difference terms by m normalises the weights so they all fall within the interval [-1, 1].

The quantity $W[X]$ calculated by RELIEF is itself an approximation of

$$W[A] = P(different\ value\ of\ A | nearest\ instance\ from\ different\ class) - \\ P(different\ value\ of\ A | nearest\ instance\ from\ same\ class) \qquad (2.3.5.11)$$

which estimates the information gain associated with the variable X. $W[X]$ is used to assign weights to variables which may then be used to carry out variable selection.

The original RELIEF algorithm is limited to scenarios with two groups. It is also not robust to significant levels of noise in data or missingness. Relief-F is an extension of RELIEF developed by Kononenko (1994) that is robust to noisy data and missingness and which can be applied to scenarios with more than 2 groups.

Relief-F differentiates itself from the original RELIEF algorithm by selecting one nearest miss from each of the different classes (RELIEF selects a single nearest miss). The contribution of all of the nearest misses to the calculation of $W[X]$ is averaged and then weighted using the prior probability of each group. When using RELIEF-F $W[X]$ is calculated according to

$$W[X] := W[X] - \frac{diff(A,R,H)}{m} + \sum_{G \neq group(R)} \left[P(C)\ x\ \frac{diff(A,R,M(G))}{m} \right] \qquad (2.3.5.12)$$

G identifies the group which observation R is being compared to. $P(C)$ is the prior probability of each class.

Kononenko compared the RELIEF-F extension to RELIEF-E. RELIEF-E is a simpler extension of RELIEF which selects a nearest miss from one of the different groups available. For this comparison Kononenko simulated 4 datasets. All of the variables in these simulated datasets were binary. The simulated datasets A and B had 3 and 4 groups respectively. Each of the groups in the first two datasets had equal prior probabilities. Each of the datasets A and B were composed of 3 random variables and 3 informative variables for each pair of groups. Thus dataset A (3 groups) had 12 binary attributes while dataset B (4 groups) had 21 attributes. A third dataset (C) was created by replacing each informative attribute in dataset A with 3 binary attributes.

The performance of RELIEF-F and RELIEF-E were assessed by calculating the linear correlation coefficient between the intended information gain of the variables and the factual information gain. The intended information gain was calculated using the probabilities that the simulated datasets were based on. The factual information gain was the value $W[X]$ (which is an estimate of the information gain). A high correlation between the intended and factual information indicated that the RELIEF method being used was effective at estimating the information gain associated with a particular variable. A low correlation indicated that the RELIEF method being used was less effective at estimating information gain. Datasets were simulated with up to 20 % noise. Two scenarios were investigated using either dependent or independent variables. In all scenarios the RELIEF-F method achieved high correlations with that of the RELIEF-E method.

In his article Kononenko (1994) acknowledges that there can be differences between the factual information gain and the intended information gain. In his work this is attributed to the random number generation aspect of data simulation. This is a potential limitation of the RELIEF-F metric as it implicitly assumes the factual information gain and the intended information gain will be highly correlated. The RELIEF-F metric is available for use as part of the `FSelector` package and is implemented as part of the comparison study presented in Chapter 4.

In summary, the information gain based filter methods are intuitive methods due to their interpretation in terms of information. Some are univariate (Todorov & Setchi, 2014, Kononenko, 1994) and some are multivariate (Yu & Liu, 2003; Zhang & Zhang, 2011).

2.3.6 Filter variable selection methods based on t-scores

Filter methods can also be based on measures that are known as t-score metrics. Ahdesmaki and Strimmer (2010) describe a method of variable selection which utilises correlation-adjusted t-scores as summary scores representing the discriminating ability of a particular variable. Linear discriminant analysis (LDA) assumes a mixture model for the p-dimensional data with multivariate probability density function,

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g) \qquad (2.3.6.1)$$

Where each of G groups is represented by a multivariate normal density, $f(x|g)$. Then LDA is used to estimate a feature weight vector ω_g which is then used to calculate the correlation-adjusted t-scores.

For this mixture model, given the a priori mixing weights π_g, the posterior probability of membership of group g given some value x is calculated according to;

$$Pr(g|x) = \frac{\pi_g f(x|g)}{f(x)} \qquad (2.3.6.2)$$

Ahdesmaki and Strimmer define the LDA discriminant score $d_g(x)$ as;

$$d_g(x) = log\{Pr(g|x)\} \qquad (2.3.6.3)$$

Once terms that are constant across groups have been dropped this equation becomes;

$$d_g^{LDA}(x) = \mu_g^T \Sigma^{-1} x - \frac{1}{2} \mu_g^T \Sigma^{-1} \mu_g + log(\pi_g)$$

Σ here is the common variance-covariance matrix. μ_g represents the group mean which is replaced with the pooled mean μ_{pool} the centred score can be interpreted as a log posterior ratio that is equivalent to the discriminant score d_g^{LDA}. The discriminant score can then be simplified to

$$\Delta_g^{LDA}(x) = \omega_g^T \delta_g(x) + log(\pi_g) \qquad (2.3.6.4)$$

The term ω_g is the feature weight vector and is calculated according to;

$$\omega_g = R^{-\frac{1}{2}} V^{-\frac{1}{2}} \left(x - \frac{\mu_g + \mu_{pool}}{2} \right) \qquad (2.3.6.5)$$

where V is the diagonal matrix of variances and R is the correlation matrix. James-Stein-type shrinkage rules are then applied to R, V and π_g. The weight vector ω_g is used to calculate the correlation-adjusted vector of t-scores τ_g^{adj} as follows

$$\tau_g^{adj} = \left(\frac{1}{n_g} - \frac{1}{n} \right)^{\frac{1}{2}} \omega_g = R^{-\frac{1}{2}} \tau_g \qquad (2.3.6.6)$$

The vector τ_g contains the variable-specific t-scores between the mean of group g and the pooled mean. These values are used to calculate summary scores S_k for each variable which reflect the impact of each of them on group discrimination.

$$S_k = \sum_{g=1}^{G} \left(\tau_{k,g}^{adj} \right)^2 \qquad (2.3.6.7)$$

Where k refers to the k^{th} variable and g refers to the g^{th}group. The scores produced in this way are de-correlated t-scores. Using these scores the variables can be ranked in terms of their discriminatory potential. Since the goal here is to select variables which can be used to train a classifier the false non-discovery rate (FNDR) is used to set a threshold for variable selection. The FNDR can be estimated using the summary score S_k for each variable and then it can be used to identify that subset of variables which does not contribute any information in terms of group separation. Once this subset of variables is identified it is removed from the full set of variables and the remaining variables are used for group predictions.

Ahdesmaki and Strimmer (2010) analysed a gene expression dataset described in Singh et al., (2002). The dataset contains expression measurements for 6,033 genes from 102 patients. 52 of the patients had been diagnosed with cancer the remaining 50 had not (i.e. this was the non-disease group). Variable selection was carried out for two scenarios one assuming correlations of zero and one assuming non-zero correlations. Under the assumption of zero correlations using a FNDR threshold of 0.2 with diagonal discriminant analysis (DDA) 166 genes were selected. Prediction errors were then estimated for this selection as well as selections made using LDA and the false discovery rate. Their method was compared to the empirical Bayes estimate method of Efron (2008) and the nearest shrunken centroids algorithm (PAM). The selections were assessed by calculating the prediction error of each method's selections. The lowest error rate was found for the selections made using LDA and FNDR at 0.0550. This method chose 131 genes as being relevant to discrimination. The DDA-FNDR method selected 166 genes with a prediction error of 0.0640. These prediction errors were lower than those for any of the other methods used in this comparison including DDA using all 6,033 genes which had a prediction error of 0.3327.

In summary, the use of cat scores as a filter variable selection method offers a multivariate assessment of the importance of each variable to the task of discriminating between groups. The use of a priori mixing weights may not be the optimal choice depending on how the weights are chosen. It may be the case that using a posteriori weights might lead to more accurate assessments of each gene.

2.3.7 Probabilistic approach to variable selection

Filter methods of variable selection method may also be based on probabilistic approaches. Liu and Setiono, (1996) describe such an approach. Their method is a type of Las Vegas algorithm, which is a form of probabilistic algorithm that is unbounded in relation to the resources used to identify the correct answer and is guaranteed to return a correct answer i.e. the set of variables that best discriminate between two groups (Brassard and Bratley, 1996). It selects sets of variables at random from the full set of variables. Once a subset S of variables has been chosen it is compared to the current best subset of variables. If the number of variables in the two subsets is equal then the inconsistency of both sets is compared. A smaller inconsistency is considered better. Where the inconsistency of the new proposed subset is less than that of the current subset then the new proposed subset is identified as the best current selection.

The method of Liu and Setiono requires a prespecified threshold γ of the inconsistency criterion. Inconsistency between two observations is identified when they are identical except in their class memberships. The inconsistency of the data described by a particular variable or set of variables is calculated by counting the total number of inconsistencies in the data and subtracting from this the largest number of instances of a particular class label (i.e. the class label with the highest frequency).

To assess the performance of this method several real and simulated datasets were used (Liu and Setiono, 1996). The simulated datasets were CorrAL (6 binary variables, one of which is correlated to the class label 75 % of the time), Monk1, Monk2, Monk3 (each of these datasets contains 6 variables, Monk1 and Monk3 require 3 variables only to describe the target concepts, Monk 2 requires all 6 variables, Monk3 is also specified to contain some noise) and Parity5+5 (10 variables, 5 of these are irrelevant). The real datasets are Vote (16 variables, 300 training instances, 135 test instances), Credit (15 variables, 490 training instances, 200 test instances), Labor (16 variables, 40 training instances, 17 test instances) and Mushroom (22 variables, 8,124 instances of which 1,000 are randomly selected for testing).

For the simulated data Liu and Setiono (1996) were able to evaluate the performance of the variable selection method using knowledge of the relevant variables. Since the relevant variables from the real datasets were unknown it was necessary to evaluate the selections by calculating estimates of their performance. Using the method of Liu and Setiono the relevant variables were selected from the simulated datasets. For the real and simulated datasets the performance of selections was evaluated using the learning algorithms ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993). Comparing the performance estimates before and after variable selection a drop in the error rate from 37.5 % to 20.8 % for both ID3 and C4.5 was observed so the variable selection improved the classification.

In summary, the inconsistency measure used in this method satisfies the definition of a filter method and is relatively easy to interpret. However, the actual mechanism of the selection method involves generating multiple random subsets of variables and then analysing these to determine the optimal subset. This aspect of the method is similar to how wrapper (see Section 2.4) and embedded methods operate in that it involves the analysis of multiple variable subsets. The drawback of this is that the time required to identify the optimal subset of variables increases with the total number of variables.

2.4 Wrapper methods for variable selection

Another category of variable selection methods are wrapper methods (Kohavi & John, 1997; Xiong et al., 2001). The main principle of wrapper methods is using a specific classifier to evaluate every possible subset of variables for its discriminatory potential in order to identify the optimal subset of variables. If there are p potential discriminatory variables this leads to 2^p number of subsets, and hence 2^p runs of classification to evaluate the discrimination potential which can be time consuming and impractical. The advantage is that wrapper methods are multivariate so that the relationships between variables are taken into account. However, as wrapper methods evaluate all possible variable subsets the computational overhead is considerably larger than competing methods particularly as the number of variables increases leading to a combinatorial explosion and an increase in the time taken to evaluate each variable subset. An outline of a general approach to wrapper variable selection follows.

The typical format of a wrapper algorithm to select the set of the best discriminatory variables is as follows. Let us assume a set $D = \{S, L\}$ where S is the set of p variables and L is a vector containing the appropriate group labels. A set S' is a pre-defined initial variable subset, let θ represent the stopping criterion. S_{opt} is the optimal subset of variables. The wrapper algorithm initialises by evaluating the initial variable subset S', producing a value of φ which is a measure of the performance of the variable subset S' that is calculated using a classifier. In each round of variable selection by the wrapper algorithm the value of φ is re-calculated for each new subset of variables which is generated by the algorithm. If this new value is greater than the previous one the new subset replaces the previous one as the optimal variable subset S_{opt}. This process continues until the stopping criteria θ is reached.

There are many implementations of wrapper methods, depending on the type of chosen classification method. Kirapech-Umpai and Aitken (2005) describe a wrapper method employing an evolutionary algorithm classifier for variable selection from large microarray datasets. The method they describe is an example of a genetic algorithm which maintains a list of predictors which are effective at classification between groups. This initial list is generated randomly from the complete set of variables and is refined with deletions and additions as different combinations of variables are applied to the classification task. Petracoin et al. (2002) describe a wrapper method employing an iterative search algorithm classifier applied to the task of identifying markers which may be used to identify ovarian cancer. In their work mass spectroscopy data is collected on cancerous and non-cancerous samples. This data is then subjected to a combination of clustering and genetic algorithms designed to identify those proteomic patterns which can be used to differentiate between cancerous and non-cancerous samples. Xiong et al. (2001) outline the use of a wrapper method to identify biomarkers in gene expression studies. Using a subset of variables they use the classification accuracy of three classifiers, (Fischer linear discriminant analysis, support vector machine and the logistic regression (LR) model) as the assessment criteria. Inza et al. (2004) describe the use of a wrapper method which employs the classifier that will be used with a particular model. A subset of variables is applied to the task of classification using the chosen classifier and the performance of this subset is assessed using the leave-one-out-cross-validation (LOOCV) technique which returns percentage accuracies.

In summary, the advantage of wrapper methods is that they are all multivariate methods and that they evaluate all possible variable subsets and thus identify the best possible subset for discriminating between groups. Conversely the main disadvantage of wrapper methods is that this is a non-deterministic polynomial time (NP) hard problem and when there are a large number of variables these methods have large computational overheads. Another disadvantage is that they are classifier-specific so the variable subset identified using a particular classifier may not work as well with an alternative classification method.

2.5 Embedded methods for variable selection

Another type of variable selection method is the embedded methods. They are a hybridisation of both filter and wrapper methods. The principle is to extend a specific classifier into a framework that facilitates both classification and variable selection in one step. In other words the variable selection is embedded into the classifier. This is facilitated by calculating a contribution score (also called variable importance) of each variable for the classification of the subjects. The classification is therefore executed only once.

Depending on the particular selection strategy employed these estimates of variable importance can be used to remove the worst-performing variables from the variable set and then re-evaluate the remaining variables or to keep the best-performing variables from the variable set and add them to the final (optimal) variable selection. Because these methods do not evaluate every possible variable subset the computational overhead is considerably reduced compared to wrapper methods.

Typically, the embedded methods assume a relationship between group membership and the variables. Zakharov and Dupont (2011) assume a logistic model between the group membership and the predictors. They apply a logistic regression model to the task of variable selection as part of an embedded approach. Initially their method ranks the available variables univariately using a modified t-test. The modified t-statistic is calculated according to the formula

$$t_j = \frac{\mu_{j+} - \mu_{j-}}{\sqrt{\frac{\sigma_{j+}^2}{m_+} + \frac{\sigma_{j-}^2}{m_-}}} \qquad (2.5.1)$$

Here μ_{j+} and μ_{j-} are the average values of the feature j for the positively and negatively labelled examples specified by m_+ and $m_{-,}$ while σ_{j+} and σ_{j-} are the standard deviations. Once calculated the values of this t-statistic are used to rank the variables. This ranking of the variables is then normalised to produce a probability distribution vector. The probability distribution vector is applied as a feature sampling probability from which variables are selected. Logistic regressions carried out using subsets of these variables and the predictive performance of these regressions are used to refine the probability distribution vector of the variables until the optimal subset is identified. The authors apply their method to produce a regularised logistic regression model which may be suited to the selection of genes from microarray data. A drawback of this method is that the initial rankings which are used to construct feature sampling probabilities are univariate in nature. However the use of a standard t-test in the ranking of variables is far less computationally demanding than alternative ranking metrics and the method benefits from its use.

Ju and Brasier (2013) describe an embedded method based on random forests (RF). Multiple trees are grown by randomly selecting variables at each split. By allowing the trees to develop fully (i.e. no pruning of any trees) the authors minimise bias in the final trees. Multiple forests are grown in an iterative process which removes a proportion of the worst performing variables at each iteration. Diáz-Uriarte and Alvarez de Andres (2006) describe the use of random forests for variable selection from gene microarray data for differentiating between cancerous and non-cancerous samples. The optimal variable subset for discriminating between groups is associated with the forest that has the best performance. Jiang et al. (2004) also describe the use of random forests to select genes from microarray data. Ensemble method selections which are made using random forests are an aggregate of multiple classification trees (500 is a common number of trees in a random forest) and it may be expected that this will produce a final selection which is an accurate representation of the discrimination potential of each variable. At the same time the “random” element means that, particularly for high dimensional data no two forests are ever the same making it necessary to carry out considerable validation of results to confirm their accuracy.

Tang and Mao (2007) describe a method for carrying out variable selection from datasets containing mixtures of continuous and nominal variables. This method is termed the mixed feature selection algorithm. For a dataset containing mixtures of continuous and nominal variables the variables are split into continuous and ordinal variables in the first step. Each of the continuous (or ordinal) variables is added to the set of previously selected variables S and the performance of this new subset is assessed. For continuous variables this is done using the Mahalanobis distance. For ordinal variables this is done using the SU.

Having identified the optimal continuous and ordinal variable from the remaining candidate variables each of these two variables is individually added to S. The performance of the resultant dataset(s) is then assessed by estimating the classification error probability. This is calculated according to;

$$P_e(X,Z) = \sum_{i=1}^{N} p(z_i) P_e(X|z_i) \qquad (2.5.2)$$

This gives us the error probability for a mixed feature set (X,Z) when the data could belong to one of G groups y_j where $j = 1,2,\dots,G$. The multi-nomial variable z contains N possible distinct values representing value-combinations of the nominal feature Z. The maximum likelihood estimate of $p(z_i)$ is the ratio of the frequency of sample associated with the value-combination z_i, (denoted as n_i), to the number of samples.

$$p(z_i) = \frac{n_i}{n} \qquad (2.5.3)$$

In order to calculate the conditional error probability $P_e(X|z_i)$ it is necessary to decompose the variable space into a set of mutually exclusive feature subspaces based on the multi-nomial variable z. Following decomposition of the variable space the conditional error is estimated using k-nearest neighbours.

$$\hat{P}_e^{KNN}(X|z_i) = 1 - \frac{1}{n_i} \sum_{l=1}^{n_i} \frac{k_{gl}}{k} \qquad (2.5.4)$$

k_{gl} is defined as the number of subjects that belong to the class g and k is the number of nearest neighbours. The best variable between the optimal continuous and ordinal variables identified in the first step is now identified using the estimated error probability.

The mixed feature selection algorithm was assessed using simulated and real datasets. One simulated scenario comprised a single continuous variable and a single nominal variable both of which were relevant to discrimination. Of the remaining 2 scenarios in one the nominal variable was irrelevant to discrimination and in the other the continuous variable was irrelevant to discrimination.

The mixed feature selection algorithm was compared to the Relief algorithm (Kononenko, 1994), the correlation based feature selection (CFS) method (Hall, 1999) and a method using the generalised Mahalanobis Distance of Bar-Hen and Daudin (1995). Each method was applied to the task of variable selection from each of 4 real datasets. The Cleveland heart disease dataset consists of observations on 303 subjects of 7 nominal variables and 6 continuous variables. The Australian credit screening dataset consists of observations on 690 subjects of 9 nominal variables and 6 continuous variables. The Horse colic dataset consists of observations on 368 subjects of 15 nominal variables and 7 continuous variables. The Pittsburgh bridges dataset consists of observations on 108 subjects of 8 nominal variables and 3 continuous variables.

For each of the simulation scenarios the values of the estimated error probability (Tang and Mao, 2006), Mahalanobis distance, generalised Mahalanobis distance and the SU were calculated. Within each scenario the sample size and number of levels for the nominal variable were varied to assess the sensitivity of each metric. The estimated error probability exhibited similar performance to the Mahalanobis distance in terms of its response to changes in sample size and number of levels. The generalised Mahalanobis distance out-performed both the estimated error probability and the Mahalanobis distance. However the best performance was achieved by the SU metric which showed the greatest difference between the differing sample sizes and number of levels within each scenario.

For the Cleveland heart disease, Australian credit screen and Pittsburgh bridges datasets the mixed feature selection algorithm was out-performed by each of the competing methods used in the study. In particular Relief-F and the CFS method had larger estimated classification performance. I can only assume that the Relief-F and CFS methods were better able to evaluate the discriminatory potential of the variables in these datasets and so the resulting selections performed better than those made using the mixed feature selection algorithm. For the horse colic dataset all 4 methods achieved similar performance in terms of number of variables selected and classification performance estimates.

The most obvious issue with the work presented by Tang and Mao is that their method is out-performed by other methods applied to 3 out of 4 real datasets and only achieves equivalent performance to these same methods in the fourth dataset. However, even when analysing results for simulated datasets the method of Tang and Mao appears to be more responsive to relevant continuous variables. When the number of nominal levels is increased there is a clear increase in the value of the estimated error probability when paired with a relevant continuous variable. When paired with an irrelevant continuous variable this increase almost disappears. The error bars for each metric are also overlapping in all scenarios with the exception of the SU estimates which are clearly separated in all scenarios. The effort required to implement the method of Tang and Mao does not seem to justify the resultant performance.

The method of Doquire and Verleysen (2011) uses MI to carry out variable selection from datasets containing mixtures of continuous and nominal variables. They refer to this as the hybrid-MI method. Continuous and nominal variables are first identified and separated. Sets of continuous variables and sets of nominal variables are then ranked independently using estimated MI. The performance of the first ranked variable in each list is then estimated and the variable which performs better using a specific classifier is added to the list of selected variables and removed from the relevant continuous or nominal variable list. The algorithm compares the top-ranked variables from each list at each step selecting the best performing variables until the lists have been exhausted (i.e. all variables have been selected). In this way the algorithm uses a filter approach (to produce the ranked lists of continuous and nominal variables) and an embedded approach (to assess subsets of the ranked variables using a specific classifier).

For continuous variables the MI is estimated using the principle of nearest neighbour methods. The MI is calculated according to

$$\widehat{H}(X) = -\Psi(k) + \Psi(N) + log(c_d) + \frac{d}{n}\sum_{n=1}^{N} log\big(\varepsilon_x(n,k)\big) \qquad (2.5.5)$$

In this formula k is the number of nearest neighbours, X is a random variable and N is the number of observations of the variable X. c_d is the volume of a unitary ball of dimension d. $\varepsilon_x(n,k)$ is twice the distance from the nth observation x to its kth nearest neighbour. Ψ is the digamma function.

For nominal variables the MI is estimated using the minimal-Redundancy maximal-Relevance criterion (mRmR). Denoting S as the set of indices of variables which have already been selected the mRmR criterion for a variable i which has not yet been selected is calculated using the formula

$$mRmR(\mathfrak{f}_i) = I(\mathfrak{f}_i;Y) - \frac{1}{|S|}\sum_{j\epsilon S} I\big(\mathfrak{f}_i;\mathfrak{f}_j\big) \qquad (2.5.6)$$

It should be noted that all estimates of mRmR are bivariate. The terms $\mathfrak{f}_i$ and $\mathfrak{f}_j$ are the probability distributions of the predictor variables and the response variable respectively.

Doquire and Verleysen compare their method to the method of Hu et al. (2008) using both simulated and real datasets. In brief the method of Hu et al. operates by selecting those variables which share their class label with a sufficient number of neighbours. Neighbours in this method are defined as two points whose nominal attributes are identical when one is amongst the k-nearest neighbours of the other. Alternatively two points may be considered neighbours if the values of their continuous variables are within a sufficient distance of each other. As Doquire and Verleyson observe the effectiveness of the method has the potential to vary depending on how “neighbours” are defined.

The simulated dataset contained two nominal variables each possessing two levels with equal prior probability of a point belonging to one level or the other. The dataset also contained two continuous variables uniformly distributed over the range $[0;1]$. The only relevant variable in this dataset was X_3 and this was ranked first in only 28 out of 50 repetitions using the method of Hu et al. and 50 out of 50 repetitions using the method of Doquire and Verleyson.

The real datasets analysed were titled Heart, Hepatitis, Australian credit and Contraception. The Heart dataset contained 6 continuous and 7 nominal variables measured on 270 samples across 2 groups. The Hepatitis dataset contained 6 continuous and 13 nominal variables measured on 80 samples across 2 groups. The Australian credit dataset contained 6 continuous variables and 8 nominal variables measured on 690 samples across 2 groups. The Contraception dataset contained 2 continuous and 7 nominal variables measured on 1,473 samples across 3 groups. Two different classifiers were used, a naïve Bayes classifier and a 5-nearest neighbours classifier.

The performance of each classifier was estimated using selections made based on the methods of Hu et al. and Doquire and Verleyson. Performance was assessed by estimating the error rate associated with the trained classifiers. Lower error rates were observed using each classifier with selections made by the Hybrid-MI method from the Hepatitis and Contraception datasets. For the Heart and Australian credit datasets the differences in error rate were not as great.

There are several issues with the work of Doquire and Verleyson. First a point of clarification; Doquire and Verleyson refer to their method as using a wrapper approach to select variables from the ranked lists of continuous and nominal variables. This is not accurate. This part of their method uses a specific classifier to select variables in the process of training the classifier which satisfies the definition of an embedded (not wrapper) method. The primary difference being that every possible variable subset is not evaluated in the Hybrid-MI method. The second problem is that the Hybrid-MI method does not select a subset of variables. Instead it first ranks all variables then trains a classifier in a specific order based on the performance of the ranked variables. Ultimately what is produced is a classifier which uses all of the variables in a set order. There is no stopping criterion by which the user may determine that the optimal subset has been selected. The third issue relates to the results of the comparison of the Hybrid-MI method with the method of Hu et al. There is a clear disparity in the performance difference between the two methods based on the proportion of nominal variables comprising the test datasets. The hepatitis and contraception datasets have the largest proportion of nominal variables (13 nominal to 6 continuous and 7 nominal to 2 continuous respectively) and the best performance of the Hybrid-MI method is observed for these datasets. The performance of the Hybrid-MI method appears to be skewed in favour of datasets containing a large proportion of nominal variables. Finally the calculation of mRmR for nominal variables is bivariate. While this is not as limiting as a univariate context it is also not as comprehensive as a multivariate context.

Guyon et al. (2002) describe a method of carrying out variable selection which utilises a support vector machine (SVM) classifier. The main operating principle of SVMs is that they project data into higher-dimensional spaces. Once a suitable higher dimensional space has been identified which is capable of separating the groups comprising the dataset the optimal hyperplane can be calculated (this is the hyperplane which separates the data with the maximal margin (Cortes and Vapnik, 1995)). The optimal hyperplane is defined by the weight vector w which is a linear combination of the variables used to train the classifier (Guyon et al., 2002).

$$w = \sum_k \alpha_k y_k x_k \qquad (2.5.7)$$

In this equation y_k is the group label (either -1 or +1 in the 2-group scenario), x_k is an n-dimensional vector used to train the classifier and α_k is the weight associated with this vector. When a SVM is being trained it only uses a subset of the training examples. The vectors comprising this subset are those which are closest to the decision boundary and which lie on the margin (these are called support vectors). As such the weights associated with most vectors are zero (Guyon et al. 2002). Any vector which has a non-zero weight is a support vector.

The method of Guyon et al. is a recursive feature elimination (RFE) method. A SVM is first trained using the available data. The weight vector w is then computed. In order to rank each of the candidate variables the weight calculated for each is squared to produce the ranking criterion c.

$$c_i = (w_i)^2 \qquad (2.5.8)$$

Here the subscript i indicates that we are considering a single variable (i.e. we are no longer considering the vector of all variables). Each time the SVM classifier is trained in this way the variable which has the smallest ranking criterion is removed from the set of candidate variables.

The method of Guyon et al. (2002) was assessed using two datasets composed of gene expression vectors collected from DNA micro-arrays. The first dataset consisted of 72 bone marrow samples from individuals with acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML). Altogether there were 47 ALL samples and 25 AML samples. The expression levels of 7,129 genes were measured on each sample. The second dataset consisted of 22 samples of healthy colon tissue and 40 samples of colon cancer. The expression levels of 2,000 genes were measured on each sample.

Guyon et al. compared their SVM-RFE method to a method developed by Golub et al. (1999) which used correlations between variables and groups to rank variables and make selections. In the method of Golub et al. (1999) a classifier was trained using weights based on the correlation coefficients. Both classifiers gave identical results for a particular set of genes. However the SVM-RFE method selected smaller numbers of genes which also achieved better performance than selections made using the method of Golub et al. (1999).

The comparison of SVM-RFE and the method of Golub et al. indicated that a linear SVM classifier and the classifier of Golub et al. exhibited equivalent performance for a given set of genes. Conversely the identity of the particular genes selected had a significant impact on the performance of the two classifiers. In this respect the smaller number of genes (8, 16) selected by the SVM-RFE method exhibited equivalent performance to those selected by the method of Golub et al. when using both classifiers and the leukemia dataset. Equivalent performance was also observed using both classifiers with genes selected using the method of Golub et al. which selected 64 genes. For the colon cancer dataset the overall performance of the SVM classifier on selections made using SVM-RFE (8 genes) and the method of Golub et al. (8 genes) was better than that observed using the classifier of Golub et al. (32 genes for SVM-RFE, 16 genes using the method of Golub et al.).

There are a number of shortcomings with the work presented by Guyon et al. the first of which is that pre-processing of datasets appears to be essential to the effective application of their method however, only very general details on this pre-processing are provided. The SVM-RFE method also requires the specification of at least two parameters (referred to as “soft margin parameters) but the details of how the optimal parameters values were determined in this work are not presented. In respect of the comparison of SVM-RFE to the method of Golub et al. the comparison of the two methods was achieved by analysing the performance of two different classifiers (linear SVM and the classifier of Golub et al.) using the gene selections made by the SVM-RFE method and the method of Golub et al.. The SVM classifier used in the comparison is linear as is the classifier used as part of the SVM-RFE method. It seems prudent to question whether the results presented are subject to some form of over-fitting. Would the same performance have been observed had a different form of classifier been used? Finally it must be noted that while the SVM-RFE method achieved similar performance to the method of Golub et al. using smaller numbers of genes this was most likely at a considerable additional cost in terms of the time and computational resources required (relative to the method of Golub et al.).

In summary the main advantage of embedded methods is that they are multivariate in nature. In addition, relative to wrapper methods the computational workload is greatly reduced when using embedded methods. A disadvantage is that depending on the classifier being used embedded methods can be quite complex and require considerable tuning of parameters to achieve optimal performance. In this thesis I use the `varselRF` method employing random forests as part of the comparison study presented in Chapter 4.

2.6 Information evaluation in signal-processing theory and SNR measures

A relevant concept for classification is the concept of the signal-to-noise (SNR) ratio, because we want to find variables that carry the largest amount of information about the groups of interest (e.g. why one patient is in group 0 and another is in group 1). The SNR concept is old and is defined as the amount of signal divided by the amount of noise within a given system (Bérubé and Wu, 2000; Kaiser et al., 1998; Czanner et al., 2015). Using this definition there are multiple versions of the SNR in existence based on the particular types of signal and the type of noise.

Examples of the application of the SNR include its use in researching how background noise affects speech understanding (Maamor & Billings, 2016). Jiang et al. (2016) describe identifying the minimal SNR which facilitates stabilization of single-input single-output linear time-invariant systems. Another example of application of SNR is the comparison of magnetic resonance imaging (MRI) images obtained using different hardware and protocols (Welvaert & Rosseel , 2013).

The work of Bérubé and Wu (2000) shows how the concept of the SNR is applied to statistical process control. The signal is the output of the production process and the noise is anything which introduces variation into the production process. The basic premise in statistical process control is that by quantifying and segmenting different sources of variation in the production process it is possible to reduce the level of variation between production runs. In this way a larger proportion of final products will be within the acceptable margins of deviation when compared to the desired product specification. In the design of the production process control factors are sources of variation which the experimenter can control while noise factors are those sources of variation which cannot be controlled. The signal-to-noise ratio used in this paper was originally described by Taguchi (1991),

$$SNR = log\frac{\mu^2}{\sigma^2} \qquad (2.6.1)$$

In this formula μ is the value of some production characteristic Y and σ is the variation in the production process associated with noise factors. These parameters are estimated using the measured values of Y and the associated variances. The calculation of the SNR facilitates segregation of control factors into adjustment and non-adjustment factors. Adjustment factors are those which affect neither the mean values of y nor the SNR value. Non-adjustment factors are those which affect the variability measured by the SNR. Successful identification of adjustment and non-adjustment factors allows the experimenter to identify the optimal levels of the control factors.

Bérubé and Wu (2000) compare the SNR to alternative methods which use the standard deviation. Additive and multiplicative models of the behaviour of noise factors within the system were used. Comparisons were carried out across a range of case studies favouring different variables in each case. Each of the scenarios was implemented by tweaking the ratios between the variable regression coefficients and non-adjustment factor functions measuring dispersion effects. The authors conclude that using this SNR where the standard deviation of the response Y is linearly proportional to the mean the correct identification of the control factors and their optimal values is possible. However, any deviation from this condition means that identification of the control factors and their optimal values can no longer be assured

Welvaert and Rosseel (2013) present a definition of the SNR which is designed for use with functional magnetic resonance imaging (fMRI) images. Normal somatic activity within the subject (e.g., vascular and respiratory activity) is recorded as noise in the fMRI image. Activity within the fMRI system itself may also contribute to noise levels. A special form of temporal SNR (tSNR) is calculated utilising knowledge of the mean signal over time. This SNR is calculated according to the formula;

$$tSNR = \frac{\bar{S}}{\sigma_N} \tag{2.6.2}$$

In this formula $\bar{S}$ is the magnitude of the activation signal and σ_N is the standard deviation of the noise signal. The activation signal measured in the course of an fMRI is the energy released by nuclei in the organic structures being scanned (Brown et al., 2007). More specifically in the presence of a magnetic field oscillating at the appropriate resonance frequency these nuclei are capable of absorbing energy from the field. As electrons in the nuclei return to their ground, unexcited state (i.e. release the energy they had previously absorbed) the associated energy release can be detected by a detector coil in the fMRI apparatus. The resonance frequency is proportional to the magnetic field of the nucleus. This fact is exploited to produce a signal that is a mixture of multiple different frequencies all peculiar to the specific region from which they originated, essentially a topographical map of the structure of interest. A limitation of this SNR definition is that the baseline measurements are highly dependent on the specific scanning parameters which are used to collect the fMRI data. Consequently this SNR is not ideally suited to the analysis of stimulus-response fMRI data. This is because the SNR calculated may not allow for fluctuations in the activation signal which arise due to the task being carried out.

Recently a SNR extension to generalised linear models for general point processes of individual neuronal electrical discharges was proposed (Czanner et al., 2015). The signal is the stimulus that makes neurons to emit electrical discharges. The noise is the thermal noise in neurons and the network activity of the neighbouring neurons. The work of Czanner et al (2015) extends the SNR into systems where the noise can be described by a Poisson distribution. This is achieved by replacing the sums-of-squares by deviances. The deviance is the difference in log-likelihood between the considered model and a saturated model. A random variable with Poisson distribution has the mean equal to the standard deviation. Hence the SNR of Czanner et al. (2015) implicitly allows the variance of the noise to vary with the mean of the signal.

To develop their SNR Czanner et al (2015) first demonstrate that standard SNR computations are valid estimates of the ratio of expected prediction errors (EPEs). In turn the ratio of EPEs may be estimated using the ratio of sum of squares of residuals for linear Gaussian models with covariates. To extend the SNR to generalized linear models (GLMs) the sum of squares in the EPEs are replaced by the residual deviances. The final SNR developed for neuronal systems is calculated for a single neuron according to;

$$\bar{\hat{S}}NR_S = \frac{Dev(n,\hat{\beta}_0,\hat{\beta}_H)-Dev(n,\hat{\beta})-dim(\hat{\beta}_0)-dim(\hat{\beta}_H)+dim(\hat{\beta})}{Dev(n,\hat{\beta})+dim(\hat{\beta})} \qquad (2.6.3)$$

where the term Dev is the residual deviance (Nelder & McCullagh, 1989). In this equation $\hat{\beta}_H$ is the parameter vector describing the history of neural activity, $\hat{\beta}_0$ is the intercept term taken from the GLM, (i.e. its' value relates to the resting potential of the neuron), and dim is the number of parameters associated with each β term. The relationship between the SNR and the coefficient of determination R^2 is also demonstrated in this work (Czanner et al., 2015). It is also shown that because Cohen's effect size (Cohen, 1992) is calculated using the coefficient of determination there is a link between Cohen's effect size and the SNR. The Cohen's effect size is calculated according to the formula

$$\mathrm{f}^2 = \frac{R^2}{1-R^2} \qquad (2.6.4)$$

Here R^2 is the explained variance.

To our knowledge there is no literature using the SNR in a scenario where the signal is defined as membership of one of two groups (i.e. healthy vs. disease) and the system measurements are the measured values of the potential predictor variables.

2.7 Discussion

In this chapter I have provided an overview of several existing methods of variable selection from amongst the categories of filter, wrapper and embedded methods. I have also provided an explanation of the method of PCA and an introduction to the concept of a signal-to-noise ratio.

The goal of variable selection is to identify a set of those variables with the strongest ability to discriminate between groups. In this way a parsimonious approach to model-building and prediction can be achieved. PCA is an alternative method which attempts to capture the information accounted for by variables and produce new variables based on the original variables. The information is measured by variance. However PCA is limited by ignoring the group membership, i.e. the group membership is not utilised in the process of finding the set of discriminatory variables.

I have reviewed several different filter methods of variable selection in this chapter. A large variety of metrics exist which are employed in filter methods. Most popular are metrics based on information gain, distance correlations and t-scores. While there is considerable potential for information capture with these different metrics a drawback is that many of them do not consider relationships between variables. In this way many filter methods are univariate in nature.

I also provided an overview of wrapper and embedded methods. Both wrapper and embedded methods are multivariate methods in the sense that they do consider relationships between variables, which is an advantage compared to filter methods. However what makes embedded and wrapper methods effective is that they evaluate a large number of possible variable subsets in order to identify the optimal subset. Unfortunately the higher the dimensionality of a dataset becomes the less practical wrapper or embedded methods become.

Finally, since I was concerned about the selection of a set of variables that contain the most information about the groups, I introduced the general concept of measuring the information and in particular I focused on the signal-to-noise ratio (SNR). SNR has been applied with a variety of purposes including as a means of regulating production by controlling sources of variation.

In the light of these points the specific objectives of this thesis are;

Objective 1: Propose improvements to the existing variable selection methods and create a novel algorithm for multivariate normal data.

- i. Evaluate the properties of the new algorithm in simulated data.
- ii. Compare the novel algorithm with the existing variable selection methods.
- iii. Evaluate the ability of the novel algorithm to handle non-normal data.
- iv. Evaluate the ability of the novel algorithm to deal with data missingness and determine the most suitable means of imputation for use with the algorithm.
- v. Evaluate imputation techniques for missing data: using mean values, multiple imputation and alternative methods.

Objective 2: Extend the novel algorithm to imaging data by incorporating the spatial correlations

Objective 3: Application of the novel algorithm to real datasets for the purpose of variable selection. Four datasets are considered:

- i. The *Diabetic REtinopathy: FUnctional and Structural study*, (DREFUS) dataset: Section 6.2, 27 variables, 2 groups; early diabetic retinopathy and no diabetic retinopathy, imbalanced groups, missingness.
- ii. The Individual Risk-based Screening for Diabetic Retinopathy (ISDR) dataset: Section 6.3, 28 variables, 2 groups; Referable DR and Non-referable DR, imbalanced groups, missingness.
- iii. The MRet dataset: Section 6.4, 81 variables, 2 groups; Survival and Death, imbalanced groups, missingness.
- iv. The Keratoconus dataset: Section 6.5, 17 variables, 2 groups; Healthy and Keratoconus, balanced groups, no missingness.

2.7.1 DREFUS

The ultimate clinical goal of DREFUS (Harding et al., 2010) was to elucidate the relationships between functional and structural variables, if the relationship depends on the level of DR, and which variables (or set of variables) can best discriminate between the DR stages. This is important in clinical settings because it can help to identify the measurements that should be used to find eyes that are at risk of having DR. The current gold standard is fluorescein angiography (FA) which is used to determine the 4 stages, but this is an expensive and invasive technique. Therefore the clinical importance of DREFUS was to evaluate less invasive and less expensive measurements that could differentiate between the 4 stages of DR. Measurements that are particular to the DREFUS dataset include microperimetry (MP), multifocal electroretinogram (mfERG), oscillatory potential (OP), cholesterol and HbA1c.

Perimetry is a technique used to quantify the visual field. Perimetry measurement can be complicated in individuals with visual impairment which compromises their visual fixation ability. MP uses infrared light to illuminate the fundus. By utilising infrared light projection MP addresses the issue of impaired fixation as light projection is independent of fixation. Thus MP facilitates an examination of the fundus topography which has a high correlation between fundus details and light sensitivity (Rohrschneider *et al.*, 2008). Electroretinography uses an electrode placed on the surface of the eye to monitor electrical responses of the retina to light stimulation. Full-field electroretinography assesses the response of the entire retina. Multi-focal electroretinography assesses the response of the retina at multiple different locations which allows a topographical mapping of the functionality in the central 40^{o} – 50^{o} of the retina (Lai *et al.*, 2007). Oscillatory potentials (OP) are High-frequency, low-amplitude wavelets occurring on the ascending limb of the b-wave (associated with photopic and scotopic responses) with frequencies between 100 and 160 Hz. Diminished OPs are associated with conditions such as familial exudative vitreoretinopthy (Ohkubo and Tanino, 1987) and diabetic retinopathy (Kizawa *et al.*, 2006).

2.7.2 ISDR

Another ophthalmic application where variable selection is needed is for discriminating between referable sight threatening diabetic retinopathy and non-referable sight threatening retinopathy in subjects with diabetes. DR is a progressive disease of the retina which causes blindness. Early and late stage DR is asymptomatic however, late stage DR can result in blindness if not treated. Digital photography is effective at screening for sight-threatening diabetic retinopathy (STDR). In England it is recommended that individuals with diabetes over the age of 12 are screened annually. While screening is integral to the early detection of STDR the costs of annual screening to the NHS are considerable. The ISDR dataset came from a study called “Introducing personalised risk based intervals in screening for diabetic retinopathy: development, implementation and assessment of safety, cost-effectiveness and patient experience” (Harding et al, 2011) which is referred to as the ISDR study. The motivation behind the ISDR study was to develop individual risk-based screening protocols thereby eliminating the need for annual screening for those with lower risk and increasing the frequency for those with high risk. Variables measured as part of this study included glycated haemoglobin (HbA1c), cholesterol, systolic and diastolic blood pressure.

Glycated haemoglobin (HbA1c) is associated with the β-N-1-deoxy fructosyl component of haemoglobin. HbA1c is taken as being representative of the average levels of blood glucose over a period of 120 days. Dysregulation of blood glucose levels is associated with diabetes and diabetic retinopathy (Long *et al.*, 2017). Cholesterol is a form of lipid found in the buman body. It is partitioned into high density lipoprotein (HDL) and low density lipoprotein (LDL), tests often measure total cholesterol which is taken as being equivalent to LDL level. LDL is typically considered to be harmful and therefore the lower the level of LDL (i.e. Cholesterol) the better the health outcome for the individual. Individuals suffering from diabetes are at increased risk of dyslipidemia, (abnormal levels of lipid in the blood), and it is has been demonstrated that LDL levels are associated with the presence of hard exudates found in patients with DR (Chang and Wu, 2013). Blood pressure is the pressure exerted by blood flowing through veins and arteries on the walls of the vessels. Blood pressure is measured as a combination of systolic (associated with contraction of the cardiac muscle pumping blood from the chambers of the heart to the rest of the body) and diastolic (pressure when the cardiac muscle is relaxed i.e. between contractions) pressure. Elevated blood pressure (hypertension) and reduced blood pressure (hypotension) are associated with many pathologies as they can indicate problems with the cardiovascular system, blockages or other problems. In relation to diabetic retinopathy hypertension has been shown to damage the retinal capillary endothelial cells of individuals with diabetes as well as a slowing of the rate of progression of DR in a study assessing the impact of blood pressure on DR (Klein and Klein, 2002).

2.7.3 MRet

The MRet dataset comes from two work packages in a Wellcome Trust funded Programme Grant entitled, “The retinal microvasculature in cerebral malaria in African children (MRet).” SP Harding, RS Heyderman, AG Craig, PS Hiscott, ME Molyneux, TE Taylor, S Kampondeni, NAV Beare, P Knox, M Mallewa, Y Zheng. (092668/Z/10/Z). Plasmodium falciparum, a unicellular protozoan species, is the causative agent of malaria in humans. The plasmodium pathogen is spread by mosquitoes with the disease being widespread in sub-saharan Africa (Hoffman *et al.*, 2015). The parasite matures inside the host in two phases; the exoerythrocytic phase (within infected liver cells) and the erythrocytic phase (infected red blood cells) (Bledsoe, 2005). In order to avoid destruction by the immune system the parasite expresses adhesive proteins on the surface of infected red blood cells which cause them to adhere to blood vessel walls. The sequestration of red blood cells in this way can lead to blockage of microvasculature and cerebral malaria if the blood-brain barrier is breached (Renia *et al.*, 2012). The retina is a portion of the central nervous system (CNS) which may be analysed without the need for invasive techniques. This allows for the possibility of assessing the retina for the presence of features such as vessel abnormality or haemorrhaging which are known to be associated with paediatric cerebral malarial (MacCormick, 2014). Impaired perfusion in retinal microvasculature has been observed in children with celebral malaria (Beare *et al.*, 2009). Hence imaging in this dataset was focused on regions considered likely to exhibit capillary non-perfusion. Additional clinical variable measured include age, sex, weight, serum lactate and respiratory data.

2.7.4 Keratoconus

Keratoconus is a disorder of the eye leading to progressive thinning of the cornea. Progressive thinning can lead to the cornea assuming an atypical morphology, often a cone-shape which leads to visual impairment (Tur *et al.*, 2017). The keratoconus dataset that I studied contains measurements taken on the eyes of healthy individuals and individuals with keratoconus in St. Paul's Eye Unit, corneal clinics. Each measurement was taken in triplicate for each subject. The dataset consists of measurements of 17 variables from 60 patients. The eye is the unit of analysis and from each patient, only one eye from each patient was used. The variables collected in this study relate to keratometry (the curvature of the cornea) and pachymetry (the thickness of the cornea). Keratometric measurements are provided for the corneal front and back. Taken together keratometry and pachometry measurements give the dimensions of the corneal structure and allow clinicians to determine the presence or absence and extent of keratoconus. Comparison of the corneal dimensions for healthy eyes and keratoconic eyes can help determine which measurements are most effective at identifying individuals with keratoconic eyes.

Chapter 3 presents theoretical details of the relevant existing statistical methods and will also give details of the developed statistical methods.

Chapter 3. Methodology: a multivariate SNR metric for variable selection in classification

3.1 Introduction

In my literature review (Chapter 2) I identified Hotelling T^2 as an important multivariate discriminatory metric that has an intuitive interpretability and that has been used previously in a filter method for variable selection for classification. However, a disadvantage of T^2 is that it assumes equality of covariance matrices across the groups, which is often not satisfied in clinical datasets (Table 3.1.1). In the signal processing literature there is no satisfactory SNR metric that allows different covariances across groups. Hence there is a need to propose a generalisation of Hotelling T^2. The goal of this chapter is to propose an extension of T^2 into a new metric that may be better suited for a scenario where covariance matrices differ across groups, and to investigate ways to use the new metric in a variable search algorithm.

Table 3.1.1: Summary of the known and developed novel concepts explained in Chapter 3. Legend: The text in italic represents the two of the three contributions of this thesis.

Method	**Assumption**	
	Equal variance-covariance matrices	Unequal variance-covariance matrices
Discriminant analysis	Linear discriminant analysis (LDA) [Section 3.2.1 and 3.3]	Quadratic discriminant analysis (QDA) [Section 3.2.2 and 3.3]
Discriminatory metric	T^2 [Section 3.4]	*Develop a novel metric [Section 3.5]* *[Section 3.6 on how T2 and SNR relate]*
Variable selection algorithm (filter method)	MFS-T2 by Lu *et al* (2005), stopping criteria based on p-value [Section 3.7]	*Develop a novel algorithm, propose suitable stopping criteria* *[Section 3.8, 3.9]*

The structure of Chapter 3 follows the structure of Table 3.1 above. First, I present the details of linear discriminant analysis (Section 3.2) and quadratic discriminant analysis (Section 3.3). Then I will present an overview of Hotelling's T^2 statistic (Section 3.4) and explain how it relates to linear discriminant analysis. I then describe the proposed novel extension of the T^2 metric, which will be called the signal-to-noise ratio (SNR) (Section 3.5) and which allows for heterogeneous variance-covariance matrices across groups. I will also show how Hotelling's T^2 statistic and the SNR are related (Section 3.6) and how I propose to employ SNR in a novel variable selection method called the MFS-SNR algorithm (Section 3.8). Finally, I will explain the structure and principles of the MFS-SNR algorithm as well as the stopping criterion used by the algorithm, the validation of selection results and how the number of simulations was chosen for my simulation studies (Section 3.8). In Section 3.9 I will discuss the computational challenges of the work presented in my thesis. Finally I summarise Chapter 3 in Section 3.10.

3.2 Discrimination methods

3.2.1 Linear discriminant analysis

Linear discriminant analysis (LDA) was originally developed by Ronald A. Fisher (Fisher, 1936). The goal of LDA is to discriminate between groups. It does this by finding the most accurate data representation in a lower dimensional space. It finds a projection to a line such that samples from different groups are well separated.

LDA implicitly assumes that the variance-covariance matrices are identical across the populations of interest. LDA also implicitly assumes that data are normally distributed.

In many situations we will not have knowledge of the population means and covariance matrices. In the absence of this information we can estimate the means and covariance matrices using our sample. The variance-covariance matrices of the sample groups of interest are combined to produce a single pooled variance-covariance matrix. The classification rule is then to allocate a subject $\boldsymbol{x_0}$ to group 1 if;

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2) \geq ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{3.2.1.1}$$

otherwise x_0 is assigned to group 2. If we assume that the product of the ratios of misclassification costs and prior probabilities is equal to 1 then the natural logarithm of this term will be 0. The classification rule then reduces to a comparison between the scalar variable

$$\hat{\boldsymbol{y}} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_{pooled}^{-1} \boldsymbol{x} = \hat{\boldsymbol{a}}' \tag{3.2.1.2}$$

which is evaluated at x_0, and the number

$$\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{3.2.1.3}$$

$$= \frac{1}{2}(\bar{\boldsymbol{y}}_1 + \bar{\boldsymbol{y}}_2)$$

where

$$\overline{\boldsymbol{y}}_1 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\boldsymbol{S}_{pooled}^{-1}\boldsymbol{x}_1 = \widehat{\boldsymbol{a}}'\overline{\boldsymbol{x}}_{\mathbf{1}} \tag{3.2.1.4}$$

and

$$\overline{\boldsymbol{y}}_2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\boldsymbol{S}_{pooled}^{-1}\boldsymbol{x}_2 = \widehat{\boldsymbol{a}}'\overline{\boldsymbol{x}}_2. \tag{3.2.1.5}$$

As this classification rule is based on sample data it is an estimate of the expected cost of misclassification (ECM). New observations can then be assigned to populations 1 or 2. This determination is made based on whether the new observation falls to the left or the right of the midpoint between the two means $\overline{\boldsymbol{y}}_1$ and $\overline{\boldsymbol{y}}_2$.

As with any classification method there may be subjects that are incorrectly classified. Two misclassifications scenarios can occur. A patient from group 2 can be incorrectly classified as being from group 1, and a patient from group 1 can be incorrectly classified as from group 2. Incorrect classification of new patients to groups is a problem. In clinical settings the need to assign patients to either a disease or non-disease group is clear, however the same need exists in other non-clinical settings For example we may be trying to assign a customer to a group that will or will not buy a product, assigning organisms to one species group or another, determining whether properties are solvent or distressed etc. The need to correctly assign subjects to the correct class is obvious in clinical settings where incorrect classification may have undesirable consequences.

The costs of misclassification may be the same across groups, but they can also differ. While the costs of misclassification may not be as severe in non-clinical settings there is still a need to ensure correct classification of new cases. The lack of accuracy of a classification method when classifying a particular subject may be a result of incomplete understanding of the relationships, which may exist between measured variables, it can also be due to unmeasured variables or due to noise or measurement error. A further complication is that where the causal mechanism for certain conditions is not fully understood the variables selected as being informative may not be the optimal choice for discriminating between groups.

To account for costs of misclassification, the classification of subjects into groups can be achieved using a method which seeks to minimise the expected cost of misclassification (ECM). For this explanation $\mathbf{S}_{1,}$ $\mathbf{S}_2$ and $\bar{x}_1$, $\bar{x}_2$ are the sample variance-covariance matrices and vectors of sample means associated with populations π_1 and π_2 respectively. In this method two regions R_1 and R_2 which minimise the ECM are defined as,

$$R_1\colon \left(\frac{f_1(x)}{f_2(x)}\right) \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{3.2.1.6}$$

$$R_2\colon \left(\frac{f_1(x)}{f_2(x)}\right) < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{3.2.1.7}$$

where $f_1(x)$ and $f_2(x)$ are the multivariate probability density functions for populations in groups 1 and 2 respectively (Johnson & Wichern,2008). Figure 3.2.1.1 shows how the two regions R_1 and R_2 might appear. Similarly $c(1|2)$ is a cost of misclassifying a subject from group 2 into group 1, and $c(2|1)$ is the cost of misclassification of a patient from group 1 into group 2. The probabilities p_1 and p_2 are the prior probabilities of a patient being in group 1 and 2, respectively. If one assumes that the population mean vectors μ_1, μ_2 and covariance Σ are known for two populations 1 and 2 (i.e. assume equal variance-covariance matrices.), then the discriminant analysis creates the regions R_1 and R_2 as two sets of values for the vector x. If a subject with vector x belongs to region R_1 then that patient will be classified as belonging to population 1, alternatively a subject with vector x that belongs to region R_2 will be classified as belonging to population 2. The multivariate normal densities for populations 1 and 2 may be expressed as,

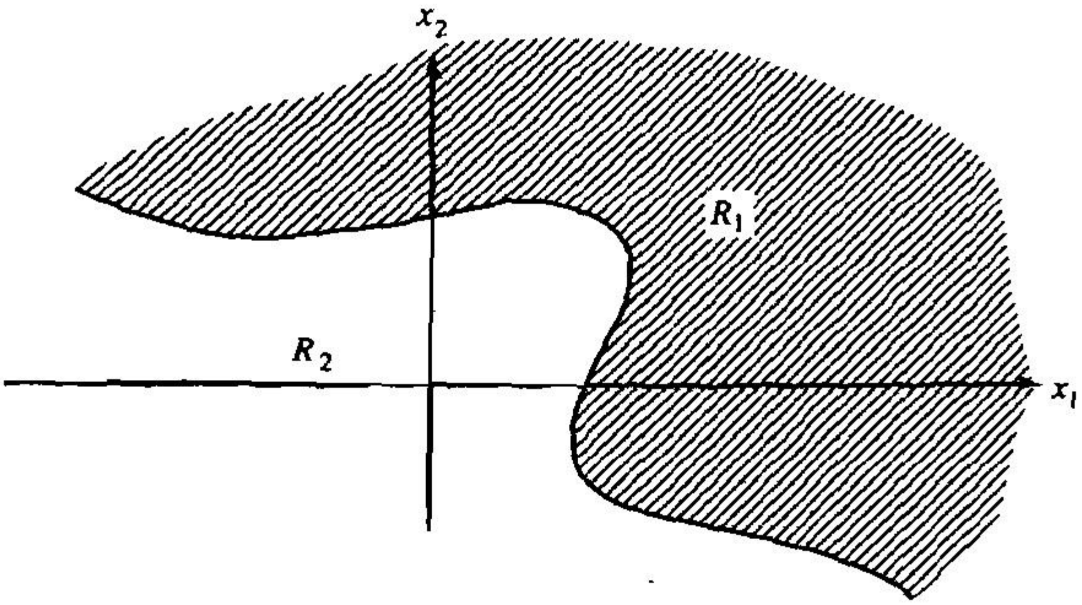
$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \tag{3.2.1.8}$$

In this equation i takes the values 1 or 2 for populations 1 or 2. The regions R_1 and R_2 are then expressed as

$$R_1\colon exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{3.2.1.9}$$

$$R_2\colon exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{3.2.1.10}$$

Figure 3.2.1.1 Regions R_1 and R_2 for two populations defined by variables X1 and X2



Source: Johnson & Wichern, Applied multivariate statistical analysis, 2008

Taking the natural logarithm of all of the terms in equations 3.2.1.9 and 3.2.1.10 we get the classification regions;

$$R_1: (\boldsymbol{x}-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \geq ln\left\{\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\} \qquad (3.2.1.11)$$

$$R_2: (\boldsymbol{x}-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) < ln\left\{\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\} \qquad (3.2.1.12)$$

where x is the vector of the subject's measurements. This rule is based on the population parameters which will not be known in most cases. Thus sample estimates of the population parameters are used instead. This rule constitutes the linear discriminant method.

LDA does assume implicitly that data are normally distributed. When testing hypotheses a violation of the normality assumption can invalidate the results of the hypothesis test. In my work LDA is not used to test hypotheses (e.g. Fischer, 1936). It is used to obtain estimates of the expected performance of variable selections (as measured by the PCC). The assumption of normality is not necessary when using LDA in this manner. However, it should be noted that because LDA does assume normality it may not be sufficiently robust to deviations from normality when used to estimate PCC.

The same LDA solution can be obtained via a so-called Mahalanobis approach (Lattin *et al*, 2003). The Mahalanobis approach identifies those points which are equidistant from the group means. These points then serve as a discriminant boundary between the two groups. A covariance-adjusted distance is calculated between each point and the mean of each of the groups. The covariance-adjusted squared distance between any point x' and group 1 is calculated according to

$$D_1^2 = (x - \boldsymbol{x}_1)'\boldsymbol{S}_p^{-1}(x - \boldsymbol{x}_1). \tag{3.2.1.11}$$

Here $\boldsymbol{S}_p$ is the pooled within-group variance-covariance matrix of the random vector $\boldsymbol{x}$.The covariance-adjusted squared distance between any point $\boldsymbol{x}'$ and group 2 is calculated according to

$$D_2^2 = (x - \boldsymbol{x}_2)'\boldsymbol{S}_p^{-1}(x - \boldsymbol{x}_2). \tag{3.2.1.12}$$

The locus of equidistant points is identified by equating Equations 3.3.11 and 3.3.12 and solving for $\boldsymbol{x}$. This gives us the discriminant boundary between our two groups.

3.2.2 Quadratic Discriminant Analysis

A natural extension of LDA (Section 3.2.1) is quadratic discriminant analysis (QDA). QDA is designed to accommodate scenarios where the variance-covariance matrices are heterogeneous. When the variance-covariance matrices are heterogeneous across groups it is not possible to simplify the multivariate probability densities as was done for LDA (Johnson & Wichern, 2008).

For a scenario where the variance-covariance matrices across the groups are not homogeneous the quadratic discriminant rule is to allocate x_0 to group 1 if

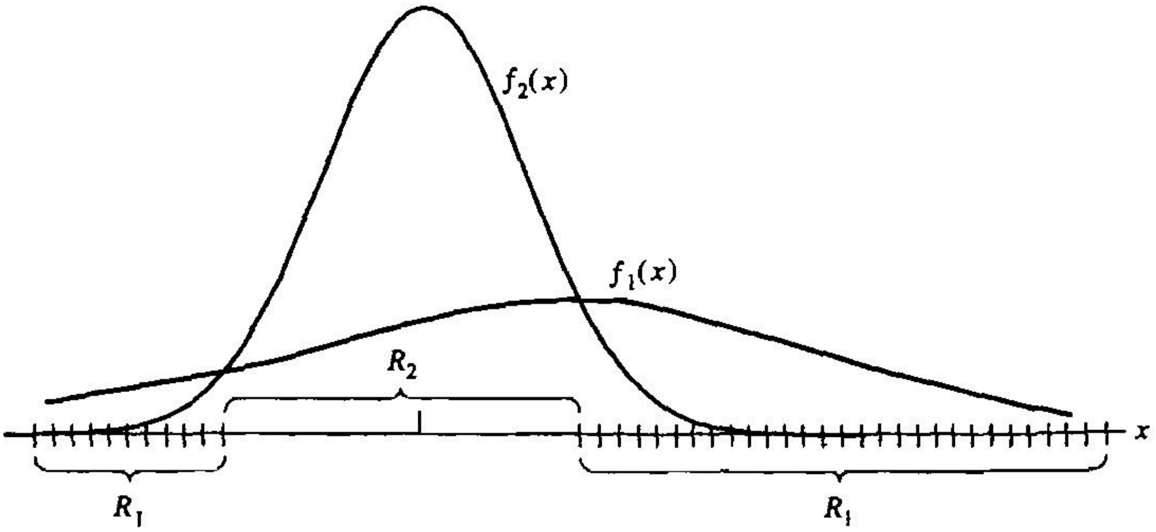
$$-\frac{1}{2}\boldsymbol{x}_0'(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x}_0 + (\bar{\boldsymbol{x}}_1'\boldsymbol{S}_1^{-1} - \bar{\boldsymbol{x}}_2'\boldsymbol{S}_2^{-1})\boldsymbol{x}_0 - k \geq ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{3.2.2.1}$$

otherwise x_0 is assigned to group 2. The term k is;

$$k = \frac{1}{2}ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\bar{\boldsymbol{\mu}}_1 - \bar{\boldsymbol{\mu}}_2'\boldsymbol{\Sigma}_2^{-1}\bar{\boldsymbol{\mu}}_2) \tag{3.2.2.2}$$

Figure 3.2.2.1 shows the rejection regions for two normally distributed samples when equal costs of misclassification and prior probabilities have been assumed. Using a quadratic classification rule results in a region R_1, which exists as two disjoint sets of points.

Figure 3.2.2.1 Regions R₁ and R₂ for tow normally distributed groups with heterogeneous variance-covariance matrices



Source: Johnson & Wichern, Applied multivariate statistical analysis, 2008

In the event that the data are not normally distributed it may be necessary to apply a normalising transformation to the data. A quadratic classification rule may be applied without transforming the data. However the greater the departure of the data from a normal distribution the less reliable the performance of a quadratic classification rule.

When different variance-covariance matrices are substituted into the multivariate probability densities (Section 3.2.1) the classification regions R_1 and R_2 are expressed as;

$$R_1: -\frac{1}{2}\mu'(\Sigma_1^{-1} - \Sigma_2^{-1})\mu + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\mu - k \geq ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (3.2.2.3)$$

$$R_2: -\frac{1}{2}\mu'(\Sigma_1^{-1} - \Sigma_2^{-1})\mu + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\mu - k < ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (3.2.2.4)$$

where $\boldsymbol{\Sigma}_1$ is the variance-covariance matrix of group 1, $\boldsymbol{\Sigma}_2$ is the variance-covariance matrix of group 2, $\bar{\mu}_1$ and μ_2 are the mean values associated with populations 1 and 2 respectively.

Equations 3.2.2.2 and 3.2.2.2 defining regions R_1 and R_2 when the variance-covariance matrices are non-homogeneous are very different to equations 3.2.1.9 and 3.2.1.10 defining the same regions when the variance-covariance matrices are homogeneous. The reason for this is that the classification regions are defined by the quadratic term $-\frac{1}{2}\boldsymbol{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}$ and when the variance-covariance regions are non-homogeneous this term cannot be cancelled out. This leaves us with more complex equations for the classification regions.

The quadratic classification rule assigns a new case with measurements x_0 to population 1 if

$$-\frac{1}{2}\boldsymbol{x}_0'(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x}_0 + (\overline{\boldsymbol{x}}_1'\boldsymbol{S}_1^{-1} - \overline{\boldsymbol{x}}_2'\boldsymbol{S}_2^{-1})\boldsymbol{x}_0 - k \geq ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{3.2.2.5}$$

otherwise the new patient with measurements $\boldsymbol{x}_0$ is assigned to population. In this equation the population variance-covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ and mean values $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are replaced with the sample population variance-covariance matrices S_1 and $\boldsymbol{S}_2$ and mean values $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

3.3 Discriminatory performance of a classifier

Whether LDA or QDA are used the performance of the trained classifier can be assessed using metrics such as the probability of correct classification (PCC), the area under the ROC curve (AUROC), positive predictive value (PPV) and negative predictive value (NPV).

The general formulae for PCC is

$$PCC \;=\; Prob(\,assign\ to\ Group\ 1\,|\,xi\ is\ from\ Group\ 2)\ x\ Prob(\,is\ from\ Group\ 2)$$

$$+\,Prob(assign\ to\ Group\ 2\,|\,xi\ is\ from\ Group\ 1)\ x\ Prob(\,is\ from\ Group\ 1)\ \text{(3.3.1)}$$

LDA achieves the highest PCC when prior probabilities are equal to the proportions of the two samples i.e. $P(Group\ 1) = n_1/n\ and\ P(Group\ 2) = n_2/n.$ This assumes that costs of misclassification are the same (Cox, 2005).

Furthermore, when using LDA and assuming equal costs of misclassification and normality of the data, the actual PCC for classification of patients to one of two groups has been shown to be equal to the formula

$$PCC = \phi\left(\frac{\Delta}{2}\right) \tag{3.3.2}$$

where ϕ is the cumulative distribution function (CDF) of the standard univariate normal distribution and Δ is the square root of the Mahalanobis distance (Section 3.2.1) between the two groups (Dunn & Varady, 1966). Equation 3.3.2 is very important, as it gives the connection between the PCC, LDA and Hotelling's T^2 statistic.

In the work presented in this thesis the PCC is estimated by summing the number of correctly classified cases and dividing this by the total number of cases to be classified. If we have a contingency table such as

Table 3.3.1 Outline of a contingency table

		Truth	
		Group 1 (e.g. disease)	Group 2 (e.g. no disease)
Assignment (results of classification)	Group 1	True Positive (TP)	False Positive (FP)
	Group 2	False Negative (FN)	True Negative (TN)

The PCC can be calculated according to:

$$PCC = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.3.3}$$

Knowledge of the true group assignments and those made by some classifier can be used to plot a receiver's operating characteristic curve (ROC). Using data presented as in Table 3.3.1 above the sensitivity and specificity may be calculated. Sensitivity (also known as the true positive rate) is calculated according to

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP+FN}\,. \tag{3.3.4}$$

Here P stands for number of positives. Similarly the specificity (also known as the true negative rate) is calculated according to

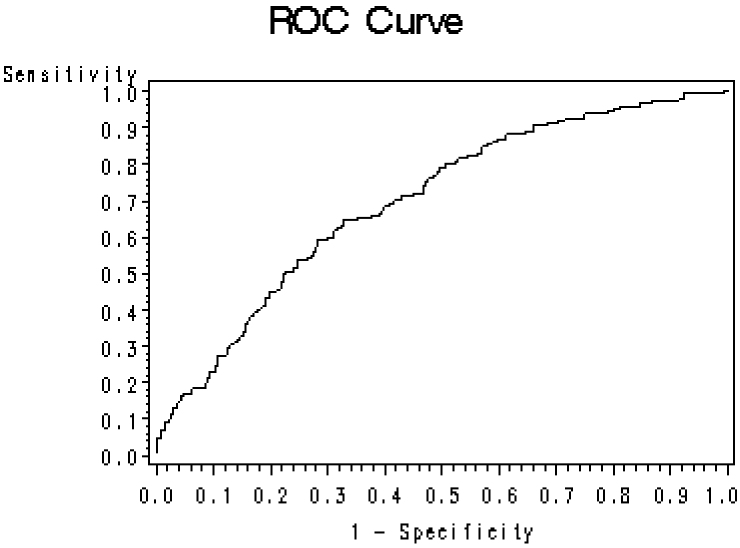
$$Specificity = \frac{TN}{N} = \frac{TN}{TN+FP} \tag{3.3.5}$$

Here N stands for number of negatives.

The sensitivity and specificity values tell us how likely a given classifier is to correctly identify positive or negative cases, respectively. The ROC curve can be constructed by plotting sensitivity on the y-axis against (1 − specificity) which is also known as the false positive rate on the x-axis, i.e., the true positive rate is plotted against the false positive rate for a range of thresholds. Figure 3.3.1 shows an example of an ROC curve. When discriminating between two groups a threshold is selected to determine membership of one group or the other. This threshold is what separates the two regions

R_1 and R_2 when using LDA or QDA. If we vary the threshold separating the two groups the sensitivity and specificity will change accordingly. It is this change in sensitivity and specificity values that gives each ROC curve its particular shape. The closer the curve is to the upper left corner the larger the area under the curve.

Figure 3.3.1 An example of an ROC curve



Source: Hosmer & Lemeshow, Applied logistic regression, 2001

PPV and NPV are measures of the proportion of positive and negative results that are actually true positives and true negatives, respectively. PPV is calculated according to the formula

$$PPV = \frac{TP}{TP+FP} \tag{3.3.6}$$

While NPV is calculated according to

$$NPV = \frac{TN}{TN+FN} \tag{3.3.7}$$

True positive (negative) occurs when a patient is identified as positive (negative) for some condition and they are truly positive (negative) for this condition (where the ‘true’ status is determined using a suitable test or gold standard accepted for the condition in question). False positive (negative) occurs when a patient is identified as positive (negative) for some condition but they are not positive (negative) for this condition. The total number of positives or negatives includes both true and false positive or true and false negative test results.

LDA and QDA both assume that data are normally distributed. LDA assumes homogeneity of variance-covariance matrices across groups which makes it unsuitable for situations where homogeneity does not apply. QDA does not assume homogeneity of variance-covariance matrices which means it can accommodate heterogeneous variance-covariance matrices across groups. However QDA is less robust to deviations from normality than LDA (Johnson and Wichern, 2008). In a scenario where variance-covariance matrices are heterogeneous across groups and data are known to be normally distributed QDA is the best choice for carrying out discriminant analysis. However if data deviate considerably from a normal distribution LDA may be the better option (Johnson and Wichern, 2008).

3.4 Hotelling's T² statistic

Linear discriminant analysis (Section 3.2.1) is linked with a statistic called Hotelling's T^2 (Hotelling, 1931) which is a multivariate extension of the Student's t-statistic. In this section I will introduce Hotelling's T^2 statistic and explain the link to the Student's t-statistic.

Let us start with a **univariate scenario** and assume X is a random variable with mean μ and variance σ^2. Also assume we have a random sample of values of X summarised in the sample mean $\bar{x}$ and let n and s be the sample size and standard deviation respectively. Then the univariate t-statistic;

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \qquad (3.4.1)$$

can be used to compare the mean value μ of a variable to some hypothetical mean μ_0. This statistic can then be squared giving us

$$t^2 = \frac{(\bar{x}-\mu_0)^2}{s^2/n}. \qquad (3.4.2)$$

This equation can be re–written as

$$t^2 = n(\bar{x}-\mu_0)(s^2)^{-1}(\bar{x}-\mu_0) \qquad (3.4.3)$$

Next, let us assume a **multivariate scenario**. Assume that $\boldsymbol{X}$ is a random vector, with mean vector μ, covariance matrix $\boldsymbol{\Sigma}$ and dimension p. We then replace the difference between the sample mean $\bar{x}$ and our hypothetical mean μ_0 in the equation above with the sample mean and hypothetical mean vectors $\bar{\boldsymbol{x}}$ and $\boldsymbol{\mu_0}$. We also replace the inverse of the sample variance s^2 with the variance-covariance matrix $\boldsymbol{S}$ to produce Hotelling's T^2 statistic.

$$T^2 = n(\bar{\boldsymbol{x}}-\boldsymbol{\mu_0})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu_0}) \qquad (3.4.4)$$

For a sufficiently large n Hotelling's T^2 statistic is approximately chi-square distributed with p degrees of freedom (Hotelling, 1931). If the sample variance-covariance matrix $\boldsymbol{S}$ is replaced with the population variance-covariance matrix $\boldsymbol{\Sigma}$ then Hotelling's T^2 statistic is chi-square distributed with p degrees of freedom for any sample size (Hotelling, 1931).

The Hotelling T^2 statistic (Eq 3.4.1) is applicable to scenarios where we have just one sample or one group. Obviously in many cases we have at least 2 groups and we are interested in testing hypotheses which compare the samples that comprise both of these groups. To that end the statistic presented above is modified by replacing the hypothetical mean with the mean of the second group (Hotelling, 1931). One of the assumptions we make at this stage is that the two groups have the same variance-covariance matrices which we estimate using the pooled variance-covariance matrix. This pooled variance-covariance matrix $\boldsymbol{S_p}$ is calculated according to

$$\boldsymbol{S_p} = \frac{(n_1-1)\boldsymbol{S_1}+(n_2-1)\boldsymbol{S_2}}{n_1+n_2-2} \tag{3.4.5}$$

where $\boldsymbol{S_1}$and $\boldsymbol{S_2}$are the variance-covariance matrices of the two groups of interest and n_1, n_2 are the sample sizes of groups 1 and 2 respectively. The two-sample Hotelling's T^2 statistic is:

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\boldsymbol{\overline{x}_1} - \boldsymbol{\overline{x}_2})'\boldsymbol{S}_p^{-1}(\boldsymbol{\overline{x}_1} - \boldsymbol{\overline{x}_2}). \tag{3.4.6}$$

When using Hotelling's T^2 statistic for statistical inference we assume that the two groups of interest are normally distributed, that the subjects from each population of interest were sampled independently and that the two groups of interest have the same variance-covariance matrix.

There is a direct relationship between Hotelling's T^2 statistic and the linear discrimination method of Fisher (1936) that we described in Section 3.2.1. In linear discriminant analysis, under the assumption of priors being equal to the proportions of the sample, that data are normally distributed, and equal costs of misclassification the PCC for classification of patients can be proven (Anderson, 1951) to be equal to

$$PCC = \phi\left(\frac{\Delta}{2}\right), \tag{3.4.7}$$

where ϕ is the CDF of the standard univariate normal distribution and Δ is the square root of the Mahalanobis distance between the two groups (Dunn & Varady, 1966)

$$D = (\boldsymbol{\overline{x}}_1 - \boldsymbol{\overline{x}}_2)'\boldsymbol{S}_p^{-1}(\boldsymbol{\overline{x}}_1 - \boldsymbol{\overline{x}}_2). \tag{3.4.8}$$

From equations 3.4.3 and 3.4.5 we see that Hotelling's T^2 statistic is equal to the Mahalanobis distance up to a constant

$$T^2 = \frac{n_1 n_2}{n_1+n_2} D \,. \tag{3.4.9}$$

Moreover, from Eq. 3.4.4 it follows

$$PCC = \phi\left(\left(\frac{n_1+n_2}{n_1 n_2}\right)^{1/2} \frac{T}{2}\right) \tag{3.4.10}$$

which is an important identity showing the connection of Hotelling's T^2 statistic with the PCC in linear discriminant analysis for a multivariate case, i.e. when X is a random vector. Equality 3.4.5. holds only if data are normally distributed. It means that as the value of Hotelling's T^2 statistic increases we should also observe an increase in the associated PCC value, and vice-versa

3.5 Proposed SNR statistic

In this section, I develop an extension of the Hotelling's T^2 statistic from Equation 3.2.3. This extension is inspired by the concept of signal-to-noise ratio (SNR) from information theory. The SNR concept is old and is defined as the amount of signal divided by the amount of noise within a given system (Berube & Wu 2000; Kaiser *et al,* 1998). As I described in Section 2.6 there are multiple versions of the signal-to-noise ratio, each uses a different definition of how to measure the amount of signal and noise. Examples of the application of the SNR include its use in researching how background noise effects speech understanding (Maamor & Billings, 2016), identifying the minimal SNR which facilitates stabilization of single-input, single-output, linear time-invariant systems (Jiang *et al*, 2016), the comparison of magnetic resonance imaging (MRI) images obtained using different hardware and protocols (Welvaert & Rosseel, 2013), and extension to generalised linear models for general point processes in neuroscience (Czanner *et al*, 2015). In this work I present an extension of the SNR concept which can be used to evaluate the discriminatory potential of random variables to discriminate between groups.

The SNR is typically defined as

$$SNR = \frac{amount\ of\ signal}{amount\ of\ noise},$$

where signal and noise are usually estimated via the minimisation such as via least-squares principle.

In our scenario of discrimination between two groups, **the signal is the group membership** and the goal is to find how much information about this signal is contained in the realisations of the random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ i.e. in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This can be estimated via mean differences such a $\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}}$ and $\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}}$, where $\overline{\overline{\boldsymbol{x}}}$ is a vector of sample means across both groups, $\overline{\boldsymbol{x}}_1$ is a vector of sample means in group 1 and $\overline{\boldsymbol{x}}_2$ is a vector of sample means in group 2.

In a **univariate scenario**, the SNR can be defined as follows

$$SNR = \frac{\sum_{Group\ 1} (\bar{x}_1 - \bar{\bar{x}})^2 + \sum_{Group\ 2} (\bar{x}_2 - \bar{\bar{x}})^2}{\sum_{Group\ 1} (\bar{x}_1 - x_i)^2 + \sum_{Group\ 2} (\bar{x}_2 - x_i)^2} \tag{3.5.1}$$

where the denominator is essentially the pooled sample variance

$$s_p^2 = \frac{\sum_{Group\ 1} (\bar{x}_1 - x_i)^2 + \sum_{Group\ 2} (\bar{x}_2 - x_i)^2}{n_1 + n_2 - 2} \tag{3.5.2}$$

This last equation assumes same variances in both groups and a univariate scenario.

This motivates the following proposed extension of **the SNR for a multivariate heteroscedastic scenario** which will be used in the work presented in later chapters:

$$SNR = (\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}})' \boldsymbol{S}_1^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}}) n_1 + (\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}})' \boldsymbol{S}_2^{-1} (\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}}) n_2 \tag{3.5.3}$$

where 1 and 2 refer to two hypothetical groups 1 and 2, $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the sample mean vectors for groups 1 and 2 respectively, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are estimates of the variance-covariance matrices of groups 1 and 2 respectively, the terms n_1 and n_2 are the sample sizes of the groups 1 and 2, and $\overline{\overline{\boldsymbol{x}}}$ represents the overall sample mean vector across groups for all the random variables:

$$\overline{\overline{\boldsymbol{x}}} = \frac{n_1 \overline{\boldsymbol{x}}_1 + n_2 \overline{\boldsymbol{x}}_2}{n_1 + n_2}. \tag{3.5.4}$$

The interpretation of the SNR (Equation 3.5.1) ratio is intuitive in that the larger the mean differences from the overall mean vector and the smaller the spread of the observations in each of the two groups, the larger the SNR and therefore the better the discriminatory performance we can expect from the given variable or set of variables. Conversely the smaller the SNR the poorer the discriminatory performance we can expect from the given variable or set of variables.

In general a higher spread around the group mean vector the less well is the group defined. Where this leads to overlap between two groups there will be a loss in discriminatory performance. This will be reflected in a lower SNR. However if a sufficiently large separation exists between two groups this is represented by high mean difference standardised with the covariance matrix. It is the combination of the variances, co-variances and mean difference values that determines the SNR for a particular set of variables.

Each term in the SNR is multiplied by the size of the relevant group. This corrects the SNR for any imbalances between group sizes by giving the smaller group a lower weight and giving the larger group a larger weight.

Under the condition of homogeneous variance-covariance matrices (hence the common variance-covariance matrix is estimated by a pooled variance-covariance matrix across groups) it will be demonstrated in Section 3.6 that the SNR is equivalent to Hotelling's T^2 statistic.

3.6 Proof of equivalency of Hotelling's T² statistic and SNR when variance-covariance matrices are homogeneous

Here I demonstrate analytically how SNR and T^2 are related. Assume we have 2 groups 1 and 2 of unequal sizes such that $n_1 \neq n_2$ and $n_1 + n_2 = n_T$, n_T being the total sample size.

First, let us assume **a univariate homoscedastic scenario** i.e. let *X* be a univariate random variable with equal variances in both groups. The SNR from Equation 3.5.1 can be written as

$$SNR = \frac{(\bar{x}_1 - \bar{\bar{x}})^2}{s_1^2} n_1 + \frac{(\bar{x}_2 - \bar{\bar{x}})^2}{s_2^2} n_2 \qquad (3.6.1)$$

where $\bar{x}_1$ is the sample mean for group 1, $\bar{x}_2$ is the sample mean for group 2 and $\bar{\bar{x}}$ is the overall average calculated across both groups according to

$$\bar{\bar{x}} = \frac{1}{n_T}(n_1\bar{x}_1 + n_2\bar{x}_2) \qquad (3.6.2)$$

The terms s_1^2 and s_2^2 are sample variances. In the assumed scenario of equal variances they can be replaced by s_P^2 the pooled or common sample variance:

$$s_P^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}, \qquad (3.6.3)$$

where k is the total number of groups. Then SNR (Eq. 3.6.1) now becomes

$$SNR = \frac{(\bar{x}_1 - \bar{\bar{x}})^2}{s_P^2} n_1 + \frac{(\bar{x}_2 - \bar{\bar{x}})^2}{s_P^2} n_2 \quad . \qquad (3.6.4)$$

Substituting in the formula the overall mean $\bar{\bar{x}}$ with $\bar{x}_1$ and $\bar{x}_2$ as appropriate we have:

$$SNR = \frac{\left(\bar{x}_1-\frac{n_1}{n_T}\bar{x}_1-\frac{n_2}{n_T}\bar{x}_2\right)^2}{s_p^2} n_1 + \frac{\left(\bar{x}_2-\frac{n_1}{n_T}\bar{x}_1-\frac{n_2}{n_T}\bar{x}_2\right)^2}{s_p^2} n_2 \qquad (3.6.5)$$

which expands to

$$SNR = \frac{n_1}{s_p^2}\left(\frac{n_2}{n_T}\bar{x}_1 - \frac{n_2}{n_T}\bar{x}_2\right)^2 + \frac{n_2}{s_p^2}\left(\frac{n_1}{n_T}\bar{x}_2 - \frac{n_1}{n_T}\bar{x}_1\right)^2 \qquad (3.6.6)$$

$$SNR = \frac{n_1 n_2^2}{s_p^2 n_T^2}(\bar{x}_1 - \bar{x}_2)^2 + \frac{n_2 n_1^2}{s_p^2 n_T^2}(\bar{x}_1 - \bar{x}_2)^2$$

After simplifying we are left with

$$SNR = \frac{(\bar{x}_1-\bar{x}_2)^2}{s_p^2}\frac{(n_1+n_2)n_1n_2}{(n_T)^2}. \qquad (3.6.7)$$

On the other hand, the Mahalanobis distance D is calculated as

$$D \quad = \frac{(\bar{x}_1-\bar{x}_2)^2}{S_p^2} \qquad (3.6.8)$$

hence from Eq 3.6.7, 3.6.8, 3.4.3 the SNR can be expressed as follows

$$SNR = D \quad \frac{n_1n_2}{n_T} = T^2. \qquad (3.6.9)$$

Therefore we can conclude that in a univariate case and under the condition of equal variances across groups the SNR is equivalent to the Hotelling's T^2 statistic.

Now, let us assume **a multivariate homoscedastic scenario** i.e. let ***X*** be a multivariate vector of dimension *px1* and let the variance-covariance matrices differ across groups. The SNR (Equation 3.5.1) is defined as

$$SNR = (\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}})'\boldsymbol{S}_1^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}})n_1 + (\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}})'\boldsymbol{S}_2^{-1}(\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}})n_2 \qquad (3.6.10)$$

where $\overline{\boldsymbol{x}}_1$ is the vector of sample means for group 1, $\overline{\boldsymbol{x}}_2$ is the vector of sample means for group 2 and $\overline{\overline{\boldsymbol{x}}}$ is the vector of overall means calculated across both groups according to

$$\overline{\overline{\boldsymbol{x}}} = \frac{1}{n_T}(n_1\overline{\boldsymbol{x}}_1 + n_2\overline{\boldsymbol{x}}_2). \qquad (3.6.11)$$

The matrices $\boldsymbol{S}_1^2$ and $\boldsymbol{S}_1^2$ are estimates of the population variance-covariance matrices (based on samples taken from the population) for groups 1 and 2. Let $\boldsymbol{S}_p^2$ be the pooled covariance matrix:

$$\boldsymbol{S}_P = \frac{n_1\boldsymbol{S}_1 + n_2\boldsymbol{S}_2}{n_T - 2} \tag{3.6.12}$$

Then the SNR (Eq. 3.6.11) now becomes

$$SNR = (\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}})' S_P^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\overline{\boldsymbol{x}}}) n_1 + (\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}})' S_P^{-1} (\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}}) n_2 \tag{3.6.13}$$

Substituting in the formula the overall mean $\overline{\boldsymbol{x}}_2 - \overline{\overline{\boldsymbol{x}}}$ with $\bar{X}_1$ and $\bar{X}_2$ as appropriate we have:

$$SNR = \left[\overline{\boldsymbol{x}}_1 - \left(\frac{n_1\overline{\boldsymbol{x}}_1 + n_2\overline{\boldsymbol{x}}_2}{n_T}\right)\right]' S_P^{-1} \left[\overline{\boldsymbol{x}}_1 - \left(\frac{n_1\overline{\boldsymbol{x}}_1 + n_2\overline{\boldsymbol{x}}_2}{n_T}\right)\right] n_1$$

$$+ \left[\overline{\boldsymbol{x}}_2 - \left(\frac{n_1\overline{\boldsymbol{x}}_1 + n_2\overline{\boldsymbol{x}}_2}{n_T}\right)\right]' S_P^{-1} \left[\overline{\boldsymbol{x}}_2 - \left(\frac{n_1\overline{\boldsymbol{x}}_1 + n_2\overline{\boldsymbol{x}}_2}{n_T}\right)\right] n_2 \tag{3.6.14}$$

which expands to

$$SNR = \left[\left(\frac{n_2\overline{\boldsymbol{x}}_1 - n_2\overline{\boldsymbol{x}}_2}{n_T}\right)\right]' S_P^{-1} \left[\left(\frac{n_2\overline{\boldsymbol{x}}_1 - n_2\overline{\boldsymbol{x}}_2}{n_T}\right)\right] n_1 + \left[\left(\frac{n_1\overline{\boldsymbol{x}}_2 - n_1\overline{\boldsymbol{x}}_1}{n_T}\right)\right]' S_P^{-1} \left[\left(\frac{n_1\overline{\boldsymbol{x}}_2 - n_1\overline{\boldsymbol{x}}_1}{n_T}\right)\right] n_2 \tag{3.6.15}$$

We can simplify this to get

$$SNR = \frac{n_2^2 n_1}{n_T^2} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}_p^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) + \frac{n_1^2 n_2}{n_T^2} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}_p^{-1} (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1) \tag{3.6.16}$$

which becomes

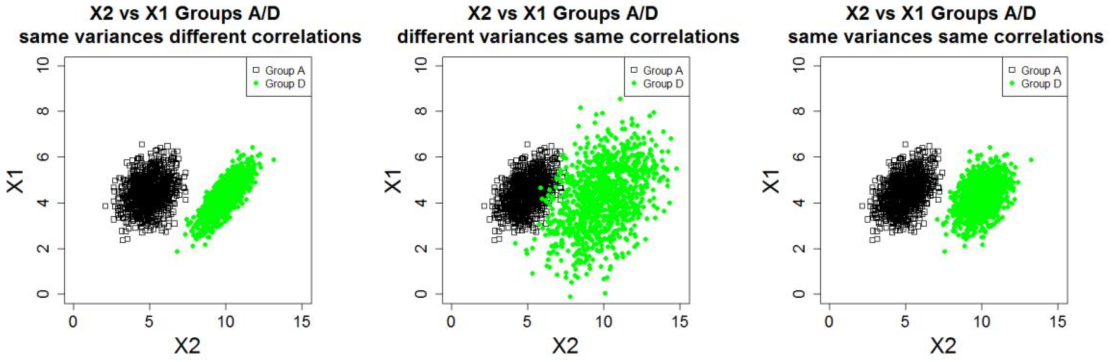
$$SNR = \frac{n_1 n_2}{n_T} \left[\frac{n_2}{n_T} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}_P^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) + \frac{n_1}{n_T} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}_P^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)\right] \tag{3.6.17}$$

Comparing this to the formula for Hotelling's T^2 statistic (Eqn. 3.4.3) it is evident that the two are identical. Thus the equivalency of the SNR and Hotelling's T^2 statistic is proven in both univariate and multivariate contexts when the variance-covariance matrices are homogeneous across the outcome groups.

It should be noted that while we do not assume equality of group variances this does allow for a scenario where the variances are equal. In a scenario where the variances are equal (i.e. common) the SNR calculated is the actual SNR. However in a scenario where the variances are not equal (so that we are using the pooled estimate) then the SNR calculated is also an estimate of the true value.

To give further intuition into how Hotelling's T^2 statistic and the SNR work I simulated several scenarios. Figure 3.6.1 below presents bivariate plots for two groups where the correlations and variances between the two variables are different across the groups. Table 3.6.1 shows the SNR values and Hotelling's T^2 statistics associated with the set containing variables X_1 and X_2 under each of the correlation and variance scenarios. From the bivariate plots in Figure 3.6.1 we see that changes in the correlation and variance parameters have an effect on the dispersion of the data points. From Table 3.6.1 it is also evident that the SNR values and Hotelling's T^2 statistics change in response to the correlation and variance parameters. Note that the SNR values are larger than the Hotelling's T^2 statistics in the scenarios with different variances or correlations. Conversely when the variance and correlation properties are homogeneous across groups both metrics are very similar (the difference between them is less than 1 %). This is expected given the theoretical equivalence of the two metrics under the condition of variance-covariance matrix homogeneity across groups. Hotelling's T^2 statistic appears relatively unresponsive (when compared to SNR) to changes in the correlation parameter while both metrics are affected by changes in the variance parameter.

Figure 3.6.1 Bivariate plots of X1 vs. X2 across simulated groups under three different scenarios



Legend: (left) same variances and different correlations, (middle) different variances and same correlation; and (right) same variances and same correlations.

Table 3.6.1 SNR values, Hotelling's T^2 statistics, means, variances and correlations for the variable set containing X1 and X2 for each of the variance and correlation scenarios.

Scenario	Hotelling's T^2 statistic	SNR	Mean (x_{1A},X_{1D})	Mean (x_{2A},X_{2D})	Variance (x_{1A},X_{1D})	Variance (x_{2A},X_{2D})	Correlation $(X_{1A}\text{-}X_{2A}, X_{1D}\text{-}X_{2D})$
Same variance different correlations	23,271.4	33,320.8	4.5, 4.3	5, 10	0.5, 0.5	0.75, 0.75	0.2, 0.8
Same correlations different variances	8,196.9	12,538.9	4.5, 4.3	5, 10	0.5, 2	0.75, 3	0.4, 0.4
Same variances and correlations	20,359.8	20,370.3	4.5, 4.3	5, 10	0.5, 0.5	0.75, 0.75	0.4, 0.4

3.7 A variable selection algorithm that is based on Hotelling's T^2

Here I describe how Hotelling's T^2 has been used previously for variable selection in the context of classification. In the paper entitled "Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays" Lu *et al* (2005) outline a method for identifying groups of differentially expressed genes (DEGs) from microarray data. The method outlined in this paper utilises Hotelling's T^2 statistic as a multivariate discrimination index implemented as part of a multiple forward selection algorithm. This algorithm serves as the inspiration for the MFS algorithm that I will propose and describe in Section 3.8.

Lu's method assumes a classification for two groups. They utilise Hotelling's T^2 statistic as an index of the discriminatory potential of a given gene or set of genes. They use Hotelling T^2 statistic in an algorithmic procedure starting with an empty set of discriminatory variables and all genes in the set of potential selection candidates. First Hotelling's T^2 statistic is calculated for each of the genes in the dataset, individually. The gene which is associated with the largest Hotelling's T^2 statistic is then identified and added to the discriminatory set. The same calculations are then carried out using the selected gene(s) and each of the genes which are still selection candidates. The gene which elicits the largest increase in the Hotelling's T^2 statistic when analysed as part of a subset including the previously selected genes is added to the selection in each case. The algorithm continues selecting genes in this way until one of the stopping criteria are satisfied.

The algorithm works as follows (Lu et al., 2005):

Step 1. T^2 statistics are calculated for each of the genes in the dataset and the gene j_1 with the largest T^2 statistic is identified. This T^2 statistic is denoted T_{j1}. j_1 is added to the set S.

Step 2. The p-value associated with T_{j1} is compared to the predefined significance level α. If this p-value is less than α, the T^2 statistic is calculated for 2 genes; j_1 and one of the remaining unselected genes. The largest T^2 statistic $T_{j1,j2}$ is identified and the gene $j_2$added to the set S.

Step 3. The p-values associated with T_{j1} and $T_{j1,j2}$ are compared. If the p-value for $T_{j1,j2}$ is less than the p-value for T_{j1} step 2 is repeated for $T_{j1,j2}$ and $T_{j1.j2,j3}$.

Step 4. Repeat step 3 until the p-value for $T_{\mathrm{j1,\dots,jn-1,jn}}$ is larger than the p-value for $T_{\mathrm{j1,\dots,jn-1}}$ or until the number of genes in S is larger than $n_1 + n_2 - 2$ ($n_1$and n_2 being the sizes of groups 1 and 2).

Step 5. Remove genes $\mathrm{j}_1,\dots,\mathrm{j}_{\mathrm{n-1}}$ from the set of candidate genes and repeat steps 1-4..

Two stopping criteria are employed by Lu et al. One criterion considers the number of genes that have been selected and terminates the algorithm if this number is larger than $n_1 + n_2 - 2$ (here n_1 and n_2 are the sample sizes of the two groups of interest). The second criterion assesses the significance of each set of DEGs using p-value. During each round of selection the p-value associated with Hotelling's T^2 statistic of the current set of variables is calculated. Before the next round of selection begins, this p-value is compared to the predefined significance level α. If the p-value is less than this significance level the algorithm continues and another round of selection takes place. However if the p-value is larger than this significance level the algorithm terminates. In subsequent rounds of selection the p-values are compared to those calculated in previous rounds of selection i.e. selection only continues if the p-value shrinks or remains the same. In this way the authors are essentially reasoning that the selection is "more" significant as variables are added and the p-value decreases. Any genes which have been selected up to the point of termination are designated as a group of DEGs.

Their method is validated using the spike-in HGU95 dataset produced by Affymetrix and gene expression data from the work of Chen et al. (2002) who analysed gene expression patterns in human liver cancers. From the analysis of the Affymetrix datasets Lu et al. demonstrated that their Hotelling's T^2 method produced fewer false positives when compared to selections made using a method based on t-tests to identify DEGs. For selections made from the data of Chen et al. the Hotelling's T^2 method identified more DEGs in pathways related to hepatocellular carcinoma (HCC) and more DEGs with second-degree associations to HCC (i.e. found in the same pathway(s) as genes directly linked to HCC). In addition, the Hotelling's T^2 method identified several novel cancer-associated genes that were highly expressed in the HCC tumours analysed by Chen et al.

The Hotelling's T^2 statistic described in the work of Lu et al. assumes homogeneity of variance-covariance matrices across groups. This assumption is not a valid assumption particularly in a study analysing gene expression. In any effort to identify genes which can differentiate between two groups the objective is to identify genes which have different expression levels in the groups of interest. Thus we are actively seeking to identify genes which have different expression levels and therefore different variance-covariance properties across the groups of interest. This point can be extended to other kinds of data. The article of Lu et al. states that a pooled variance-covariance matrix is used in the calculation of Hotelling's T^2 statistic. The assumption of homogeneity of the variance-covariance matrices across groups is integral to the calculation of the pooled variance-covariance matrix. As the authors have not addressed this in their article I must conclude that the assumption is implicitly maintained. This raises a question as to the suitability of Hotelling's T^2 statistic as a metric of discriminatory potential both in a general sense and in the specific context of the work presented in this article. I have identified this as a problem which I will address in my thesis.

3.8 Proposed variable selection algorithm that is based on SNR

Here I describe how I propose to use the SNR metric for variable selection in classification. In order to apply the SNR to the task of variable selection a suitable algorithmic framework is required. The idea is that the algorithm can use the SNR to assess the potential of different subsets of variables for discriminating between groups and ultimately identify the optimal subset for this task. For this purpose I proposed and implemented a multiple forward selection (MFS) algorithm as an extension on the work of Lu et al (2005) by utilising the SNR (Section 3.5).

The proposed algorithm works as follows:

Step 1. The algorithm starts with an empty set of selected discriminatory variables.

Step 2. The SNR is calculated for every variable in the dataset using Eq. 3.5.1. The variable with the highest SNR is found and it is added to the set of discriminatory variables provided its PCC (calculated using QDA with cross-validation – equation 3.3.3) is above the minimum specified threshold. This dataset is then called S_1 and its PCC is called $PCC_{1.}$ We assume we have selected j variables in the set, S_j, of selected discriminatory variables via comparison of SNR values. Let PCC_j be the overall PCC of the set S_j.

Step 3. We consider each non-selected variable X and we calculate the SNR of the following set: $\{S_j, X\}$.

Step 4. We find the variable X^* that gives the largest SNR in Step 3. Then we calculate the PCC for the $\{S_j, X^*\}$ and call it PCC_{j+1}. If the difference between PCC_{j+1} and PCC_j is greater than or equal to the specified threshold then we update the set of selected discriminatory variables by including the variable X^* in it and call it S_{j+1} and its PCC as PCC_{j+1}, otherwise we stop the algorithm. If the difference between PCC_{j+1} and PCC_j is greater than the specified threshold we go back to Step 3 and Step 4 (where j becomes j+1).

I implemented this algorithm in the R program. The full details are in the Appendix.

3.9 Computational challenges

In order to ensure that a variable selection algorithm is effective at carrying out the task of variable selection several computational challenges had to be addressed: a suitable stopping criterion as well as an appropriate method for estimating the discriminatory performance of selected variables are required. It was also necessary to identify the optimal number of simulations to run when assessing the performance of the algorithm and comparing it to other methods. These issues are addressed in Sections 3.9.1, 3.9.2 and 3.9.3 below.

3.9.1 Stopping criteria

The stopping criterion is an important element of each variable selection algorithm. In the original formulation of the MFS algorithm (Lu et al., 2005) each time a variable (say X) is considered for discrimination a p-value is calculated to compare the current subset of discriminatory variables (say S) with the new subset of discriminatory variables containing the new variable $\{S, X\}$ (Lu et al., 2005). In the context of this method a p-value greater than the chosen significance level is considered relevant to discriminating between the outcome groups. If the selection of variable X is no longer considered significant at a significance level of 0.05 (i.e. p-value of S is < p-value of $\{S, X\}$) then the algorithm terminates and returns the subset of selected variables chosen up to that point (i.e. S).

I have studied the use of p-value as a stopping criterion in my MFS-T2 algorithm. I observed, at fixed values of mean vectors and covariance matrices, that as more variables capable of discriminating between the groups of interest are selected the p-value tends to become smaller. When a variable with no discriminatory potential is selected there will be no change in the p-value associated with this selection (when compared to the p-value from the previous round of selection) and so the algorithm terminates. The assumption is that all discriminating variables have been selected at this point and it is appropriate for the algorithm to terminate and output the identities of the selected variables. However, analysis of the Hotelling's T^2 statistics calculated reveal that this termination is premature and further rounds of variable selection are necessary (Table 3.9.3.3).

In an effort to evaluate the use of the p-value as a stopping criterion a dataset of 20 variables was simulated. The variables were assumed to be normally distributed and no correlations were specified between any variables. The dataset had the following discrimination structure:

- $X_1 \dots X_{10}$ are all discriminating variables with decreasing discriminatory potential going from X_1 to X_{10},
- $X_{11} \dots X_{20}$ are all non-discriminating variables.

The p-values, Hotelling T^2 statistics and PCC estimates were then calculated for expanding subsets of variables. The results for a subset of the variables are presented in table 3.7.3.3 below. From the table we can see that a p-value $<10^{-6}$ is calculated for the first variable. As more variables are added to the selection the p-values, Hotelling T^2 statistics and PCC estimates were re-calculated. While the Hotelling T^2 statistics and PCC estimates continued to increase with each variable addition the p-value remained at $<10^{-6}$. These results indicate that the p-value may not be a suitable stopping criterion when carrying out variable selection as it may fail to capture the discriminatory potential of the variables being analysed.

Table 3.9.1.1 Hotelling T^2 statistics, PCC estimates and p-values for subsets of normally distributed, uncorrelated variables

Variable subset	Hotelling T^2 statistic	PCC	p-value
X1	8,251	97.6	$<10^{-6}$
X1, X2	12,235	99.2	$<10^{-6}$
X1, X2, X3	15,144	99.7	$<10^{-6}$
X1, X2, X3, X4	16,826	99.9	$<10^{-6}$
X1, X2, X3, X4, X5	18,176	99.9	$<10^{-6}$

Therefore I propose a new stopping criteria based on the change in the PCC value between each round of variable selection which is used in the current version of the algorithm. It is based PCC values of sets of S and $\{S, X\}$. The difference between the two PCC values is calculated and then compared to a prespecified PCC change threshold.

3.9.2 Estimation of the probability of correct classification

It is important to evaluate the performance of the set of selected variables. This is referred to as validation and several methods of validation exist. These can be split into two general categories; internal and external validation. Internal validation uses the same dataset to train and validate a classifier by splitting the dataset into two portions one of which is used for training a classifier and the other for evaluating that classifier. Splitting the data in this way ensures that training and evaluation are not carried out using the same data. This helps to minimise over-fitting. External validation uses separate training and validation datasets. Unlike internal validation these datasets are collected separately and therefore may have different properties in terms of their variances and correlations. As such external validation can give researchers information regarding the generalisability of a classifier’s performance when using a particular set of variables. The choice of what kind of validation to use will be heavily influenced by the availability of data. If at least two datasets are available it may be possible to carry out external validation. Failing that if the existing dataset is sufficiently large it may be possible to split this dataset into training and validation portions and carry out internal validation.

The algorithm that I propose estimates PCC during each round of variable selection for a subset of variables comprising the current variable under consideration and all of the variables selected in previous rounds. In this thesis PCC is calculated using LDA (Section 3.2.1) for the MFS-T2 algorithm and QDA (Section 3.2.2) for the MFS-SNR algorithm. For each algorithm the PCC is calculated using internal leave-one-out cross-validation (LOOCV).

Once all subjects have been classified a contingency table showing the true/false positive/negative prediction frequencies is constructed and the PCC can be calculated from this table (see Section 3.3). Using the classification results presented in Table 3.9.2.1 PCC is calculated as;

$$PCC = \frac{33+44}{100} = \frac{77}{100} = 0.77 \tag{3.9.2.1}$$

Table 3.9.2.1 Contingency table showing hypothetical classification results

		Truth	
		Group 1	Group 2
Assignment (results of classification)	Group 1	True Positive (TP) 33	False Positive (FP) 17
	Group 2	False Negative (FN) 6	True Negative (TN) 44

LOOCV facilitates the validation of variable selections when a second dataset is not available. The estimation of PCC using LOOCV is also relatively quick compared to other hold-out methods (such as k-fold cross validation) where the computations can quickly become intractable. One potential disadvantage of LOOCV relates to the underlying quality of the dataset (Breiman & Spector, 1992, Kohavi, 1995, Efron & Tibshirani 1997). For example if there are duplicate observations (eg. if multiple subjects have the same age, or height) in the dataset then LOOCV will be less effective. This is because the presence of duplicate measurements the classifier will be asked to classify observations it has already seen. In other words the presence of duplicate mesaurements can introduce an element of overfitting to LOOCV. There is also a risk that LOOCV is not sufficiently robust to small changes in the dataset which may cause large changes in the estimates obtained.

3.9.3 Deciding on the size of simulations

Simulated data are used extensively in this thesis. The reason for this is that it is important to find out the statistical properties of the developed methods, i.e. of SNR and MFS-SNR. It is difficult to derive their properties analytically, hence I use computer simulations (Burton et al., 2006, Crowther and Lambert, 2012). Computationally it is possible to generate any number of simulated datasets with the properties we require. Since we also have prior knowledge of the optimal variable selections from these datasets we can better evaluate the performance of the MFS-SNR algorithm. However an important question is what is the appropriate number of simulations?

Where simulations are being used to obtain estimates of some statistical parameters, it is sometimes possible to calculate associated statistical parameters which reflect the precision of the estimate obtained through simulation (Ritter et al., 2011). Ritter et al. describe an example using a cognitive appraisal and subtraction task where subjects are given a random number and asked to repeatedly subtract another number from this while speaking aloud the result each time. Mistakes are recorded and the subject is asked to correct them. Subjects assess the task beforehand with the objective of the test being to link the type of appraisal with the subject's physiological response. A cognitive model was created designed to simulate the number of subtractions made. Using this model Ritter et al. demonstrate that as the number of simulation runs is increased the true average is approached. Using such an estimate it is possible to determine the optimal number of simulations to be run. One possibility for determining the optimal number of simulations is to use the standard error of the mean (SEM) (Ritter et al., 2011). This can be calculated according to;

$$SEM = \sqrt{\frac{Variance}{N}}. \tag{3.9.3.1}$$

where the term Variance refers to the variance of the parameter being estimated in the simulations. In my simulations this is the frequency with which each of the relevant variables (i.e. of those that are relevant for the discrimination) is selected. If we have an estimate of the standard deviation, (or if we have some prior knowledge about what this value should be for example from previous simulations), we can substitute this into equation 3.9.3.1. We can then solve the equation for N which is the number of simulations we will need to run to achieve the desired standard deviation.

Alternatively power calculations may be used to estimate the optimal number of simulations. Statistical power (δ) may be calculated according to (Ritter et al., 2011);

$$\delta = effect\ size * \sqrt{\frac{N}{2}}. \qquad (3.9.3.2)$$

We can use the standard deviation as the effect size with this equation. We can then set the power to an appropriate value and determine the number of simulations required to achieve this by solving Eq. (3.9.3.2) for N (Ritter *et al*, 2011).

In determining the optimal number of simulations it is also necessary to consider whether the process being studied is deterministic or stochastic. If it is deterministic then the outcome is known and so it may not be necessary to run the simulation more than once. However where the process being studied is stochastic a single simulation will be insufficient to effectively assess the process.

For the work presented in this thesis it would not have been appropriate to use the mean value or any other statistic to determine the optimal number of simulations as this would risk simulations being too similar. The SNR could not be used for the same reason. The only alternative was to determine the optimal number of simulations by analysing the convergence of the selection frequencies for the variables that are known to be discriminating as the number of simulations increases and identifying a number of simulations above which frequencies remained static.

In simulating data for the purpose of assessing the MFS-SNR algorithm I extracted the mean and variance/co-variance parameters of the discriminating variables from the real variables Cholesterol, HbA1c and mfERG Central Amplitude of the DREFUS dataset. I chose to simulate the following discrimination structure:

- X_1 and X_2 are 2 discriminating variables i.e. they each have some discriminatory strength when used alone,
- X_3 is a non-discriminating variable which is however relevant to discrimination when added to the 2 discriminating variables i.e. this variable does not carry any information about the two assumed disease groups but is highly correlated with variables X_1 and X_2 and hence can explain some uncertainty around X_1 and X_2 and so may improve the discrimination,
- $X_4 \dots X_{10}$ are non-discriminating and irrelevant variables i.e. they do not carry any information that discriminates between the two disease groups when used alone or when used with other discriminating variables.

Knowing that the simulated variables X_1, X_2 and X_3 were the optimal variables for discriminating in this dataset it was possible to assess the performance of the MFS-SNR algorithm.

To determine the upper limit of how many simulations were needed to accurately assess the performance of the MFS-SNR algorithm I initially ran 10,000 simulations at varying group sizes. I observed that at group sizes of $n_1 = n_2 = 1{,}000$ the variables X_1, X_2 and X_3 were selected in 100 % of simulations by the MFS-SNR algorithm. At group sizes of $n_1 = n_2 = 100$ i observed a negligible drop in the selection frequencies of X_1 and X_2 while the selection frequencies of X_3 dropped by 9 %. At group sizes of and $n_1 = n_2 = 20$ the selection frequencies dropped to 84 %, 69 % and 38 % for X_1, X_2 and X_3 respectively. I then dropped the number of simulations to 1,000 and repeated the same scenarios to investigate what impact a lower number of simulations would have on the selection frequencies at each of the group sizes. I observed that the profile of changes in selection frequencies remained unaltered at 1,000 simulations with the frequencies at each group size showing almost identical changes to those observed for 10,000 simulations. These results are presented in Tables 3.9.3.1 and 3.9.3.2 below. On the basis of these results subsequent simulations were capped at 1,000.

Table 3.9.3.1 Selection frequencies and PCC ranges (maximum-minimum) for all variables simulated 10,000 times for group sizes of 20, 100 and 1,000.

Variable	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Range
Frequency (n=20)	84	69	38	10	10	9	10	10	10	10	75
Frequency (n=100)	100	99	91	10	10	10	9	10	11	11	91
Frequency (n=1,000)	100	100	100	11	11	11	11	12	11	11	89

Table 3.9.3.2 Selection frequencies (in %) and PCC ranges (maximum-minimum) for all variables simulated 1,000 times for group sizes of 20, 100 and 1,000.

Variable	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Range
Frequency (n=20)	84	68	40	12	10	10	9	10	11	9	75
Frequency (n=100)	100	99	91	10	10	11	9	9	12	11	90
Frequency (n=1,000)	100	100	100	12	13	10	11	11	11	11	89

Considering the ranges of the selection frequencies for 1,000 and 10,000 simulations there is a difference of just 1 % at group sizes of $n_1 = n_2 = 100$ while the ranges are identical at group sizes of $n_1 = n_2 = 20$ and $n_1 = n_2 = 1{,}000$. Therefore 1,000 simulations is deemed sufficient to comprehensively assess the performance of the MFS-SNR algorithm.

3.10 Discussion

In this chapter I described the main statistical methods used in this thesis: the existing methods and the methods that I derived. The existing methods were: Hotelling T^2, LDA, QDA and methods of evaluating the quality of classification (AUROC and PCC). Then I described the methodological developments of this thesis: the proposed SNR and the proposed algorithm for the variable selection. I also prosed a more suitable stopping criteria for the variable selection algorithm.

In this chapter I have demonstrated that the existing T^2 statistic and the new SNR statistic are equivalent under the condition of homogeneity of variance-covariance matrices across groups. The SNR (Section 3.3) is a metric of the discriminatory potential of a given variable or set of variables. This means the SNR statistic is a suitable candidate for use as part of a variable selection algorithm.

In this chapter, I have also presented a summary of the variable-selection work by Lu et al. (2005). They used Hotelling's T^2 statistic to identify groups of differentially expressed genes. The variable selection algorithm presented in that work served as an inspiration for early versions of the MFS algorithm I developed in the course of the work presented in this thesis. Having appraised the work of Lu et al. I then presented an outline of my proposed MFS algorithm utilising the SNR.

As part of the variable selection process it is necessary for the MFS algorithm (as for any forward selection algorithm) to estimate performance of variable subsets. I have described how PCC estimates are calculated for each variable subset by the MFS algorithm using the method of internal validation, LOOCV. I also described the stopping criterion of the MFS-SNR algorithm. It is based on the relationship between the change in estimated PCC after each round of selection and the minimum change specified when calling the algorithm. Without an appropriate stopping criterion the algorithm will produce a list where the variables are ranked from best to worst in terms of their ability to classify cases to the correct group. The inclusion of a stopping criterion based on a minimum increase in performance specified by the user ensures that a set of variables is selected which will meet the user's requirements. This criterion will be used in future chapters.

In this chapter I also looked into how many simulated datasets I will need for my simulation studies. The results presented in tables 3.9.3.1 and 3.9.3.2 show that the difference in selection frequencies is negligible between 10,000 simulations and 1,000 simulations. The ranges for the selection frequencies are almost identical for all group sizes simulated for both 10,000 and 1,000 simulations.

Based on these results the performance of the MFS-SNR algorithm appears to be consistent when simulations are run 1,000 times. Hence in Chapters 4 and 5 I will simulate data 1,000 times to evaluate how my algorithm works on simulated data.

In the next chapter I will apply SNR to the task of variable selection for discrimination using computer simulated data in a multivariate normal scenario. I will present the results of a comparison of the MFS-SNR algorithm to several existing methods of variable selection. These include univariate and multivariate filter methods and embedded methods.

CHAPTER 4. A comparison of the variable selection performance of the MFS-SNR algorithm with existing alternative methods using simulated normal data

4.1 Introduction

In Chapter 2 I reviewed existing methods of variable selection for classification, then I proposed the novel SNR metric (Section 3.5). I also demonstrated that the SNR is better suited to summarizing the discriminatory strength of a set of variables in scenarios where the variance-covariance matrices are different across groups compared to Hotelling's T^2 statistic which assumes homogeneity of variance-covariance matrices across groups. Then I introduced the MFS-SNR algorithm (Section 3.8), which employs the SNR to find the best set of discriminatory variables and I discussed its computational details (Section 3.9). As a next step it is necessary to see how the MFS-SNR algorithm performs relative to the MFS-T2 algorithm and other variable selection algorithms. It is not possible to do such a comparison analytically and hence, in this chapter, I did it via computer simulations.

In order to compare the MFS-SNR algorithm to existing variable selection methods I carried out a comparison study using simulated data. It was envisaged that my algorithm could be used for clinical data, hence in the simulations I set up the parameters to be inspired by real clinical datasets. The simulations were inspired by the values of means and covariance's of real clinical datasets from the area of ophthalmology.

The aim of this chapter is to evaluate the performance of the MFS-SNR algorithm in multivariate normal computer-simulated scenarios; and to compare it to several existing variable selection methods.

The structure of this chapter follows the standard structure recommended for simulation studies (Burton at al., 2006). Section 4.2 describes the data-generating mechanisms, Section 4.3 lists the chosen methods for comparison, Section 4.4 presents the performance criteria, Section 4.5 gives results of simulations where I applied the existing variable selection methods as well as the MFS-SNR methods; and finally Section 4.5 outlines the conclusions based on this simulation study.

4.2 Data-generating mechanisms

This section describes how I chose the data-generating mechanisms i.e. how the data were simulated. First, for the purpose of this simulation study, I assumed a discriminatory problem with two groups and a multivariate normal distribution for the potential discriminatory variables.

Next, I decided that in the simulations the effect of the following components should be investigated:

- Sample size
    - equal vs. unequal sample size
    - small, medium vs. large sample size
- Covariance structures
    - unequal covariance across groups
    - Small and large differences in correlation between variables across groups.

Next, a decision had to be made on the number of variables. If a large number of variables was chosen then it might be challenging to make conclusions from simulations due to the complexity of correlations, on the other hand if a very small number of variables it might not be possible to make conclusions due to the correlations being too simple. In the existing literature there is a high level of diversity in the composition of simulated datasets with respect to the number of variables they contain. For example Bolòn-Canedo et al. (2013) describe 11 datasets containing anywhere from 6 to more than 4,000 variables. Therefore I chose to simulate a dataset consisting of ten variables, which are referred to as $X_{1,\dots,}X_{10}$.

Next, I decided on the discrimination structure i.e. on the relationship between the grouping variable (with two levels or groups) and the 10 continuous variables. This choice was motivated by the discrimination structures common in clinical studies: there can be a variable that is non-discriminatory when used alone and which is highly correlated with the discriminatory variables (Guyon & Elisseeff, 2003). Therefore I chose to simulate the following discrimination structure:

- $X_1$and X_2 are 2 discriminating variables i.e. they each have some discriminatory strength when used alone,
- X_3 is a non-discriminating variable (when used alone) which is however relevant to discrimination when added to the 2 discriminating variables i.e. this variable does not carry any information about the two assumed disease groups but this variable is highly correlated with the variables X_1 and X_2 and hence can explain some uncertainty around X_1 and X_2 and may improve the discrimination,
- X_4 … X_{10} are non-discriminating and irrelevant variables i.e. they do not carry any information about the two assumed disease groups when used alone and they do not carry any information about the two assumed disease groups when added to the two discriminating variables.

Next, the means, variances and correlations for the multivariate normal distribution of the 10 continuous variables were chosen. In deciding about these values, it was important that the simulated data have properties similar to an existing clinical dataset. So the simulations were based on the values of means and covariance's from a real clinical dataset from the area of ophthalmology. Specifically, the first two discriminating variables X_1 and X_2 were based on the parameters of the variables HbA1c and mfERG Central density from the DREFUS study (Chapter 6). The non-discriminating variable X_3 was based on the variable Cholesterol from the DREFUS study (Chapter 6). Then the variables X_4 to X_{10} were simulated to be independent of each other, and independent of the first three variables, thus representing noise or redundant information. For the purpose of the simulations we assumed groups 0 and 1 had mean vectors

$$\mu_0 = (8.6, 57.1, 4.5, 1, 1, 1, 1, 1, 1, 1)$$

and

$$\mu_1 = (7.2, 77.1, 4.3, 1, 1, 1, 1, 1, 1, 1).$$

with the 10x10 variance-covariance matrices

$$\Sigma_0 = \begin{bmatrix} \Sigma_{0,1:3,1:3} & \Sigma_{0,1:3,4:10} \\ \Sigma_{0,4:10,1:3} & \Sigma_{0,4:10,4:10} \end{bmatrix}$$

and

$$\Sigma_1 = \begin{bmatrix} \Sigma_{1,1:3,1:3} & \Sigma_{1,1:3,4:10} \\ \Sigma_{1,4:10,1:3} & \Sigma_{1,4:10,4:10} \end{bmatrix}$$

where $\Sigma_{0,1:3,1:3}$ is a 3x3 variance-covariance matrix of the random vector (X_1, X_2, X_3), $\Sigma_{0,1:3,4:10}$ is a 3x7 variance-covariances matrix between the random vectors (X_1, X_2, X_3), and $(X_4, \dots, X_{10})$, $\Sigma_{0,4:10,4:10}$ is a 7x7 variance-covariance matrix of the random vector $(X_4, \dots, X_{10})$, and $\Sigma_{1,4:10,1:3}$ is a 7x3 transpose of the matrix $\Sigma_{A,1:3,4:10}$, in group 0. The variance-covariance matrix of group 1 is defined analogously.

The estimated variance-covariance matrices from the DREFUS dataset (Chapter 6) for HbA1c, mfERG Central Density and Cholesterol, are

$$\hat{\Sigma}_{0,1:3,1:3} = \begin{bmatrix} 2.4 & 11.2 & 0.9 \\ 11.2 & 340.8 & 6.9 \\ 0.9 & 11.2 & 1.25 \end{bmatrix}$$

and

$$\hat{\Sigma}_{1,1:3,1:3} = \begin{bmatrix} 1.7 & 15.0 & 0.5 \\ 15.0 & 1250.6 & -3.2 \\ 0.5 & -3.2 & 0.8 \end{bmatrix}.$$

The corresponding estimated correlation matrices were

$$\widehat{\mathrm{Corr}}_{0,1:3,1:3} = \begin{bmatrix} 1.00 & 0.39 & 0.53 \\ 0.39 & 1.0 & 0.33 \\ 0.53 & 0.33 & 1.0 \end{bmatrix}$$

$$\widehat{\mathrm{Corr}}_{1,1:3,1:3} = \begin{bmatrix} 1.0 & 0.33 & 0.42 \\ 0.33 & 1.0 & -0.1 \\ 0.42 & -0.1 & 1.0 \end{bmatrix}$$

Furthermore in setting up the variance-covariance matrices of $\Sigma_{1,4:10,1:3}$, $\Sigma_{1,4:10,1:3}$ and $\Sigma_{1,4:10,1:3}$ we used the following specifications, which are same in all our simulations

$$\Sigma_{1,1:3,4:10} = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \end{bmatrix}$$

$$\Sigma_{1,4:10,4:10} = \begin{bmatrix} 1\,0\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,0\,1 \end{bmatrix}$$

$$\Sigma_{1,4:10,1:3} = transpose\left(\Sigma_{,1:3,4:10}\right).$$

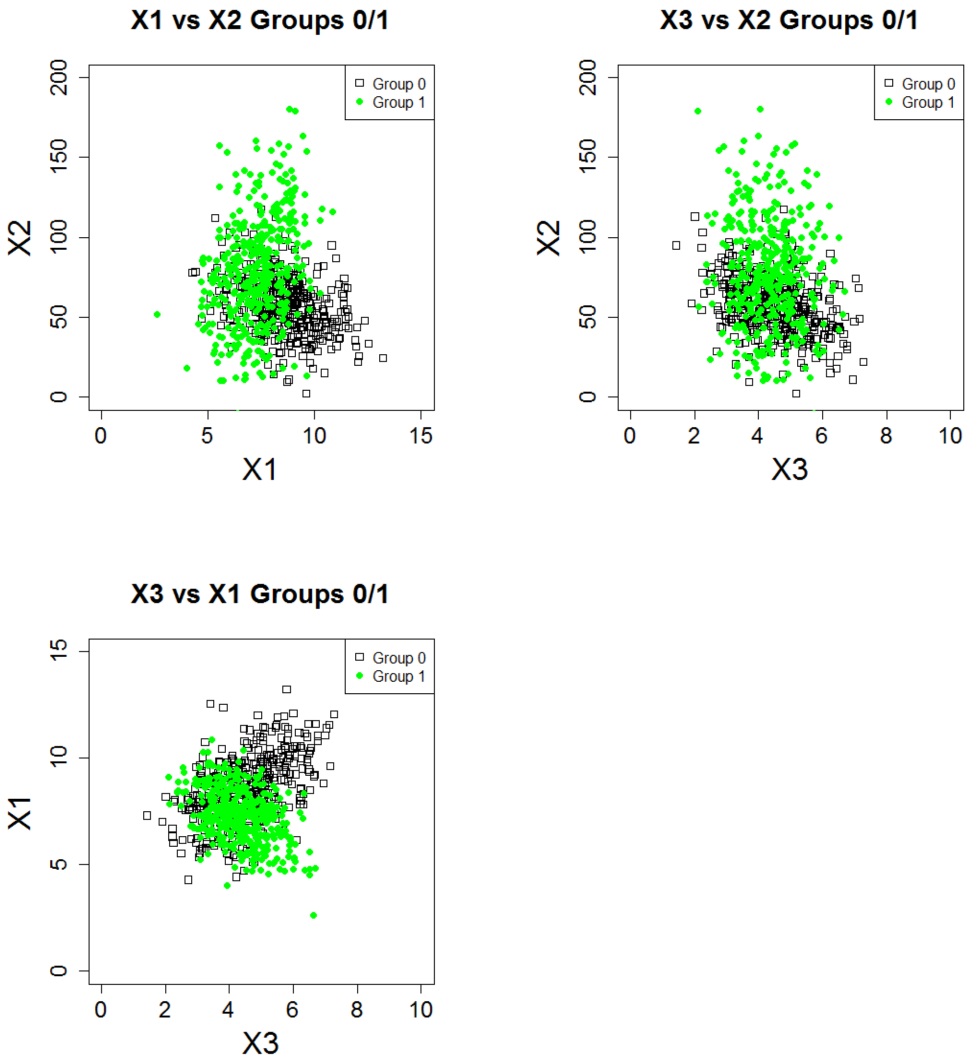
All the above considerations led to 12 data-generating scenarios. In each scenario the means are as defined above (see μ_0 and μ_1), correlation for group 0 is as defined above ($\widehat{\mathrm{Corr}}_{0,1:3,1:3}$); and correlation matrices $\Sigma_{1,4:10,4:10}, \Sigma_{1,1:3,4:10}$ are as defined above. The variances were kept the same as above for each of two groups. The other parameters that were modified in simulations are the sample sizes and the correlations. The parameters of the 12 scenarios are presented in Table 4.2.1.

Table 4.2.1 Group sizes and variance-covariance matrices for each of 12 simulation scenarios

	$n_0 = n_1 = 40$	$n_0 = n_1 = 400$	$n_0 = n_1$ $= 1{,}000$	$n_0 = 50,$ $n_1 = 150$
$Corr_{1,1:3,1:3} = \begin{bmatrix} 1.0 & 0.33 & 0.42 \\ 0.33 & 1.0 & 0.1 \\ 0.42 & 0.1 & 1.0 \end{bmatrix}$	Scenario 1	Scenario 4	Scenario 7	Scenario 10
$Corr_{1,1:3,1:3} = \begin{bmatrix} 1.0 & 0.33 & 0.42 \\ 0.33 & 1.0 & 0.1 \\ 0.42 & 0.1 & 1.0 \end{bmatrix}$	Scenario 2	Scenario 5	Scenario 8	Scenario 11
$Corr_{1,1:3,1:3} = \begin{bmatrix} 1.0 & 0.33 & 0.9 \\ 0.33 & 1.0 & 0.1 \\ 0.9 & 0.1 & 1.0 \end{bmatrix}$	Scenario 3	Scenario 6	Scenario 9	Scenario 12

Next, I wanted to see how the simulations visualised on scatter plots i.e. if scatterplots would reveal the discrimination structure. However it is impossible to visualise the simulated data from a 10-dimensional distribution on a single scatter plot. So to get an idea of how the groups 0 and 1 were separated I created several 2-D scatter plots of the first three variables. Plots of the simulated data are presented in Figure 4.2.1.1 below. Each dot represents one simulated patient. There is some overlap between the two groups in the plots but there is also some degree of discrimination between the two groups, which is typical for clinical datasets as seen in Chapter 6.

Figure 4.2.1 Bivariate plots of X_1, X_2, X_3, for $n_0 = n_1 = 400$ using Cor(X_1/X_3)=0.42 in Group 1 and Cor(X_1/X_3)=0.11 in Group 1 (Scenario 4)



4.3 Comparators

Next I had to decide which existing variable selection methods should be compared to my MFS-SNR algorithm. As mentioned in sections 2.3-2.5 there are many variable selection methods categorised as filter, wrapper and embedded. Guyon and Elisseeff (2003) have written a comprehensive introduction to variable selection methods. Additional reviews are available which provide examples of each of these types of variable selection methods e.g. in Pacheco et al. (2006), Saeys et al. (2007), Chandrashekar & Sahin (2014) and Karper (2014). The MFS-SNR algorithm is a multivariate filter method of variable selection. Filter methods function by calculating an index (i.e. summary statistic) for each variable which is designed to reflect that variables' discrimination ability. Filter methods are often (though not always) univariate which means they do not consider correlations or dependencies which may exist between variables. While this is generally considered to be a limitation of these types of methods it can also prove advantageous as it means the computational overhead is considerably lower than that of competing methods. Filter methods are also independent of the classifier algorithm however, whether this is advantageous or not must be ascertained on a case-by-case basis.

In this comparison study I compared the MFS-SNR algorithm with 6 existing methods of variable selection. I chose to compare MFS-SNR with the following methods:

- MFS algorithm based on Hotelling's T^2 statistic, (MFS-T2) which is a type of multivariate filter variable selection method. Including this version of the MFS algorithm in the simulation study facilitated assessment of the difference in classification performance arising from the ability of the SNR to accommodate heterogeneous variance-covariance matrices.
- Three univariate filter methods using chi-square statistics (Chi-squared), information gain (Info. Gain) and the Relief-F algorithm are also included in the simulation study. The MFS-SNR algorithm is a multivariate filter method, the purpose of including several univariate filter methods was to assess the performance of the MFS-SNR algorithm relative to existing univariate filter methods.
- A multivariate filter method using a SVM classifier is also included in the study (using the R package `e1071`). The SVM classifier is used to evaluate each of the variables. This is different to the MFS-SNR algorithm in that the SNR assesses variables without the need to train a classifier. Including the SVM-driven method facilitated assessment of the performance of the SNR relative to a SVM classifier.
- The method `varselRF` is an embedded method using RFs. It is the final method included in this study. Embedded methods operate by evaluating a large number of the potential variable subsets to identify the optimal subset. The purpose of including this method in the study was to assess the performance of the MFS-SNR relative to an embedded method of variable selection.

As a multivariate filter method the MFS-SNR algorithm evaluates individual variables before making variable selections.

The variable selections using filter methods based on chi-squared statistics, information gain and the relief-F algorithm are implemented in the `FSelector` package in R (Romanski & Kotthoff, 2014) that were downloaded from R Project website (R Core Team, 2015). For each method weights are calculated for each variable in a given simulated dataset. These weights reflect the importance of each variable to the task of discriminating between the groups of interest. Variables with values greater than zero are selected.

I implemented a multivariate filter method employing a SVM in this work in R. The SVM classifier was taken from the `e1071` package in R. I wrote a script in R which utilised the SVM classifier to carry out variable selection. It is based on the following idea; for each simulated dataset a classifier is first constructed using the `svm` function with the full set of variables and a PCC value is calculated. Each variable is then removed in turn and the classifier re-trained using the partial set of 9 remaining variables. A PCC value is calculated for each of these classifiers. The difference between the "full" PCC value and the "partial" PCC values for each variable is taken as a measure of that variables' importance to discrimination. The least important variable is removed (i.e. the variable with the smallest contribution to discrimination as measured by its effect on PCC). This process is repeated until a list ranking the variables by their performance from best to worst is produced. The 3 best performing variables are chosen for each simulated dataset i.e. the algorithm assumes a priori that 3 variables are to be chosen.

I implemented an embedded method of variable selection employing RFs using the `varSelRF` and `randomForest` packages in R. The `varSelRF` function uses the `randomForest` function to grow a random forest containing 5,000 trees. The idea of this variable selection algorithm is the following; at each branching point a number of variables are chosen randomly and the data is split (or branches) at that point using whichever of these randomly chosen variables produces the most homogeneous groups. The number of variables to be randomly selected at each branch point is determined based on the total number of variables. Multiple random forests are grown starting with the full set of variables and removing a proportion (0.2) of the variables at each iteration before growing a new forest. The final set of variables chosen are those associated with the forest that has the smallest number of variables as well as an error rate that is within u standard errors of the minimum error rate of all of the forests. The algorithm assumes priory priori that 3 variables are to be chosen.

4.4 Performance measures

Next it was important to choose suitable performance measures for MFS-SNR and other variable selection methods. The aim of the simulation study was to see if the variable selection methods select the correct variables for classifications. Hence I used the following performance measures:

- For each variable (out of $X_1, \dots, X_{10}$) the probability of being selected by a particular variable selection method was recorded, in %.

- For each variable selection method I evaluated the selection performance across all 1,000 simulations. The chosen criteria of performance are the probability of correct classification (PCC) and area under the receiver-operating curve (AUC) calculated using linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Each selection was validated externally using the validation dataset produced during each round of simulation.
- Computational speed. To compare the time taken for each method to carry out variable selection the time (in minutes and seconds) was measured for each method. This measurement was only made for $n_0 = n_1 = 40$ and correlations between $X_1$and X_3 set to 0.9 (Scenario 2).

In order to ensure that the estimates of the performance measures were unbiased, in each simulated scenario I simulated the multivariate normal data 1,000 times i.e. 1,000 datasets were simulated for each scenario. Then I split each simulated data using in a 50:50 ratio to produce a training dataset and a validation dataset. Next I applied each of the considered variable selection methods (see Section 4.3) to each of the simulated training datasets and I evaluated the performance measures. Further, computational details are described in Chapter 3 (Section 3.9).

4.5 Results of comparison of the MFS-SNR algorithm to existing variable selection methods in simulated data

The main goal of this comparison was to see how well MFS-SNR performed at the task of selecting the correct variables in 12 simulated scenarios (Section 4.2) relative to the alternative variable selection methods used in this study (Section 4.3) with respect to the performance measures (Section 4.4).

The scenarios that were selected are challenging in the sense that they assume complex correlation structure among the first three variables. This is significant because when two variables are highly correlated it is possible to use the value of one variable to predict the value of the other. These correlations were chosen to mimic real datasets, but also to see how current variables selection methods work in complex structures. One complexity is in the fact that there are variables X_1 and X_3 that are highly correlated, but variable X_3 has no discriminatory potential. This means that each of the two highly correlated variables reduces uncertainty about the other variable. When the correlation between two variables is sufficiently high this reduction in uncertainty can enhance the discriminatory potential of both variables. In a scenario where one of the variables has limited ability to discriminate between two groups on its' own a sufficiently high correlation with the other variable (which has a greater ability to discriminate between the groups of interest) will increase its' utility as it can enhance the discriminatory potential of the variable to which it is correlated. This is why variable X_3 in the simulated dataset is important to discriminating between the groups and therefore should be part of any variable selection. I therefore expected that by construction, variables X_1, X_2 and X_3 should be selected by variable selection methods.

The results presented below relate to 12 different scenarios with varying assumed correlations and group sizes. For each of the altered correlations a parallel change in the selection frequencies for the variables X_1, X_2 and X_3 reflecting the changing correlations is expected.

4.5.1 Variable selection in the presence of a small difference between the correlation matrices of the groups.

In this subsection I present the results of variable selection when the difference in correlation matrices between groups is small, specifically when $Corr(X_1, X_3) = 0.42$. This corresponds to scenarios 1, 4, 7 and 10 in Tables 4.5.1.1, 4.5.1.2, 4.5.1.3 and 4.5.1.4 below for $n_0 = n_1 = 40$, $n_0 = n_1 = 400$, $n_0 = n_1 = 1{,}000$ and $n_0 = 50$, $n_1 = 150$ respectively. In each of these scenarios the correlations between X_1 and X_3, and between X_2 and X_3 were 0.42 and 0.1 respectively for group 1. These correlations were inspired by the parameters from the DREFUS dataset (e.g. HbA1c, mfERG central amplitude and cholesterol in Chapter 6) and hence may be viewed as a baseline against which the other correlation scenarios may be compared.

In scenario 1 with low sample size ($n_0 = n_1 = 40$) the algorithms MFS-SNR and Relief-F selected the correct variables X_1 , X_2 and X_3 with the highest frequencies (Table 4.5.1.1) among all considered algorithms. While the frequency of choosing the correct variable X_3 was higher for Relief-F than for MFS-SNR, Relief-F also incorrectly selected non-discriminating variables with higher probabilities approaching 50 %. Random Forest selected non-discriminating variables with a lower frequency than MFS-SNR, the same is not true for MFS-T2 and SVM both of which selected non-discriminating variables with higher frequency than MFS-SNR. The SVM achieved the highest AUC and PCC values (Table 4.5.1.1), however this is at the cost of a higher frequency of non-discriminating variable selection. Furthermore SVM also failed to identify the importance of X_3 in a majority of simulations (75%). RF also achieved higher PCC and AUC values than MFS-SNR but failed to identify the importance of X_3 in a majority of simulations (77 %). The remaining filter methods (based on information gain and chi-squared statistics), failed to identify the importance of X_3 and had lower performance estimates. However these methods did select the non-discriminating variables with the lowest frequencies.

In scenario 4 with medium sample size ($n_0 = n_1 = 400$) the algorithms Relief-F, MFS-SNR, SVM and RF selected the correct variables X_1, X_2 and X_3 with the highest frequencies (Table 4.5.1.2). Relief-F however also selected the non-discriminating variables with higher frequencies than MFS-SNR, SVM and RF. MFS-T2 as well as the filter methods using information gain and chi-square statistics selected all 3 discriminating variables in less than 50 % of cases. PCC and AUC estimates of selection performance were similar for all methods. Relative to MFS-SNR the computationally expensive SVM and RF methods have selected non-discriminating variables with lower frequencies.

In scenario 7 with large sample size ($n_0 = n_1 = 1{,}000$) all algorithms except MFS-T2 selected the correct variables X_1, X_2 and X_3 with the highest frequencies (Table 4.5.1.3) of greater than 80 %. The Relief-F algorithm selected the non-discriminating variables with the highest frequencies of any method while RF and MFS-SNR selected the non-discriminating variables with similar frequencies. PCC and AUC estimates of selection performance were similar for all methods.

In scenario 10 with imbalanced sample sizes, ($n_0 = 50$, $n_1 = 150$) univariate filter methods using chi-square statistics and information gain failed to select X_3 in a majority of simulations. Relief-F selected X_1, X_2 and X_3 with frequencies over 80 % however it also selected non-discriminating variables with frequencies approaching 70 %. The MFS-T2 and MFS-SNR algorithms, SVM and Random Forest-based methods all selected X_1 and $X_2$with frequencies above 80 %. The MFS-T2 algorithm selected X_3 with a frequency of 29 % while the MFS-SNR algorithm, SVM and Random Forest-based methods selected X_3 with frequencies approaching 50 %.

Table 4.5.1.1 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n_0 = n_1 = 40$ (Scenario 1).

	Selection frequencies of variables												
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	57	57	5	2	2	2	2	2	2	2	64	66	68	71
Info. Gain	57	57	5	2	2	2	2	2	2	2	64	66	68	71
Relief F	87	86	65	46	44	47	48	46	46	46	64	66	67	70
MFS-T2	83	63	14	13	12	13	15	14	13	14	66	65	68	70
MFS-SNR	83	70	39	10	9	11	12	9	11	9	67	66	68	71
SVM	86	75	25	12	14	13	15	19	19	22	67	69	72	76
RF	92	82	23	8	7	8	9	8	8	9	65	68	69	73

Table 4.5.1.2 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n_0 = n_1 = 400$ (Scenario 4).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	100	100	41	0	0	0	0	0	0	0	66	69	73	79
Info. Gain	100	100	41	0	0	0	0	0	0	0	66	69	73	79
Relief F	93	93	79	54	54	52	56	56	53	57	65	69	72	78
MFS-T2	100	99	18	12	13	13	11	12	12	12	65	69	73	78
MFS-SNR	100	99	97	9	10	11	11	10	11	11	65	70	73	80
SVM	100	100	89	1	1	1	2	1	2	2	66	70	73	81
RF	100	100	95	6	7	7	8	8	8	8	65	70	73	81

Table 4.5.1.3 Selection frequencies (in %) and AUC/PCC estimates (in %) for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n_0 = n_1 = 1,000$ (Scenario 7).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	100	100	94	0	0	0	0	0	0	0	65	70	73	81
Info. Gain	100	100	94	0	0	0	0	0	0	0	65	70	73	81
Relief F	94	92	81	58	56	58	59	58	58	54	65	69	72	79
MFS-T2	100	100	20	10	10	9	10	9	11	9	65	69	73	78
MFS-SNR	100	100	100	14	12	10	12	11	11	8	65	70	73	81
SVM	100	100	100	0	0	0	0	0	0	0	65	70	73	81
RF	100	100	100	12	13	10	12	10	11	11	65	70	73	81

Table 4.5.1.4 Selection frequencies (in %) and AUC/PCC estimates (in %) for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n_0 = 50$, $n_1 = 150$ (Scenario 10).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	82	80	8	0	0	0	1	0	0	0	58	59	80	82
Info. Gain	82	80	8	0	0	0	1	0	0	0	58	59	80	82
Relief F	96	90	82	69	70	71	69	70	69	70	58	59	80	81
MFS-T2	100	85	29	12	12	13	14	13	13	12	58	60	81	83
MFS-SNR	91	84	53	10	8	9	9	8	9	8	58	60	81	82
SVM	100	80	49	5	7	8	9	12	13	17	58	60	82	84
RF	100	94	45	5	5	5	5	5	5	4	58	60	82	84

4.5.2 Variable selection in the presence of a large difference between the correlation matrices of the groups. A case of high correlation between the variables X_1 and X_3.

In this subsection I present the results of variable selection when the difference in correlation matrices between groups is large, specifically when $Corr(X_1, X_3) = 0.9$. This corresponds to scenarios 2, 5, 8 and 11 in Tables 4.5.2.5, 4.5.2.6, 4.5.2.7 and 4.5.2.8 below for $n_0 = n_1 = 40$, $n_0 = n_1 = 400$, $n_0 = n_1 = 1{,}000$ and $n_0 = 50$, $n_1 = 150$ respectively. In each of these scenarios the correlations between X_1 and X_3 and between X_2 and X_3 were 0.9 and 0.1, respectively, in group 1. Relative to Section 4.3.1 the correlation between X_1 and X_3 has more than doubled while the correlation between X_2 and X_3 stays the same.

When comparing variable selections in high $X_1$and X_3 correlation (Scenarios 2, 5, 8 and 11) to medium X_1 and X_3 correlation (Scenarios 1, 4, 7 and 10) I found differences. In scenario 2 with low sample size ($n_0 = n_1 = 40$) Relief-F and MFS-SNR selected X_1, X_2 and X_3 with high frequencies (Table 4.5.2.1). Both of these methods had similar PCC/AUC values whether calculated using LDA or QDA. However MFS-SNR selected non-discriminating variables with a lower frequency than Relief-F. Selections made using SVM included X_3 with a much lower frequency than MFS-SNR but had similar PCC values to MFS-SNR. This was likely a result of the higher frequency of non-discriminating variable selection by this method relative to MFS-SNR. MFS-SNR selections achieved similar performance with a more parsimonious selection while selecting the discriminating variables with high probability. The remaining filter methods (based on information gain and chi-squared statistics), failed to identify the importance of X_3 and had lower performance estimates. The same is also true for MFS-T2. RF Achieves higher PCC/AUC value than MFS-T2 but it also selects X_3 with a much lower frequency than MFS-SNR.

In scenario 5 with medium sample size ($n_0 = n_1 = 400$) X_1, X_2 and X_3 were selected with the highest frequencies by the Relief-F algorithm, MFS-SNR, SVM and RF (Table 4.5.2.2). The selection of non-discriminating variables was highest for the Relief-F algorithm and lowest for SVM and RF. MFS-T2 as well as the filter methods using information gain and chi-square statistics selected all 3 discriminating variables in less than 50 % of cases. PCC and AUC estimates of selection performance were similar for MFS-SNR, SVM and RF which is expected as their selection frequencies are so similar.

In scenario 8 with large sample size ($n_0 = n_1 = 1{,}000$) X_1, X_2 and X_3 were selected with a frequency of greater than 90 % by all methods except MFS-T2. The Relief-F algorithm selected the non-discriminating variables with the highest frequencies followed by MFS-SNR and MFS-T2 which both selected the non-discriminating variables at similar frequencies and RF which selected them with lower frequencies. The selection of non-discriminating variables by SVM was negligible at this group size. PCC and AUC estimates of selection performance were similar for all methods except MFS-T2. This can be attributed to the similar selection frequencies of the variables X_1, X_2 and X_3 and the lack of any discriminatory ability in the non-discriminating variables (i.e. because variables X_4 to X_{10} were non-discriminating their contribution to the overall performance is negligible and therefore even at high frequencies of selection they have negligible impact on classification performance).

In scenario 11 with imbalanced sample sizes ($n_0 = 50$, $n_1 = 150$), the univariate filter methods using chi-square statistics and information gain failed to select X_3 in a majority of simulations (Table 4.5.2.1). The selection frequencies for X_1 and X_2 also fell for both of these methods relative to the correlations in Section 4.5.1. Relief-F selected X_1, X_2 and X_3 with frequencies over 80 % however it also selected non-discriminating variables with frequencies approaching 70 %. The MFS-T2 and MFS-SNR algorithms selected X_1 and X_2 with frequencies above 80 %. The selection frequency for X_3 by the MFS-T2 algorithm more than doubled in this correlation scenario. The SVM and Random Forest-based methods selected X_2 with a reduced frequency, with drops of 23 and 40 %, respectively. Conversely the selection frequency for X_3correctly increased (almost doubled) for MFS-SNR, SVM and Random Forest in this correlation scenario.

Table 4.5.2.1 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X_1 and X_2 in group 1 and $n_0 = n_1 = 40$ (Scenario 2).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	61	59	5	2	2	2	2	2	2	2	64	65	69	71
Info. Gain	61	59	5	2	2	2	2	2	2	2	64	65	69	71
Relief F	93	86	79	43	43	45	45	45	43	42	64	71	70	80
MFS-T2	84	60	16	12	11	13	14	14	13	13	64	66	69	72
MFS-SNR	84	79	81	11	12	12	13	10	11	10	64	72	70	82
SVM	92	63	51	10	11	9	15	14	17	19	67	72	72	81
RF	94	78	45	7	6	6	8	7	8	7	66	71	70	78

Table 4.5.2.2 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X_1 and X_2 in group 1 and $n_0 = n_1 = 400$ (Scenario 5).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	100	100	42	1	0	0	0	0	0	0	65	70	74	82
Info. Gain	100	100	42	1	0	0	0	0	0	0	65	70	74	82
Relief F	98	91	96	51	51	48	51	52	51	53	65	72	73	87
MFS-T2	100	99	32	12	13	12	13	15	12	14	65	70	73	81
MFS-SNR	100	100	100	12	13	12	12	12	11	14	65	73	74	89
SVM	100	96	100	3	6	2	3	5	4	1	66	73	74	89
RF	100	94	100	3	3	2		3	3	2	65	73	73	89

Table 4.5.2.3 Selection frequencies (in %) and AUC and PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X_1 and X_2 in group 1 and $n_0 = n_1 = 1,000$ (Scenario 8).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	100	100	94	0	0	0	0	0	0	0	65	73	74	88
Info. Gain	100	100	94	0	0	0	0	0	0	0	65	73	74	88
Relief F	99	91	96	52	48	51	54	55	54	52	65	73	74	88
MFS-T2	100	100	50	12	14	11	13	12	12	13	65	71	74	83
MFS-SNR	100	100	100	13	14	12	12	13	13	13	65	73	74	89
SVM	100	100	100	0	0	0	0	0	0	0	65	73	74	89
RF	100	100	100	1	2	3	3	2	3	3	65	73	74	89

Table 4.5.2.4 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n_0 = 50$, $n_1 = 150$ (Scenario 11).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	83	81	8	0	0	0	1	0	0	0	58	59	80	82
Info. Gain	83	81	8	0	0	0	1	0	0	0	58	59	80	82
Relief F	99	89	96	68	68	68	68	67	67	68	58	61	84	90
MFS-T2	100	85	73	14	15	15	15	14	16	16	58	61	84	89
MFS-SNR	92	94	92	17	14	17	16	14	16	16	58	62	84	91
SVM	100	57	92	2	3	4	6	8	12	16	58	62	85	91
RF	100	54	87	1	1	1	1	1	1	1	58	62	85	91

At the larger group sizes of 400 and 1,000 differences in selection frequencies are negligible. However when the group sizes are 40 I observed an increase in the selection frequency for X_3 for MFS-SNR, SVM and RF. The largest increase is seen for MFS-SNR. For each of these methods an increase in the performance estimates is also observed with the largest increase for MFS-SNR. The change in selection frequency of X_3 for the MFS-T2 algorithm is negligible. The changes in the performance estimates for MFS-T2 are also negligible. For the univariate filter methods any changes in the selection frequencies and performance estimates are negligible. At the imbalanced group sizes of $n_0 = 50$, $n_1 = 150$ the frequency of X_3 selection by the Relief-F algorithm increases by 16 %. For the univariate filter methods minor changes are noticed under these altered correlation conditions. For MFS-SNR the frequency of selection for X_3 almost doubles while the frequency for X_2 increases by 10 %. Changes in the selection frequency of X_1 are minor. For MFS-T2 the selection frequency of X_2 more than doubles while the frequencies for X_1 and X_3 are unchanged. SVM and Random Forest both exhibit a drop in the selection frequency of X_2. At the same time the selection frequency for X_3 almost doubles for both methods.

4.5.3 Variable selection in the presence of a large difference between the correlation matrices of the groups. A case of high correlation between the variables X_2 and X_3.

In this subsection I present the results of variable selection when the difference in correlation matrices between groups is large, specifically when Corr(X_2, X_3)=0.9. This corresponds to scenarios 3, 6, 9 and 12 in Tables 4.5.3.1, 4.5.3.2, 4.5.3.3 and 4.3.3.4 below for $n_0 = n_1 = 40$, $n_0 = n_1 = 400$, $n_0 = n_1 = 1{,}000$ and $n_0 = 50$, $n_1 = 150$ respectively. In each of these scenarios the correlations between $X_1$and X_3 and X_2 and X_3 were 0.42 and 0.9, respectively, in group 1. Relative to section 4.5.1 the correlation between X_1 and X_3 was unchanged while the correlation between X_2 and X_3 was increased by a factor of 9.

When comparing the variable selection in low X_2 and X_3 correlation (Scenarios 2, 5, 8 and 11) to medium high X_2 and X_3 correlation (Scenarios 3, 6, 9 and 12) we found differences. In scenario 3 with low sample size ($n_0 = n_1 = 40$) the discriminating variables X_1, X_2 and X_3 were selected with the highest frequencies by the Relief-F algorithm (Table 4.3.3.1). However the Relief-F algorithm also selected the non-discriminating variables with frequencies approaching 50 %. All of the other methods selected X_1, X_2 and X_3 together with frequencies of less than 50 %. The selection frequencies for non-discriminating variables were similar for MFS-T2, MFS-SNR and SVM but lower for RF. There were differences in the performance estimates AUC and PCC calculated for MFS-T2, MFS-SNR, SVM and RF. These differences can be explained by the differences in selection frequencies of the discriminating variables for each of these methods.

In scenario 6 with medium sample size ($n_0 = n_1 = 400$). The discriminating variables X_1, X_2 and X_3 were selected with the highest frequencies by the Relief-F algorithm, MFS-T2, MFS-SNR, SVM and RF (Table 4.5.3.2). The highest selection frequency for the non-discriminating variables was for the Relief-F algorithm followed by MFS-T2, MFS-SNR and RF. Performance estimates were similar for MFS-T2 and MFS-SNR which can be explained by the similar selection frequencies for variables X_1, X_2 and X_3. Performance estimates for SVM and RF are higher due to the higher frequency of selection of the variables X_1, X_2 and X_3.

In scenario 9 with large sample size ($n_0 = n_1 = 1{,}000$) selection frequencies for the discriminating variables are very similar for all methods (Table 4.5.3.3). The selection frequencies for the non-discriminating variables were largest for the Relief-F algorithm, followed by random forests and then MFS-T2 and MFS-SNR which had very similar selection frequencies for the non-discriminating variables. AUC and PCC performance estimates were very similar for all methods at this group size due to the similar selection frequencies for X_1, X_2 and X_3 by each method.

In scenario 12 with imbalanced sample sizes ($n_0 = 50$, $n_1 = 150$) the univariate filter methods using chi-square statistics and information gain failed to select X_3 in a majority of simulations. The selection frequencies for X_1 and X_2 also fell for both of these methods relative to the scenarios with medium and low correlations between X_1 and X_3 and X_2 and X_3 respectively. Relief-F selected X_1, X_2 and X_3 with frequencies over 80 % however it also selected non-discriminating variables with frequencies approaching 70 %. MFS-T2, MFS-SNR, SVM and Random Forest selected X_1 and $X_2$with frequencies above 80 %. For MFS-SNR the selection frequency for X_3 increased by 3 % relative to section 4.5.1. MFS-T2, SVM and Random Forest all selected X_2 with increased frequency in this correlation scenario. Differences in performance estimates for all methods arise out of differences in the frequencies of selection for X_3.

Table 4.5.3.1 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X_1 and X_3 in group 1 and $n_0 = n_1 = 40$ (Scenario 3).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	57	54	5	2	2	2	2	2	2	2	64	66	68	71
Info. Gain	57	54	5	2	2	2	2	2	2	2	64	66	68	71
Relief F	87	86	75	46	44	49	46	46	44	46	65	69	68	74
MFS-T2	84	61	26	12	12	13	15	13	12	13	65	67	68	71
MFS-SNR	81	67	40	12	11	14	13	12	13	12	65	67	68	73
SVM	87	84	45	92	8	10	12	13	14	17	67	71	73	79
RF	90	85	36	9	8	8	9	7	7	9	66	69	70	75

Table 4.5.3.2 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X_1 and X_3 in group 1 and $n_0 = n_1 = 400$ (Scenario 6).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	100	100	42	0	0	0	0	0	0	0	66	70	74	81
Info. Gain	100	100	42	0	0	0	0	0	0	0	66	70	74	81
Relief F	93	95	91	52	52	51	56	55	54	54	66	71	74	83
MFS-T2	100	99	86	11	10	12	11	11	11	14	66	72	74	85
MFS-SNR	100	100	95	11	12	14	12	12	13	15	66	72	75	86
SVM	100	100	100	0	0	0	0	0	0	0	67	72	75	86
RF	100	100	100	3	3	4	3	3	2	4	67	72	75	86

Table 4.5.3.3 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods for a correlation of 0.9 between X_1 and X_3 in group 1 and $n_0 = n_1 = 1,000$ (Scenario 9).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	100	100	94	0	0	0	3	0	0	0	66	72	75	86
Info. Gain	100	100	94	0	0	0	3	0	0	0	66	72	75	86
Relief F	93	94	93	56	55	58	58	58	57	54	66	71	74	84
MFS-T2	100	100	96	14	16	13	14	14	13	12	66	72	75	86
MFS-SNR	100	100	100	14	14	13	12	13	13	13	66	72	75	86
SVM	100	100	100	0	0	0	0	0	0	0	66	72	75	86
RF	100	100	100	46	47	46	47	43	47	48	66	72	75	86

Table 4.5.3.4 Selection frequencies (in %) and AUC/PCC (in %) estimates for MFS-SNR, MFS-T2 and selected filter and embedded methods with unadjusted correlations in group 1 and $n_0 = 50$, $n_1 = 150$ (Scenario 12).

	Selection frequencies of variables													
Selection method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Chi-squared	82	80	8	0	0	0	0	0	0	0	58	59	80	82
Info. Gain	82	80	8	0	0	0	0	0	0	0	58	59	80	82
Relief F	95	94	95	70	70	69	70	70	70	71	58	60	80	85
MFS-T2	99	86	32	11	10	13	10	11	9	8	58	60	81	84
MFS-SNR	90	85	56	12	12	10	12	12	12	11	58	60	80	85
SVM	99	87	66	4	4	5	8	8	8	10	59	61	82	87
RF	95	97	75	2	3	1	1	2	3	2	59	61	82	88

At the larger group sizes of 400 and 1,000 differences in selection frequencies were negligible. At group sizes of 40 the change in selection frequency of X_3 was negligible for MFS-SNR while the selection frequency almost doubled for MFS-T2, SVM and RF. There was a parallel increase in the performance estimates for MFS-T2, SVM and RF though the increases were smaller than those observed when the correlation between X_1 and X_3 was doubled. For the univariate filter methods any changes in the selection frequencies and performance estimates were negligible. At the imbalanced sample sizes ($n_0 = 50$, $n_1 = 150$) changes in selection frequencies for the univariate filter methods were negligible relative to section 4.3.1. MFS-SNR, MFS-T2, SVM and Random Forest all selected X_1 and X_2 with frequencies of over 80 %. The selection frequencies for X_3 also increased for MFS-SNR, MFS-T2, SVM and Random forest by 3 %, 3, %, 17 % and 30 % respectively. Minor changes to the selection frequencies for the non-discriminating variables were observed for MFS-SNR, MFS-T2, SVM and Random Forest.

4.6 Discussion

The MFS-SNR algorithm is a multivariate filter method which I implemented (in Chapter 3) as a paradigm to select the best set of variables for classification. In this chapter I compared it with six common variable selection methods in simulated datasets. I considered various group sizes and various correlation structures as both of these aspects can impact the performance of the variable selection methods. It is also important to consider the time and computational space required by each method when analysing the results. For example filter methods of variable selection generally (see Chapter 2) have lower computational requirements as they do not analyse a large number of variables (relative to embedded and wrapper methods) in the process of variable selection. This is in contrast to embedded and wrapper methods of variable selection which analyse large numbers of variable subsets (or all the variables) in the course of variable selection.

4.6.1 Effect of sample size

When I used medium and large group sizes of 400 and 1,000 subjects per group all the variable selection methods gave comparable results. The selection frequencies of variables X_1, X_2 and X_3 for all of the methods correctly increased. At group sizes of 1,000 all methods (with the exception of MFS-T2) selected variables X_1, X_2 and X_3 with frequencies of over 90 %. This improved performance is not surprising as larger groups will be more robust to the presence of outlying values and any metrics calculated using these larger groups will more accurately reflect the discriminatory potential of variables. The MFS-T2 algorithm is the exception to this observation and this is due to the underlying assumption of variance-covariance matrix homogeneity across groups. This assumption is not valid in these simulations which accounts for the poorer performance of the MFS-T2 algorithm at larger group sizes.

The MFS-SNR algorithm has out-performed or worked at least as well as the competing filter methods in all 12 simulation scenarios. For the Relief-F algorithm, information gain and chi-square statistic-based methods this is in part due to the univariate nature of these methods compared to the multivariate nature of the SNR and the MFS-SNR algorithm. While the MFS-T2 algorithm shares the mechanism of the MFS-SNR algorithm Hotelling's T^2 statistic assumes homogeneity of variance-covariance matrices which is an invalid assumption in these simulations. Hence the MFS-T2 algorithm fails to perform as well as the MFS-SNR algorithm.

The MFS-SNR algorithm and the SVM and RF-based methods exhibit similar performance in all scenarios though the non-discriminating variable selection frequency is generally lower for the SVM method. In part this is due to the way in which selections were made using SVM. This method requires the user to specify the number of variables to be chosen, which was set to three. In comparison the MFS-SNR algorithm continues to select variables until its' stopping criteria is reached which is based on the change in PCC (see Section 3.9.1). Also RF requires the user to specify the number of variables to be chosen and it selects the variables associated with the best-performing forest (i.e. neither the RF method nor the MFS-SNR algorithm have a cap on the number of variables selected). If the SVM-based method did not have its' selections capped in this way the selection frequencies for the non-discriminating variables would be higher than those presented in this work.

4.6.2 Computing time considerations

It is important to note that the final variable subset is not the only consideration when looking at the relative performance of each of the methods, for example, while the performance of the MFS-SNR algorithm and the methods based on SVM and RF are comparable (assuming Normality of data) the computational requirements and the time required to make the selections are not.

A comparison of the time involved in carrying out variable selection was also undertaken. The length of time taken to run a script simulating 1,000 datasets and carrying out variable selection from each of them was recorded. Group sizes were $n_0 = n_1 = 40$ and correlations between X_1 and X_3 were set to 0.9 (Scenario 2). All methods were run using the same script for simulating data and for evaluating the performance of the selected variables. The only difference in each case was the specific variable selection method being used. Therefore the difference in the time taken can be attributed to differences in the methods being used. Shortest computation time is for selections made using chi-square statistics. This took 38 seconds to run. The longest computation time is for selections made using random forests as part of an embedded method. This took 8 minutes and 50 seconds to run. The MFS-SNR algorithm took 2 minutes 36 seconds, to run all 1000 simulations.

Comparing MFS-SNR, SVM and Random Forest methods the overall results and performance were similar however the computational time required was greater for SVM and Random Forest-based methods (4:48 and 8:50, min:sec, respectively). The time required for the MFS-T2 algorithm was less than that for the MFS-SNR algorithm (2:07 compared to 2:36, min:sec, respectively). However the MFS-T2 method failed to select the variable X_3 in a majority of simulations. While the univariate filter methods using chi-square statistics and information gain were considerably faster than MFS-SNR (each of the univariate methods required less than 1 minute) they also failed to select the variable X_2 in a majority of simulations. The Relief-F method not only took longer than MFS-SNR and the alternative univariate filter methods it also selected the non-discriminatory variables with higher frequencies than any other method. In a real-world scenario the number of variables is likely to be much higher with a parallel increase in the computation time required for methods based on SVM and RF. While the MFS-SNR algorithm would also experience an increase in computational overhead this would still be less than that of the methods using SVM and RF.

Embedded methods (such as the RF method used in this study) and wrapper methods analyse large numbers of variable subsets to identify the optimal subset for discriminating between groups. This analysis involves training a classifier and estimating the performance of that classifier for each variable subset to identify the optimal. This is an NP-hard problem (a non-deterministic problem whose solution time is upper bounded by a polynomial expression) the solution of which rapidly becomes impractical. In contrast the MFS-SNR algorithm identifies the optimal subset by selecting the variables which cause the largest increase in the SNR. While a QDA classifier is trained for each subset the total number of subsets is smaller than for embedded or wrapper methods and so this is not an NP-hard problem.

Table 4.6.2.1 Run-time (mins:secs) for each variable selection method applied to 1,000 simulated datasets under the conditions of scenario 2

Selection method	Chi-squared	Info. Gain	Relief F	MFS-T2	MFS-SNR	SVM	RF
Run-time	0:38	0:42	7:02	2:07	2:35	4:48	8:50

4.6.3 Effect of correlation differences across groups

In the simulation study I intentionally added the following complexity of correlations: the variable X_3 was simulated to be non-discriminating but also have high correlation with the discriminating variables X_1 and X_2. This caused X_3 to have a role in enhancing overall discrimination when used with X_1 and X_2. The purpose of including variable X_3 in this simulation study was to determine whether or not each of the variable selection methods being compared could also identify that role for X_3. I found that all considered the variable selection methods were able to identify this role for X_3 in larger sample sizes but at the smallest group size of 40 the MFS-SNR algorithm clearly out-performs the competing variable selection methods.

Where correlations are increased between X_3 and X_1 or X_2 i.e. where a large difference in correlation between two groups is introduced, it is expected that X_3 is chosen more often as it explains better the uncertainty of of X_1 and X_2. In these simulations I observed that the largest changes in selection frequency occur for X_3 (changes do occur for selection frequency of X_1 and X_2 however these changes are negligible). For an increase in the correlation between X_1 and X_3 from

0.42 to 0.9 the largest change in selection frequency of X_3 is observed for MFS-SNR and this is matched by the largest increase in performance estimates. For an increase in correlation between X_2 and X_3 from 0.1 to 0.9 the change in selection frequency of X_3 is smaller for MFS-SNR than either SVM or RF. The change in performance estimates is also smaller for MFS-SNR than for SVM and RF. However it must be noted that while the proportional change in selection frequencies are larger for SVM and RF the actual selection frequency for X_3 is very similar to that for MFS-SNR.

In summary, these simulation results demonstrate that having very different correlation matrices across groups impacts on the frequency of variable selection in all considered variable selection algorithms. In that context when the correlations between X_1, X_2 and X_3 are increased the selection frequency of X_3 also increases as was expected. The variable X_3 was chosen not to be discriminatory when used alone, but it was constructed to improve the discrimination of X_1 (or X_3) if there is strong enough correlation. The increased frequency of X_3 selection is matched by increased performance estimates. However the SVM and RF methods only begin to select X_3 with appreciably high frequencies when the correlations are increased whereas the MFS-SNR algorithm correctly selects X_3 even with small differences in correlations. Thus the MFS-SNR algorithm is better able to identify the role of X_3 in enhancing the discriminatory potential of other variables, in the considered simulation scenarios.

4.7 Conclusion

In conclusion, in the considered scenarios, the MFS-SNR algorithm performs at least as well as the best competing considered methods which are SVM and RF across a range of correlations relationships and group sizes. The MFS-SNR algorithm is also more effective at identifying the role of the non-discriminating variable X_3 in enhancing the overall classification performance when used with other variables X_1 and X_2. While the SVM and RF methods offer better performance (than MFS-SNR) in some cases this is at the cost of increased computational complexity and also at the cost of needing to specify the number of variables to be chosen. The MFS-SNR algorithm took 2 minutes 36 seconds while the SVM and RF methods took 4 minutes 48 seconds and 8 minutes 50 seconds, respectively, to run all 1000 simulations for scenario 2.

One of the drawbacks of these simulations is that a multivariate normal distribution was assumed. Chapter 5 investigates the properties of the MFS-SNR algorithm when selecting from amongst a set of variables whose distribution deviates from normality.

Chapter 5. Performance of the MFS-SNR algorithm for non-normal data

Chapter 4 investigated the properties of the MFS-SNR variable selection algorithm on simulated normal data. Normality is implicitly assumed by the MFS-SNR algorithm, however the assumption of normality is often not valid for real data. The aim of this chapter is to examine whether the MFS-SNR algorithm is robust to deviations from normality, i.e. if it would still choose the correct variables for discrimination.

This chapter is organised as follows. First I introduce the problem (Section 5.1), then I describe the methods of simulating the data (Section 5.2). Finally I evaluate the MFS-SNR algorithm using simulated non-normal data (Section 5.3).

5.1 Aim of simulation

The aim is to examine if the MFS-SNR algorithm is robust to deviations from normality. The problem of variable selection from data that are not normally distributed is an active area of research.

In Chapter 2 I summarised variable selection methods and discussed their suitability for data that are not normally distributed. Therefore here I just mention the main points. Filter variable selection methods that are based on mutual information are suitable for use with data that are not normally distributed (e.g. Doquire & Verleyson, 2011; Todorov & Setchi, 2014) as the definition does not make any assumption on the joint probability density function. Another variable selection method is the Method of Tang and Mao (2007) who use factor level combinations for ordinal and nominal variables. Finally, filter variable selection methods based on importance scores are suitable for variable selection from non-normal distributed variables (e.g. Tang et al., 2007). The importance score measures the strength of the association between the outcome of interest and the variable in question. Only those variables with an importance score above a certain threshold are retained during the variable selection procedure. Pavlidis et al. (2001) address the issue on datasets containing quantitative gene expression data and qualitative phylogenetic data. They train a heterogeneous kernel based on both quantitative and qualitative data, (intermediate integration). Alternatively they train two kernels on the quantitative and qualitative data respectively then amalgamate the discriminant values of these two kernels to produce the final discriminants (late integration). Bar-hen and Daudin (1995) generalize the Mahalanobis distance so that it may be calculated for mixtures of quantitative and qualitative data. This generalised Mahalanobis distance is a summation of the distance contributions of both the quantitative and qualitative variables. Wilson and Martinez (1997) describe an extension to the value difference metric (VDM) to produce the heterogeneous value difference metric (HVDM). The VDM is designed to estimate distance values between the levels of nominal variables. This estimation is based on the correlation between nominal levels and outcomes. The HVDM adds the ability to calculate the normalized Euclidean distance for quantitative variables.

The approach here is to investigate the potential of the MFS-SNR algorithm (Section 3.8) for variable selection where data are not normally distributed. It is not possible to study the robustness of the MFS-SNR algorithm analytically, hence it was studied in computer simulations. I explored three scenarios representing deviation from normality: an ordinal variable with 2 categories, an ordinal variable with three categories and a log-normal distributed variable. Then I ran the MFS-SNR algorithm to see if it could identify the correct set of discriminating variables.

5.2 Data generating mechanism

In this section I describe the data-generating mechanism for the simulations, i.e. the methods used to simulate the datasets.

To make the simulation problem tractable, I considered 10 variables:

- variable X_1 was discriminatory but not normally distributed,
- variable X_2 was discriminatory and normally distributed,
- variable X_3 is not discriminatory but improves the discrimination when added to X_1 and was normally distributed
- variables $X_4, \dots, X_{10}$ were not discriminatory but assumed to follow a multivariate normal distribution.

I considered three schemes of non-normality for the variable X_1

- ordinal with two categories,
- ordinal with three categories
- log-normally distributed.

I used the following simulation strategy. I simulated ten continuous variables from a multivariate normal distribution (Section 5.2.1) and then transformed the first variable to make it either dichotomous, trichotomous or log-normal (Section 5.2.2).

I needed to decide how to simulate ordinal data. The challenge was that the absolute difference between the levels of the ordinal variable does not have a direct interpretation. For example the American Joint Committee on Cancer (AJCC) developed a colorectal cancer staging system which assigns cancers to the appropriate stage based on the prognostic severity of the disease (Hu et al., 2015). This means that when comparing stage 2 and stage 1 cancers the stage 2 cancers will have a greater prognostic severity than the stage 1 cancers. Thus as we progress up through the stages the prognostic severity of the cancer increases. However it does not increase at a constant rate and so the difference between stages 1 and 2 may not be the same as the difference between stages 2 and 3. However the algorithm MFS-SNR assumes that these differences between cancer stages have equal clinical difference i.e. the difference between stages 1 and 2 is the same as the difference between stages 2 and 3. The MFS-SNR algorithm treats the levels as a realisation of a continuous variable and calculates means and variances for each group. To avoid the assumption of equality of clinical difference between the levels of the ordinal variable, the values of the ordinal variables can be recoded. The difficulty with this approach is that we generally do not have sufficient knowledge to apply values to the levels which accurately reflect any differences between the levels. While this is obviously a more important consideration with ordinal data where there is some intrinsic order to the levels of the variable, it is also important for nominal data especially when attempting to use nominal data with the MFS-SNR algorithm.

Several recoding schemes exist which can be applied to ordinal or nominal data levels (Bishop et al., 1975). Three of the most commonly applied methods are dummy coding, effects coding and contrast coding. If we assume c levels of an ordinal or nominal variable, then the general protocol is to encode the ordinal variable into c-1 dummy variables. Dummy coding involves assigning one level to be the reference or control level which is assigned a zero value for each of the dummy variables. Each of the remaining levels is then encoded relative to the reference level using either dichotomous or polychotomous variables as appropriate. For example assuming we have a factor with 3 levels X, Y and Z we could assign level Z as our reference level with a value of 0. There are 3 levels in total so we will create two dummy variables one for each of the levels X and Y which we call D_x and D_y. The coding scheme for these variables is presented in Table 5.2.1.

Table 5.2.1 Dummy variable coding for a factor with 3 levels; X, Y and Z. Level Z used as the reference level.

Level	D_x	D_y
X	1	0
Y	0	1
Z	0	0

Effects coding is similar to dummy coding in that c-1 dummy variables will be created (assuming c levels of an ordinal or nominal variable). A value of -1 is assigned to the reference level when using effects coding. For example assuming we have a factor with 3 levels X, Y and Z we could assign level Z as our reference level with a value of -1 and create two variables D_x and D_y for the levels X and Y. The coding scheme for these variables is presented in Table 5.2.2.

Table 5.2.2 Effect variable coding for a factor with 3 levels; X, Y and Z. Level Z used as the reference level.

Level	D_x	D_y
X	1	0
Y	0	1
Z	-1	-1

In contrast coding the scheme allows the researcher to test specific hypotheses. To test these hypotheses the contrast coding scheme assigns coefficient values to variables which are orthogonal. The sum of the contrast coefficients must also equal zero. For example assuming we have a factor with 3 levels X, Y and Z we could assign level Z as our reference level with a value of -2 and create two variables D_x and D_y for the levels X and Y. The coding scheme for these variables is presented in Table 5.2.3. Note that by assigning the codes in this manner we are sequestering data according to the comparisons we wish to carry out.

Table 5.2.3 Contrast variable coding for a factor with 3 levels; X, Y and Z. Level Z used as the reference level.

Level	D_x	D_y
X	1	0
Y	0	1
Z	-2	-2

5.2.1 Simulated data

I simulated a dataset consisting of ten variables. The first three variables are motivated by the variables HbA1c, mfERG Central Amplitude and Cholesterol from the DREFUS dataset (Chapter 6). I assumed two groups for discrimination, 0 and 1. The mean vectors considered for the 10 variables for each group were

$$\mu_0 = (8.6, 57.1, 4.5, 1, 1, 1, 1, 1, 1, 1)$$

$$\mu_1 = (7.2, 77.1, 4.3, 1, 1, 1, 1, 1, 1, 1)$$

and the variance-covariance matrices

$$\Sigma_{0,1:3} = \begin{bmatrix} 2.4 & 11.2 & 0.9 \\ 11.2 & 340.8 & 6.9 \\ 0.9 & 11.2 & 1.25 \end{bmatrix}$$

$$\Sigma_{1,1:3} = \begin{bmatrix} 1.7 & 15.0 & 0.5 \\ 15.0 & 1250.6 & -3.2 \\ 0.5 & 3.2 & 0.8 \end{bmatrix}.$$

The correlation matrices for the variables Cholesterol, HbA1c and mfERG Central density in both groups were assumed to be as presented below;

$$\mathrm{Corr}_{0,1:3} = \begin{bmatrix} 1.00 & 0.39 & 0.53 \\ 0.39 & 1.0 & 0.33 \\ 0.53 & 0.33 & 1.0 \end{bmatrix}$$

$$\mathrm{Corr}_{1,1:3} = \begin{bmatrix} 1.0 & 0.33 & 0.42 \\ 0.33 & 1.0 & 0.1 \\ 0.42 & -0.1 & 1.0 \end{bmatrix}$$

The rest of the components of the variance-covariance matrices Σ_0 and Σ_1 were set to be as follows

$$\Sigma_{1,1:3,4:10} = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \end{bmatrix}$$

$$\Sigma_{1,4:10,4:10} = \begin{bmatrix} 1\,0\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,0\,1 \end{bmatrix}$$

$$\Sigma_{1,4:10,1:3} = transpose\left(\Sigma_{1,1:3,4:10}\right).$$

I assumed the following simulation scenarios

- Scenario 1: $n_0 = n_1 = 100$, no transformation of X_1

- Scenario 2: $n_0 = n_1 = 500$, no transformation of X_1
- Scenario 3: $n_0 = n_1 = 100$, dichotomisation of X_1
- Scenario 4: $n_0 = n_1 = 500$, dichotomisation of X_1
- Scenario 5: $n_0 = n_1 = 100$, trichotomisation of X_1
- Scenario 6: $n_0 = n_1 = 500$, trichotomisation of X_1
- Scenario 7: $n_0 = n_1 = 100$, log-normal transformation of X_1
- Scenario 8: $n_0 = n_1 = 500$, log-normal transformation of X_1

where n_0 and n_1 are the group sizes.

For each simulation scenario I recorded the frequency of each variables selection as well as estimates of the selection performance across 1,000 simulations. The measures of accuracy considered were PCC and AUC both of which were calculated using LDA and QDA. During each round of data simulation the data was split into two portions. The first portion was the training portion and was used to carry out variable selection and the training of classifiers. The second portion of data was the validation portion. Following training of the classifiers the variable selections were validated using the validation portion of the simulated data. In this way each set of results was validated externally using the validation portion of the data.

5.2.2 The simulation of dichotomised, trichotomised and log-normal transformed data

I assumed three data generation schemes where data were not multivariate normal. In order to simulate the data, in each scheme I first simulated datasets from a multivariate Normal distribution. Then I subjected the values for variable X_1 to a transformation producing a non-normal distribution.

In scenarios 3 and 4 I dichotomised the variable X_1. To dichotomise X_1 the median (7.9) across groups 0 and 1 was used to assign values to level 0 or 1. All subjects with values of X_1 less than or equal to 7.9 were assigned to level 0 and their X_1 value was set to 0. All subjects with values of X_1 greater than or equal to 7.9 were assigned to level 1 and their X_1 values were set to 1. Hence in summary this transformation can be written as

$$X_{1,dich} = \begin{cases} 0 \; if \; X_1 < 7.9 \\ 1 \; if \; X_1 \geq 7.9 \end{cases}$$

In scenarios 5 and 6 we trichotomised X_1. To trichotomise X_1 the 33rd and 66th percentiles were calculated. All subjects with values of X_1 less than or equal to the 33rd percentile were assigned to level 1 and their X_1 value was set to 1. All subjects with values of X_1 greater than or equal to the 66th percentile were assigned to level 2 and their X_1 value was set to 2. Finally all subjects with values of

X_1 between the 33rd and 66th percentiles were assigned to level 3 and their X_1 value was set to 3. Hence in summary this can be written as

$$X_{1,trich} = \begin{cases} 1\ if\ X_1 \leq 33^{rd} \\ 2\ if\ X_1 \geq 66^{th} \\ 3\ if\ 33^{rd} < X_1 < 66^{th} \end{cases}$$

In scenarios 7 and 8 I subjected the variable $X_1$to a logarithmic transformation i.e.

$$X_{1,log} = log(X_1)$$

5.3 Evaluation of MFS-SNR algorithm and of variable selection in simulated non-Normal data

For each simulated scenario I aimed to evaluate the robustness of MFS-SNR via two sets of performance measures:

- the frequency with which MFS-SNR selected correctly the discriminatory variables (X_1, X_2 and X_3) and selected incorrectly the non-discriminatory variables ($X_4, \dots, X_{10}$),
- PCC and AUC.

The results for all of the scenarios are presented in Table 5.3.1.

Table 5.3.1 Variable selection frequencies and performance estimates for dichotomised and trichotomised variables for groups sizes of $n_0 = n_1 = 100$ and $n_0 = n_1 = 500$ from scenarios 1, 2, 3, 4, 5, 6, 7, and 8

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	AUC (LDA)	AUC (QDA)	PCC (LDA)	PCC (QDA)
Scenario 1 (n=100)	98	93	98	11	11	11	12	11	10	11	65	73	73	87
Scenario 2 (n=500)	100	100	100	12	13	11	13	14	14	12	65	73	74	89
Scenario 3 (n=100)	89	93	56	9	8	7	8	9	10	9	63	68	69	75
Scenario 4 (n= 500)	100	100	100	12	10	11	12	12	13	11	64	69	70	79
Scenario 5 (n= 100)	48	98	26	8	9	9	8	8	9	8	60	64	65	71
Scenario 6 (n-500)	97	100	85	7	5	7	6	7	8	8	61	67	67	77

Scenario 7 (n=100)	92	91	71	18	19	17	21	18	20	17	62	66	65	69
Scenario 8 (n=500)	97	100	96	22	21	24	22	22	22	20	61	67	66	67

First I looked at the effect of dichotomisation and trichotomisation at smaller and larger sample sizes i.e. scenarios 3, 4, 5 and 6. There are several findings from this. The variables X_3 and X_1 were more frequently selected when X_1 was dichotomised than when X_1 was trichotomised. The selection frequency for X_2 at group sizes of 100 was slightly better (higher) when X_1 was trichotomised. Selection frequencies for the non-discriminating variables $(X_4, \ldots, X_{10})$ were similar for both scenarios (when X_1 is dichotomised or trichotomised) at group sizes of 100. AUC and PCC performance estimates were higher for the dichotomised data at this group size. When the group size was increased to 500 the selection frequencies for X_1 and X_2 were similar for both scenarios however the selection frequency for X_3 was higher when X_1 is dichotomised. AUC and PCC performance estimates were also higher when X_1 was dichotomised at the larger group size of 500. Worth noting is that for group sizes of 500 the selection frequency for non-discriminating variables

Next, I evaluated the MFS-SNR algorithm across sample sizes 100 and 500 when X_1 was log-normal-transformed. This corresponds to scenarios 7 and 8. There are several findings. The AUC and PCC estimates at the smaller group size of 100 were similar to those of the dataset containing the trichotomised variable. However the selection frequencies for the non-discriminating variables were larger for the dataset containing the log-normal variable than either dataset containing the dichotomised or trichotomised variable. At the larger group size of 500 the selection frequencies of the discriminating variables X_1, X_2 and X_3 were close to those of the dataset containing the dichotomised variable however the selection frequencies for the non-discriminating variables were larger than those for either the dichotomised or trichotomised data. The PCC and AUC estimates were lower at the larger sample size than for the datasets containing the dichotomised or trichotomised variable.

Next, I investigated the performance of the MFS-SNR algorithm for multivariate normal data i.e. for scenarios 1 and 2. At the smaller group size the selection frequencies of the discriminating variables X_1 and X_3 for the untransformed data were lower than those for the dataset containing the dichotomised variable but higher than the dataset containing the log-normal transformed variable. The selection frequencies of X_3 and X_1 for the dataset containing the log-normal transformed data were higher than for the dataset containing the trichotomised variable. However, the selection frequency for X_1 was slightly lower for the dataset containing the log-normal transformed variable.

The AUC and PCC estimates for the untransformed data were larger than for any of the datasets containing transformed variables at both group sizes. Selection frequencies for both discriminating and non-discriminating variables increased slightly at the larger group size of 500.

5.4 Discussion

In this chapter, the aim was to evaluate the ability of the MFS-SNR algorithm to choose the correct discriminating variables when data are non-normally distributed. To make this problem computationally tractable and comparable to the multivariate normal case (Chapter 4) I made similar assumptions to the previous chapter (Chapter 4), but i recoded the variable X_1 into a variable that is not normally distributed using three transformations (one transformation at a time): dichotomisation, trichotomisation and log-normal transformation. The conclusions can be summarised into three points below.

My first conclusion is that at the smaller sample size of $n_0 = n_1 = 100$ for the data with X_1 being categorical (with two or three categories) or log-normal the MFS-SNR algorithm chose discriminating and non-discriminating variables less frequently compared to when the data are multivariate normal. This was expected as the transformation may have caused a loss of information. Consequently for the classification performance estimates were also lower relative to the scenario of multivariate normal data. PCC-QDA was lower by 12 % for the dataset containing the dichotomised variable, 16 % for the dataset containing the trichotomised variable and 18 % for the dataset containing the log-normal-transformed variable. PCC-LDA dropped by 3.7 % for the dataset containing the dichotomised variable, 7.8 % for the dataset containing the trichotomised variable and 7.7 % for the dataset containing the Log-Normal transformed variable. AUC-QDA estimates were lower by 0.05 for the dataset containing the dichotomised variable, 0.09 for the dataset containing the trichotomised variable and 0.07 for the dataset containing the Log-Normal transformed variable, AUC-LDA dropped by 0.02 for the dataset containing the dichotomised variable, 0.05 for the dataset containing the trichotomised variable and 0.03 for the dataset containing the Log-Normal transformed variable at the smaller group sizes. This loss of precision of variable selection and reduction in PCC and AUC was expected at the smaller sample size of $n_0 = n_1 = 100$, because as indicated above dichotomisation and trichotomisation can lead to loss of information.

The second conclusion is that at larger group sizes of $n_0 = n_1 = 500$ and for datasets containing a variable which has been subjected to dichotomising, trichotomising and log-normal transformations the performance of the MFS-SNR algorithm was approaching the same performance as when using the original multivariate normal data –in terms of variable selections and PCC and AUC. In data with

X_1 being an ordinal variable with three categories the MFS-SNR algorithm exhibited an improvement (drop) in non-discriminating variable selection frequencies. In data with X_1 being log-normal the MFS-SNR algorithm showed an improvement (increase) in the frequency of discriminating variable selection at larger sample size but also showed a worsening (increase) in the frequency of non-discriminating variable selection.

The third conclusion relates to performance of the MFS-SNR algorithm across sample sizes, i.e. between $n = 100$ vs $n = 500$. When data were multivariate normal (Chapter 4) the difference in selection frequencies and performance of MFS-SNR across different group sizes was relatively small across the sample sizes. In non-normally distributed data there was a difference in selection frequencies and performance of MFS-SNR. The largest increase in performance estimates as group sizes increase from n=100 to n=500 was seen when the datasets included a trichotomised variable, followed by the dataset containing the dichotomised variable, and then by the dataset containing the log-Normal-transformed variable. The MFS-SNR algorithm gave the worst-performance for the dataset containing the trichotomised variable. This may be due, at least in part, to the larger variance likely associated with the trichotomised variable as a result of the larger number of levels compared to the dichotomised data. The log-normal transformation causes the smallest decrease in discriminating variable selection frequency however it almost doubles the selection frequencies of the non-discriminating variables.

In summary, the results indicate that, in the considered scenarios with ordinal data (dichotomous and trichotomous), the MFS-SNR algorithm was able to make the correct variable selections. I have also demonstrated that re-coding of variables was a viable approach with the MFS-SNR algorithm. There was a loss of information after each of the transformations resulting in loss of selection performance relative to the untransformed data, which was expected. However this is in part a result of using the values for a continuous variable to assign subjects to levels (i.e. an alternative means of assigning subjects to levels might produce a distribution of levels across groups better reflecting the underlying information content of the variable - this would be expected with a well-defined ordinal variable).

In the next chapter I present a comprehensive analysis of the performance of the MFS-SNR algorithm when selecting variables from four real ophthalmic datasets.

Chapter 6. Application to real data

6.1 Introduction

In Chapters 4 and 5 computer simulations were used to investigate the performance of the MFS-SNR variable selection algorithm and to compare it to a set of existing variable selection algorithms. The advantage of computer simulations is that we can make assumptions about the relationship between the variables and groups, hence we know which variables are discriminatory and should be selected by the variable selection algorithm in the simulation. A disadvantage of computer simulations is that while simulated data are based on real data in as much as is possible they represent a particular assumed scenario and hence their generalisability to real data is questionable. Therefore in order to further characterise the performance of the selection algorithm it is necessary to apply it to real datasets, which is the aim in this chapter.

In real datasets there are several challenges in selecting the best set of variables for discrimination. Some of these challenges can be difficult to mimic in simulated data. One challenge is that the true underlying multivariate probability distribution is unknown. Hence in simulated data we make assumptions about the distribution. Missingness in real data is also a problem especially if data are not missing at random. Real data can have imbalanced groups i.e. unequal number of subjects across groups. A large number of variables are measured in many clinical studies in an attempt to identify variables relevant to the underlying disease groups. However, not all of the measured variables may be useful for assigning new observations to the appropriate groups and the presence of confounding relationships between relevant and irrelevant variables can further complicate the task of variable selection. Another challenge of real data is the complexity of multivariate correlations because correlations and variances may vary across groups. These last two challenges are the main challenges tackled in this thesis and also in this chapter.

In order to obtain a comprehensive assessment of how the new variable selection algorithm (Chapter 3, Section 3.8) works it is necessary to apply it to the task of variable selection from real datasets. Therefore this chapter will study the performance of the new algorithm in four real datasets. The four datasets were chosen to represent a spectrum of challenges in ophthalmology. Each of the datasets is unique in terms of whether there is a significant proportion of missingness, how balanced the data are between the groups of interest and the composition and type of data present i.e. if data normally distributed or not, if data are longitudinal, if data continuous or ordinal or nominal.

This chapter is structured as follows. I present the application of the novel MFS-SNR algorithm to the task of variable selection from diabetic maculopathy data (Section 6.2), diabetic retinopathy data (Section 6.3), malarial retinopathy data (Section 6.4) and keratoconus data (Section 6.4). Each of the sections 6.2-6.4 first describes the dataset, the clinical significance, relevant information on data collection, specific data challenges faced, the description of how the variable selection algorithms MFS-SNR and MFS-T2 were used, the resulting best set of variables that give the best discrimination and the evaluation of the discrimination of the selected variables. Then Section 6.5 discusses the statistical methodological and clinical findings.

6.2 Application of methods to discriminate between disease stages in diabetic retinopathy [DREFUS dataset]

6.2.1 Introduction

Here I applied the MFS-SNR algorithm (Section 3.8) to data from diabetic retinopathy (DR). DR can be split into 4 categories: diabetes and no DR, early DR, late DR and ischaemic maculopathy. This thesis used data from a clinical study where clinicians collected 27 variables on healthy patients with no DR and on patients with 4 levels of DR. The study is called *Diabetic REtinopathy: FUnctional and Structural study*, (DREFUS) (Harding et al., 2010). The ultimate clinical goal of DREFUS was to elucidate the relationships between functional and structural variables, if the relationship depends on the level of DR, and which variables (or set of variables) can best discriminate between the DR stages. This is important in clinical settings because it can help to identify the measurements that should be used to find eyes that are at risk of having DR. The current gold standard is fluorescein angiography (FA) which is used to determine the 4 stages, but this is an expensive and invasive technique. Therefore the clinical importance of DREFUS was to evaluate less invasive and less expensive measurements that could differentiate between the 4 stages of DR. This chapter will look at all the measurements and use the MFS-SNR and MFS-T2 algorithms to find the variables that best discriminate between two stages: diabetes with no DR and early DR.

6.2.2 Methods

The data that I use here come from the DREFUS study with 27 continuous variables measured on 36 patients. They are

- Functional variables (measuring function of the eye, retina or patient):

    - Multifocal electroretinogram (mfERG) measurements: mfERG Central density, mfERG Ring 2 density, mfERG Ring 3 density, mfERG Central Latency, mfERG Ring 2 latency, mfERG Ring 3 Latency
    - Oscilatory potential (OP) measurements: OP Sum amplitude, OP1 amplitude, OP2 amplitude, OP3 amplitude, OP4 amplitude, OP1 implicit, OP2 implicit, OP3 implicit, OP4 implicit, OP5 implicit, OP6 implicit, OP7 implicit, OP8 implicit and Oct CFD
    - Microperimetry (MP) measurements: MP1 Central Total, MP1 Ring 2 Total, MP1 Ring 3 Total
- Clinical variables: Cholesterol, HbA1c (glycated haemoblobin measured by venepuncture), Blood Pressure (BP) systolic, BP diastolic

First, univariate analyses were carried out on all the variables for each of the two groups. For each variable I calculated Hotelling's T^2 statistic, the associated p-value and the SNR. PCC estimates were calculated for each variable using both QDA and LDA with LOOCV. Mean values for each variable in the early and no DR groups were also calculated as well as standard errors associated with these mean values. The Shapiro-Wilks test was applied to each variable to test for normality. On the basis of the results of the Shapiro-Wilks test the parametric 2-sample t-test or the non-parametric Wilcoxon signed rank test was applied to each variable to study the differences in the measurements between the two groups. The results of the Shapiro-Wilks test were also used to determine whether the parametric Bartlett test or the non-parametric Fligner test (Conover *et al*, 1981) was applied to each variable to test the null hypothesis of variance-covariance matrix homogeneity across the groups. This analysis identified which variables had the strongest potential to discriminate between the two groups in a univariate context.

Then I performed multivariate variable selection using the MFS-T2 and MFS-SNR algorithm (Chapter 3) and discussed the variables selected to discriminate across the groups of no DR vs early DR.

6.2.3 Results

The results of the univariate analysis of the variables in the DREFUS dataset are presented in Table 6.2.3.1 below. Univariate analysis of the variables in the DREFUS dataset showed that the largest T^2 statistics and SNR values were associated with HbA1c, then mfERG Central density and then mfERG Ring 3 latency (T^2=7.6, 4.8 and 2.0, SNR=8.7, 4.9 and 1.8, respectively) indicating that they were the strongest discriminating variables, when considered univariately. The PCC estimates for HbA1c, mfERG central density and mfERG Ring 3 latency are all approximately 70 % (PCC-LDA=69.0, 72.2 and 69.4, PCC-QDA = 69.4, 72.2 and 69.4%, respectively). Results of Fligner or Bartlett tests on these three variables indicate that the variances are not significantly different across the two groups (p=0.5, 0.1 and 0.7). This explains why the T^2 statistics and SNR values have similar values for these three variables as well as multiple other variables in the DREFUS dataset. The variables mfERG Ring 2 density and mfERG Ring 3 density are the exceptions. The results of the Fligner test indicate that the variances of mfERG Ring 2 density and mfERG Ring 3 density are different across the two groups ($p=3x10^{-2}$ and $6x10^{-2}$). However the T^2 statistics and SNR values of mfERG Ring 2 density and mfERG Ring 3 density are less than one and the PCC estimates are 61.1 and 61.1%. These values are lower than those for HbA1c, mfERG Central density and mfERG Ring 3 latency. The results of the Wilcoxon signed rank test and the 2-sample t-test indicate that the only variables which are not significantly different across the groups are Cholesterol, BP systolic, BP diastolic, OP2 amplitude and OP3 amplitude. Excepting HbA1c, mfERG Central density and mfERG Ring 3 latency the PCC estimates for the remaining variables are all below 64% while the T^2 statistics and SNR values are all less than 2.0. Therefore using a univariate approach (and threshold of PCC≥64% and SNR≥2) the variables HbA1c, mfERG Central density and mfERG Ring 3 latency appear to the optimal selections for discrimination.

Table 6.2.3.1. Univariate analysis of all 27 variables in the early DR and no DR groups

Variable (units of measurement)	Group no DR	Group with early DR	Test of normality	Test of association with group	Test of covariance homogeneity	Discrimination			
	Mean, standard deviation	Mean, standard deviation	Shapiro-Wilks test, p-value	Wilcoxon signed rank test or 2-sample t-test, p-value	Fligner or Bartlett test, p-value	T2 statistic	PCC (LDA-CV)	SNR	PCC (QDA-CV)
Cholesterol (mmol/dl)	4.5 ± 1.1	4.3 ± 0.9	0.1	0.7	0.3	0.2	64	0.2	63.9
HbA1c (mmol/mol)	8.6 ± 1.6	7.2 ± 1.3	0.1	$7x10^{-4}$	0.5	7.6	69	8.7	69.4
BP systolic (mm/Hg)	134.8 ± 17.1	132.1 ± 16.8	0.3	0.7	1	0.2	64	0.2	61.1
BP diastolic (mm/Hg)	77.4 ± 8.6	75.2 ± 7.7	0.5	0.4	0.7	0.6	61	0.6	61.1
MP1 Central total	75.8 ± 16.8	80.3 ± 15.4	0.02	$1.7x10^{-7}$	0.6	0.6	63.9	0.7	63.9
MP1 Ring 2 total	317.7 ± 35	325.5 ± 27.2	$1.2x10^{-5}$	$1.7x10^{-7}$	0.2	0.5	63.9	0.6	63.9
MP1 Ring 3 total	358.6 ± 71.8	360.6 ± 46.3	$3.3x10^{-6}$	$1.8x10^{-7}$	0.5	0.009	61.1	0.01	61.1
mfERG Central density (nV/o2)	57.7 ± 18.5	77.1 ± 35.4	0.01	$1.8x10^{-7}$	0.1	4.8	72.2	4.9	72.2
mfERG Ring 2 density (nV/o2)	30.5 ± 10.6	33.4 ± 11.5	$2x10^{-3}$	$1.8x10^{-7}$	0.03	0.6	61.1	0.6	58.3
mfERG Ring 3 density (nV/o2)	19.2 ± 7.3	19.2 ± 6.9	$7.4x10^{-5}$	$1.8x10^{-7}$	0.06	$5x10^{-4}$	61.1	$5x10^{-4}$	61.1
mfERG Central latency (ms)	38.8 ± 2.4	38.4 ± 2.7	$4x10^{-4}$	$1.6x10^{-7}$	0.7	0.2	63.9	0.2	58.3
mfERG Ring 2 latency (ms)	33.5 ± 1.9	33.9 ± 2.0	$8x10^{-4}$	$1.5x10^{-7}$	0.4	0.5	61.1	0.5	61.1
mfERG Ring 3 latency (ms)	32.3 ± 1.5	33.1 ± 2.0	0.001	$1.5x10^{-7}$	0.7	2.0	69.4	1.8	69.4
OP Sum of amplitude	74.2 ± 31.8	76.4 ± 34.8	0.5	$8.9x10^{-16}$	0.7	0.04	63.9	0.03	61.1
OP1 amplitude	13 ± 5.1	15.5 ± 8.2	0.02	$1.8x10^{-7}$	0.5	1.3	61.1	1.2	63.9
OP2 amplitude	31.8 ± 19.7	31.4 ± 17.9	0.3	0.9	0.7	0.005	63.9	0.005	63.9
OP3 amplitude	19.2 ± 9.1	17 ± 10	0.6	0.5	0.7	0.5	63.9	0.4	61.1
OP4 amplitude	13.4 ± 16.5	12.5 ± 11.7	8.2x10-9	$1.8x10^{-7}$	1.0	0.03	61.1	0.04	61.1
OP1 Implicit time	13.6 ± 4.6	12.5 ± 3.5	$1.1x10^{-5}$	$1.6x10^{-7}$	0.7	0.6	63.9	0.8	61.1
OP2 Implicit time	18.7 ± 3.4	18.2 ± 1.6	$1.2x10^{-10}$	$1.1x10^{-7}$	1	0.2	61.1	0.5	25
OP3 Implicit time	22.7 ± 3.3	22.5 ± 1.5	$1x10^{-9}$	$1.2x10^{-7}$	0.9	0.06	61.1	0.2	38.9
OP4 Implicit time	26.6 ± 3.3	26.5 ± 1.7	$3.4x10^{-7}$	$1.5x10^{-7}$	0.9	0.01	61.1	0.02	44.4
OP5 Implicit time	31.4 ± 3.8	31.2 ± 2.8	$5.4x10^{-4}$	$1.6x10^{-7}$	0.8	0.02	61.1	0.02	61.1
OP6 Implicit time	36.4 ± 4.9	35.8 ± 4.3	0.02	$1.7x10^{-7}$	1	0.2	63.9	0.2	63.9
OP7 Implicit time	41.8 ± 5.6	41.6 ± 5.1	0.01	$1.7x10^{-7}$	0.9	0.01	63.9	0.01	63.9
OP8 Implicit time	46.1 ± 5.3	46.9 ± 5.4	0.07	$1.7x10^{-7}$	0.7	0.2	63.9	0.2	61.1
Oct CFT	275.1 ± 27.9	269.4 ± 25.2	0.1	$2.2x10^{-16}$	0.7	0.4	61.1	0.4	61.1

Then I carried out multivariate variable selection of the variables for discrimination between no DR and early DR using the MFS-T2 and MFS-SNR algorithms. The results are presented in tables 6.2.3.2 and 6.2.3.3 below. The MFS-T2 algorithm selected HbA1c and mfERG Central density with a combined PCC estimate of 75.0%. The MFS-SNR version of the selection algorithm (Table 6.2.3.3) selected both of these variables however it also selected Cholesterol from the remaining variables with a combined PCC estimate of 83.3%. The maximum PCC for the SNR-based selection was therefore 83.3 %. In summary, to discriminate between the diabetes and no DR group and the early DR group the MFS-SNR algorithm selected three variables (HbA1c and mfERG Central density and Cholesterol) achieving 83.3% PCC while MFS-T2 algorithm chose only two variables (HbA1c and mfERG Central density) achieving 75.0% PCC.

Table 6.2.3.2 Multivariate selections using the MFS-T2 algorithm for discrimination between the early DR and no DR groups

Variable selected	T2 statistic	PCC (LDA-CV)
HbA1c	7.6	69.4
mfERG Central density	11.8	75.0

Legend: LDA-CV = PCC values calculated using LDA and with leave-one-out cross-validation. PCC estimates in each row are calculated for the variable in that row and all previous rows.

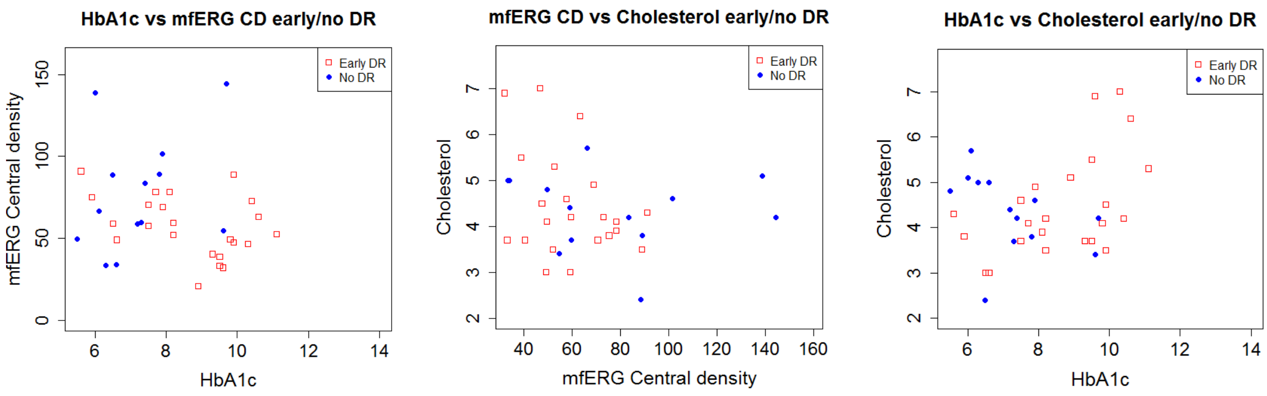
Table 6.2.3.3 Multivariate selections using the MFS-SNR algorithm for discrimination between the early DR and no DR groups

Variable selected	SNR	PCC (QDA-CV)
HbA1c	8.7	69.4
mfERG Central density	15.4	75.0
Cholesterol	19.0	83.3

Legend: QDA-CV = PCC values calculated using LDA and with leave-one-out cross-validation. PCC estimates in each row are calculated for the variable in that row and all previous rows.

In an effort to get further insight into the selected variables and the group separation achieved using these variables I created a series of bivariate plots. Bivariate plots of the variables identified by the selection algorithms as having the best discriminating ability between the early and no DR groups are presented in Figure 6.2.3.1 below. When HbA1c is plotted against mfERG central density or Cholesterol there is some separation of the early and no DR groups, with some degree of overlapping patients between the two groups. When mfERG central density is plotted against Cholesterol there is no clear separation of the early and no DR groups. It is evident that the MFS-SNR algorithm has successfully identified this performance-enhancing role of Cholesterol whereas the MFS-T2 algorithm has failed to do so. From the plots alone it is hard to understand why Cholesterol improves the discrimination so much (from 75.0 to 83.3%) hence next we looked into further analyses.

Figure 6.2.3.1 Pairwise bivariate plots of Cholesterol, mfERG Central density and HbA1c across the early DR and no DR groups



Legend: Correlation of HbA1c with Cholesterol in the early and no DR groups is 0.5 and -0.4, respectively, and these correlations are different across the groups (p=0.02, Bartlett test). mfERG CD is mfERG Central Density

What follows is an explanation of the selection of Cholesterol by the MFS-SNR algorithm despite Cholesterol having no discriminatory potential if used alone. The plot of mfERG Central density against Cholesterol shows poor separation between the group with early DR and the group with no DR. Looking at the univariate data for Cholesterol in Table 6.2.3.1 it is evident from the p-value of the 2-sample t-test (p=0.7) that there is no significant difference in the values of Cholesterol across the two groups. The p-value for the Bartlett test indicates that the variances of Cholesterol are homogeneous across the two groups (p=0.3). When considering HbA1c and Cholesterol together the p-value of the Bartlett test indicates that the variances are heterogeneous across the two groups (p=0.01). It should be noted that the Hotelling's T^2 statistic and SNR values for Cholesterol are amongst the lowest for the subset of 27 continuous variables (T^2=0.2, SNR=0.2). Looking at the univariate data for HbA1c it has the largest values for either Hotelling's T^2 statistic or the SNR. From the plots of HbA1c against Cholesterol there is some separation between the two groups. What also is also apparent is that Cholesterol is highly negatively correlated with HbA1c in the Group with Diabetes and no DR, (R=-0.4, p=0.2) and highly positively correlated with the Early DR Group (R=0.5, p=0.01), i.e. the covariance matrices differ across groups (p=2x10^{-2}, Bartlett test). What we are seeing is that Cholesterol is not associated with either of the two groups but is highly correlated with one of the discriminators HbA1c and it can enhance the discrimination achieved by HbA1c because of this correlation. This is why the final PCC estimate of the MFS-SNR selections was 83.3 % while for the MFS-T2 selections it was only 75 %. However, the correlations of Cholesterol with HbA1c differ across the groups hence the MFS-T2 algorithm did not recognise the potential of Cholesterol. This is because the calculation of Hotelling's T2 statistics uses pooled covariance matrices.

6.2.4 Discussion

I have applied the MFS-T2 and MFS-SNR algorithms to find the set of discriminatory variables for discrimination between the group with diabetes and no DR and the group with early DR. The MFS-T2 algorithm chose 2 variables (HbA1c and mfERG central density) achieving PCC of 75.0% while the MFS-SNR algorithm chose 3 variables (HbA1c, mfERG Central density and Cholesterol) achieving better discrimination with an estimated PCC of 83.3%. The selection of HbA1c and mfERG Central density is in agreement with the univariate analysis of the variables. The additional variable chosen by MFS-SNR is Cholesterol. Cholesterol was chosen despite having limited discriminatory potential by itself. It was chosen by MFS-SNR because it has a strong negative correlation with HbA1c in the group with diabetes and no DR, and it has strong positive correlation with HbA1c in the group with early DR. Hence addition of Cholesterol increases the discriminatory strength of HbA1c. MFS-T2 did not select Cholesterol, because Cholesterol's correlation varies across the groups and MFS-T2 pools the covariances. This demonstrates the advantage of MFS-SNR over MFS-T2 in this dataset.

In this study I encountered several challenges posed by the real DREFUS dataset. Real data are often not normally distributed, imbalanced in terms of group sizes and have different variances across the two disease groups. The variable selection results for the MFS-T2 and MFS-SNR algorithms show that the MFS-SNR algorithm is better able to handle data with these properties than the MFS-T2 algorithm. Looking at the variable selections for discriminating between the early DR vs no DR groups (Table 6.2.3.3) it is evident that the SNR is a superior measure of the discrimination ability of a variable than Hotelling's T^2 statistic under these conditions. In conclusion the MFS-SNR algorithm is better suited to choosing the optimal selection of variables with which to achieve discrimination between two groups in this dataset. This is also supported by considering the bivariate plots which show the group separation achieved using different pairwise combinations of variables (see Figure 6.2.3.1).

The bivariate plots and the correlation analyses in both disease groups support the selections made by the MFS-SNR algorithm but they also show that different levels of separation can be achieved using different subsets of variables. Due to MFS-SNR's superior ability to handle conditions of differing group variances the MFS-SNR algorithm is better able to identify the best discriminating variable subset from the complete set of variables. From the DREFUS data the MFS-SNR algorithm has also identified the ability of cholesterol to enhance the discriminatory ability of other variables which represents a novel clinical finding.

6.3 Application of methods to predict conversion to sight threatening diabetic retinopathy [ISDR]

6.3.1 Introduction

Another ophthalmic application where variable selection is needed is for discriminating between referable sight threatening diabetic retinopathy and non-referable sight threatening retinopathy in subjects with diabetes. DR is a progressive disease of the retina which causes blindness. Early and late stage DR is asymptomatic however, late stage DR can result in blindness if not treated. Digital photography is effective at screening for sight-threatening diabetic retinopathy (STDR). In England it is recommended that individuals with diabetes over the age of 12 are screened annually. While screening is integral to the early detection of STDR the costs of annual screening to the NHS are considerable.

I used data from a study called "Introducing personalised risk based intervals in screening for diabetic retinopathy: development, implementation and assessment of safety, cost-effectiveness and patient experience" (Harding et al, 2011) which is referred to as the ISDR study. The motivation behind the ISDR study was to develop individual risk-based screening protocols thereby eliminating the need for annual screening for those with lower risk and increasing the frequency for those with high risk.

The objective in this section was to use the variable selection algorithms MFS-T2 and MFS-SNR to identify those variables which could discriminate between referable STDR and non-referable STDR. Therefore I applied the new variable selection algorithms MFS-SNR (Section 3.8) and MFS-T2 to the task of discriminating between STDR and no STDR.

6.3.2 Methods

I used data from the observational longitudinal study performed in Liverpool called "Introducing personalised risk based intervals in screening for diabetic retinopathy (ISDR): development, implementation and assessment of safety, cost effectiveness and patient experience" (RP-DG-1210-12016) In this study patients with diabetes were invited to annual visits to the screening programme. At each visit several variables were collected. Colour fundus digital images of both retinas were taken and then graded. This grading of retinal structures was used to assign patients to a risk category. Additional data are collected from general practitioners (GPs): Cholesterol, HbA1c levels, systolic and diastolic blood pressure. Each patient was assigned a risk grade based on the progression of their diabetic retinopathy as seen from the images. The risk values associated with each stage are presented in Table 6.3.2.1 below. Patients with a risk of 2 were assigned to the referable STDR group. All other patients, (risk scores of 0, 1 or 1.5), were assigned to the non-referable STDR group.

Table 6.3.2.1 Definition of risk scores assigned to each patient and definition of the two groups used in our analyses

Risk score	Explanation	Group
0	No DR in either eye	Non-referable STDR
1	Background DR in one eye	
1.5	Background DR in both eyes	
2	Pre-proliferative DR, proliferative DR or maculopathy present	Referable STDR

The ISDR dataset available for the work in my thesis contains observations for 28 variables on 16,228 patients. Efforts were made by the ISDR research team to match information from GPs to the dates of patients' visits to the clinic. Unfortunately due to the nature of data collected in primary care there was considerable missingness within this dataset. It was assumed that only data from the last 3 years of each patients' history were needed for discrimination. Those patients (10,956) who did not have data on at least 3 years were excluded. These steps produced a dataset containing 5,198 individuals in the non-referable STDR group and 74 individuals in the referable STDR group.

Of the 5,272 patients in the dataset only 74 were in the referable STDR group. This is a large imbalance between the group sizes and would have an impact on variable selections. In order to carry out variable selection I sampled 370 patients without replacement from the non-referable STDR group using R's `sample` function (R core team). The referable STDR group contained 74 patients, 370 patients were sampled from the non-referable STDR group to ensure that the ratio of disease group patients to non-disease group patients was maintained at 5:1. I then created a data frame composed of STDR patients and non-STDR patients and passed it to the MFS-T2 and MFS-SNR algorithms for variable selection. This resampling selection procedure was carried out for 10 repeated random selections, 100 repetitions and 1,000 repetitions using the MFS-SNR and MFS-T2 algorithms. The selection frequencies for each variable in the ISDR set as well as the final PCC estimates are reported.

The final analysable dataset considered for variable selection and discrimination contains the variables:

My_LD, My_Sex, t0Age, T1risk, t2score, t0HbA1c, t1HbA1c, t1HbA1c, t0Chol, t1Chol, t1Chol, t0SP, t1SP, t2SP, t0DP, t1DP, t2DP

Table 6.3.2.2 Definition of variables in dataset for discrimination between referable and non-referable STDR

Variable name	Definition
my_Sex	Gender
t0Age	Age at diagnosis of diabetes
my_LD	Diabetes type (I or II)
t1risk	Risk score assigned at previous visit (1 year ago)
t2risk	Risk score assigned prior to last visit (2 years ago)
t0HbA1c	HbA1c at the last visit (at time 0)
t1HbA1c	HbA1c at the previous visit (1 year ago)
t2HbA1c	HbA1c at 2 years prior to last visit (2 years ago)
t0Chol – chol at visit 0	Cholesterol at the last visit (at time 0)
t1Chol – chol at visit “-1”	Cholesterol at the previous visit (1 year ago)
t2Chol – chol at visit “-2”	Cholesterol at 2 years prior to last visit (2 years ago)
t0SP – Systolic BP, at visit 0	Systolic pressure at the last visit (at time 0)
t1SP – Systolic BP, at visit “-1”	Systolic pressure at the previous visit (1 year ago)
t2SP – Systolic BP, at visit “-2”	Systolic pressure at 2 years prior to last visit (2 years ago)
t0DP – Diastolic BP, at visit 0	Diastolic pressure at the last visit (at time 0)
t1DP – Diastolic BP, at visit “-1”	Diastolic pressure at the previous visit (1 year ago)
t2DP – Diastolic BP, at visit “-2”	Diastolic pressure at 2 years prior to last visit (2 years ago)

The discrimination results depend on the prior probabilities. First I calculated sensitivity and specificity for all pairs of prior probabilities from (0.01, 0.99) to (0.99, 0.01). Then for each pair of sensitivity and specificity the value of d was calculated using the equation below. The optimal prior probabilities were then identified as those that yield the smallest value of d

(Kumar & Indrayan, 2011). Optimal priors were identified in this way for the full set of variables using a resampled non-disease group. LDA was used for calculation of optimal priors for use with the MFS-T2 algorithm. QDA was used in calculation of optimal priors for use with the MFS-SNR algorithm. The equation for the optimal priors is the following:

$$d = \sqrt{(1 - specificity)^2 + (1 - sensitivity)^2} \qquad (6.3.2.1)$$

I then carried out all further calculations and classifications using these optimal prior probability values. For the MFS-T2 algorithm all estimates use LDA. The optimal prior probabilities identified for use with the MFS-T2 algorithm were (0.64, 0.36). For the MFS-SNR algorithm all estimates use QDA. The optimal prior probabilities identified for use with the MFS-SNR algorithm are (0.65, 0.35).

The variables t1risk and t2score are ordinal variables whereas the other variables analysed in this chapter (and this dataset) are continuous. It was necessary to further investigate the role of t1risk and t2score in discriminating between STDR and no STDR. Two dummy variables where created for each of t1risk and t2score based on the risk scores 0, 1 and 1.5. The dummy variables were assigned based on the table below. t1risk was replaced with the dummy variables X_1 and X_2. t2score was replaced by the dummy variables X_3 and X_4. The dummy variables were assigned using the coding scheme presented in table 6.3.2.3 below.

Table 6.3.2.3 Coding scheme for dummy variables X1, X2, X3 and X4

	t1risk		t2score	
Risk score	X_1	X_2	X_3	X_4
0	0	0	0	0
1	1	0	1	0
1.5	0	1	0	1

Then an exploratory analysis was carried out on all the variables univariately i.e. one variable at a time. For each variable Hotelling's T^2 statistic and the associated p-values were calculated as well as the SNR. PCC estimates were calculated for each variable using both QDA and LDA with LOOCV. Mean values for each variable in the referable and non-referable ISDR groups were also calculated as well as standard errors associated with these mean values. The Shapiro-Wilks test was applied to each variable to test for normality. On the basis of the results of the Shapiro-Wilks test the parametric 2-sample t-test or the non-parametric Wilcoxon signed rank test was applied to each variable to study the differences in the measurements between the two groups. The results of the Shapiro-Wilks test were also used to determine whether the parametric Bartlett test or the non-parametric Fligner test (Conover *et al*, 1982) was applied to each variable to test the null hypothesis of variance-covariance matrix homogeneity across the groups. This analysis identified which variables had the strongest potential to discriminate between the two groups in a univariate context.

6.3.3 Results

Exploratory analysis of the ISDR data was carried out. The results are presented in Tables 6.3.3.1 and 6.3.3.2 below.

Table 6.3.3.1 Univariate analysis of all variables in the background DR and proliferative DR groups

Variable (units of measurement)	Group with Non-referable STDR	Group with Referable STDR	Test of normality	Test of association with group	Test of covariance homogeneity	Discrimination			
	Mean, standard deviation	Mean, standard deviation	Shapiro-Wilks test, p-value	Wilcoxon signed rank test or 2-sample t-test, p-value	Fligner or Bartlett test, p-value	T2 statistic	PCC (LDA-CV)	SNR	PCC (QDA-CV)
my_LD	2.0 ± 0.2	1.8 ± 0.4	2.2x10^{-16}	1.9x10^{-8}	1.8x10^{-8}	34	84.2	18.9	84.2
my_Sex	1.6 ± 0.5	1.7 ± 0.5	2.2x10^{-16}	0.3	0.3	1.2	83.3	1.2	83.3
t0Age	66.3 ± 11.2	60.6 ± 15.1	1.0x10^{-6}	0.007	0.007	14	83.8	10	83.6
t1risk	0.3 ± 0.5	0.9 ± 0.6	2.2x10^{-16}	2.2x10^{-16}	2x10^{-4}	86.4	85.1	65	85.1
t2risk	0.3 ± 0.5	0.9 ± 0.7	2.2x10^{-16}	1.5x10^{-15}	3.3x10^{-6}	75.7	82.4	57	82.4
t0HbA1c	53.6 ± 14.5	63.3 ± 19.8	2.2x10^{-16}	1.9x10^{-5}	2.3x10^{-4}	24.5	82.2	17	82.4
t1HbA1c	54.5 ± 14.5	65.7 ± 22.7	2.2x10^{-16}	2.5x10^{-6}	0.001	30	83.1	18.6	81.1
t2HbA1c	53.8 ± 14.3	62.9 ± 20.2	2.2x10^{-16}	2x10^{-4}	3.1x10^{-5}	21.4	83.1	14.6	81.5
t0Chol	4.1 ± 1.0	4.1 ± 1.0	9.9x10^{-11}	0.7	0.6	7.3x10^{-4}	83.3	7x10^{-4}	83.3
t1Chol	4.1 ± 1.0	4.2 ± 1.0	2.9x10^{-13}	0.2	0.1	1.6	83.3	1.5	82.4
t2Chol	4.2 ± 1.1	4.2 ± 1.1	3.6x10^{-14}	0.8	0.8	0.06	83.3	0.05	83.3
t0SP	130.6 ± 14.4	133.8 ± 15.6	2.7x10^{-9}	0.1	0.2	3.1	83.3	2.8	82.7
t1SP	130.7 ± 13.7	132.7 ± 12.8	1.4x10^{-6}	0.3	0.4	1.3	83.3	1.4	83.3
t2SP	132.2 ± 13.2	131.2 ± 14.9	8.0x10^{-4}	0.6	0.1	0.4	83.3	3.4	83.3
t0DP	73.2 ± 9.6	72.9 ± 9.8	5.0x10^{-4}	0.9	1	0.04	83.3	4.2	83.3
t1DP	73.7 ± 9.2	74.7 ± 9.0	9.1x10^{-5}	0.4	0.4	0.8	83.3	7.8	83.3
tt2DP	74.7 ± 8.8	73.7 ± 9.6	4.0x10^{-5}	0.3	0.7	0.8	83.3	7.3	83.3

The results of the Wilcoxon signed rank test for the variables measuring Cholesterol, Systolic and Diastolic pressure indicated that there were no significant differences in these variables across the two groups. Similarly the results of the Fligner test for these variables indicated that their variances were not significantly different across the two groups. The means and standard deviations were also very similar across both groups for these variables. PCC estimates for the variables measuring Cholesterol, Systolic and Diastolic pressure were all approximately 83 %. The T^2 statistic and SNR values were also lower for these variables compared to my_LD, t0Age, t1risk, t2risk, t0HbA1c, t1HbA1c and t2HbA1c.

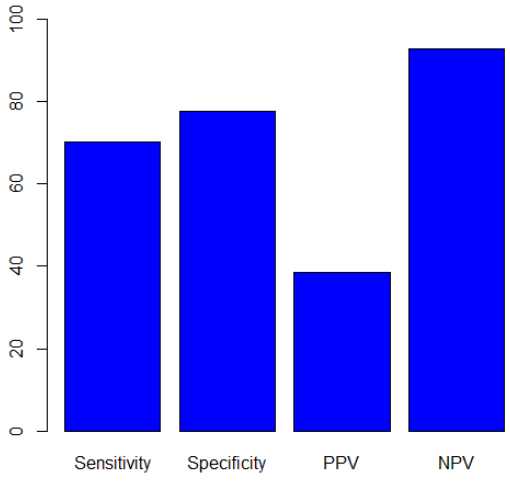
Considering the variables my_LD, my_Sex, t0Age, t1risk, t2risk, t0HbA1c, t1HbA1c and t2HbA1c the variable “my_Sex” had the lowest T^2 statistic and SNR values of 1.2. The variable t1risk had the largest T^2 statistic and SNR values of 86.4 and 65 respectively. The results of the Fligner test indicated that the variances were different across the two groups for my_LD, t0Age, t1risk, t2risk, t0HbA1c, t1HbA1c and t2HbA1c (p=1.8x10^{-8}, 0.007, 2x10^{-4}, 3.3x10^{-6}, 2.3x10^{-4}, 0.001, 3.1x10^{-5}) . my_Sex is the exception with the results of the Fligner test indicating that variances were the same for this variable across the two groups (p=0.3). The Wilcoxon test results also indicated that the variables my_LD, t0Age, t1risk, t2risk, t0HbA1c, t1HbA1c and t2HbA1c were significantly different across the two groups (p=1.9x10^{-8}, 0.007, 2.2x10^{-16}, 1.5x10^{-15}, 1.9x10^{-5}, 2.5x10^{-6} and 2x10^{-4}). Again, my_Sex is the exception with the result of the Wilcoxon signed rank test indicating that there was no significant difference in this variable across the two groups (p=0.3). The largest PCC estimates were also associated with t1risk (85.1 %).

Table 6.3.3.2 and Figure 6.3.3.1 below present the results for variable selections from the ISDR data carried out using the MFS-T2 algorithm for 1,000 samples taken without replacement. The optimal priors used with these selections were (0.64, 0.36). The MFS-T2 algorithm selected only one variable which is t1risk. The PCC estimates calculated for 10, 100 and 1,000 samplings were 76.5, 76.6 and 76.3 respectively. t1risk was identified as the most important variable for predicting a patient's progression to a risk score of 2. At all three sampling frequencies t1risk was chosen in almost 100 % of samplings. Though the selection frequencies for variables other than t1risk were higher than those for the MFS-SNR algorithm these frequencies were still negligible relative to those for t1risk. Similar patterns to the sensitivity, specificity, PPV and NPV values were observed. Again these results were suspected to be caused by the larger proportion of non-disease patients present in the data.

Table 6.3.3.2 Selection frequencies for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, frequencies are only shown for selected variables

my_LD	t1risk	t2risk	t1HbA1c
22	977	24	2

Figure 6.3.3.1 Average diagnostic estimate values for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset

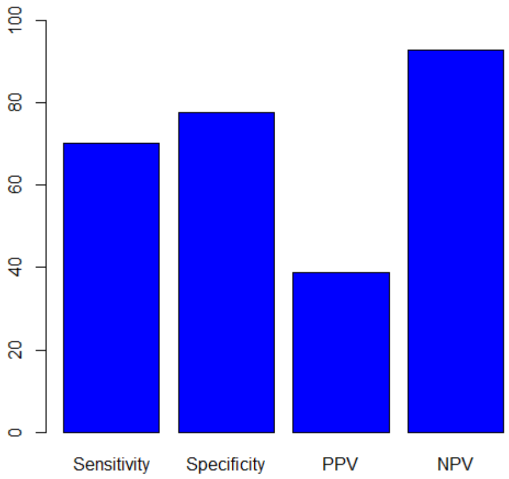


Next, I used the multivariate MFS-SNR algorithm to find out what variables offer the best discrimination between the non-referable STDR group and the referable STDR group. Table 6.3.3.3 and Figure 6.3.3.2 below present the results for 1,000 samples taken without replacement. The optimal priors used with these selections were (0.65, 0.35). The PCC estimates calculated for 10, 100 and 1,000 samplings were 76.5, 75.9 and 76.4 respectively. It is clear from the selection frequencies that t1risk is the most important variable for predicting a patients' progression to a risk score of 2. At all three sampling frequencies t1risk was chosen in almost 100 % of samplings. Other variables were chosen with marginally increasing frequencies as the number of sampling repetitions increased but these frequencies were negligible relative to those of t1risk. While the sensitivity, specificity and NPV values were all above 70 % the PPV values were all less than 40 %. This was in large part due to the larger proportion of non-disease patients present in the data.

Table 6.3.3.3 Selection frequencies for variable selections using MFS-SNR made from each of 1,000 samples taken without replacement, frequencies are only shown for selected variables

my_LD	t1risk	t2risk	t1HbA1c	t0SP
13	1000	8	3	1

Figure 6.3.3.2 Average diagnostic estimate values for variable selections using MFS-SNR made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset



The variables t1risk and t2risk were then replaced with the dummy variables X_1, X_2 and X_3, X_4 respectively. Table 6.3.3.4 and Figure 6.3.3.3 below present the variable selections using the MFS-T2 algorithm on the dataset altered to include dummy variables X_1, X_2, X_3 and X_4 for 1,000 samples taken without replacement. The dummy variable X_2 was chosen in almost 100 % of resamplings. While the selection frequencies for other variables were slightly higher than those of the MFS-SNR algorithm they were still negligible compared to the frequencies for X_2. PCC estimates for 10, 100 and 1,000 resamplings were 84.8, 85.1 and 85.1 respectively.

Table 6.3.3.4 Selection frequencies for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, frequencies are only shown for selected variables

my_LD	X_1	X_2	X_4
112	1	977	23

Figure 6.3.3.3 Average diagnostic estimate values for variable selections using MFS-T2 made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset

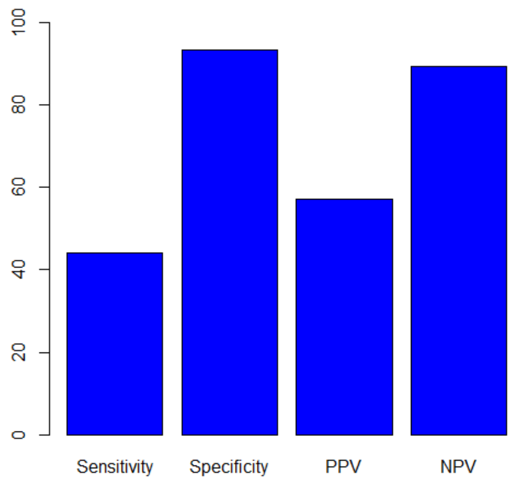
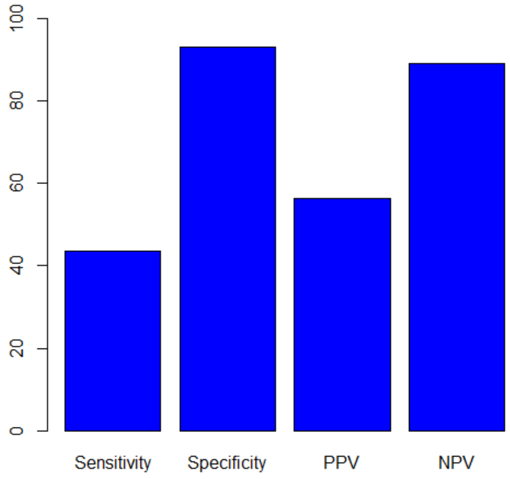


Table 6.3.3.5 and Figure 6.3.3.4 below present the variable selections using the MFS-SNR algorithm on the dataset altered to include dummy variables X_1, X_2 and X_3, X_4 for 1,000 samples taken without replacement. In terms of selection frequencies similar results were observed with X_2 being chosen in almost 100 % of samplings. Given the coding scheme used a value of 1 for X_2 is what differentiates risk scores of 0 and 1 from 1.5. While it is possible for a patient with a risk score of 1.5 to devolve to a lower risk score, the probability that they will progress to a risk score of 2 is higher than for individuals with risk scores of 0 or 1. As such X_2 contains similar information to T1risk which explains its' higher frequency of selection. PCC estimates for 10, 100 and 1,000 resamplings were 85.3, 84.9 and 85 respectively.

Table 6.3.3.5 Selection frequencies for variable selections using MFS-SNR made from each of 1.000 samples taken without replacement, frequencies are only shown for selected variables

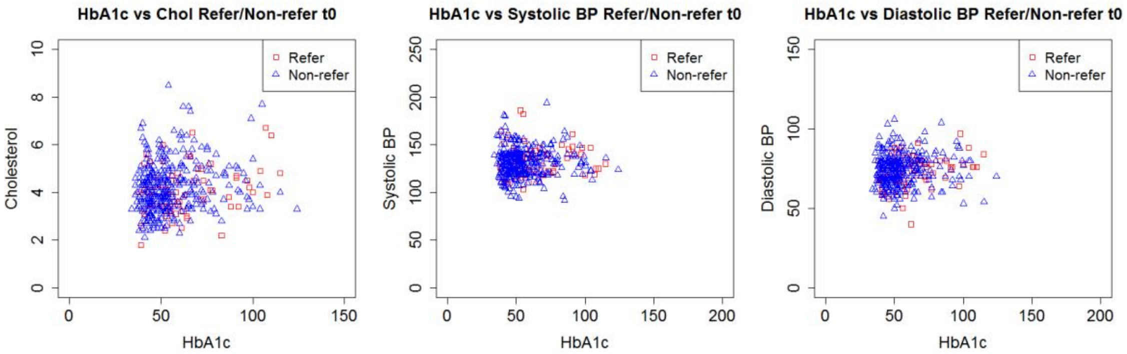
My_LD	X_1	X_2	X_3	X_4
14	12	983	2	17

Figure 6.3.3.4 Average diagnostic estimate values for variable selections using MFS-SNR made from each of 1,000 samples taken without replacement, estimates were calculated for selections made after each sampling of the full dataset



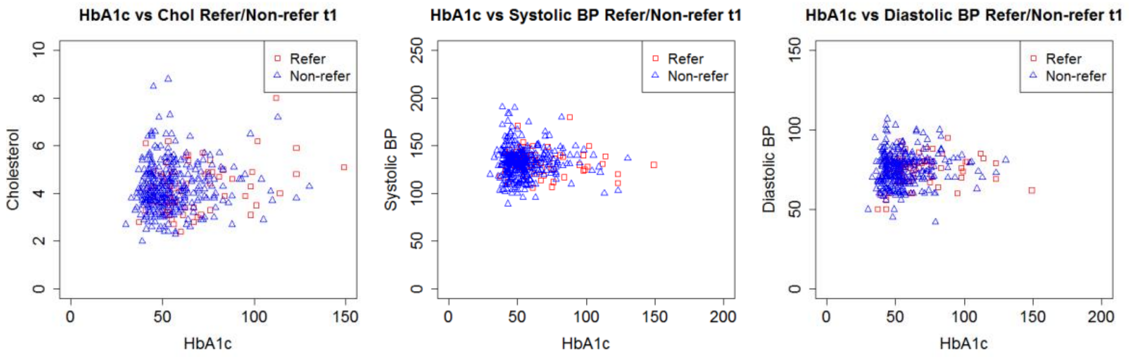
In order to understand why none of the HbA1c variables were selected by the MFS-SNR or MFS-T2 algorithms pairwise plots of HbA1c against cholesterol, systolic and diastolic blood pressure were prepared for each time point. These plots are presented in figures 6.3.2.1, 6.3.2.2 and 6.3.2.3 below.

Figure 6.3.3.1 Pairwise plots of HbA1c against Cholesterol, Systolic BP and diastolic BP for the background and proliferative groups at t0.



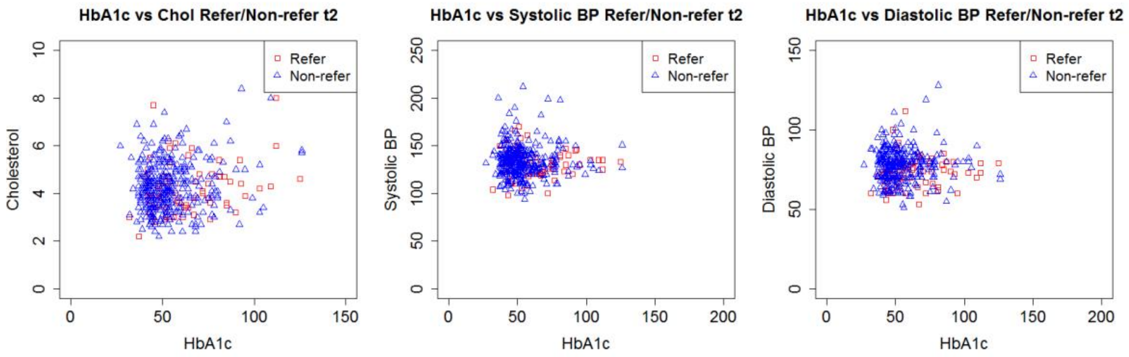
Legend: t0 is the time of each patient's last visit. Refer is the group with referable STDR, non-refer is the group with non-referable STDR.

Figure 6.3.3.2 Pairwise plots of HbA1c against Cholesterol, Systolic BP and diastolic BP for the background and proliferative groups at t1



Legend: t1 is 1 year prior to t0. Refer is the group with referable STDR, non-refer is the group with non-referable STDR.

Figure 6.3.3.3 Pairwise plots of HbA1c against Cholesterol, Systolic BP and diastolic BP for the background and proliferative groups at t2



Legend: t2 is 2 years prior to t0. Refer is the group with referable STDR, non-refer is the group with non-referable STDR.

It is evident from these plots (Figures 6.3.3.1, 6.3.3.2 and 6.3.3.3) that there is no clear separation between the background and proliferative groups in any of the plots or at any of the time points. This is in agreement with the results presented in Tables 6.3.3.2 to 6.3.3.5 in that the plots indicate the absence of a statistically significant difference for any of these variables across the two groups.

6.3.4 Discussion

I applied the MFS-T2 and MFS-SNR algorithms to the task of finding the optimal set of variables for differentiating between groups with non-referrable STDR and referrable STDR. Both the MFS-T2 and

MFS-SNR algorithms identified t1risk as the most important variable for discriminating between the two groups. This is in agreement with the results of the univariate analysis of the variables. In the univariate analysis the largest T^2 statistic and SNR are calculated for t1risk.

The imbalance between the disease and non-disease groups in the ISDR dataset complicated variable selection from this dataset. Sampling of non-disease patients in a ratio of 5:1 non-disease: disease addressed this problem to some extent facilitating the generation of a more balanced dataset from which variable selection could be carried out. I performed repeated sampling which produced bootstrapped variable selections. Unfortunately the estimates of sensitivity, specificity, NPV and PPV calculated have been influenced heavily by the imbalance in the two groups. In particular the sensitivity and PPV values fell below 50 % for the analysis of data with and without dummy variables replacing t1risk and t2risk respectively. Due to the fact that there are many times more patients in the non-referrable STDR group any incorrect classifications of patients with referable STDR have a larger impact on the calculated sensitivity and PPV values. The selection of t1risk as being the optimal variable for discriminating between the disease and non-disease groups is supported by univariate analysis of all the variables present in the dataset and the differences in these variables across the groups. Differences in the performance of the MFS-SNR and MFS-T2 algorithms are negligible for this dataset (PCC difference is 0 %).

6.4 Application of methods to discriminate between survival outcome groups in malarial retinopathy

6.4.1 Introduction

I applied the variable selection algorithms MFS-T2 and MFS-SNR to the task of discriminating between children who died after contracting cerebral malaria and children who survived with full recovery or survived with sequelae using data from two work packages in a Wellcome Trust funded Programme Grant entitled, “The retinal microvasculature in cerebral malaria in African children (MRet).” SP Harding, RS Heyderman, AG Craig, PS Hiscott, ME Molyneux, TE Taylor, S Kampondeni, NAV Beare, P Knox, M Mallewa, Y Zheng. (092668/Z/10/Z).

There are two outcome groups in this dataset :

- patients who survived with full recovery or with neurological sequelae. These are referred to as “survived”,
- patients who died.

I attempted to use the MFS-T2 and MFS-SNR algorithms to find the best set of variables for discriminating between the two outcome groups.

6.4.2 Methods

The MRet imaging dataset comprises measurements on 79 variables for 154 patients. The variables are

- 72 imaging variables of capillary non-perfusion (CNP). Each image was divided into 48 sectors and information contained in each sector was extracted. This produced the variables:
  - ○ $CNP_1..CNP_{24}$ variables give CNP in 24 sectors of macula,
  - ○ $CNP_{25}..CNP_{48}$ variables give CNP in 24 sectors of early periphery,
  - ○ $CNP_{49}..CNP_{72}$ variables were created as a sum of macula and periphery. So for example $CNP_1 + CNP_{25} = CNP_{49}$.
- Clinical variables: In addition to the imaging sector variables 7 clinical variables were measured in each patient:
  - ○ Age
  - ○ Sex
  - ○ Weight
  - ○ Serum lactate
  - ○ Respiratory data

One challenge of the dataset is that there are missing values. The missingness is particularly large in imaging data due to missing part of the image. This was due to the challenges in imaging comatose children with roving eye movements. Only 10 (out of 154) children had complete data on all imaging and clinical variables.

This missingness complicated the calculation of the optimal prior probabilities for the MRet dataset. In the presence of missingness it was not possible to train a classifier using LDA or QDA, because of the implicit assumption that each patient has complete data which is how they are implemented in R. The only alternative offered by the LDA or QDA functions was to limit the analysis to complete data i.e. omit any patient with at least one missing value. Unfortunately using only the complete cases to train the classifier was not a viable option as this left us with data on only 10 children which was insufficient.

To obtain the optimal prior probabilities, another option was to carry out imputation of missing values. Any imputation is based on assumptions drawn from the non-missing data. If the optimal priors are calculated based on a subset of the variables found in the dataset those priors will favour that variable subset during the variable selection process.

With this limitation in mind, to obtain the prior probabilities, I carried out the imputation of the full dataset using the `norm` and `norm.predict` functions from the `mice` package in an attempt to identify the optimal priors. Attempts to identify the optimal priors using QDA failed due to an insufficient number of patients in the death group. Optimal priors were identified using LDA however these priors are limited in applicability to the MFS-T2 algorithm (which uses LDA to estimate PCC values). The optimal priors identified using LDA were applied to the task of variable selection using the MFS-T2 algorithm. Variable selections from the imputed dataset were also carried out using the MFS-T2 and MFS-SNR algorithms.

Additional statistics were extracted from the imaging data as potential discriminators. From the available data on sectors 1 to 24 the averages, variances, maximum values, minimum values and ranges were extracted. Variable selection from these statistics was carried out using the MFS-T2 and MFS-SNR algorithms. Variable selection was also carried out from datasets containing combinations of these statistics, the clinical variables and sectors 1 to 24 using the MFS-T2 and MFS-SNR algorithms.

First, univariate analyses were carried out on all the variables for each of the two groups. For each variable I calculated Hotelling's T^2 statistic the associated p-value and the SNR. PCC estimates were calculated for each variable using both QDA and LDA with LOOCV. Mean values for each variable in the survival and death groups were also calculated as well as standard errors associated with these mean values. The Shapiro-Wilks test was applied to each variable to test for normality. On the basis of the results of the Shapiro-Wilks test the parametric 2-sample t-test or the non-parametric Wilcoxon signed rank test was applied to each variable to study the differences in the measurements between the two groups. The results of the Shapiro-Wilks test were also used to determine whether the parametric Bartlett test or the non-parametric Fligner test (Conover *et al*, 1982) was applied to each variable to test the null hypothesis of variance-covariance matrix homogeneity across the groups. This analysis identified which variables had the strongest potential to discriminate between the two groups in a univariate context.

6.4.3 Results

Exploratory analysis of the MRet data was carried out. The differences between the variables across the survival and death groups were analysed. The results of this analysis are presented in Table 6.4.3.1 below.

Table 6.4.3.1 Univariate analysis of all variables in the survival and death groups.

Variable (units of measurement)	Survival group	Death group	Test of normality	Test of association with group	Test of covariance homogeneity	Discrimination			
	Mean, standard deviation	Mean, standard deviation	Shapiro-Wilks test, p-value	Wilcoxon signed rank test or 2-sample t-test, p-value	Fligner or Bartlett test, p-value	T2 statistic	PCC (LDA-CV)	SNR	PCC (QDA-CV)
Age	41.2 ± 22.2	50.8 ± 30.2	5x10^{-7}	0.2	0.2	3.4	84.4	2.3	81.8
Sex	1.5 ± 0.5	1.4 ± 0.5	2.2x10^{-16}	0.4	0.5	0.8	84.4	0.8	84.4
Resp	46.2 ± 11.7	46.9 ± 11.8	3.2x10^{-6}	0.7	0.7	0.1	84.4	0.1	83.8
Wt	12.6 ± 3.9	13.5 ± 4.3	6.1x10^{-5}	0.3	0.8	1.1	84.4	0.9	84.4
Comasc	1.5 ± 0.6	1.1 ± 0.7	4.3x10^{-15}	0.007	0.3	8.2	84.4	7.3	79.9
Ahct	19.2 ± 5.9	19.4 ± 5.9	0.03	0.9	0.4	0.03	84.4	0.03	84.4
Admlact	7.0 ± 4.6	9.9 ± 3.5	1.2x10^{-6}	0.002	0.2	8.9	84.4	13.5	84.4
CNP_{1}	0.1 ± 0.07	0.1 ± 0.05	2.9x10^{-5}	0.7	0.3	1.1x10^{-4}	85.2	2x10^{-4}	85.2
CNP_{2}	0.1 ± 0.06	0.1 ± 0.05	4x10^{-4}	0.4	0.1	0.6	84.5	0.9	84.5
CNP_{3}	0.1 ± 0.06	0.1 ± 0.04	2x10^{-4}	0.2	0.06	1.3	84.9	2.1	84.9
CNP_{4}	0.1 ± 0.06	0.1 ± 0.04	0.03	0.4	0.02	0.8	85.4	1.6	85.4
CNP_{5}	0.2 ± 0.07	0.2 ± 0.05	0.007	0.7	0.03	1.6x10^{-5}	84.6	2.5x10^{-5}	84.6
CNP_{6}	0.2 ± 0.08	0.2 ± 0.06	2x10^{-4}	1.0	0.7	0.07	84.6	0.09	84.6
CNP_{7}	0.2 ± 0.09	0.2 ± 0.06	2x10^{-9}	0.8	0.4	0.3	85.1	0.5	85.1
CNP_{8}	0.2 ± 0.08	0.2 ± 0.06	0.001	0.8	0.3	0.1	86.7	0.2	86.7
CNP_{9}	0.2 ± 0.09	0.2 ± 0.06	1.9x10^{-5}	0.5	0.3	0.5	85.7	0.9	85.7
CNP_{10}	0.2 ± 0.1	0.3 ± 0.1	4x10^{-5}	0.3	0.7	1.5	84.6	1.2	83.8
CNP_{11}	0.2 ± 0.1	0.3 ± 0.1	9x10^{-4}	0.05	0.6	3.0	83.3	2.9	83.3
CNP_{12}	0.2 ± 0.1	0.3 ± 0.1	2x10^{-4}	0.4	0.2	0.2	85.7	0.4	85.7
CNP_{13}	0.2 ± 0.1	0.2 ± 0.1	2x10^{-5}	0.2	0.2	0.4	83.6	0.8	83.6
CNP_{14}	0.2 ± 0.1	0.2 ± 0.1	7x10^{-4}	0.6	0.6	0.2	83.6	0.2	83.6
CNP_{15}	0.2 ± 0.1	0.2 ± 0.1	0.001	0.06	0.8	2.5	83.5	2.9	83.5
CNP_{16}	0.2 ± 0.1	0.2 ± 0.1	1.1x10^{-7}	0.1	1.0	1.4	84.1	1.6	84.1
CNP_{17}	0.2 ± 0.1	0.2 ± 0.1	1x10^{-7}	0.8	0.7	0.1	83.8	0.1	83.8
CNP_{18}	0.2 ± 0.1	0.2 ± 0.1	1.2x10^{-7}	0.6	0.5	0.3	84.4	0.3	83.7
CNP_{19}	0.2 ± 0.1	0.2 ± 0.1	3.6x10^{-6}	0.6	0.4	0.2	84.6	0.2	84.6
CNP_{20}	0.2 ± 0.1	0.1 ± 0.06	1.4x10^{-7}	0.5	0.06	0.5	84.1	0.8	84.1
CNP_{21}	0.2 ± 0.1	0.2 ± 0.1	5.9x10^{-5}	0.7	0.09	0.7	84.5	1.0	84.5
CNP_{22}	0.2 ± 0.1	0.2 ± 0.1	2.8x10^{-7}	0.5	0.03	1.3	84.6	2.6	84.6
CNP_{23}	0.2 ± 0.1	0.1 ± 0.06	0.05	0.8	0.4	0.2	84.6	0.2	84.6
CNP_{24}	0.1 ± 0.05	0.1 ± 0.04	3x10^{-4}	0.8	0.7	0.4	84.1	0.6	84.1
CNP_{25}	0.06 ± 0.04	0.06 ± 0.04	2.5x10^{-5}	0.8	0.7	0.04	80.8	0.04	80.8
CNP_{26}	0.06 ± 0.04	0.07 ± 0.05	5.3x10^{-8}	0.3	1.0	1.3	82.8	1.0	80.6
CNP_{27}	0.05 ± 0.05	0.05 ± 0.03	2.8x10^{-11}	0.3	0.4	0.2	82.6	0.2	82.6
CNP_{28}	0.04 ± 0.04	0.06 ± 0.04	5.5x10^{-7}	0.1	0.9	1.7	82.3	1.7	80.6
CNP_{29}	0.08 ± 0.07	0.06 ± 0.04	2.8x10^{-9}	0.5	0.5	1.2	84.6	2.4	84.6
CNP_{30}	0.09 ± 0.08	0.09 ± 0.07	9.8x10^{-11}	0.7	0.9	0.06	84.7	0.06	84.7
CNP_{31}	0.1 ± 0.08	0.1 ± 0.1	1.3x10^{-9}	0.2	0.6	0.5	84.1	0.4	83.2
CNP_{32}	0.1 ± 0.08	0.1 ± 0.08	4.1x10^{-9}	0.9	1.0	0.02	84.8	0.02	84.9
CNP_{33}	0.1 ± 0.08	0.1 ± 0.05	1x10^{-6}	0.3	0.2	0.06	86.9	0.2	86.9
CNP_{34}	0.1 ± 0.08	0.1 ± 0.05	2.x10^{-4}	0.7	0.1	0.002	86.6	0.004	86.6
CNP_{35}	0.1 ± 0.09	0.2 ± 0.1	1x10^{-4}	0.006	0.9	8.0	85.0	7.8	83.3
CNP_{36}	0.1 ± 0.08	0.2 ± 0.1	0.002	0.05	0.4	4.3	87.7	3.3	87.7
CNP_{37}	0.1 ± 0.07	0.2 ± 0.1	0.001	0.02	0.1	9.0	82.9	6.0	84.2
CNP_{38}	0.1 ± 0.08	0.1 ± 0.07	5.9x10^{-5}	0.03	0.8	3.1	83.0	3.9	84.0

CNP_{39}	0.1 ± 0.08	0.1 ± 0.06	1.9x10^{-9}	0.3	0.6	0.2	83.7	0.3	83.7
CNP_{40}	0.1 ± 0.06	0.1 ± 0.08	2.8x10^{-6}	0.2	0.4	3.4	80.7	2.4	82.6
CNP_{41}	0.1 ± 0.07	0.1 ± 0.09	1.6x10^{-10}	0.4	0.2	1.6	80.8	11.1	80.0
CNP_{42}	0.08 ± 0.06	0.07 ± 0.05	9.6x10^{-10}	1.0	0.5	0.2	81.9	0.3	81.9
CNP_{43}	0.06 ± 0.05	0.07 ± 0.05	7.9x10^{-6}	0.4	0.8	0.5	81.5	0.5	80.8
CNP_{44}	0.05 ± 0.05	0.07 ± 0.06	7.4x10^{-12}	0.07	0.3	2.1	81.2	1.9	80.5
CNP_{45}	0.06 ± 0.05	0.07 ± 0.05	4.8x10^{-7}	0.4	0.7	0.1	81.3	0.2	81.3
CNP_{46}	0.06 ± 0.05	0.08 ± 0.05	3x10^{-6}	0.1	0.6	1.5	79.0	1.5	77.1
CNP_{47}	0.004 ± 0.006	0.003 ± 0.005	1.3x10^{-12}	1.0	0.3	0.2	78.4	0.2	78.4
CNP_{48}	0.004 ± 0.007	0.006 ± 0.008	2.2x10^{-12}	0.4	0.05	0.9	77.1	0.8	74.0
CNP_{49}	0.09 ± 0.05	0.1 ± 0.04	1.3x10^{-5}	0.3	0.9	0.6	80.9	0.8	81.7
CNP_{50}	0.1 ± 0.04	0.1 ± 0.04	0.008	0.9	0.3	0.005	83.1	0.005	83.1
CNP_{51}	0.1 ± 0.04	0.1 ± 0.04	5x10^{-4}	0.9	0.3	0.09	83.5	0.1	83.5
CNP_{52}	0.1 ± 0.04	0.1 ± 0.03	0.06	1.0	0.06	0.02	83.3	0.03	83.3
CNP_{53}	0.1 ± 0.06	0.1 ± 0.03	5.9x10^{-6}	0.5	0.06	0.1	84.0	0.3	84.0
CNP_{54}	0.1 ± 0.07	0.1 ± 0.06	9.2x10^{-6}	0.6	0.7	0.007	84.5	0.009	84.4
CNP_{55}	0.1 ± 0.07	0.1 ± 0.06	1.1x10^{-5}	0.7	0.5	0.3	83.8	0.3	83.8
CNP_{56}	0.1 ± 0.06	0.1 ± 0.06	2.5x10^{-5}	0.7	1.0	0.07	84.7	0.08	84.7
CNP_{57}	0.1 ± 0.06	0.1 ± 0.04	0.2	0.8	0.2	0.02	86.6	0.04	86.6
CNP_{58}	0.2 ± 0.07	0.2 ± 0.05	0.5	0.2	0.2	1.1	86.4	2.0	86.4
CNP_{59}	0.2 ± 0.08	0.3 ± 0.1	0.006	0.02	0.2	7.6	84.5	5.0	82.8
CNP_{60}	0.2 ± 0.1	0.2 ± 0.1	0.4	0.06	0.9	5.2	87.7	4.7	86.2
CNP_{61}	0.2 ± 0.1	0.2 ± 0.1	2.0	0.03	0.5	5.2	82.7	4.9	80.0
CNP_{62}	0.2 ± 0.1	0.2 ± 0.04	9x10^{-4}	0.3	0.2	0.4	84.0	0.8	84.0
CNP_{63}	0.1 ± 0.07	0.2 ± 0.1	3.5x10^{-5}	0.07	0.9	1.7	82.4	2.3	83.3
CNP_{64}	0.1 ± 0.06	0.2 ± 0.1	4.8x10^{-5}	0.03	0.2	6.5	81.5	4.9	81.5
CNP_{65}	0.1 ± 0.07	0.2 ± 0.1	1.6x10^{-9}	0.4	0.7	0.5	82.3	0.4	80.5
CNP_{66}	0.1 ± 0.06	0.1 ± 0.06	4x10^{-9}	0.7	0.2	0.08	82.8	0.08	82.0
CNP_{67}	0.1 ± 0.05	0.1 ± 0.05	3.1x10^{-6}	0.7	1.0	0.2	82.4	0.2	81.6
CNP_{68}	0.1 ± 0.05	0.1 ± 0.05	5.9x10^{-5}	0.6	0.5	0.4	81.9	0.4	81.1
CNP_{69}	0.1 ± 0.05	0.1 ± 0.05	0.002	0.3	0.6	0.4	81.4	0.4	81.4
CNP_{70}	0.1 ± 0.05	0.1 ± 0.05	0.5	0.4	0.8	0.9	79.2	0.9	79.2
CNP_{71}	0.08 ± 0.03	0.08 ± 0.04	0.5	0.6	0.3	0.4	78.6	0.4	78.6
CNP_{72}	0.06 ± 0.02	0.06 ± 0.03	0.4	0.4	0.4	0.8	76.8	0.7	74.7

The majority of the variables had homogeneous variances (Flinger test, p>0.05), except CNP_3, CNP_4, CNP_5, CNP_{20}, CNP_{21}, CNP_{22}, CNP_{48}, CNP_{52} and CNP_{53}. The majority of the variables were not associated with outcome (Wilcoxon rank sum test, p>0.05) except Comasc, Admlact, CNP_{11}, CNP_{15}, CNP_{35}, CNP_{36}, CNP_{37}, CNP_{38}, CNP_{44}, CNP_{59}, CNP_{50}, CNP_{61}, CNP_{63} and CNP_{64}.

In univariate analysis the discriminatory strength of all variables was between 76 and 86 %, i.e. for each variable separately. Considering the SNR values the largest value was associated with Admlact followed by (in descending order), CNP_{41}, CNP_{35} and Comasc (SNR=13.5, 11.1, 7.8 and 7.3). Considering the T^2 statistics the largest value was associated with CNP_{37} followed by (in descending order) Admlact, Comasc and CNP35 (T^2=9.0, 8.9, 8.2, 8.0).

Next I attempted a multivariate analytical approach to variable selection. The aim was to find the selection of variables that best discriminates between death and survival of retinal malaria patients. Tables 6.4.3.2 and 6.4.3.3 below present the results for variable selections made from the complete malarial retinopathy imaging dataset using the MFS-T2 and MFS-SNR algorithms respectively. The selections made by the MFS-T2 algorithm had an associated PCC estimate of 89.5% compared to a PCC estimate of 84.4% for the selections made using the MFS-SNR algorithm. However the MFS-T2 algorithm selected the sector variable CNP_{37}, admlact and resp to achieve this PCC. In contrast the MFS-SNR algorithm selected only Admlact.

Table 6.4.3.2 Multivariate selections by the MFS-T2 algorithm for discriminating between the death and survival groups.

Variable	T2 statistic	PCC (LDA-CV)
CNP_{37}	9.0	82.9
Admlact	22.7	85.5
Resp	31.3	89.5

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows. All 154 patients were used, 24 from the death group and 130 from the survival group.

Table 6.4.3.3 Multivariate selections by the MFS-SNR algorithm for discriminating between the death and survival groups

Variable	SNR	PCC (QDA-CV)
Admlact	13.5	84.4

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows. All 154 patients were used, 24 from the death group and 130 from the survival group.

Next I imputed the missing data and then carried out variable selection. The idea was to see how sensitive the variable selection was to missing data. Following imputation of the missing entries using the `norm` and `norm.predict` functions from the `mice` package. Missing values were imputed 5 times which is the default number for the `mice` function. Variable selection was repeated using the MFS-T2 and MFS-SNR algorithms with the imputed datasets. The results of variable selection from these datasets using the MFS-T2 and MFS-SNR algorithms are presented in Tables 6.4.3.4 and 6.4.3.5 respectively. What is interesting about the results is that following imputation both the MFS-T2 and MFS-SNR algorithms both selected Admlact only from the full imputed dataset.

Table 6.4.3.4 Multivariate selections by the MFS-T2 algorithm for discriminating between the dead and survivor groups following imputation of the malarial retinopathy dataset

Variable	T2 statistic	PCC (LDA-CV)
Admlact	8.8	84.4

Legend: The imputation was done using the norm and norm.predict functions from the mice package. PCC estimates in each row are calculated for the variable in that row and all previous rows. All 154 patients were used, 24 from the death group and 130 from the survival group.

Table 6.4.3.5 Multivariate selections by the MFS-SNR algorithm for discriminating between the death and survival groups following imputation of the malarial retinopathy dataset

Variable	SNR	PCC (QDA-CV)
Admlact	13.5	84.4

Legend: The imputation was done using the norm and norm.predict functions form the mice package. PCC estimates in each row are calculated for the variable in that row and all previous rows. All 154 patients were used, 24 from the death group and 130 from the survival group.

The imputed datasets were then used to identify the optimal priors. Optimal priors could only be identified for use with the MFS-T2 algorithm and the variable selections made using these priors are presented in Tables 6.4.3.6 and 6.4.3.7 below. While the variable selections were different for each of the datasets produced using the `norm` and `norm.predict` functions from the `mice` package Admlact is the first variable selected in each case. Also noteworthy is the fact that the changes in PCC estimates following the selection of additional variables were negligible demonstrating that the largest proportion of the discriminatory potential of the selections lies with the Admlact variable.

Table 6.4.3.6 Multivariate selections by the MFS-T2 algorithm for discriminating between the death and survival groups using optimal priors identified for the dataset produced following imputation of the malarial retinopathy dataset

Variable	T2 statistic	PCC (QDA-CV)
Admlact	8.8	84.4
Comasc	16.2	84.4
Age	20.2	84.4
CNP_{40}	25.2	84.4

Legend: The imputation was done using the norm function from the mice package. PCC estimates in each row are calculated for the variable in that row and all previous rows. All 154 patients were used, 24 from the death group and 130 from the survival group.

Table 6.4.3.7 Multivariate selections by the MFS-T2 algorithm for discriminating between the death and survival groups using optimal priors identified for the dataset produced following imputation of the malarial retinopathy dataset.

Variable	T2 statistic	PCC (LDA-CV)
Admlact	8.8	84.4
Comasc	16.2	84.4
CNP_{38}	24.2	85.7

Legend: The imputation was done using the norm.predict function from the mice package. PCC estimates in each row are calculated for the variable in that row and all previous rows. All 154 patients were used, 24 from the death group and 130 from the survival group.

To investigate the information content of the retinal image sectors a separate analysis of CNP_1 to CNP_{24} was carried out. The analysis was limited to CNP_1 to CNP_{24} as the region covered by these sectors was expected to contain a relatively larger proportion of capillary non-perfusion. Averages, variances, minimum values, maximum values and ranges were extracted for CNP_1 to CNP_{24}. Variable selection was then carried out from datasets composed of from combinations of these statistics, the clinical variables and CNP_1 to CNP_{24}. These results are presented below. CNP_{11} was selected by both the MFS-T2 and MFS-SNR algorithms from a dataset containing just CNP_1 to CNP_{24} and a dataset consisting of CNP_1 to CNP_{24} and the statistics for CNP_1 to CNP_{24}. From a dataset containing just the clinical variables Admlact was selected. Similarly Admlact alone was selected from datasets consisting of the clinical variables and statistics for CNP_1 to CNP_{24}, a dataset containing the clinical variables and CNP_1 to CNP_{24} and a dataset containing the clinical variables, the statistics for CNP_1 to CNP_{24} and CNP_1 to CNP_{24}. Lastly from a dataset containing the statistics of CNP_1 to CNP_{24} only the variances were selected. Variable selection from amongst the statistics for CNP_1 to CNP_{24} using the MFS-T2 algorithm was not possible due to the occurrence of a singularity error. These results are presented in Tables 6.3.3.8 and 6.3.3.9 below.

Table 6.4.3.8 Variable selected by MFS-T2, T² statistics and PCC estimates for selections made from combinations of image sectors 1-24, statistics extracted from sector variables and clinical variables.

Variable(s) selected	T2 statistic	PCC (LDA-CV)	Candidate variables considered in the variable selection
CNP_{11}	3	83.3	Sectors 1-24
Admlact	8.8	84.4	Sector statistics + clinical variables
Admlact	8.8	84.4	Sectors 1-24 + clinical variables
CNP_{11}	3	83.3	Sectors 1-24 + sector statistics
Admlact	8.8	84.4	Sectors 1-24+sector statistics+clinical variables

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows.

Table 6.4.3.9 Variable selected by MFS-SNR, SNR values and PCC estimates for selections made from combinations of image sectors 1-24, statistics extracted from sector variables and clinical variables.

Variable(s) selected	SNR	PCC (QDA-CV)	Candidate variables considered in the variable selection
CNP_{11}	2.9	83.3	Sectors 1-24
Admlact	13.5	84.4	Sector statistics + clinical variables
Admlact	13.5	84.4	Sectors 1-24 + clinical variables
CNP_{11}	2.9	83.3	Sectors 1-24 + sector statistics
Admlact	13.5	84.4	Sectors 1-24+sector statistics+clinical

			variables
Sector variances	0.7	84.4	Sector statistics

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows.

6.4.4 Discussion

I applied the MFS-T2 and MFS-SNR algorithms (Section 3.8) to find the variables for discriminating between groups of children who died after contracting cerebral malaria and children who survived with full recovery or survived with neurological sequelae. The MFS-SNR algorithm consistently identified Admlact as important for discriminating between the two groups. This result was repeated following imputation of missing values, analysis of different subsets of variables and variable selection from a dataset including statistics extracted from the original image variables CNP_1 to CNP_{24}. Conversely the MFS-T2 algorithm only identified the importance of Admlact following imputation of missing values.

The large proportion of missingness in the malarial retinopathy imaging data complicated efforts to carry out variable selection. After considerable work attempting to address the missingness issue with imputation as well as optimal prior identification and attempting to create new variables using the available data on CNP_1 to CNP_{24} two things became clear. Firstly, admlact is the most important variable for achieving optimal discrimination between the groups survival vs death in this dataset. Secondly the MFS-SNR algorithm identified the importance of Admlact alone without any additional editing of the dataset. In contrast the MFS-T2 algorithm selected Admlact alone only after imputation of the dataset to address the high proportion of missingness. When using the optimal prior probabilities the MFS-T2 algorithm selected several variables in addition to Admlact but the PCC did not increase by much. This is evidence of the superior ability of the SNR to assess the discriminatory potential of variables even in cases of significant missingness as the MFS-SNR algorithm made the correct selections without any additional editing or imputation of the dataset.

The univariate analysis of the malarial retinopathy dataset identified additional variables (CNP_{35}, CNP_{37}, CNP_{41} and Comasc) as potential candidates for the optimale subset of variables. The MFS-SNR algorithm selects Comasc only after imputation of missing values and does not select any of the other variables CNP_{35}, CNP_{37} or CNP_{41}. The MFS-T2 algorithm selects CNP_{37} prior to imputations. After imputation of missing values the MFS-T2 algorithm selects Comasc. These results indicate that the SNR is more robust to missingness than the T^2 statistic.

6.5 Application of methods to discriminate between disease stages in keratoconus

6.5.1 Introduction

This section applies the proposed MFS-SNR variable selection algorithm (Section 3.8) to the task of discriminating between healthy eyes and eyes with keratoconus.

The keratoconus dataset that I studied contains measurements taken on the eyes of healthy individuals and individuals with keratoconus in St. Paul's Eye Unit, corneal clinics. Each measurement was taken in triplicate for each subject. The dataset consists of measurements of 17 variables from 60 patients. The eye is the unit of analysis and from each patient, only one eye from each patient was used.

6.5.2 Methods

I created two distinct analysable datasets from the keratoconus data. The first dataset contained the first measurement taken for each variable on each individual. The second dataset contained the average of all 3 measurements of each variable for each individual. This was done to see how the precision of measurement will affect the variable selection. Both these datasets were then passed to the MFS-T2 and MFS-SNR algorithms and variable selection was carried out.

I carried out exploratory analysis on all the variables univariately i.e. one variable at a time. For each variable the Hotelling's T^2 statistic and associated p-value were calculated as well as SNR values. PCC estimates were calculated for each variable using both QDA and LDA with LOOCV. The mean values for each variable in healthy and keratoconic eye groups were also calculated as well as standard errors associated with these mean values. The Shapiro-Wilks test was applied to each variable to test for normality. On the basis of the results of the Shapiro-Wilks test the parametric 2-sample t-test or the non-parametric Wilcoxon signed rank test was applied to each variable. On the basis of the results of the Shapiro-Wilks test the parametric Bartlett test or the non-parametric Fligner test was applied to each variable to test the null hypothesis of variance-covariance matrix homogeneity across the groups.

6.5.3 Results

Exploratory analysis of the keratoconus datasets was carried out. The results of these analyses are presented in Tables 6.5.3.1 and 6.5.3.2 below.

Table 6.5.3.1 Univariate analysis of all variables in the healthy and keratoconus groups for a dataset using the averages of each variable

Variable (units of measurement)	Healthy group	Keratoconus group	Test of normality	Test of association with group	Test of covariance homogeneity	Discrimination			
	Mean, standard deviation	Mean, standard deviation	Shapiro-Wilks test, p-value	Wilcoxon signed rank test or 2-sample t-test, p-value	Fligner or Bartlett test, p-value	T2 statistic	PCC (LDA-CV)	SNR	PCC (QDA-CV)
K1 flat corneal front	42.6 ± 1.1	45.9 ± 4.7	$1.4x10^{-6}$	0.02	$1.5x10^{-5}$	14.2	77.0	72.4	80.0
K2 steep corneal front	43.4 ± 1.1	49.6 ± 5.5	$2.5x10^{-7}$	$4.1x10^{-8}$	$3.1x10^{-5}$	37.5	82.0	207.2	82.0
K mean corneal front	43 ± 1.1	47.7 ± 5	$2.4x10^{-7}$	$1.4x10^{-5}$	$2.2x10^{-5}$	25.4	82.0	136.8	78.0
Flat axis corneal front	116.2 ± 67.5	78.2 ± 58.6	$1.5x10^{-6}$	0.01	0.8	5.4	67.0	5.5	63.0
Steep axis corneal front	82.6 ± 21.8	105.1 ± 38.2	0.03	0.06	$2x10^{-3}$	7.8	65.0	10.5	65.0
Astigm corneal front	0.8 ± 0.5	3.8 ± 1.8	$1.5x10^{-5}$	$5.4x10^{-9}$	$3.2x10^{-6}$	77.7	90.0	282.4	88.0
K1 flat corneal back	-6.1 ± 0.2	-6.6 ± 1.3	$1.1x10^{-7}$	0.003	$7x10^{-7}$	4.9	78.0	61.4	83.0
K2 steep corneal back	-6.4 ± 0.2	-7.5 ± 1.0	$9.6x10^{-7}$	$1.4x10^{-7}$	$9.8x10^{-5}$	39.7	83.0	194.7	88.0
K mean corneal back	-0.2 ± 1.6	0.3 ±1.9	$2.6x10^{-13}$	$8.3x10^{-6}$	$6x10^{-4}$	1.0	53.0	1	63.0
Flat axis corneal back	105.3 ± 68.0	72.4 ± 64.0	$1.5x10^{-6}$	0.08	0.3	3.7	65.0	3.8	65.0
Steep axis corneal back	86.9 ± 13.0	102.3 ± 30.1	$4.2x10^{-5}$	0.04	0.008	6.6	62.0	12.5	63.0
Astigm corneal back	0.3 ± 0.1	0.8 ± 0.4	$2.6x10^{-6}$	$2.3x10^{-7}$	$5x10^{-5}$	40.2	87.0	140.6	85.0
Pachymetry Apex (um)	545.4 ± 26.7	462.4 ± 47.0	0.02	$9.2x10^{-9}$	0.006	70.7	83.0	95.8	85.0
Pachymetry Thinnest (um)	539 ± 28.0	450.3 ± 45.8	0.03	$2.1x10^{-9}$	0.04	82.0	83.0	103.4	87.0
Pachymetry X	-0.03 ±0.7	-0.1 ± 0.6	$1x10^{-4}$	1.0	0.4	0.1	53.0	0.1	48.0
Pachymetry Y	-0.5 ± 0.3	-0.5 ± 0.2	0.8	0.5	0.07	0.08	57.0	0.09	55.0

Note that for the keratoconus dataset there is no missingness and the groups are balanced in terms of their sizes. According to the results of the Fligner tests the variances were different across the two groups for all variables except Flat axis corneal front, Flat axis corneal back and Pachymetry X (p=0.8, 0.3 and 0.4). The Wilcoxon signed rank test results indicated that the groups were significantly different for all variables except Steep axis corneal front, Flat axis corneal back, Pachymetry X and Pachymetry Y (p=0.06, 0.08, 1.0, 0.5). While the mean values for Flat axis corneal front, Flat axis corneal back and Pachymetry X appeared to be different across the groups the standard deviations are more than 50 % of the means. The largest SNR (282.4) was associated with the variable Astigm corneal front. The largest PCC estimate (88 %) was also associated with Astigm corneal front. The mean and standard deviation values indicate that there should be minimal overlap across the two groups for the variable Astigm corneal front.

Looking at the SNR values and PCC estimates the largest values were (in descending order) associated with Astigm corneal front, k2 steep corneal front, K2 steep corneal back, Astigm corneal back and K mean corneal front (SNR=282.4, 207.2, 194.7, 140.6 and 136.8, PCC=88.0, 82.0, 88.0, 85.0 and 78.0). Looking at the T^2 statistics and PCC estimates the largest values were (in descending order) associated with Pachymetry thinnest, Astigm corneal front, Pachymetry apex, Astigm corneal back and K2 steep corneal back (T^2=82.0 77.7, 70.7, 40.2 and 39.7, PCC=83.0, 90.0, 83.0, 87.0 and 83.0).

It is worth noting that the SNR values were all larger than the corresponding Hotelling T^2 statistic values for most of the variables in this dataset. This is in keeping with expectations as the variances appeared to be significantly different across the two groups. The SNR can accommodate this heterogeneity of variances across groups whereas the Hotelling T^2 statistic cannot.

Table 6.5.3.2 Univariate analysis of all variables in the healthy and keratoconus eyes groups for a dataset using the first measurement of each variable

Variable (units of measurement)	Healthy group	Keratoconus group	Test of normality	Test of association with group	Test of covariance homogeneity	Discrimination			
	Mean, standard deviation	Mean, standard deviation	Shapiro-Wilks test, p-value	Wilcoxon signed rank test or 2-sample t-test, p-value	Fligner or Bartlett test, p-value	T2 statistic	PCC (LDA-CV)	SNR	PCC (QDA-CV)
K1 flat corneal front	52.6 ± 1.1	45.9 ±4.7	1.1x10^{-6}	1.5x10^{-3}	2x10^{-5}	12.1	73.0	57.0	77.0
K2 steep corneal front	43.4 ± 1.2	49.6 ± 5.5	2.6x10^{-7}	4.1x10^{-8}	8.2x10^{-5}	35.5	82.0	174.2	83.0
K mean corneal front	43.0 ± 1,1	47.7 ± 5.0	2.1x10^{-7}	1.4x10^{-5}	8.3x10^{-5}	23.2	82.0	113.4	80.0
Flat axis corneal front	116.2 ± 67.5	78.2 ± 58.6	1x10^{-6}	0.01	0.9	4.8	65.0	4.8	63.0
Steep axis corneal front	82.6 ± 21.8	105 ± 38.2	0.03	0.06	0.03	6.7	68.0	7.5	65.0
Astigm corneal front	0.8 ± 0.5	3.8 ± 1.8	1.1x10^{-5}	5.4x10^{-9}	1.3x10^{-6}	71.5	90.0	258.4	88.0
K1 flat corneal back	-6.1 ± 0.2	-6.6 ± 1.3	1.3x10^{-5}	0.003	4.6x10^{-7}	11.0	75.0	62.8	80.0
K2 steep corneal back	-6.4 ± 0.2	-7.5 ± 1.0	1.3x10^{-6}	1.5x10^{-7}	3x10^{-4}	35.7	83.0	148.4	83.0
K mean corneal back	-0.2 ± 1.6	0.3 ± 1.9	4x10^{-13}	8.3x10^{-6}	0.002	1.0	53.0	1.0	62.0
Flat axis corneal back	105.3 ± 68	72.4 ± 64.0	4.5x10^{-8}	0.08	0.6	2.5	63.0	2.6	63.0
Steep axis corneal back	86.9 ± 13	102.3 ± 30.1	5x10^{-4}	0.04	0.02	5.6	63.0	8.2	63.0
Astigm corneal back	0.3 ± 0.1	0.8 ± 0.4	4.2x10^{-6}	2.3x10^{-7}	1x10^{-4}	35.8	85.0	129.4	88.0
Pachymetry Apex (um)	545.4 ± 26.7	462.4 ± 47.0	0.05	9.2x10^{-9}	0.02	65.9	85.0	85.8	85.0
Pachymetry Thinnest (um)	539 ± 28	450.3 ± 45.8	0.1	5.9x10^{-12}	0.03	80.9	87.0	97.8	87.0
Pachymetry X	0.03 ±0.7	-0.1 ± 0.6	1.6x10^{-4}	1.0	0.5	0.2	53.0	0.2	50.0
Pachymetry Y	-0.5 ± 0.2	-0.5 ± 0.3	0.2	0.5	0.04	0.3	47.0	0.3	48.0

According to the results of the Fligner tests the variances were different across the two groups for all variables except flat axis corneal front, flat axis corneal back and Pachymetry X (p=0.9, 0.6 and 0.5. Similarly the Wilcoxon signed rank test results indicated that the groups were significantly different for all variables except Steep axis corneal front, Pachymetry X and Pachymetry Y (p=0.06, 0.08, 1.0 and 0.5). While the mean values for Flat axis corneal front, Flat axis corneal back and Pachymetry X appeared to be different across the groups the standard deviations were more than 50 % of the means. The largest SNR ratio (282.4) was associated with the variable Astigm corneal front. The largest PCC estimate (88 %) was also associated with Astigm corneal front. The mean and standard deviation values indicated that there should be no overlap across the two groups for the variable Astigm corneal front.

Looking at the SNR values and PCC estimates the largest values were (in descending order) associated with Astigm corneal front, k2 steep corneal front, K2 steep corneal back, Astigm corneal back and K mean corneal front (SNR=258.4, 174.2, 148.4, 129.4 and 113.4, PCC=88.0, 83.0, 83.0, 88.0 and 80.0). Looking at the T^2 statistics and PCC estimates the largest values were (in descending order) associated with Pachymetry thinnest, Astigm corneal front, Pachymetry apex, Astigm corneal back and K2 steep corneal back (t^2=80.9, 71.5, 65.9, 35.8 and 35.7, PCC= 87.0, 90.0, 85.0, 85.0 and 83.0).

As was observed for the dataset using the averages of each variable the SNR values are larger than the Hotelling's T^2 statistics for each variable. This is expected as the variances are significantly different across the two groups and only the SNR is capable of accommodating this heterogeneity leading to larger SNR values.

The MFS-SNR and MFS-T2 algorithms were used to carry out variable selection from a dataset comprised of the average values of each of the variables in the Keratoconus dataset. Tables 6.5.3.3 and 6.5.3.4 present the selection results for both the MFS-T2 and MFS-SNR algorithm from a dataset containing the average values of each of the variables in the Keratoconus dataset. At first glance it appeared that the MFS-T2 algorithm had made more parsimonious selections however, using all of the variables selected by the MFS-T2 algorithm a maximum PCC estimate of just 95 %. Using just the first 4 variables chosen by the MFS-SNR algorithm a PCC of 100 % is calculated. This is expected given that the results of the univariate analysis indicate that the variance-covariance matrices are not homogeneous across the two groups.

Table 6.5.3.3 Variable selection by the MFS-T2 algorithm from a dataset composed of the variable averages from the Keratoconus data

Variable	T2 statistics	PCC (LDA-CV)
Pachymetry thinnest	82	83.3
Astigma corneal front	128	93.3
Steep axis corneal front	169.1	93.3
PachymetryY	175.8	93.3
Pachymetry X	180.4	95

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows.

Table 6.5.3.4 Variable selection by the MFS-SNR algorithm from a dataset composed of the variable averages from the Keratoconus data

Variable	SNR	PCC (QDA-CV)
Astigma corneal front	282.4	88.3
Pachymetry Apex	357.5	91.7
K2 steep corneal back	419.6	96.7
Steep axis corneal front	516.3	100
Flat axis corneal front	566.6	100
Pachymetry Thinnest	605.1	100

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows.

The MFS-SNR and MFS-T2 algorithms were used to carry out variable selection from a dataset comprised of the first measurements for each of the variables in the Keratoconus dataset. Tables 6.5.3.5 and 6.5.3.6 below present the selection results for both the MFS-T2 and MFS-SNR algorithm from a dataset containing the first measurements for each of the variables in the Keratoconus dataset. It appeared that the MFS-T2 algorithm had made more parsimonious selections however using just the first 3 variables selected by the MFS-SNR algorithm the PCC estimate is 95 %. Using 6 of the 7 variables selected by the MFS-T2 algorithm the PCC estimate reaches 93.3 % which is still lower than the corresponding MFS-SNR estimate. This is not unexpected as the results of the univariate analysis indicate that the variance-covariance matrices are not homogeneous across the two groups which favours the use of the MFS-SNR algorithm.

Table 6.5.3.5 Variable selection by the MFS-T2 algorithm from a dataset composed of the first measurement for each variable from the Keratoconus data. PCC estimates in each row are calculated for the variable in that row and all previous rows.

Variable	T2 statistics	PCC (LDA-CV)
Pachymetry thinnest $[\mu m]$	80.9	86.7
Astigma corneal front	124.4	91.7
Steep axis corneal front	157.6	93.3
Flat axis corneal back	161	93.,3
K1 flat corneal back	164.2	93.3
K2 steep corneal back	174.2	93.3
K mean corneal front	218.4	96.7

Table 6.5.3.6 Variable selection by the MFS-SNR algorithm from a dataset composed of the first measurement for each variable from the Keratoconus data

Variable	SNR	PCC (QDA-CV)
Astigma corneal front	2587.4	88.3
Pachymetry thinnest	325.6	91.7
Steep axis corneal front	382.8	95
Flat axis corneal front	475.2	95
Flat axis corneal back	508.8	95
K1 flat corneal back	527.9	96.7
Pachymetry X	547.7	96.7

Legend: PCC estimates in each row are calculated for the variable in that row and all previous rows.

6.5.4 Discussion

The MFS-T2 and MFS-SNR algorithms were applied to the task of identifying the optimal subset of variables for discriminating between groups of healthy eyes and keratoconic eyes.

The search for the best discriminatory variables using both the MFS-T2 and MFS-SNR algorithms indicated that the average values of 3 measurements of each of the variables contained a larger amount of information for discriminating between the two groups than the first measurements of each variable. This is not surprising because by averaging we are removing noise or uncertainty due to measurement error. This was further demonstrated by the fact that both the MFS-T2 and MFS-SNR algorithms achieved higher estimates of PCC with fewer variables when making selections using the mean values of the 3 measurements. Fewer variables were selected from the dataset using the average values than those made using the first values of each variable.

Next, it was found that the measurements of Astigma corneal front, Pachymetry thinnest and Steep axis corneal front were identified as being most discriminatory by both versions of the algorithm when selecting from either the first measurements or the average of three measurements. While the MFS-SNR algorithm selected more variables or the same number of variables as the MFS-T2 algorithm larger PCC estimates were achieved with smaller numbers of variables by the MFS-SNR algorithm (see Tables 6.5.3.3-6.5.3.6).

The MFS-SNR algorithm selected a smaller set of variables than the MFS-T2 algorithm. The selections made by the MFS-SNR algorithm also had a higher PCC estimate. The results of the univariate analyses indicate that the variance-covariance matrices are not homogeneous for either the dataset based on variable averages or the dataset based on the first measurements of each variable. On this basis the results for the MFS-SNR selections were expected to be more accurate in terms of optimality for discriminating between groups. This is because the SNR is designed to accommodate heterogeneous variance-covariance matrices whereas Hotelling's T^2 statistic is not. Similarly the PCC estimates calculated using QDA by the MFS-SNR algorithm are expected to be more accurate as QDA is suited to heterogeneous variance-covariance matrices whereas LDA is not.

6.6 Conclusions

In this chapter the aim was to assess the performance of the MFS-T2 and MFS-SNR algorithms when applied to the task of variable selection from four real clinical datasets.

In the first dataset, the univariate analysis of the variables in the DREFUS dataset indicated that HbA1c and mfERG central density were important to discriminating between the group with DR and the group without DR. Both the MFS-T2 and the MFS-SNR algorithms selected HbA1c and mfERG central density. However the MFS-SNR algorithm also selected cholesterol and achieved a larger estimated PCC. Bivariate plots of the selected variables indicated that when cholesterol was paired with HbA1c better separation of groups was achieved than when cholesterol was paired with mfERG central density. These results indicate that the MFS-T2 algorithm failed to recognise the ability of cholesterol to improve the discrimination when added to HbA1c and mfERG central density. In contrast the SNR - as used by the MFS-SNR algorithm - did respond to the improved discriminatory potential arising as a result of the inclusion of the variable cholesterol. This dataset therefore highlighted the importance of the MFS-SNR algorithm over the MFS-T2 algorithm in scenarios where two phenomena occur: first, covariances differ across groups and second, there is a variable that cannot discriminate when used alone, but is highly correlated with other variables and hence can increase the discriminatory strength of those variables. The key statistical methodological finding is that the MFS-T2 algorithm failed to recognise important variable(s) that can improve classification. The key clinical finding is that cholesterol, HbA1c and mfERG central intensity is the set of variables that has the strongest discriminatory strength to differentiate between diabetic eye with no DR and diabetic eye with early DR.

The second dataset was the ISDR dataset. The results of a univariate analysis of the ISDR data indicated that the variables t1risk and t2risk were important to discriminating between the non-referrable STDR and referrable STDR groups. Both multivariate MFS-T2 and MFS-SNR algorithms selected t1risk from the ISDR dataset but did not choose t2risk due to high correlation with t1risk. The inclusion of t2risk would not provide any new information about the groups. The large imbalance between the non-referable STDR and referable STDR groups made calculating performance estimates difficult. This was the reason why all of the variables appeared to have similar performance when considering the PCC estimates from the univariate analyses. Despite this, both the MFS-T2 and MFS-SNR algorithms identified t1risk as the optimal variable for differentiating between the non-referable STDR and referable STDR groups.

The third dataset was MRet. Univariate analysis of the variables in the MRet dataset suggested variables Admlact, Comasc, CNP_{35}, CNP_{37} and CNP_{41} as potential candidates for discriminating between groups of individuals who died and individuals for survived in those diagnosed with malarial retinopathy. The MFS-SNR algorithm identified the variable Admlact as the most important variable for discriminating between the two groups. In contrast the MFS-T2 algorithm selected CNP_{37} and Admlact to achieve similar performance in terms of PCC. Following imputation of missing values, both the MFS-SNR and MFS-T2 algorithms selected Admlact only. The key statistical methodological finding in this work is that the SNR is more robust to the effects of missing data than Hotelling's T^2 statistic. The key clinical finding is that Admlact has the most discriminatory strength to didderentiate between patients who died and survived.

The fourth dataset was keratoconus. Univariate analysis of the variables in the keratoconus dataset suggested Astigm corneal front, Pachymetry thinnest, K2 steep corneal front, K2 steep corneal back, Astigm corneal back and K mean corneal front as potential candidates for discriminating between groups of individuals with healthy eyes and individuals with keratoconus. Also worth noting is that the variances were not homogeneous across the two groups for most of the variables. The MFS-T2 and MFS-SNR algorithms selected multiple variables from each of the keratoconus datasets. However the MFS-SNR selections were more parsimonious achieving higher performance estimates. The first selection by the MFS-T2 algorithm from both datasets was Thinnest um. The first selection by the MFS-SNR algorithm from both datasets was Astigm corneal front. Astigm corneal front had larger PCC estimates in the univariate analysis so it was reasonable to expect it to be the variable with the largest discriminatory potential. The fact that only the MFS-SNR algorithm selected Astigm corneal front first is a result of the SNR's ability to accommodate heterogeneous variance properties across the groups of interest.

All four datasets analysed in this chapter were selected purposefully as they cover a wide spectrum of properties. These include imbalanced group sizes, missingness and mixtures of both normal and non-normal data. Sometimes the MFS-SNR algorithm outperformed the MFS-T2 algorithm, such as in DREFUS dataset. The MFS-SNR algorithm selected Admlact from the MRet dataset. This indicates that the SNR ratio is robust to the presence of missingness in the MRet dataset. The SNR had larger values than Hotelling's T^2 statistic when analysing variables from the keratoconus dataset. This reflects the ability of the SNR to accommodate heterogeneous variance and co-variance properties across groups. This leads to a variable selection by the MFS-SNR algorithm which achieves higher PCC estimates than the corresponding MFS-T2 algorithm selections.

Chapter 7 will present a discussion of all of the results outlined in this thesis and discuss advantages and disadvantages of the proposed approach as well as highlighting possible future work.

Chapter 7. Conclusions and further work

7.1 Introduction

In classification it is of the utmost importance that new subjects are assigned to the correct groups. The reason for this is obvious in clinical settings where misclassification may result in a patient not receiving important treatment. However the correct classification of subjects is also important in non-clinical settings. The challenge is that in order to achieve accurate classification it is necessary to identify those variables which have the greatest potential for discriminating between the groups of interest. This challenge is an important and unsolved statistical problem.

Where datasets have high dimensionality the large number of variables can make traditional discriminant analysis methods impractical. The existence of high correlations between explanatory variables is also important as these correlations can enhance the discriminatory performance of variables (by including variables not directly relevant to discrimination). On the other hand the existence of high correlations between two explanatory variables and the outcome variable can make one of the explanatory variables redundant in the presence of the other. The difficulty in identifying the optimal variable subsets is also compounded by the fact that methods such as LDA make invalid assumptions about data such as whether or not it is normally distributed or whether variance-covariance matrices are homogeneous across groups.

My aim in this thesis was to review existing variable selection methods then to propose and explore a new method for variable selection in simulations and in real datasets. The MFS-SNR algorithm I have presented in this thesis is multivariate in nature, makes no assumptions about the underlying data and can accommodate high dimensional datasets without the need to exhaustively analyse every single variable or subset. A summary of the work I have completed is provided below.

7.2 Summary of findings

7.2.1 Summary of findings from literature review

As is clear from my literature review there exist a large number of variable selection methods developed for identifying those variables best suited to differentiating between groups of interest. I presented my literature review at the 10th Tartu Conference on Multivariate Statistics, June/July 2016, "Recent advances in multivariate filter methods of variable selection for discrimination".

The existing variable selection methods are split into three general categories; filter methods, embedded methods and wrapper methods. Filter methods (Section 2.3) utilise a data summary metric which attempts to estimate the discriminatory potential of a variable or set of variables. The estimated values are then used to rank variables and this ranking may be used to carry out variable selection. The principal advantage of filter methods is that they are fast because they are not evaluating large numbers of variable subsets (relative to wrapper and embedded methods). Another advantage is that the variable selections are not classifier-dependent meaning that they may be generalised for use with multiple different classifiers. The main disadvantages of some of the filter methods are that they are univariate, failing to take account of correlations between variables and they are not guaranteed to identify the optimal subset of variables for discriminating between the groups of interest (Guyon and Elisseef, 2003; Chandrashekar and Sahin, 2014; Saeys et al., 2005)

I have also reviewed wrapper methods (Section 2.4) and embedded methods (Section 2.5) for variable selection in classification. Wrapper methods operate by evaluating every possible subset of variables that may be drawn from a particular dataset. Embedded methods embed the variable selection process into the training of the classifier. Both wrapper and embedded methods have high computational requirements, considerably higher than for filter methods and they both assume a particular classifier.

From my review it became clear that the filter methods are popular for their independence of classifier and low computational need. A disadvantage of some existing filter methods is that they often do not take into account possible heterogeneity of variance-covariance matrices across the groups. The novelty in my review is that it highlighted a very important connection of filter methods to the signal processing literature, namely to signal-to-noise ratio. This then led to the proposal of a novel version of signal-to-noise ratio filter metric and to the novel version of the SNR-based forward search algorithm, which I investigated in subsequent chapters.

7.2.2 Summary of findings on the SNR metric

The proposed SNR metric is a generalisation of Hotelling's T^2 for a multivariate discriminatory scenario where variance-covariance metrics are heterogeneous. I gave a theoretical proof that SNR reduces to Hotelling's T^2 when variance-covariance matrices are homogeneous. In simulations I saw that SNR was always bigger or equal to Hotelling's T^2 except when variance-covariance matrices are homogeneous across groups, as expected. In real data, relative to Hotelling's T^2 statistic the SNR metric was more robust to issues such as missingness, imbalanced groups sizes and mixtures of data types.

The SNR assumes multivariate normality only implicitly. In other words, just like Hotelling's T^2 statistic, the SNR metric can be calculated for data that are not multivariate normal, but is expected to perform most optimally for multivariate normal data.

7.2.3 Summary of findings on implementation of the MFS-SNR algorithm

In this thesis I also studied and proposed a new stopping criterion for the MFS algorithm. The MFS-SNR algorithm operates by identifying the variable with the largest SNR value in the first round of variable selection. This variable is then added to the set of selected variables and the PCC value associated with this selection is estimated. In subsequent rounds of selection the SNR value associated with previously selected variables is calculated with each of the remaining candidate variables. The candidate variable which elicits the largest change in the SNR is then added to the set of selected variables and the PCC value associated with the new subset is estimated. The original MFS algorithm was created by Lu et al. (2005) and it uses p-value as the stopping criterion. However, my MFS algorithm (Section 3.8) does not use the p-value associated with a particular variable or set of variables as the stopping criterion. Instead I use the change in estimated PCC between each round of selection as the stopping criterion (Section 3.9.1). In simulations I observed that the stopping criteria based on p-value leads to premature termination of the selection algorithm resulting in a smaller number of selected variables. In particular, the number of selected variables depends on the sample size of the groups. This problem was rectified by using PCC as the stopping criteria.

7.2.4 Summary of findings from simulated data

I have proposed and implemented a novel SNR-based multivariate filter method of variable selection called the MFS algorithm (Section 3.8) and compared it to several existing variable selection methods. I have used simulated normal data (Chapter 4) and non-normal data (Chapter 5), as well as 4 real ophthalmological datasets (Chapter 6). These included three univariate filter methods using chi-square statistics, information gain and the Relief-F algorithm, a multivariate filter method utilising a SVM classifier and an embedded method using random forests.

The MFS-SNR method (a multivariate filter method) was better at the variable selection task than univariate filter methods, as expected, in multivariate normal simulations. All simulated datasets were composed of ten variables (Sections 4.2 & 5.2) with two variables that can discriminate when used alone, one non-discriminating variable that can enhance the performance of the discriminating variables. In simulations of normal data I found that the MFS-SNR algorithm outperformed the univariate filter methods in all 12 scenarios. For the filter methods using chi-square statistics, information gain and the Relief-F algorithm this difference in performance is attributed to the univariate nature of each of these methods. Essentially each of these methods failed to take sufficient account of correlations between the variables X_1, $X_2$and X_3.

The multivariate MFS-SNR filter method was also found to be better than the multivariate MFS-T2 filter method, in all simulated multivariate normal scenarios i.e. when variance-covariance matrices are heterogeneous. The poor performance of the MFS-T2 algorithm (relative to the MFS-SNR algorithm) can be attributed to the inability of Hotelling's T^2 statistic to accommodate heterogeneous variance-covariance matrices. The MFS-SNR algorithm is a multivariate method and so it took proper account of the correlations between the variables X_1, X_2 and X_3 and unlike Hotelling's T^2 statistic the SNR metric can accommodate heterogeneous variance-covariance matrices. Thus the superior performance of the MFS-SNR algorithm was not surprising.

I also found that MFS-SNR performed at least as well as computationally intensive methods like SVM and RF in the simulated multivariate normal scenarios. In other words MFS-SNR was comparable in terms of variable selection frequencies and performance estimates across all 12 scenarios. Though non-discriminating variable selection frequencies were generally lower for the multivariate filter SVM method I attribute this, at least in part, to the existence of a cap on the number of variables selected by this method. However, the embedded RF method and the multivariate filter SVM method took longer than the MFS-SNR method to return selections (2:36, 4:48 and 8:50, min:sec, respectively) and had greater computational requirements. All 3 methods identified the importance of X_3 in enhancing the discriminatory performance of X_1 and X_2.

A very important property of MFS-SNR is that it does not require the user to specify the number of variables to be chosen i.e. it does not require number of selected variables a priori. The number of variables was not required a priori by any of the filter methods. However this was required by SVM and RF methods.

In scenarios where the assumption of multivariate normality was violated the MFS-SNR algorithm still selected all 3 discriminating variables (although the selection frequencies fell). In the three simulated scenarios of non-normal data the MFS-SNR algorithm (Chapter 5) showed worse variable selection performance than in scenarios with normally distributed data. At group sizes of $n = 500$ the performance of the MFS-SNR algorithm was similar to when using normally distributed data. The best performance was observed for log-normal transformed data followed by dichotomised and then trichotomised data. The MFS-SNR algorithm still proved capable of identifying the importance of X_3 in addition to X_1 and X_2 following transformation of the variable X_1. I attribute the loss in performance (at least in part), observed when compared to normally distributed data, to loss of information caused by the transformation of the data.

In summary, the analysis of the performance of the MFS-SNR algorithm in each of the simulated scenarios described demonstrated that the MFS-SNR algorithm is capable of selecting those variables with the greatest discriminatory potential:

- whether data are normally distributed or not
- over a range of group sizes
- when groups are imbalanced.

The MFS-SNR algorithm achieves similar performance to the RF method without the need to analyse 5,000 variable subsets, selecting the optimal subset of variables in a quarter of the time if took the RF method. The multivariate filter SVM method had a smaller workload than the RF method however it still took nearly twice as long as the MFS-SNR algorithm to identify the optimal subset of variables. The MFS-SNR algorithm achieved similar performance to the SVM method without this computational burden as the SNR metric is capable of quantifying the discriminatory potential of a variable without training and evaluating a classifier. It is also not necessary to have separate training and validation data when using the MFS-SNR algorithm. The MFS-SNR algorithm also functions without any tuning parameters, and without a priori knowledge of the number of the selected variables. The SNR metric is multivariate so correlations between variables are considered by the MFS-SNR algorithm.

As part of the stopping criterion the user must specify the minimum PCC change they wish to see after each variable selection. However, this is part of the stopping criterion and has nothing to do with the variable selection process (i.e. the order of variable selection is not changed by altering the minimum PCC change).

In the simulated scenarios I studied it is evident that whether data are normally or not normally distributed the MFS-SNR algorithm is still capable of identifying the variables with the greatest discriminatory potential. Similarly for the real datasets that were analysed the MFS-SNR algorithm identified the optimal subset of variables from each dataset regardless of whether variables were normally distributed or not. Based on these results the SNR metric does not appear to make any assumptions about the distribution of the data (however it must be noted that this may not be generalisable to every dataset).

Lastly, the MFS-SNR algorithm does not exhaustively analyse every possible variable subset in the course of identifying the optimal subset. In the simulated scenarios the MFS-SNR algorithm achieves similar performance to the RF method despite the fact that the RF method is an embedded method which takes a brute force approach to variable selection.

7.2.5 Summary of findings from real data

Simulated data is useful for analysis in research principally because the parameters of the data can be arranged to suit the needs of the researcher. However simulated data can never fully replicate the complexity of real data. Therefore it was necessary to assess the performance of the MFS-SNR and MFS-T2 algorithms when carrying out variable selection from real datasets. Four ophthalmological datasets were used in this work, each with different types of data and missingness:

- the DREFUS dataset: Section 6.2, 27 variables, 2 groups; discrimination of early DR and no DR; data challenges of imbalanced groups and missingness,
- the ISDR dataset: Section 6.3, 28 variables, 2 groups; discrimination of referable STDR and non-referable STDR; data challenges of imbalanced groups and missingness,
- the MRet dataset: Section 6.4, 81 variables, 2 groups; discrimination of subjects with respect to outcome survival and death; data challenges of imbalanced groups and missingness,
- the Keratoconus dataset: Section 6.5, 17 variables, 2 groups; discrimination of corneas healthy and keratoconus; data had balanced groups and there was no missingness.

Each dataset was first analysed univariately to identify the optimal variables for discriminating between each group. This was then compared with the variable selection by multivariate methods. These 4 datasets covered a broad spectrum of data quality with imbalanced group sizes, varying levels of missingness and mixtures of normal and non-normal data.

First I looked into finding the best set of variables to discriminate between no and early diabetic retinopathy in subjects with diabetes, using the DREFUS dataset. Both the MFS-SNR and MFS-T2 algorithms identified HbA1c and mfERG central density as being important for discriminating between the early DR and no DR groups. However, only the MFS-SNR algorithm identified a role for cholesterol in enhancing the discriminatory performance of HbA1c and mfERG central density. The selection of cholesterol also represents a novel clinical finding. The selection of HbA1c and mfERG central density was consistent with the findings from the univariate analysis of the dataset. The identification of cholesterol by the MFS-SNR algorithm and not by the MFS-T2 algorithm was expected. The correlation between cholesterol and HbA1c differed across the groups and Hotelling's

T^2 statistic assumes homogeneity of variance-covariance matrices. This is the reason cholesterol was not selected by the MFS-T2 algorithm.

Next, I investigated what is the best set of variables for discrimination between non-referable STDR and referable STDR using the ISDR dataset. Both the MFS-SNR and MFS-T2 algorithms identified t1risk, (risk score assigned at previous visit), as being important for discriminating between the Referable STDR and Non-referable STDR groups. The two groups were especially imbalanced in this dataset (74 patients in the Referable STDR group and 5,198 patients in the Non-referable STDR group) which made it difficult to obtain accurate performance estimates. However the univariate analysis corroborated the selection of t1risk as the optimal variable for discriminating between the Referable STDR and Non-referable STDR groups. Furthermore the variances and covariances were homogeneous across groups which is also consistent with the fact that MFS-SNR and MFS-T2 selected the same discriminatory variable.

I then investigated what is the best set of variables to discriminate between death and survival outcomes in subjects with malarial retinopathy, using retinal capillary non-perfusion imaging data in the MRet dataset. The MFS-SNR algorithm identified the variable Admlact (serum lactate) as being important for discriminating between the groups of individuals who died and individuals for survived in those diagnosed with malarial retinopathy. The MFS-T2 algorithm only selected Admlact following imputation of missing values in the dataset. This demonstrated that the SNR is more robust to the effects of missing data than the MFS-T2 algorithm. The selection of Admlact was consistent with the findings from the univariate analysis of the dataset.

Lastly, I studied what is the set of the best discriminatory variables to differentiate between the normal and eyes with keratoconus, using the keratoconus dataset. This dataset was unique amongst those analysed in this thesis because both groups were balanced and there was no missing data. It should also be noted that the variance-covariance matrices were heterogeneous across the two groups. The original dataset was used to produce two datasets. The first was based on the average measurement of each variable and the second using just the first measurement of each variable (all variable measurements had been taken in triplicate). The MFS-SNR and MFS-T2 algorithms selected different variables from each dataset. However the MFS-SNR selections achieved better performance estimates with fewer variables than the MFS-T2 selections, and Astigm corneal front was the first variable selected by the MFS-SNR algorithm from both datasets. I attribute the better performance of the MFS-SNR algorithm with the keratoconus dataset to the ability of the SNR metric to accommodate heterogeneous variance-covariance matrices across the 2 groups, something which Hotelling's T^2 statistic is not capable of.

7.4 Recommendations for practice

In this thesis methods for variable selection have been considered and their extensions have been proposed and studied. On the basis of these analyses and the results presented these methods may be recommended for research.

7.4.1 Discriminatory metric

When researchers want to summarise the discriminatory strength of a collection of variables they need to think carefully about the choice of appropriate discrimination metric. It is recommended to first do a set of univariate exploratory analyses, check of normality and check of equality of variance-covariance matrices across groups. Then in scenarios of unequal variance –covariance matrices I have demonstrated that the SNR is better able to assess the discriminatory potential of variables (univariately or multivariately) than Hotelling's T^2 statistic.

7.4.2 Variable selection for classification

When researchers want to do variable selection for classification then it is recommended to think carefully and strategically about the type of variable selection method. There are several points to consider, which I describe below:

First it is important to think what type of classifier will be used with the dataset being studied. If only one classifier is to be used then it may be appropriate to use the relevant wrapper or embedded method as it may be the most optimal. Conversely if the objective is to identify the optimal subset of variables for discrimination regardless of the classifier used then a filter method may be a better choice because filters are independent of classifier.

Another criterion is to look into the size of data (in terms of number of variables) and computational resources. If the number of variables is very large and computing resources limited then a filter method may be more appropriate.

Another criterion is the distribution of the data, especially, if it is normal and if variance-covariance matrices are the same across the groups. If the distribution of the data is normal and variance-covariances are homogeneous then the MFS-SNR and MFS-T2 algorithms give the same results. But since the MFS-T2 algorithm estimates fewer parameters this method is preferred over the MFS-SNR algorithm in scenarios of equal covariance matrices or scenarios when the difference is small.

7.5 Recommendations for future research

In this thesis I have proposed and implemented a novel multivariate filter method of variable selection; the MFS-SNR algorithm, which utilises the novel SNR metric, an extension of Hotelling's T^2 statistic. I have demonstrated the ability of the MFS-SNR algorithm to effectively carry out variable selection from data which are normally distributed and non-normally distributed as well as from datasets of varying quality in terms of group sizes, missing data and containing mixtures of variables. I have also demonstrated that the novel MFS-SNR algorithm performs at least as well as several existing methods of variable selection.

The proposed SNR metric and the MFS-SNR algorithm are simple and ready to be used. This thesis has highlighted areas where further research is recommended particularly with regards to the discriminatory metric and variable search algorithms.

7.5.1 Future research on the discriminatory SNR metric

As the SNR calculated in my work is based on a sample it is an estimate of the true SNR value. This means that its statistical properties need to be investigated. The SNR estimate is obtained by using the data twice: once to estimate means and covariances and a second time to evaluate SNR. This means there is a possible bias in the estimates calculated using sample data when compared to the actual (population) value. One possible means of addressing this problem is through regularisation of the SNR metric as has been done for Hotelling's T^2 statistic (Chen et al., 2011) or introducing a bias correction (Czanner et al., 2015).

Future work could be done on estimating the variance-covariance matrix for the SNR metric. For example in imaging data the fact that the data are spatially correlated may be exploited by imposing a specific suitable correlation structure e.g. an autocorrelated structure.

7.5.2 Future research on MFS-algorithms for variable selection in classification

The MFS-SNR (and MFS-T2) algorithm(s) use a forward selection paradigm. It would be constructive to investigate how extending the algorithm to use backward selection or stepwise selection might impact the performance of the algorithm. The forward selection mechanism is a valid and common approach but adding backward or stepwise selection would make the MFS algorithm more versatile and permit the assessment of alternative variable subsets that may not be identified using forward selection.

If the MFS-SNR algorithm was used to carry out variable selection from datasets for which the optimal variable subset was already identified (i.e. had been determined by another researcher) and the findings of both analyses were in agreement it would enhance the characterisation of the MFS-SNR algorithm that I have presented in this thesis. In particular datasets with properties not already addressed in my work would extend the use of the MFS-SNR algorithm (for example a dataset containing a significant proportion of ordinal and nominal variables).

The discriminating variables I simulated to assess the MFS-SNR algorithm and compare it to other methods of variable selection were based on the parameters of variables from the DREFUS dataset and specified to be normally distributed. Since these variables were based on the properties of variables from the DREFUS dataset I had an informed expectation as to the performance I would observe for the MFS-SNR algorithm. It would be productive to simulate a dataset containing variables not based on any data which the MFS-SNR algorithm has previously been exposed to (essentially a “blind” analysis from the point of view of the MFS-SNR algorithm). Simulations could also be expanded to include larger numbers of variables, different types of variables and mixtures of both normal and non-normally distributed data.

The change in PCC estimates was proposed and used here as the stopping criterion of the MFS-SNR and MFS-T2 algorithms. While the relationship between Hotelling’s T^2 statistic and LDA has been established the same relationship between the SNR metric and QDA has not been demonstrated. I have used QDA in the MFS-SNR algorithm because both the SNR metric and QDA can accommodate heterogeneity of the variance-covariance matrices. However, as no relationship between the SNR metric and QDA has been proven it is possible there is a difference in the discriminatory potential (as measured by the SNR metric) and the expected performance (as estimated using QDA). Altering the MFS-SNR algorithm to use a hybrid of the SNR metric and QDA estimates may capture this discrepancy and improve the overall performance of the MFS-SNR algorithm.

7.6 Conclusions

In this thesis I have proposed an extension of Hotelling’s T^2 statistic which does not assume that variance-covariance matrices are homogeneous across groups; the novel SNR. I then applied the SNR to the task of variable selection as part of a forward selection algorithm; the MFS-SNR algorithm.

Using simulated datasets I have demonstrated that the MFS-SNR algorithm is capable of selecting the relevant discriminatory variables whether data are normally distributed or not. In the simulated scenarios the MFS-SNR algorithm performed at least as well as competing methods.

When used to carry out variable selection from 4 real clinical datasets encompassing a range of common quality issues the MFS-SNR algorithm successfully identified the optimal variable sets for discrimination from each of the datasets. These selections were consistent with the results of conventional statistical analysis of each of the datasets.

In conclusion the MFS-SNR algorithm utilising the SNR is a novel multivariate filter method of variable selection. It addresses the limitations of existing filter methods by being multivariate and accommodating heterogeneity of variance-covariance matrices. It has been shown to perform at least as well as alternative methods in simulated scenarios and this performance has been achieved without the large computational requirements associated with embedded and wrapper methods.

Bibliography

Ahdesmaki M, Strimmer K. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics* 2010. 4:1, p.503-519.

Aiello L.P., Avery R.L., Arrigg P.G., Keyt B.A., Jampel H.D., Shah S.T., Pasquale L.R., Thieme H., Iwamoto M.A., Park J.E., Nguyen H.V., Aiello L.M., Ferrara N., King G.L., Vascular endothelial growth factor in ocular fluid of patients with diabetic retinopathy and other retinal disorders, *The New England Journal of Medicine* 1994. 331, p.1480-1487.

Allison M., Is personalized medicine finally arriving?, *Nature Biotechnology* 2008. 26, p.509-517.

Anderson T.W., Classification by Multivariate Analysis, *Psychometrika* 1951. 16, p.16-50.

Andrews JL, McNicholas PD. Variable Selection for Clustering and Classification, *Journal of Classification* 2014. 31, p.136-153.

Aronson J.K., Biomarkers and surrogate endpoints, British Journal of Pharmacology 2005. 59(5), p.491-494.

Axel-Siegel, Herscovici Z., Gabbay M., Mimouni K., Weinberger D., Gabbay U., The relationship between diabetic retinopathy, glycemic control, risk factor indicators and patient education, *The Israel Medical Association Journal* 2006. 8(8), p.523-526.

Barhen A., Daudin J.J., Generalization of the Mahalanobis distance to the mixed case, *Journal of Multivariate Analysis* 1995. 53(2), p.332-343.

Beare N.A.V., Harding S.P., Taylor T.E., Lewallen S., Molyneux M.E., Perfusion abnormalities in children with cerebral malaria and malarial retinopathy, *The Journal of Infectious Diseases* 2009, 199, p.263-271.

Bérubé J., Wu C.F.J., Signal-to-noise ratio and related measures in parameter design optimization: An overview, *Sankhyā: The indian Journal of Statistics, Series B* 2000. 62(3), p.417-432.

Bledsoe G.H., Malaria primer for clinicians in the United States, *Southern Medical Journal* 2005, 98(12).

Bolòn-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A., Benítez J.M., Herrera F., A review of microarray datasets and applied feature selection methods, *Information Sciences* 2014. 282, p.111-135.

Bouhamed H., Lecroq T., Rebai A., New filter method for categorical variables selection, *IAENG International Journal of Computer Science* 2012. 9(3), p.10-19.

Brassard G., Bratley P., *Fundamentals of Algorithms*, Prentice Hall, New Jersey, 1996.

Brusco MJ, Steinley D. Exact and approximate algorithms for variable selection in linear discriminant analysis. *Computational Statistics and Data Analysis* 2011. 55, p.123-131.

Breiman L, Spector P., Submodel selection and evaluation in regression. The X-random case, *International Statistical Review* 1992. 60(3), p.291-319.

Brown G.G., Perthen J.E., Liu T.T., Buxton R.B., A primer on functional magnetic resonance imaging, *Neuropsychology Review* 2007. 17(2), p.107-125.

Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression, *Source Code for Biology and Medicine* 2008. 3:17.

Burton A., Altman D.G., Royston P., Holder R.L., The design of simulation studies in medical statistics, *Statistics in Medicine* 2006. 25(24), p.4279-4292.

Caruana R., Freitag D., Greedy attribute selection, *Proceedings of the 11th International Conference on machine Learning* 1994. p.28-36.

Caveney E.J., Cohen O.J., Diabetes and biomarkers, *Journal of Diabetes Science and Technology* 2011. 5(1), p.192-197.

Chandrashekar G, Sahin F. A survey of feature selection methods. *Computers and Electrical Engineering* 2004> 40, p.16-28.

Chang Y.-C., Wu W.-C., Dyslipidemia and diabetic retinopathy, *The Review of Diabetic Studies* 2013, 10(2-3),p.121-132.

Chavent M., Kuentz V., Liquet B., Saracco J., ClustofVar, R Package Version 1.1.

Chen LS, Paul D, Prentice RL,Wang P. A regularised Hotelling's T^2 test for pathway analysis in proteomic studies, *Journal of the American Statistical Association* 2011. 16(496), p.1345-1360.

Chen X., Cheung S.T., So S., Fan S.T., Barry C., Higgins J., Lai K.M., Ji J., Dudoit S., Nq I.O., Van de Rijn M., Botstein D., Brown P.O., Gene expression patterns in human liver cancers, *Molecular Biology of the Cell* 2002. 13(6), p.1929-1939.

Cohen J., A power primer, *Psychological Bulletin* 1992, 112(1), 155-159.

Conover W.J., Johnson M.E., Johnson M.M., A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data, *Technometrics* 1981. 23(4), p.351-361.

Cortes C., Vapnik V., Support vector networks, *Machine Learning* 1995. 20(3), p.273-297.

Cox TF. *An introduction to multivariate data analysis,* Wiley, 2005.

Crawford S.O., Hoogeveen R.C., Brancati F.L., Astor B.C., Ballantyne C.M., Schmidt M.I., Young J.H., Association of blood lactate with type 2 diabetes: the atherosclerosis risk in communities carotid MRI study, *International Journal of Epidemiology* 2010. 39, p.1647-1655.

Crowther M.J., Lambert P.C., Simulating biologically plausible complex survival data, *Statistics in Medicine* 2012. 32, p.4118-4134.

Czanner G., Sarma S.V., Ba D., Eden U.T., Wu W., Eskander E., Lim H.H., Temereanca S., Suzuki W.A., Brown E.M., Measuring the signal-to-noise ratio of a neuron, *Proceedings of the National Academy of Sciences of the United States of America* 2015. 112(23) p.7141-7146.

Das S., Filters, wrappers and a boosting-based hybrid for feature selection, *Proceedings of the 18th International Conference on Machine Learning* 2001. p.74-81.

Dasgupta S., Variable selection using Kullback-Liebler divergence loss, *Journal of the Indian Statistical Association* 2015. 53, p.153-174.

Diaz-Uriarte R, Alvarez de Andres S., Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* 2006. 7, 3.

Diaz-Uriarte R,, varSelRF, R Package Version 0.7-8.

Domingos P., Why does bagging work? A Bayesian account and its implications, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* 1997. p.155-158.

Doquire G., Verleysen M., An hybrid approach to feature selection for mixed categorical and continuous data, *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* 2011. p.386-393.

Dunn O.J., Varady P.D., Probabilities of correct classification in discriminant analysis, *Biometrics* 1966. 22(4) p.908-924.

Edwards A.O., Ritter III R., Abel K.J., Manning A., Panhuysen C., Farrer L.A., Complement factor-H polymorphism and age-related macular degeneration, *Science* 2005. 308, p.421-424.

Efron B., Tibshirani R., Improvements on cross-validation: The .632+ bootstrap method, *Journal of the American Statistical Association* 1997. 92(438), p.548-560.

Ekdakl M., Koski T., Bounds for the loss in probability of correct classification under model based approximation, *Journal of Machine Learning Research* 2006. 7, p.2449-2480.

Fan J., Lv J., Sure independence screening for ultrahigh dimensional feature selection (with discussion), *Journal of the Royal Statistical Society, Series B* 2008. 70, p.849-911.

Fisher R.A., The use of multiple measurements in taxonomic problems, Annals of Eugenics 1936. 7(2) p.179-188.

Giove T.J., Deshpande M.M., Gagen C.S., Eldred W.D., Increased neuronal nitric oxide synthase activity in retinal neurons in early diabetic retinopathy, *Molecular Vision* 2009. 15, p.2249-2258.

Golub T.R., Slonim D.K., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 1999. 286(5439), p.531-537.

Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine learning Research* 2003. 3, p.1157-1182.

Guyon I., Weston J., Barnhill S., Vapnik V., Gene selection for cancer classification using support vector machines, *Machine Learning* 2002. 46, p.389-422.

Hall M.A., Correlation-based feature selection for discrete and numeric class machine learning , *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning* 2000. p.359-366.

Hall P., Miller H., Using generalized correlation to effect variable selection in very high dimensional problems, *Journal of Computational and Graphical Statistics* 2009. 18, p.533-550.

Hallgren K.A., Conducting simulation studies in the R programming environment, *Tutorials in Quantitative Methods for Psychology* 2014. 9, p.43-60.

Hamburg M.A., Collins F., The path to personalized medicine, *The New England Journal of Medicine*, 2010. 363, p.301-304.

Harding S., Campa C., Broadbent D., Brown M.C., Beare N. AV, Briggs M.C., Criddle T., Hagan R.P., Heimann H., Pearce I.A., Sahni J.N., Stangos A.A., Zheng Y., Study protocol: *Diabetic retinopathy: Functional and structural study*, St. Paul's Eye Unit, Royal Liverpool University Hospital, 2010.

Harding SP et al. 2011. Study Protocol for NIHR Programme Grant: Introducing personalised risk based intervals in screening for diabetic retinopathy: development, implementation and assessment of safety, cost-effectiveness and patient experience. Reference: RP-PG-1210-12016.

Hoffman S.L., Vekemans J., Richie T.L., Duffy P.E., *The march toward malaria vaccines*, Vaccine 2015, 33, p.$13-D23.

Hotelling H., The generalisation of Student's ratio", *Annals of Mathematical Sciences*, 1931. 2(3), p.360-378.

Hu Q., Liu J., Yu D., Mixed feature selection based on granulation and approximation, *Knowledge-Based Systems*, 2008. 21, p.294-304.

Hu Z.D., Zhou Z.R., Qian S., How to analyse tumor stage data in clinical research, *Journal of Thoracic Disease* 2015. 7(4), p.566-575.

Huang Y., McCullage P., Black N., Harper R., Feature selection and classification model construction on type 2 diabetic patients data, *Artificial Intelligence in medicine* 2007. 41, p.251-262.

Inza I, Larranaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains, *Artificial Intelligence in Medicine* 2004. 31, p.91-103.

James C.R., Quinn J.E., Mullan P.B., John ston P.G., Harkin D.P., BRCA1, A potential predictive biomarker in the treatment of breast cancer, *Oncologist* 2007. 12(2), p.142-150.

Jiang H, Deng Y, Chen H-S, Tao L, Sha Q, Chen J, Tsai C-J, Zhang S. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 2004. 5, 81.

Jiang X.-W., Hu B., Guan Z.-H., Yu L., The minimal signal-to-noise ratio required for stability of control systems over a noisy channel in the presence of packet dropouts, *Information Sciences* 2016p. 372 579-590.

Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics* 2005. 6, p.148-158.

Johnson HE, Broadhurst D, Goodacre R, Smith AR. Metabolic fingerprinting of salt-stressed tomatoes. *Photochemistry* 2003. 62, p.919-928.

Johnson R.A., Wichern D.W., *Applied multivariate statistical analysis*, Pearson 2008.

Joussen A.M., Doehmen S., Le M.L., Koizumi K., Radetzky S., Krohne T.U., Poulaki V., Semkova I., Kociok N., TNF-α mediated apoptosis plays an important role in the development of early diabetic retinopathy and long-term histopathological alterations, *Molecular Vision* 2009. 15, p.1418-1428.

Ju H, Brasier AR. Variable selection methods for developing a biomarker panel for prediction of dengue hemorrhagic fever. *BMC Research Notes* 2013. 6, p.365-372.

Kaiser M.E., Bohlin R.C., Lindler D.J., Gilliland R.L., Argabright V.S., Kinble R.A., STIS signal-to-noise capabilities in the ultraviolet, *Publications of the Astronomical Society of the Pacific* 1998. 110(750), p.978-990.

Kaiser P.K., Prospective evaluation of visual acuity assessment: A comparison of Snellen versus ETDRS charts in clinical practice (an AOS thesis), *Transactions of the American Ophthalmological Society* 2009. 107, p.311-324.

Karper A., “Feature and variable selection in classification”, arXiv:1402.2300v1., 2014.

Kizawa J., Machida S., Kobayashi T., Gotoh Y., Kurosaka D., Changes in oscillatory potentials and photopic negative response in patients with early diabetic retinopathy, *Japanese Journal of Ophthalmology* 2006, 50, p.367-373.

Klein R., Klein B.E.K., Blood pressure control and diabetic retinopathy, *British Journal of Ophthalmology* 2002, 86(4), p.365-367.

Koenig R.J., Peterson C.M., Jones R.L., Saudek C., Lehrman M., Cerami A., Correlation of glucose regulation and haemoglobin A_{1c} in diabetes mellitus, The New England Journal of Medicine 1976. 295(8), p.417-420.

Kohavi R., A study of cross-validation and bootstrap for accuracy estimation and model selection, IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence 1995. 14(2), p.1137-1143.

Kohavi R., John G.H., Wrappers for feature subset selection, *Artificial Intelligence* 1997. 97(1-2) p.273-324.

Koller D., Sahami M., Toward optimal feature selection, *13th International Conference on Machine Learning* 2005. p.284-292.

Kononenko I., Estimating attributes: Analysis and extensions of RELIEF, *Lecture Notes in Computing Science* 1994. p.171-182.

Kumar R., Indrayan A., Receiver Operating Characteristic (ROC) curve for medical researchers, *Indian Pediatrics*, 2011. 48(4) p. 277-287.

Kumar V., Minz S., Feature Selection: A literature review, Smart Computing Review 2014. 4(3), p.211-229.

Lai T.Y., Chen W.M., Lai R.Y., Ngai J.W., Li H., Lam D.S., The clinical applications of multifocal electroretinography: a systematic review, *Survey of Ophthalmology* 2007. 52, p.61-96.

Lattin J.M., Carroll J.D., Green P.E., *Analyzing multivariate data*, Thomson Brooks/Cole, 2003.

Lazar C., Taminau J., Meganck S., Steenhoff D., Coletta A., Molter C., de Schaetzen V., Duque R., Bersini H., Nowé A., A survey on filter techniques for feature selection in gene expression microarray analysis, Transactions on Computational Biology and Bioinformatics 2012. 9, p.1106-1119.

Lee I-H, Luchington GH, Visvanathen M. A filter-based feature selection approach for identifying potential biomarkers in lung cancer. *Journal of Clinical Bioinformatics* 2011. 1, p.11-19.

Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *Journal of the American Statistical Association* 2012. 107:499, p.1129-1139.

Liu H., Setiono R., A probabilistic approach to feature selection – a filter solution, *Proceedings of the 13th International Conference on Machine Learning* 1996. p.319-327.

Long M., Wang C., Liu D., Glycated haemoglobin A1C and vitamin D and their association with diabetic retinopathy severity, *Nutrition & Diabetes* 2017, 7.

Loscalzo S., Yu L., Ding C., Consensus group stable feature selection, Proceedings of the 15th ACM SIGKDD International Conference on knowledge Discovery and Data Mining 2009. p.567-576.

Lu Y, Liu P-Y, Xiao P, Deng H-W. Hotellings T^2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics* 2005. 21(14), p.3105-3113.

Maamor N., Billings C.J., Cortical signal-in-noise coding varies by noise type, signal-to-noise ratio, age and hearing status, *Neuroscience Letters* 2017. 636 p.258-264.

MacCormick I.J.C., Beare N.A.V., Taylor T.E., Barrera V., White V.A., Hiscott P., Molyneux M.E., Dhillon B., Harding S.P., Cerebral malaria in children: using the retina to study the brain, *BRAIN* 2014, 137, p.2119-2142.

MacCormick I.J.C., Maude R.J., Beare N.A.V., Borooah S., Glover S., Parry D., Leach S., Molyneux M.E., Dhillon B., Lewallen S., Harding S.P., Grading fluorescein angiograms in malarial retinopathy. Malaria Journal 2015. 14: 367.

Mahat NI, Krzanowski WJ, Hernandez A.Variable selection in discriminant analysis based on the location model for mixed variables, *Advances in Data Analysis and Classification* 2007. 1, p.105-122.

Maugis C, Celeux G, Martin-Magniette M.-L. Variable selection in model-based discriminant analysis 2010; [research report] RR-7290, 2010. <inria-00483229>.

Meyer D., Hornik K., Weingessel A., Leisch F., Chang C.-C., Lin C.-C., e1071, R Package Version 1.6-8.

Nkiet GM. Direct variable selection for discrimination among several groups, *Journal of Multivariate Analysis* 2012. 105, p.151-163.

Ohkubo H., Tanino T., Electrophysiological findings in familial exudative vitreoretinopathy, *Documenta Ophthalmologica* 1987, 65, p.461-469.

Olaussen K.A., Dunant A., Fouret P., Brambilla E., Andre F., Haddad V., Taranchon E., Filipits M., Pirker R., Popper H.H., Stahel R., Sabatier L., Pignon J.P., Tursz T., Le Chevalier T>, Soria J.C., IALT Bio Investigators, DNA repair by ERCC1 in non-small cell lung cancer and cisplatin-based adjuvant chemotherapy, The New England Journal of Medicine 2006. 355(10), p.983-991.

Ozturk B.T., Bozkurt B., Kerimoglu H., Okka M., Kamis U., Gunduz K., Effect of serum cytokines and VEGF levels on diabetic retinopathy and macular thickness, *Molecular Vision* 2009. 15, p.1906-1914.

Pacheco J, Casado S, Nunez L, Gomez O. Analysis of new variable selection methods for discriminant analysis. *Computational Statistics and Data Analysis* 2006. 51, p.1463-1478.

Pacheco J., Casado S., Porras S., Exact methods for variable selection in principal component analysis: Guide functions and pre-selection, *Computational Statistics and Data Analysis* 2013. 57(1), p.95-111.

Pareja M, Mohib A, Birkett MA, Dufour S, Glinwood RT. Multivariate statistics coupled to generalized models reveal complex use of chemical cues by a parasitoid. *Animal Behaviour* 2009. 77, p.901-909.

Park P.J., Pagano M., Bonetti M., A nonparametric scoring algorithm for identifying informative genes from microarray data, *Pacific Symposium on Biocomputing* 2001. p.52-63.

Paul D., Bair E., Hastie T., Tibshirani R., Preconditioning for feature selection and regression in high-dimensional problems, *The Annals of Statistics* 2008. 36(4), p.1595-1618.

Pavlidis P., Weston J., Cai J., Grundy W.N., Gene functional classification from heterogeneous data, *Proceedings of the Fifth International Conference on Computational Molecular Biology* 2001. p.242-248.

Petracoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer", Mechanisms of Disease 2002. 359, p.572-577.

Quinlan J.R., Induction of decision trees, *Machine Learning* 1986. 1(1), p.8-106.

Quinlan J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rénia L., Howland S.W., Claser C., Gruner A.C., Suwanarusk R., Teo T.-H., Russell B., Ng L.F.P., Cerebral Malaria, *Virulence* 2012, 3(2), p.193-201.

Rohrschneider K., Bultmann S., Springer C., Use of fundus perimetry (microperimetry) to quantify macular sensitivity, *Progress in Retinal and Eye Research* 2008, 27, p.536-548.

Ripley B., Venables B., Bates D.M., Hornik K., Gebhart A., Firth D., MASS, R Package version 7.3-48.

Ritter F.E., Schoelles M.J., Quigley K.S., Klein L.C., Determining the number of simulation runs: Treating simulations as theories by not sampling their behaviour, *Human-in-the-loop simulations: Methods and practice* 2011. *Springer*.

Riva C.E., Basic principles of laser Doppler flowmetry and application to the ocular circulation, *International Ophthalmology* 2001. 23, p.183-189.

Romanski P., Kotthoff L., FSelector, R Package version 0.20.

Saeys Y, Inza I, Larranaga O. A review of feature selection techniques in bioinformatics, *Bioinformatics* 2007. 23:19, p.2507-2517.

Singh D., Febbo P.G., Ross K., Jackson D.G., Manola J., Ladd C., Tamayo P., Renshaw A.A., D'Amico A.V., Richie J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R., Sellers W.R., Gene expression correlates of clinical prostate cancer behaviour, *Cancer Cell* 2002. 1, p.203-209.

Strimbu K., Tavel J.A., What are biomarkers, *Current Opinion in HIV and AIDS* 2010. 5(6), p.463-466.

Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* 1993. 88(424), p.1350-1355.

Tang S, Chen L, Tsui K-W, Doksum K. Nonparametric variable selection and classification: The CATCH algorithm. *Computational statistics and data analysis* 2014. 72, p.158-175.

Tang W., Mao K.Z., Feature selection algorithm for mixed data with both nominal and continuous features, *Pattern Recognition Letters* 2007. (28) p.563-571.

Taylor J, King RD, Altmann T, Fiehn O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* 2002. 18(2), p.S241-S248.

Tibshirani R. Regressions shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* 1996. 58:1, p.267-288.

Todorov D, Setchi R. Time-efficient estimation of conditional mutual information for variable selection in classification, *Computational Statistics and Data Analysis* 2014. 72, p.105-127.

Tur V.M., MacGregor C., Jayaswal R., O'Brart D., Maycock N., A review of keratoconus: Diagnosis, pathophysiology and genetics, *Survey of Ophthalmology* 2017, 62, p.770-783.

Van Burren S., Groothuis-Oudshoorn K., Robitzsch A., Vink G., Doove L., Jolani S., Schouten R., Gaffert P., Meinfelder F., Gray B., mice, R Package Version 2.46.0.

Vera L, Acena L, Guasch J, Boqué R, Mestres M, Busto O. Discrimination and sensory description of beers through data fusion. Talanta 2011. 87, p.136-142.

Welvaert M., Rosseel Y., On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data, Public Library of Science ONE 2013. 8(11) e77089.

Weston J., Elisseeff A., Scholkopf B., Use of the Zero-Norm with linear models and kernel methods, *Journal of Machine Learning Research* 2003. 3, p.1439-1461.

WHO International Programme on Chemical Safety. Biomarkers in Risk Assessment: Validity and Validation. 2001. Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

Wickham H., reshape, R Package Version 0.8.7Wilinski A., Osowski S., Siwek K., Gene selection for cancer classification through ensemble of methods, *Lecture Notes in Computer Science* 2009. 5495, p.507-516.

Wilson D.R., Martinez T.R., Improved heterogeneous distance functions, *Journal of Artificial Intelligence Research* 1997. 6, p.1-34.

Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. *Genome Research* 2011. 11, p.1878-1887.

Xiong M, Zhao J, Boerwinkle E. Generalized T^2 test for genome association studies, *American Journal of Human Genetics* 2002. 70, p.1257-1268.

Yu L., Liu H., Feature selection for high-dimensional data: A fast correlation-based filter solution, *Proceedings of the 20th International Conference on machine Learning* 2003. p.856-863.

Zakaria A, Shakaff AYM, Masnam MJ, Ahmad MN, Adom AH, Jaafar MN, Ghani SA, Abdullah AH, Aziz AHA, Kamarudin LM, Subari N, Fikri NA. A biomimetic sensor for the classification of honeys of differential floral origin and detection of adulteration. *Sensors* 2011. 11, p.7799-7822.

Zakharov R, Dupont P. Ensemble logistic regression for feature selection. *Pattern Recognition in Bioinformatics*, *Lecture Notes in Computer Science* 2011. 7026, p.133-144.

Zhang Y., Zhang Z., Feature subset selection with cumulate conditional mutual information minimization, *Expert Systems with Applications* 2012. 39, 6078-6088.

Zhou W., Dickerson J.A., A novel class dependent feature selection method for cancer biomarker discovery, *Computers in Biology and Medicine* 2014. 47, p.66-75.

Zhu L.P., Li L., Li R., Zhu L.X., Model-free feature screening for ultrahigh dimensional data, *Journal of the American Statistical Association* 2011. 106, p.1464-1475.

Zuber V., Strimmer K., Gene ranking and biomarker discovery under correlation, Bioinformatics 2009. 25:20, p.2700-2707.

Appendix – R code

To run either the MFS-T2 or MFS-SNR algorithms the `mfs` and `stat` scripts for the relevant function must be loaded into the R workspace.

Once the two scripts have been loaded the algorithm may be called by entering:

```
mfs(input,groupfeature,labels,deltapcc)
```

The user must specify the values of the parameters `input`, `groupfeature`, `labels` and `deltapcc`. These are described below.

- `input`: this is the title of the query dataset which has already been loaded into the R workspace
- `groupfeature`: this is the identity of the grouping variable in the query dataset, it must be entered as a character string
- `labels`: this contains the identifiers for the groups of interest, must be entered as a vector of character strings
- `deltapcc`: this is the minimum PCC increase the user wishes to achieve each time a variable is added to the set of selected variables

MFS-T2 algorithm

mfs function:

```
mfs<-function(input, groupfeature, labels, deltapcc)

require(MASS)

t2<-0

indexi<-NULL
t2i<-NULL
pcci<-NULL
pccvals<-NULL
t2vals<-NULL
fi<-NULL
vm<-NULL

pcc<-0
improved<-1

groups<-input[groupfeature]

vars <- names(input) %in% names(groups)
input <- input[!vars]

vr<-seq(1,dim(input)[2],1)
```

```
while ((improved == 1) & (length(vr) >= 1)){

       for (e in vr) {

              results<-stat(input,groups,labels,vm,e)

              indexi[e]<-e
              t2i[e]<-results[1]
              pcci[e]<-results[2]

                     }

       inter<-data.frame(indexi,t2i,pcci)
       interorder<-inter[order(t2i,decreasing = TRUE),]

       if (interorder$t2i[1] > t2) {
              t2 <- interorder$t2i[1]
              vm<-c(vm,interorder$indexi[1])
              vr<-subset(vr,vr!=interorder$indexi[1])
                                   }

       if (interorder$pcci[1]-pcc < deltapcc) {
              improved<-0
                                   }

       pcc <- interorder$pcci[1]
       pccvals<-c(pccvals,interorder$pcci[1])
```

```
      t2vals<-c(t2vals,interorder$t2i[1])

                                    }


lvm<-length(vm)-1

vm<-vm[1:lvm]

t2vals<-t2vals[1:lvm]

pccvals<-pccvals[1:lvm]

selection<-input[vm]

selectionnames<-colnames(selection)

selectionpcc<-data.frame(selectionnames,t2vals,pccvals)


return(selectionpcc)


}
```

stat function:

```
stat<-function(input, groups, labels, vm, e) {

mdm<-NULL

scm<-data.frame(groups,input[vm[1:length(vm)]],input[e])
scm<-subset(scm,scm[1]==labels[1] | scm[1]==labels[2])

for (i in 2:length(scm)) {
       scm<-subset(scm,scm[i]!="NA")
                                  }

g1m<-subset(scm,scm[1]==labels[1])
g2m<-subset(scm,scm[1]==labels[2])

n1m<-dim(g1m)[1]
n2m<-dim(g2m)[1]

nm<-n1m+n2m
pm<-(dim(scm)[2]-1)

varg1m<-data.frame(g1m[2:dim(scm)[2]])
varg2m<-data.frame(g2m[2:dim(scm)[2]])

cov1m<-cov(varg1m)
cov2m<-cov(varg2m)
Sm<-((n1m-1)*cov1m+(n2m-1)*cov2m)/(n1m+n2m-2)


  for (i in 1:length(varg1m)) {

    mdm<-c(mdm,sum(varg2m[i])/n2m - sum(varg1m[i])/n1m)

  }


  meandiffm<-matrix(mdm,nrow=(pm),ncol=1)

  mahalm<-t(meandiffm) %*% solve(Sm) %*% meandiffm


  T2m<-(n1m*n2m/(n1m+n2m))*mahalm

  fm<-(n1m+n2m-pm-1)*T2m/((n1m+n2m-2)*pm)

  pvm<-1-pf(fm, pm, n1m+n2m-pm-1)


  sc2<-scm[2:length(scm)]


  testlda<-lda(sc2, scm$Study.Group,CV=TRUE)

  ct <- table(testlda$class, scm$Study.Group)

  diag(prop.table(ct, 1))

  pcc<-sum(diag(prop.table(ct)))




  results<-c(T2m,pcc)


return(results)


}
```

MFS-SNR algorithm

mfs function:

```
mfs<-function(input, groupfeature, labels, deltapcc) {

require(MASS)

indexi<-NULL
snri<-NULL
pcci<-NULL
pccvals<-NULL
snrvals<-NULL
fi<-NULL
vm<-NULL

snr<-0
pcc<-0
improved<-1

groups<-input[groupfeature]

vars <- names(input) %in% names(groups)
input <- input[!vars]

vr<-seq(1,dim(input)[2],1)

while ((improved == 1) & (length(vr) >= 1)){
        for (e in vr) {
                results<-stat(input,groups,labels,vm,e)
                indexi[e]<-e
                snri[e]<-results[1]
                pcci[e]<-results[2]
                                        }

        inter<-data.frame(indexi,snri,pcci)
        interorder<-inter[order(snri,decreasing = TRUE),]

        if (interorder$snri[1] > snr) {
                snr <- interorder$snri[1]
                vm<-c(vm,interorder$indexi[1])
                vr<-subset(vr,vr!=interorder$indexi[1])
                                        }

        if (interorder$pcci[1]-pcc < deltapcc) {
                improved<-0
                                        }

        pcc <- interorder$pcci[1]

        pccvals<-c(pccvals,interorder$pcci[1])
        snrvals<-c(snrvals,interorder$snri[1])
                                        }

lvm<-length(vm)-1
```

```
vm<-vm[1:lvm]

snrvals<-snrvals[1:lvm]

pccvals<-pccvals[1:lvm]

selection<-input[vm]

selectionnames<-colnames(selection)

selectionpcc<-data.frame(selectionnames,snrvals,pccvals)


return(selectionpcc)


}
```

stat function:

```
stat<-function(input, groups, labels, vm, e) {

md1<-NULL

md2<-NULL

VarMeans<-NULL

OverMean<-NULL

sc<-data.frame(groups,input[vm[1:length(vm)]],input[e])

sc<-subset(sc,sc[1]==labels[1] | sc[1]==labels[2])

for (i in 2:length(sc)) {

       sc<-subset(sc,sc[i]!="NA")

                                    }

g1<-subset(sc,sc[1]==labels[1])

g2<-subset(sc,sc[1]==labels[2])

n1<-dim(g1)[1]

n2<-dim(g2)[1]

n<-n1+n2

p<-(dim(sc)[2]-1)

varg1<-data.frame(g1[2:dim(sc)[2]])

varg2<-data.frame(g2[2:dim(sc)[2]])
```

```
for (i in 1:length(varg1)) {

       OverMean[i]=(sum(varg1[i])+sum(varg2[i]))/n

                                 }


cov1<-cov(varg1)

cov2<-cov(varg2)


for (i in 1:length(varg1)) {

       md1<-c(md1,(sum(varg1[i])/n1 - OverMean[i]))

                                 }


for (i in 1:length(varg2)) {

       md2<-c(md2,(sum(varg2[i])/n2 - OverMean[i]))

}


meandiff1<-matrix(md1,nrow=(p),ncol=1)

meandiff2<-matrix(md2,nrow=(p),ncol=1)


SNR<-
t(meandiff1)%*%solve(cov1)%*%(meandiff1)*n1+t(meandiff2)%*%solve(cov2)%*%(meandiff2
)*n2


sc2<-sc[2:length(sc)]

fit <- qda(sc2, sc$Study.Group,CV=TRUE)

ct <- table(sc$Study.Group, fit$class)

diag(prop.table(ct, 1))

pcc<-sum(diag(prop.table(ct)))
```

```
results<-c(SNR,pcc)

return(results)

}
```