SOME ASPECTS OF COVARIANCE REGULARISATION IN DISCRIMINANT ANALYSIS

A Thesis presented in fulfilment of the
requirements for the degree of
Doctor of Philosophy in Statistics at
Massey University, New Zealand.

John Peter Koolaard BSc.(Hons), MSc.

1997

Errata

to the thesis by J.P.Koolaard entitled "Some Aspects of Covariance Regularisation in Discriminant Analysis".

Page 1, line -10 'Prostrate' should be *prostate.*

Page 3, line 5 After the words "...for group k." add the sentence: "It is evident from expression (1.3) that all vectors in the thesis are considered as column vectors, unless stated otherwise."

Page 4, line 6 Sentence beginning on this line should read:"In effect, the $\mathbf{S}_k$ are replaced by the pooled covariance matrix, and the variance of the elements of $\mathbf{S}_p$ are smaller ..."

Page 15, line -11 "...where the pooled sample estimate..." should read: "...where the inverse pooled sample estimate...".

Page 56, lines -6 to -4 Rewrite these three lines as: "It should be noted that in his article, Friedman used robust covariance estimators in place of $\mathbf{S}_k$ and $\mathbf{S}_p$ in expressions (3.6) and (3.7). The resulting robustification of (3.6) is written as ..."

Page 76, lines -9, -8, -5 In these lines replace $\tilde{\mathbf{S}}_k$ with $\mathbf{S}_k$ and $\tilde{\mathbf{S}}_p$ with $\mathbf{S}_p$.

Page 3, line 14 Change "... expression (1.9) ..." to "... expressions (1.9) and (1.10) ...".

Page 5, line 12 To avoid any possible confusion, change "$(i, j = 1, \ldots, K(i \neq j))$" to "(for all $j(\neq i) = 1, 2, \ldots, K$)".

Page 5, line 15 Remove the word "directly".

Page 12, line 1 Change "mean" to "mean vector".

Page 14, line 8 Change "off diagonal" to "off-diagonal".

Page 17, line 10 Change "(1993)" to "(1993))".

Page 21, line 14 Change "samples of" to "samples of size".

Abstract

Statistical discriminant analysis and classification are multivariate techniques concerned with separating distinct set of objects, and with allocating new objects to previously defined populations or groups. In this process the covariance matrix plays an important role, and usually this matrix has to be estimated from sample data. In this thesis, attention is focussed on investigating the problem of (poor) estimation of the covariance structure and its effects in statistical discriminant analysis. The quality or statistical properties of these estimates usually affect the resultant classification rules which are constructed using them.

Reasons for the (usually, consistent) estimators of the covariance matrices being poor are mainly to do with the quality and/or size of the training sample in relation to the number of parameters which have to be estimated. In this thesis, we are interested in investigating this problem as it occurs in the small sample, high-dimensional situation. In particular, we are interested in the problem of covariance estimation in the situations when the sample size to dimension ratios are relatively small. The criterion used to determine the success or otherwise of various methods used to address this problem is the estimated (overall) error rate. One method of dealing with a situation which potentially results in poor estimation of the covariance matrix is to impose a prescribed (simple) structure on the covariance matrix, such as the identity matrix, or multiple of it. Another method is to make the assumption that all the groups have the same covariance matrix. The effect of such simplifying assumptions is to reduce the number of parameters to be estimated. Consequently, the (fewer) parameters are estimated with higher precision. It has been demonstrated that this may result in better statistical discriminant analysis, even if the simplifying assumptions may not be entirely correct.

Of the classification rules based on the normal distribution, the quadratic discriminant function (QDF) makes no restrictions on the population parameters, and as such is the most general of this class of classification rules. However, it is also the one most affected by poor population parameter estimates. The two common simplifying techniques mentioned earlier (i.e. imposing an identity matrix structure on the covariance matrix, or assuming a common covariance among all populations) lead to two other discriminant rules, namely, the Euclidean distance function (EDF, based on the Euclidean distance between the group means) and the

popular linear discriminant function (LDF, based on the Mahalanobis distance between the groups) respectively. The sample-based versions of these two classifiers are compared using expected error rates (conditional on a set of training data), and these expected error rates are obtained through the derivation of asymptotic expansions. The expansions are evaluated under a range of settings, defined by employing combinations of various values of dimension, group separation, and covariance structure. It is shown that the simpler sample Euclidean distance function (SEDF) performs as well as or better than the sample linear discriminant function (SLDF) under most of the settings used. Exceptions occurred when the Mahalanobis distance between populations was much greater than the Euclidean distance.

A flexible discrimination model, or rather, class of models, was developed by Friedman (1989), and called the regularised discriminant function (RDF). The sample version of the RDF (i.e. SRDF) model incorporates the general sample quadratic discriminant function (SQDF), the two previously-mentioned restricted models (SEDF and SLDF), as well as a wide range of models intermediate to these, through the use of additional "regularisation" parameters. The method employs two types of shrinkage of the covariance estimates - towards the pooled estimate on one hand, and towards a multiple of the identity matrix on the other. A separate regularisation parameter controls shrinkage to each. The training data is used in the model selection process to determine appropriate values for the regularisation parameters, through the use of cross-validation. The quality of model selection procedure which specifies a discriminant model is a crucial factor, since if it is performing well, it will result in a classification rule close to the optimal one from the class of models available.

Through large-scale simulation studies, the performance of the sample regularised discriminant function (SRDF) is investigated and it is shown that the SRDF generally leads to lower overall error rates than the standard classification rules. This is found to be largely due to the facility which allows shrinkage of the covariance matrices to sphericity, or eigenvalue regularisation. It is also found that the SEDF performs very well in relation to the SRDF for a variety of settings. Further simulation studies show that the performance of the SRDF is more sensitive to the parameter controlling shrinkage to sphericity than the one controlling covariance mixing. Also, it is found that under some circumstances, the SRDF performs better than the other classifiers even for quite large sample size to dimension ratios.

A crucial negative feature of the SRDF is its lack of scale invariance. The cause of this is eigenvalue regularisation. A modified classification rule is developed which is scale invariant, and is compared to the SRDF and the other classifiers via simulation. The modified rule omits eigenvalue regularisation, but otherwise increases sensitivity to the data by allowing for varying degrees of shrinkage to the pooled covariance for each group. It is shown that eigenvalue regularisation is generally beneficial for discrimination in medium to large dimensional problems, through its variance-reduction effect which stabilises the covariance estimates. Thus, the study concludes that scale invariance must be sacrificed in order to achieve reductions in error rate, in the absence of a suitable replacement for eigenvalue regularisation.

The use of cross-validation in the model selection process of the SRDF is also investigated, for several reasons: the computational effort involved, and the fact that it rarely leads to a unique choice of model, and often uses only a small subset of the available observations, in the model selection process. Consequently, another method for determining the optimal regularisation parameters is investigated. In particular, it is investigated whether appropriate values for the regularisation parameters can be indicated from a measure of the distance between the groups. For this purpose, the Bhattacharyya distance is chosen since it comprises a term primarily pertaining to the difference between group means, and a further term which indicates the level of disparity between group covariance structures. It is shown that the magnitudes of the various components of the Bhattacharyya distance, when considered on their own and in relation to each other, do give information as to appropriate values for the regularisation parameters. A new simulation study, as well as various case studies are presented to assess the performance of a new regularised discriminant function which uses the Bhattacharyya distance estimates between groups to select regularisation parameters for given training data. This classifier is shown to perform as well as the SRDF, and is computationally much faster since it avoids any re-sampling methods.

It is clear that most of the investigations and assessments of the various regularised discriminant rules have to be undertaken using Monte-Carlo simulation techniques, especially to estimate error rates. This is because exact analytical expressions for the unconditional error rate of the SRDF do not exist, except in certain limited circumstances. It has not been possible to obtain asymptotic expansions or some form of approximations of these error rates in a general context.

However, an approximation which can be used to calculate algebraically the error rate of the SQDF, assuming known population parameters under (other) strict conditions, is available in the literature. This approximation is used in this thesis to further examine the effects (observed in earlier simulation work) of the covariance regularisation parameters on error rates. This is the last piece of work in the thesis and, in spite of its limited extent (because of the restricted conditions of the approximations given), it largely confirms the results which were obtained from simulation experiments in the previous parts of the thesis.

Acknowledgements

I would like to thank my chief supervisor, Charles Lawoko, now at the Queensland University of Technology, for his immense help and guidance throughout this project. His office always had an 'open' door for me, and his patience with this part-time student was considerable. Likewise, I would like to thank my second supervisor, Ganesalingam, for his frequent doses of moral support. His too was an open door for me, and I was privileged to have supervisors such as these to sit with and think about the problems at hand.

The staff in the Statistics Department have treated me wonderfully, and I'd like to especially thank Professor Jeffrey Hunter, former Head of Department for his confidence in me, and his general concern and kindness towards me and my family, expressed in a number of tangible ways throughout my time in the Department.

Thanks also to Professor Dick Brook, an old golfing partner, who always kept an eye out for me, and to whom I could always go to for a chat. To Paula, Helen, Greg, Chin Diew, Ganesh and the others too, thanks.

The support of Massey University for this project must also be acknowledged. I am grateful for the support of a Massey University Doctoral Scholarship for 1995, as well as for part-time employment in the the form of Assistant Lectureships for four years from 1991 to 1994.

It has been a slow, long, hard grind all the way, and one who knows that more than all the others is my dear wife Antoinette. Hers was a constant support, including necessary admonishment at times. She had to put up with a lot from me, and I'm very grateful for it all. It was and is always a joy to come home to Antoinette and my two lovely daughters, Lydia and Natasha.

Finally, thanks and praise be to the Father of my Lord Jesus Christ, who carried me through it all, and who sustains and upholds me.

List of Additional Publications by Author Including Papers Presented at Conferences

See Appendix C

1. **Koolaard**, J. P. and Lawoko, C. R. O. (1993). Estimating error rates in discriminant analysis with correlated training observations: a simulation study. *J. Statist. Comput. Simul.* 48, 81-99.

2. **Koolaard**, J. P. and Lawoko, C. R. O. (1994). Some results on the error rates of the Euclidean and linear discriminant functions. *Proceedings of the ORSNZ/NZSA Conference, Massey University, Palmerston North, New Zealand* (August 1994). pp 327-332.

3. **Koolaard**, J. P. (1995). Covariance Shrinkage in Discriminant Analysis. Paper presented to the A. C. Aitken Centenary Conference, Dunedin, New Zealand (August 1995). [Winner of SPSS Statistics Prize for best statistics paper presented by a student.]

4. **Koolaard**, J. P., Lawoko, C. R. O. and Ganesalingam, S. (1996). Regularized discriminant (classification) analysis involving Bhattacharya distance measure. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 2, Poster, pp 35-43.

5. **Lawoko**, C. R. O., and Koolaard, J. P. (1996). Applications of regularised discriminant(classification) functions in the classification of objects: a discussion of potential applications to remote sensing. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 1, pp 177-184.

6. **Koolaard**, J. P., Ganesalingam, S. and Lawoko, C. R. O. (1996). Comparison of regularised discriminant analysis with the standard discrimination methods. Paper presented to the International Biometrics Conference (IBC '96), Amsterdam, the Netherlands (July 1996). Also submitted to the Journal of Classification.

7. **Koolaard**, J. P. and Lawoko, C. R. O. (1996). The linear and Euclidean discriminant functions: a comparison via asymptotic expansions and simulation study. *Commun. Statist.- Theory Meth.*, (To appear).

Contents

Abstract ii

Acknowledgements vi

List of Publications by Author vii

Table of Contents viii

List of Figures xi

List of Tables xv

List of Abbreviations xxi

1 Chapter 1 1

1.1 General Framework 1

1.2 Error Rates 5

1.3 Outline of Research Undertaken 6

1.4 Notation and Definitions 11

2 Chapter 2 13

2.1 Introduction 13

2.2 Literature Review 14

2.2.1 Raudys and Pikelis (1980) 14

2.2.2 Peck and Van Ness (1982) 15

2.2.3 Marco, Young and Turner (1987) 16

2.2.4 Implications of results from Marco, Young and Turner (1987) 17

2.2.5 Motivation for the present study 20

2.3 Asymptotic Expansions 21

2.4 Numerical Evaluations of Asymptotic Expansions 24
2.5 Simulation Results 33
2.6 Graphical Displays 35

3 Chapter 3 **50**
3.1 Introduction 50
3.2 Problems with Estimating Covariance Matrices 51
3.3 Regularised Estimates of Σ_k 52
3.4 Selecting Regularisation Parameter Values 56
3.5 Assessment of the SRDF 58
3.5.1 Comparison of SRDF with other classifiers 58
3.5.2 Simulations for groups with small mean differences 66
3.6 Further Model Selection Considerations for the SRDF: Breaking of Ties 69

4 Chapter 4 **75**
4.1 Introduction 75
4.2 Invariance 76
4.3 Assessing the Performance of the Modified Regularised Discriminant Function (SRDF-M) 77
4.3.1 The performance of SRDF-M when the population shapes are similar 77
4.3.2 The performance of SRDF-M when the population shapes are very different 88
4.4 Performance of the Regularised Discriminant Function in terms of the Sample Size to Dimension Ratio 96
4.4.1 Simulation study 96
4.4.2 Simulation results 97

5 Chapter 5 **109**
5.1 Introduction 109
5.2 Construction of a Model Selection Procedure Based on the Bhattacharyya Distance 111
5.2.1 Distance measures and their applications in discrimination 111
5.2.2 The Bhattacharyya distance 111

5.2.3 Behaviour of Bhattacharyya distance with regularised covariances . . . 113
5.2.4 Model selection . . . 114
5.2.5 Model selection when there are more than two groups . . . 118
5.3 Simulation Studies and Results . . . 119
5.4 Case Studies . . . 127

6 Chapter 6 **132**
6.1 Introduction . . . 132
6.2 Error Rates of the QDF in the Literature . . . 133
6.3 Computing the Error Rate for the QDF and its Derivative in the Univariate Case . . . 135
6.4 Derivative of P(1|2) in the Univariate Situation . . . 137
6.5 Error rate for the QDF: Multivariate Normal Populations . . . 139
6.6 Results . . . 141
6.6.1 Univariate populations . . . 141
6.6.2 Multivariate populations . . . 147

7 Chapter 7 – Summary Chapter **160**

8 Bibliography **165**

A Appendix A - Asymptotic Expansions for the Conditional Error Rate of the LDF **172**
A.1 Covariance matrix of the form $\Sigma = \Sigma_A$. . . . 172
A.1.1 Obtaining $\frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}}$. . . . 174
A.1.2 Obtaining $\frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i} \partial \bar{x}_{2j}}$. . . . 178
A.1.3 Obtaining $\frac{\partial^2 \Phi(.)}{\partial s_{kl} \partial s_{ij}}$. . . . 183
A.2 Covariance matrix of the form $\Sigma = \Sigma_B$. . . . 188

B Appendix B - Heuristic Algorithm for Model Selection Procedure using Bhattacharyya Distance **194**

C Appendix C - Other Publications and Conference Presentations of the Author **197**

List of Figures

2.1 $|\zeta^L|$ and $|\zeta^E|$ for $\Sigma = \Sigma_A$, and various Δ^2 and ρ values ($\rho > 0$). . . 38
2.2 $|\zeta_s^L|$ and $|\zeta_s^E|$ for $\Sigma = \Sigma_A$, and various Δ^2 and ρ values ($\rho > 0$). . . 39
2.3 $|\zeta^L|$ and $|\zeta^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho > 0$). . . 40
2.4 $|\zeta_s^L|$ and $|\zeta_s^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho > 0$). . . 41
2.5 $|\zeta^L|$ and $|\zeta^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$). . . 42
2.6 $|\zeta_s^L|$ and $|\zeta_s^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$). . . 43
2.7 ζ^L and ζ^E for $\Sigma = \Sigma_A$, and various Δ^2 and ρ values ($\rho > 0$). 44
2.8 ζ_s^L and ζ_s^E for $\Sigma = \Sigma_A$, and various Δ^2 and ρ values ($\rho > 0$). 45
2.9 ζ^L and ζ^E for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho > 0$). 46
2.10 ζ_s^L and ζ_s^E for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho > 0$). 47
2.11 ζ^L and ζ^E for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$). 48
2.12 ζ_s^L and ζ_s^E for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$). 49

3.1 The extreme points on the (λ, γ) grid, and what each represents. . . 55

4.1 Equal spherical population covariance matrices. Classifier Error Rate vs. n/p ratio. 98
4.2 Unequal spherical population covariance matrices. Classifier Error Rate vs. n/p ratio. 101
4.3 Equal, highly ellipsoidal population covariance matrices. Population mean differences concentrated in the low variance subspace. Classifier Error Rate vs. n/p ratio. 103
4.4 Equal, highly ellipsoidal population covariance matrices. Population mean differences concentrated in the high variance subspace. Classifier Error Rate vs. n/p ratio. 104
4.5 Unequal, highly ellipsoidal population covariance matrices. Population means equal. Classifier Error Rate vs. n/p ratio. 106

4.6 Unequal, highly ellipsoidal population covariance matrices. Population means unequal. Classifier Error Rate vs. n/p ratio. 107

6.1 Overall error rate (P_e) versus Lambda (λ) when the two population means and variances are similar. ($p = 1, \mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 1$) 142

6.2 Overall error rate (P_e) versus Lambda (λ) when the two population means are similar, but their variances are disparate. ($p = 1, \mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$) 143

6.3 Overall error rate (P_e) versus Lambda (λ) when the two population variances are similar, but their means are disparate. ($p = 1, \mu_1 = 0, \mu_2 = 1, \sigma_1^2 = 0.75$ and $\sigma_2^2 = 1$) 145

6.4 Overall error rate (P_e) versus Lambda (λ) when the two population means and variances are disparate. ($p = 1, \mu_1 = 0, \mu_2 = 3, \sigma_1^2 = 1$ and $\sigma_2^2 = 2$) 146

6.5 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal and spherical covariance matrices ($p = 6$). (i.e. Condition 1 in Chapter 3, Section 3.5) 147

6.6 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal and spherical covariance matrices ($p = 10$). (i.e. Condition 1 in Chapter 3, Section 3.5) 148

6.7 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal and spherical covariance matrices ($p = 20$). (i.e. Condition 1 in Chapter 3, Section 3.5) 148

6.8 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal and spherical covariance matrices ($p = 6$). (i.e. Condition 2 in Chapter 3, Section 3.5) 149

6.9 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal and spherical covariance matrices ($p = 10$). (i.e. Condition 2 in Chapter 3, Section 3.5) 150

6.10 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal and spherical covariance matrices ($p = 20$). (i.e. Condition 2 in Chapter 3, Section 3.5) 150

6.11 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace ($p = 6$). (i.e. Condition 3 in Chapter 3, Section 3.5) 151
6.12 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace ($p = 10$). (i.e. Condition 3 in Chapter 3, Section 3.5) 151
6.13 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace ($p = 20$). (i.e. Condition 3 in Chapter 3, Section 3.5) 152
6.14 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the high variance subspace ($p = 6$). (i.e. Condition 4 in Chapter 3, Section 3.5) 153
6.15 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the high variance subspace ($p = 10$). (i.e. Condition 4 in Chapter 3, Section 3.5) 154
6.16 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the high variance subspace ($p = 20$). (i.e. Condition 4 in Chapter 3, Section 3.5) 154
6.17 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with zero mean differences ($p = 6$). (i.e. Condition 5 in Chapter 3, Section 3.5) 155
6.18 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with zero mean differences ($p = 10$). (i.e. Condition 5 in Chapter 3, Section 3.5) 156

6.19 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with zero mean differences ($p = 20$). (i.e. Condition 5 in Chapter 3, Section 3.5) 156
6.20 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with non-zero mean differences ($p = 6$). (i.e. Condition 6 in Chapter 3, Section 3.5) 157
6.21 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with non-zero mean differences ($p = 10$). (i.e. Condition 6 in Chapter 3, Section 3.5) 158
6.22 Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with non-zero mean differences ($p = 20$). (i.e. Condition 6 in Chapter 3, Section 3.5) 158

List of Tables

2.1 The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "non-equivalence" with $\Sigma = \Sigma_A$. 27

2.2 The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "non-equivalence" with $\Sigma = \Sigma_B$. 28

2.3 The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "equivalence" with $\Sigma = \Sigma_A$. 30

2.4 The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "equivalence" with $\Sigma = \Sigma_B$ and positive ρ. 31

2.5 The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "equivalence" with $\Sigma = \Sigma_B$ and with negative ρ. 32

3.1 Equal Spherical Covariance Matrices. Average error rate (with standard deviation) for several discriminant functions. 63

3.2 Unequal Spherical Covariance Matrices. Average error rate (with standard deviation) for several discriminant functions. 63

3.3 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace. Average error rate (with standard deviation) for several discriminant functions. 64

3.4 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace. Average error rate (with standard deviation) for several discriminant functions. 65

3.5 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. Average error rate (with standard deviation) for several discriminant functions. 65
3.6 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. Average error rate (with standard deviation) for several discriminant functions. 66
3.7 Equal Spherical Covariance Matrices. Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.1. 67
3.8 Unequal Spherical Covariance Matrices. Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.2. 67
3.9 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace. Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.3. 67
3.10 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace. Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.4. 68
3.11 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.6. 68
3.12 Equal Spherical Covariance Matrices. Comparison of SRDF and SRDF1 error rates and regularisation parameter values. 70
3.13 Unequal Spherical Covariance Matrices. Comparison of SRDF and SRDF1 error rates and regularisation parameter values. 71
3.14 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace. Comparison of SRDF and SRDF1 error rates and regularisation parameter values. 71
3.15 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace. Comparison of SRDF and SRDF1 error rates and regularisation parameter values. 72

3.16 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. Comparison of SRDF and SRDF1 error rates and regularisation parameter values. 72
3.17 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. Comparison of SRDF and SRDF1 error rates and regularisation parameter values. 73

4.1 Equal Spherical Covariance Matrices. Average error rate (with standard deviation) and parameter values for several discriminant functions. 78
4.2 Unequal Spherical Covariance Matrices. Average error rate (with standard deviation) and parameter values for several discriminant functions. 79
4.3 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace. Average error rate (with standard deviation) and parameter values for several discriminant functions. . 80
4.4 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace. Average error rate (with standard deviation) and parameter values for several discriminant functions. . 81
4.5 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. Average error rate (with standard deviation) and parameter values for several discriminant functions. 83
4.6 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. Average error rate (with standard deviation) and parameter values for several discriminant functions. 84
4.7 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. Comparison of SRDF-M and SRDF-M1 classifiers. . . . 86
4.8 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. Comparison of SRDF-M and SRDF-M1 classifiers. . . . 87
4.9 Two equal and highly ellipsoidal covariances, one spherical covariance matrices. Mean differences in the low variance subspace: Average error rates with standard deviations. 90

4.10 Three unequal covariance matrices - one highly ellipsoidal, one moderately ellipsoidal, one spherical. Group mean differences spread equally over all subspaces: Average error rates with standard deviations. 91
4.11 Three unequal covariance matrices: Two highly ellipsoidal, one spherical. Zero group mean differences: Average error rates with standard deviations. 92
4.12 Three unequal covariance matrices: one highly ellipsoidal, two spherical. Group mean differences spread evenly over all subspaces: Average error rates with standard deviations. 94
4.13 Comparison of computation times between SRDF-M1 and SRDF. 95
4.14 Equal Spherical Covariance Matrices. Average error rates with standard deviations over a range of n/p ratios. 99
4.15 Unequal Spherical Covariance Matrices. Average error rates with standard deviations over a range of n/p ratios. 100
4.16 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace. Average error rates with standard deviations over a range of n/p ratios. 102
4.17 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace. Average error rates with standard deviations over a range of n/p ratios. 102
4.18 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. Average error rates with standard deviations over a range of n/p ratios. 108
4.19 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. Average error rates with standard deviations over a range of n/p ratios. 108

5.1 Example of (λ, γ) grid of Bhattacharyya distance values $(e_{cv}(B1, B2))$ 113
5.2 Equal, Spherical Covariance Matrices. (Two Groups) Error rate (with standard deviation) for several discriminant functions. 120
5.3 Unequal, Spherical Covariance Matrices. (Two Groups) Error rate (with standard deviation) for several discriminant functions. 121

5.4 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the Low-Variance Subspace. (Two Groups) Error rate (with standard deviation) for several discriminant functions. 121
5.5 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the High-Variance Subspace. (Two Groups) Error rate (with standard deviation) for several discriminant functions. 122
5.6 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. (Two Groups) Error rate (with standard deviation) for several discriminant functions. 122
5.7 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. (Two Groups) Error rate (with standard deviation) for several discriminant functions. 123
5.8 Equal, Spherical Covariance Matrices. (Three Groups) Error rate (with standard deviation) for several discriminant functions. 123
5.9 Unequal, Spherical Covariance Matrices. (Three Groups) Error rate (with standard deviation) for several discriminant functions. 124
5.10 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the Low-Variance Subspace. (Three Groups) Error rate (with standard deviation) for several discriminant functions. 124
5.11 Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the High-Variance Subspace. (Three Groups) Error rate (with standard deviation) for several discriminant functions. 125
5.12 Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. (Three Groups) Error rate (with standard deviation) for several discriminant functions. 125
5.13 Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. (Three Groups) Error rate (with standard deviation) for several discriminant functions. 126
5.14 Ratios of CPU times required for each method (SRDF-B/SRDF). . 127

6.1 Error rates in the case of similar population means and variances. ($\mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 1$) 143
6.2 Error rates in the case of similar population means, but variances disparate. ($\mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$) 144

6.3 Error rates in the case of similar population variances, but disparate means. ($\mu_1 = 0, \mu_2 = 1, \sigma_1^2 = 0.75$ and $\sigma_2^2 = 1$) 145
6.4 Error rates for the case of disparate population means and variances. ($\mu_1 = 0, \mu_2 = 3, \sigma_1^2 = 1$ and $\sigma_2^2 = 2$) 146

List of Abbreviations used in this Thesis

SLDF	Sample linear discriminant function.
SQDF	Sample quadratic discriminant function.
SRDF	Sample regularised discriminant function, similar to the method developed by Friedman (1989).
SRDF1	Rule based on SRDF, but where a policy of minimum regularisation (instead of maximum regularisation as with SRDF) is employed to break ties in cases where the model selection procedure does not yield a unique choice of values for the regularisation parameters.
SRDF-M	A modified regularised rule which omits the eigenvalue shrinkage parameter γ but allows for as many covariance mixing parameters (λ's) as there are groups to be discriminated between. This rule is scale-invariant, unlike the SRDF.
SRDF-M1	Similar to SRDF-M, but where a policy of minimum regularisation is employed to break ties (as for SRDF1).
SRDF-B	Regularised discriminant rule which chooses the λ and γ parameters by using information obtained from a measure of the Bhattacharyya distance between pairs of populations (of interest).
SEDF	Sample Euclidean distance function. In this thesis, this rule is formed by setting the regularisation parameters (λ and γ) in the SRDF rule both equal to one.

Chapter 1

INTRODUCTION

1.1 GENERAL FRAMEWORK

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new observations to previously defined groups. As a separatory procedure, it is often employed on a one-time basis in order to investigate observed differences when casual relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well defined rules, which can be used for assigning new objects.

An assumption underlying the use of discriminant analysis is that there is a way of correctly classifying the initial data. In other words, there must exist some variable or variables which allow the different groups to be established and definitively identified. For example, in a study of prostrate cancer, measurements from a biopsy would be used to define the groups "cancer" and "non-cancer". Or, in a study to determine if the New Zealand kiwi bird will be susceptible to rabbit calicivirus disease (RCD), a virus which kills rabbits in large numbers (to attempt to control the rabbit plague in some parts of New Zealand), tissue culture from various organs of a kiwi which has been exposed to the virus are taken and examined for the presence or absence of antibodies against the disease, and various measurements are taken.

These variables cannot be used directly to predict the group to which an individual belongs. In many instances these variables are difficult to obtain. In the

prostrate cancer example, a biopsy is not always practical (for reasons of expense and discomfort) for all patients. Rather, only those who are very likely to have the disease will be operated on. In the kiwi example, the destruction of the bird is necessary to obtain the required tissue samples which will indicate conclusively whether antibodies against the disease have been produced, and thus indicate how susceptible the kiwi is to the disease. Since the kiwi is the national emblem of New Zealand and its population is extremely small, any such experimentation would have to be limited to just one or two birds. Thus in most problems, other variables will be used which are more readily available, or less invasive and destructive. It is hoped that these variables will be sufficiently sensitive and indicatory to allow an accurate assignment to be made.

The formal purpose of discriminant analysis is to assign objects to one of several (K) populations or groups defined *a priori*. The assignment is based on a set of p measurements $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ obtained from p variables from each object. If each variable is thought of as an axis in a metric space, the observations, $\boldsymbol{x}$, are points in p–dimensional measurement space. Different groups would ideally occupy different regions in the measurement space as this would allow allocation methods to assign observations based on their locations in the space. Often the different regions overlap, and correct allocation is not possible every time. Nevertheless, it is important that the assignment of an unknown observation to a group be carried out with a small probability of misclassification (often referred to as the "error rate").

The measurements associated with the population of observations in the k^{th} group comprise a distribution of values with probability density function (pdf) $f_k(\boldsymbol{x})$, $k = 1, \ldots, K$. The optimal (Bayes) rule for allocating an observation $\boldsymbol{x}$ is arrived at through minimising the total probability of misclassification under the assumption that all group parameters are known (see for example, Seber (1984), Section 6.2.2). The rule may be written as: choose group $\hat{k}$ such that

$$\pi_{\hat{k}} f_{\hat{k}}(\boldsymbol{x}) = \max_{1 \leq k \leq K} \{\pi_k f_k(\boldsymbol{x})\}, \tag{1.1}$$

where π_k is the *a priori* probability that $\boldsymbol{x}$ belongs to the k^{th} group.

Given the commonly used assumption that the groups are normally distributed,

the following rule is obtained: assign $\boldsymbol{x}$ to group $\hat{k}$ such that

$$d_{\hat{k}}(\boldsymbol{x}) = \min_{1 \leq k \leq K} \{d_k(\boldsymbol{x})\} \tag{1.2}$$

where

$$d_k(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_k)' {\Sigma_k}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln|\Sigma_k| - 2\ln\pi_k \tag{1.3}$$

and $\boldsymbol{\mu}_k$ and Σ_k represent the mean vector and covariance matrix for group k. The quantity $d_k(\boldsymbol{x})$ is often called the discriminant score for allocation of observation $\boldsymbol{x}$ to the k^{th} group, but is sometimes referred to as the generalised distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_k$. The first term of $d_k(\boldsymbol{x})$ is the square of the well known Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_k$, while the other two terms are adjustment factors. The quantity $d_k(\boldsymbol{x}) + 2\ln\pi_k$ is called the quadratic discriminant function (QDF) since it separates the different regions in the measurement space (corresponding to different group classifications) by quadratic boundaries. In practice, the parameters $\boldsymbol{\mu}_k$ and Σ_k will not be known and may be replaced by the usual estimates $\bar{\boldsymbol{x}}_k$ and $\mathbf{S}_k$ respectively (defined in expression (1.9)). The sample discriminant rule is to assign $\boldsymbol{x}$ to group $\hat{k}$ such that

$$\hat{d}_{\hat{k}}(\boldsymbol{x}) = \min_{1 \leq k \leq K} \left\{\hat{d}_k(\boldsymbol{x})\right\}, \tag{1.4}$$

and the sample quadratic discriminant function (SQDF) is

$$\hat{d}_k(\boldsymbol{x}) + 2\ln\pi_k = (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)' {\mathbf{S}_k}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_k) + \ln|\mathbf{S}_k|. \tag{1.5}$$

The performance of the SQDF can be badly affected if the training sample size is small, and this is due especially to the instability of the estimates, $\mathbf{S}_k$. If n_k is close to or less than p, $\mathbf{S}_k$ may be singular or nearly singular and some elements of ${\mathbf{S}_k}^{-1}$ will have extremely large or infinite values, with serious consequences for expression (1.5). Various approaches to addressing the problem of not being able to obtain stable or reliable estimates of the Σ_k have been adopted. The general theme throughout this thesis deals with allocation rules which are used to counteract problems associated with the estimation of the covariance matrices and their effects on discrimination.

A common way to overcome instability in the $\mathbf{S}_k$ is to use the linear discriminant rule, which also assumes normality but with the additional assumption that all

groups have equal covariance matrices. The resulting decision boundaries between groups are linear, and the $\mathbf{S}_k$ are replaced by the pooled sample covariance matrix, $\mathbf{S}_p$ (see expression (1.11)). The resulting sample linear discriminant function (SLDF) may be written (assuming equal priors and costs of misclassification)

$$\hat{d}_k(\boldsymbol{x}) + 2\ln\pi_k = -2\boldsymbol{x}'\mathbf{S}_p{}^{-1}\bar{\boldsymbol{x}}_k + \bar{\boldsymbol{x}}_k'\mathbf{S}_p{}^{-1}\bar{\boldsymbol{x}}_k. \tag{1.6}$$

and the rule is the same as that in expression (1.4). In effect, the $\mathbf{S}_k$ are biased towards the pooled covariance matrix, but the variance of the elements of $\mathbf{S}_p$ are smaller than the variances of the corresponding elements of the $\mathbf{S}_k$. This reduction in variance enables the SLDF to out-perform the SQDF for small sample sizes even when the $\boldsymbol{\Sigma}_k$ differ (see, for example, Marks and Dunn (1974), Wahl and Kronmal (1977), Bayne et al. (1983)).

One of the simplest allocation rules adopts the approach of ignoring the covariance matrix and assigning an unknown observation $\boldsymbol{x}$ to one group on the basis of the Euclidean distance between $\boldsymbol{x}$ and each group mean, $\boldsymbol{\mu}$. The resulting (sample-based) nearest-means classifier is termed the sample Euclidean distance function (SEDF) and is written as

$$\hat{d}_k(\boldsymbol{x}) + 2\ln\pi_k = (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)'(\boldsymbol{x} - \bar{\boldsymbol{x}}_k). \tag{1.7}$$

The SEDF has been compared to other more commonly used discriminant functions including the SLDF by Raudys and Pikelis (1980) and Marco et al. (1987). It was shown to perform well in comparison, especially when the group conditional distributions are spherically normal, and when the dimensionality is large relative to the training sample size.

A different way of addressing the problems associated with estimating the $\boldsymbol{\Sigma}_k$ in expression (1.3) is to employ shrinkage techniques on the covariance estimates. James and Stein (1961), Stein et al. (1972), Efron and Morris (1976), Haff (1980) and Dey and Srinivasan (1985) sought to obtain more reliable eigenvalue estimates, correcting eigenvalue distortion present in the sample covariance matrix. Further details are given in Section 3.2 of this thesis. The approach involved seeking estimates that minimise particular loss criteria on the eigenvalue estimates. Regularisation of the covariance matrix is a similar technique that has been used for situations where an estimate $\mathbf{S}_k$ is singular or nearly singular, as can occur when the number of parameters to be estimated is similar to the number of training sample observations available. Regularisation attempts to improve an estimate by

biasing it from the estimated value to a value deemed physically plausible. An example of such a plausible value to bias the individual covariance estimates $\mathbf{S}_k$ to is $\mathbf{S}_p$. This value would be appropriate in many cases where the $\boldsymbol{\Sigma}_k$ are not greatly heterogeneous. If varying degrees of biasing to the chosen value is permitted, new (regularised) covariance estimates may be obtained to produce intermediate models between the heteroscedastic (SQDF) and homoscedastic (SLDF) models — the former being possibly too diffuse, and the latter perhaps too rigid. The effect of this is to reduce the variance of the sample estimate at the expense of potentially increasing its bias (see Friedman (1989)).

1.2 ERROR RATES

The error rate associated with the optimal or Bayes rule is the probability that a randomly selected individual from group i is misallocated to group j $(i, j = 1, \ldots, K (i \neq j))$ on the basis of the optimal allocation rule which assumes that the parameters are known. Since the optimal rule minimises the total probability of misclassification, this probability is known as the optimal error rate. It is directly related to the degree of separation between the groups.

In practice, the optimal rule and optimal error rate are not achievable and allocation rules must be constructed on the basis of available training samples. The conditional error rate of a sample-based rule is the probability, conditional on the sample, that a randomly selected individual from group i is allocated to group j $(i, j = 1, \ldots, K;\ (i \neq j))$. This is sometimes referred to as the actual error rate. The expected value of the conditional error rate (on averaging the conditional error rates over the distribution of the training sample) is termed the expected actual, or unconditional error rate (see Lachenbruch (1975), Hand (1986), Chapter 2 of this thesis, and McLachlan (1992) Section 1.10). This terminology was established by Hills (1966).

The optimal, conditional and unconditional error rates of a sample-based rule depend on the usually unknown population parameters, and as such, these error rates must be estimated in practice. Estimating techniques, whether parametric or non-parametric, are strictly functions of the sample data and they have usually been evaluated regarding their performance in estimating the conditional error rate. Glick (1978) alludes to some of the difficulties involved here:

> The task of estimating probabilities of correct classification confronts the statistician simultaneously with difficult distribution theory, questions intertwining sample size and dimension, problems of bias, variance, robustness and computation costs.

A commonly used error rate estimator is the plug-in error rate, obtained by replacing the unknown group parameters by their sample estimates in the available expressions for the conditional error rates. For a comprehensive summary of error rate estimators, including relevant references, see McLachlan (1992), Chapter 10.

Analytical results for the conditional error rates of the sample-based Bayes discriminant rules have proved difficult to obtain because of the complexity of the distributions of the various discriminant functions. A few such results have been obtained, but only for very special cases, such as for only two normal groups with equal covariance matrices. Most of the problems involving the distributions of the discriminant functions and their associated error rates have been tackled using asymptotic methods. McLachlan (1992), Chapter 4, gives a thorough summary of the available error rate results for the case of multivariate normal groups.

1.3 OUTLINE OF RESEARCH UNDERTAKEN

This thesis begins where the unpublished work of Lim (1992) left off — comparing the Euclidean and linear discriminant functions. The linear discriminant function, first proposed by Fisher (1936), is still very popular, partly due to its optimal properties when the parameters are known (Anderson, T. W. (1984)). However, since it is recognised that the SLDF is not uniformly optimal and its performance can be poor relative to other classifiers when the dimension is large relative to the training sample size (Peck and van Ness (1982)), the SEDF has been identified as a possible competing allocation procedure for discriminant analysis (Raudys and Pikelis (1980), Marco et al. (1987)). Raudys and Pikelis employed numerical integration techniques in their study, while Marco et al. demonstrated the superiority of the SEDF over the SLDF in certain conditions through a Monte Carlo simulation experiment and comparing estimated error rates (in the form of probabilities of *correct* classification).

Lim (1992) embarked on a study to compare the expected conditional error rates

(i.e. the unconditional error rates) of the SEDF and SLDF via asymptotic expansions of the error rates for the case of two multivariate normal groups. Non-trivial conditions for achieving equivalence of the SEDF and SLDF, when all parameters are known, were derived by Marco et al. (1987) and this result provides an appropriate scenario to allow a fair comparison of the two classifiers. Lim obtained the expected error rates of the SEDF and SLDF for conditions where the classifiers are not equivalent, but was unable to derive the asymptotic expansions for the conditional error rate of the SLDF under conditions of equivalence, due to the complexity of both the differentiation and evaluation of the final expression. Therefore, to enable a satisfactory comparative study of the error rates of the two classifiers, the above asymptotic expansions (under conditions of equivalence) are derived in Chapter 2 and numerically evaluated to obtain the expected error rates. Chapter 2 contains a comparative study of the SEDF and SLDF, using a different approach to that of Raudys and Pikelis (1980) and Marco et al. (1987). The conditions under which comparison is made are also broadened to include different structures for the group covariance matrix.

The focus in Chapter 3 changes from looking at rigid techniques to deal with the previously mentioned instability in the group covariance estimates in discriminant functions, to the very flexible technique of covariance regularisation. Avoiding estimation of the Σ_k, as occurs when using the SEDF, is shown in Chapter 2 (and indeed in subsequent chapters) to be a very useful procedure in a number of situations. Nevertheless, it is an extreme procedure. As a different approach, regularisation of the type devised by Friedman (1989), and described fully in Chapter 3, allows for intermediate rules between the heteroscedastic and homoscedastic models. Furthermore, it allows for intermediate rules between those based on expression (1.5) which employ covariance estimates, and those nearest-means rules which are largely based on the Euclidean distances from an unknown observation to the various group means, but perhaps weighted by a scalar based on the covariance estimates. Such a compromise is made possible by employing two separate regularisation parameters to obtain estimates of the Σ_k. Each is a continuous variable over the range [0,1]. One parameter controls shrinkage of the heteroscedastic estimates, $\mathbf{S}_k$, towards the pooled estimate, $\mathbf{S}_p$. The other parameter controls the strength of biasing towards a multiple of the identity matrix, $\mathbf{I}$. The identity matrix is used by Friedman(1989), but there is nothing special about it, and other matrices could

be used.

The selection of these parameters leads to the regularisation model, and training sample information is used to select values, which, it is hoped, are at least approximately optimal in terms of leading to discriminant models with minimal error rates. Friedman used the re-sampling technique of cross-validation (see, for example, Lachenbruch and Mickey (1968), Lachenbruch (1975)) to obtain estimates of the regularisation parameters, since appropriate values to use are unlikely to be known in advance. Rayens and Greene (1991) pointed out that this technique often may not yield a unique value, and that in such cases a "tie-breaking" policy must be implemented to select one value to use in the model. A Monte Carlo simulation study is described in Chapter 3, and the purpose is to give an indication of the effects of this action on the main criterion that is used in this thesis to assess the performance of classifiers — the estimated overall error rate. Simulation work must be relied upon and the estimated overall error rate used in any comparison of the regularised rule with the other discriminant functions, since no analytic results concerning the distribution of error rates of such regularised rules exist in the literature. Even for the QDF, in the case of unequal group means and covariances, exact expressions for the conditional probabilities of misclassification have been obtained only for the case of $p = 1$.

The regularised rule as devised by Friedman (1989), which we shall term the sample regularised discriminant function (SRDF), is not generally scale invariant. This is not a desirable characteristic of a discriminant rule and arises through the use of the regularisation parameter which allows shrinkage to the identity matrix. In Chapter 4 a modification of the SRDF is presented which removes this parameter but attempts to compensate for the loss by introducing group-conditional regularisation parameters controlling shrinkage to the pooled covariance matrix. This means that each group covariance matrix is able to be regularised to $\mathbf{S}_p$, to an extent that is appropriate for that group rather than biasing all group covariances to the same degree. This modified regularisation rule is compared to the SRDF and the other more common discrimination rules. It is found that it performs reasonably well, although in the high dimensional settings especially, where the covariance estimates suffer from high variance and bias, Friedman's original SRDF still proved superior. This shows the importance of the second type of regularisation towards the identity matrix which shrinks the eigenvalues of the $\mathbf{S}_k$ towards

equality. The effect is to dampen the variance in the high variance subspace, a procedure that clearly seems to appear to enhance discrimination.

Thus it appears that the eigenvalue shrinkage technique of the regularised rule gives this rule the edge in a number of situations over the other rules not employing the technique (See also Aeberhard et al. (1994)). However, it is assumed that this advantage would only be apparent when the sample size to dimension ratio is small, since it is in these situations that most problems involving estimation of the population parameters occur. The advantage would be expected to diminish as the sample size increases in relation to the dimension. To examine this proposition, a further simulation study was undertaken to compare the performances of the other previously introduced discriminant rules over a range of sample size to dimension ratios for a variety of simulation conditions as in the previous simulation studies. This study is also presented and discussed in Chapter 4, and results are presented. The goal is to determine if, for a given situation, there comes a point where the sample size is sufficiently large relative to the dimension such that the eigenvalue shrinkage technique of the regularised rule no longer is advantageous for discrimination. This can be ascertained by comparing the regularised rule to those rules without the eigenvalue shrinkage facility.

In Chapter 5, some of the criticisms of the model selection (i.e. regularisation parameter selection) of the SRDF are addressed. In addition to the observation made earlier that a unique choice of parameter values may not usually be available, Rayens and Greene (1991) pointed out that often the choice is determined by only a small portion of the data available. This phenomenon arises through the use of error rates (i.e. misclassification probabilities), empirically obtained from the training sample, as the criterion for choosing the regularisation parameters. In order to address these potential weaknesses, a different approach to the model selection procedure used by Friedman is considered. Friedman used the criterion of estimated error rate based on the training data, and employed the empirical technique of cross-validation to estimate the error rates. In Chapter 5, a criterion of "distance between groups" is employed to gain information from the training sample regarding appropriate values of the regularisation parameters. The goal of Friedman's model selection procedure is to choose values that lead to the formation of a discriminant rule which seeks to allocate unknown or test observations with as small an error rate as possible. Therefore, it is a direct approach to use the training

sample error rate as the criterion upon which to base the choice of regularisation parameters. On the other hand, distance measures are at best indirect indicators of the conditional error rate of a discriminant rule. In fact, exact results linking certain distance measures to error rate are available only when population parameters are known and then usually in terms of bounds on the error rate. Nevertheless, articles in the literature suggest some measures of distances between two groups that can provide information about appropriate regularisation parameter values to use for a given set of data.

The sample Bhattacharyya distance between two groups with the simplifying assumption of normality is a popular measure of similarity (see Fukunaga (1972)) and is the distance criterion used here. Under this assumption the expression comprises two terms: one which is very similar to the familiar Mahalanobis distance which primarily measures the shift in means between the groups, and one which is a measure of the covariance shift, and which involves determinants of the covariance estimates. Despite the latter term being more seriously affected by bias than the former (Fukunaga and Hayes (1989)), it gives an indication of the similarity of the group covariances and thus the appropriate degree of regularisation to the pooled covariance matrix. Similarly, if the covariance shift term dominates the mean shift term in the Bhattacharyya distance expression, it may indicate that eigenvalue shrinkage needs to be employed to reduce the variation and bias in the estimates.

Thus, while the distance measure approach is relatively crude in terms of not drawing on established analytical results but rather relying on empirical data and empirically derived "rules of thumb", it does afford advantages over the model selection procedure in Friedman's rule. Firstly, all the available training data contributes to selection of the regularisation parameters. Secondly, a unique choice of those parameters is obtained, avoiding arbitrary procedures to break ties. Finally, re-sampling techniques are avoided, thus leading to a much faster computational procedure. The discriminant rule developed is tested in a simulation study against Friedman's rule as well as the other rules used throughout this thesis for comparison. It is also extended to the case of three groups. Several case studies are also presented with real data sets incorporated as part of the comparative analysis of the various rules.

Most of the work thus far regarding Friedman's regularised rule involves comparing its estimated (conditional) error rate with that of other methods. To the best

of the author's knowledge, no exact analytic results regarding error rates are available in the literature which incorporate the effects of the regularisation parameters. Houshmand (1993) provided exact expressions for computing the probabilities of misclassification for the univariate QDF with two groups and known covariance matrices. Since the effect of introducing regularisation of the covariance matrix on error rate has only been studied via simulation experiments, it is of considerable interest to attempt to describe this effect with analytic expressions. In Chapter 6, the exact expressions given by Houshmand for the error rate of the QDF are differentiated with respect to the covariance mixing parameter. The resultant expression, after evaluation, provides information on the rate of change of the error rates with respect to the regularisation parameters. Thus analytic results can be compared with the empirical results obtained. Using the algorithms of Lau (1980) and Narula and Desu (1981) the integrals in the derivative expressions are computed, and the derivatives evaluated, for several combinations of population parameters and over the range of values of the regularisation parameter. From the limited analytical results obtained, confirmation of some results from earlier chapters is made.

1.4 NOTATION AND DEFINITIONS

Some notation that is used throughout this thesis will be established in this section, and a few well known results rewritten for convenience since they will be used extensively elsewhere in this work.

Vectors and matrices are written in bold type. The transpose, trace and determinant of a matrix $\mathbf{M}$ are denoted by $\mathbf{M}'$, $\mathrm{tr}\{\mathbf{M}\}$ and $|\mathbf{M}|$ respectively, and $\mathbf{I}$ is the identity matrix. The symbol $\phi(.)$ denotes the standard normal density function, given by

$$\phi(x) = (2\pi)^{-1/2}\exp\left\{-x^2/2\right\},$$

and the integral of $\phi(x)$ from $-\infty$ to y is denoted $\Phi(y)$, the (cumulative) normal distribution function.

The square of the Mahalanobis distance between two groups or populations Π_1 and Π_2 with means $\boldsymbol{\mu}_i$ $(i = 1, 2)$ and common covariance matrix $\boldsymbol{\Sigma}$ is

$$\Delta^2 = [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \tag{1.8}$$

where Δ is taken to be positive.

The maximum likelihood estimates of the mean and covariance matrix computed from a training sample from group k are

$$\begin{aligned} \hat{\boldsymbol{\mu}}_k &= \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_i \ (= \bar{\boldsymbol{x}}_k) \\ \text{and} \\ \hat{\boldsymbol{\Sigma}}_k &= \frac{1}{n_k}\sum_{i=1}^{n_k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)' \end{aligned} \tag{1.9}$$

respectively, where n_k is the size of the sample. The estimator $\hat{\boldsymbol{\Sigma}}_k$ is biased, so that $\boldsymbol{\Sigma}_k$ is estimated by the usual sample covariance matrix

$$\mathbf{S}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)' \tag{1.10}$$

The pooled sample covariance matrix for K samples is

$$\mathbf{S}_p = \frac{1}{N - K}\sum_{k=1}^{K} \mathbf{S}_k(n_k - 1) \tag{1.11}$$

where

$$N = \sum_{k=1}^{K} n_k. \tag{1.12}$$

The simulation experiments undertaken in this thesis were implemented using MATLAB™ (The MathWorks, Inc. (1992)). The built-in random number generators *rand* and *randn* were used to generate the synthetic data for the simulation studies in Chapters 2 through 5.

Chapter 2

COMPARISON OF THE LINEAR AND EUCLIDEAN DISCRIMINANT FUNCTIONS

2.1 INTRODUCTION

In parametric statistical discriminant analysis, the linear discriminant function (LDF), which is based on assumptions of multivariate normality and equal covariance matrices, is quite popular because of its robustness and simplicity. Clearly, there are situations when the LDF is inappropriate and related competitors like the quadratic discriminant function (QDF) or the Euclidean distance function (EDF) may be used instead; see, for example, McLachlan (1992, Chapters 3 and 5). In this chapter the particular interest is to compare the LDF with the simpler EDF via their asymptotic error rate under prescribed conditions.

In giving the background for this study, it is necessary to revise some related literature whose results motivated this study to compare the LDF with the EDF. There has been considerable interest in the literature in the relative performances of these discriminant functions. These comparisons have usually been based on various measures or estimates of error rates (probabilities of misclassifications) since direct algebraic evaluations of some of these probabilities for unknown population parameters have proved intractable. The main references are summarised below,

and these provide comparisons between the EDF and the LDF (and sometimes a few additional discriminant functions) under various conditions and assumptions.

2.2 LITERATURE REVIEW

2.2.1 Raudys and Pikelis (1980)

Raudys and Pikelis (1980) performed a comparative study of four classifiers: the sample EDF (SEDF – see expression (1.7)), the sample LDF (SLDF – see expression (1.6)), the sample QDF (SQDF – see expression (1.5)) and a variant of the SLDF for independent measurements (where the off diagonal elements of the pooled sample covariance matrix $\mathbf{S}_p$ are set to zero). The performance of each discriminant function was evaluated when allocating individuals from two spherical normal populations. A second aim of their study was to monitor the effects of training sample size, n, and dimensionality, p, on error rates. All the classifiers used are Bayes procedures for normal populations that differed only in their assumptions on the structure of the covariance matrices. The error rates were obtained through numerical integration.

An exact expression for the expected value of the probability of misclassification (conditional on the sample size) for the SLDF was derived by Sitgreaves (1961). This expected conditional error rate is the unconditional error rate for the classifier. However, Sitgreaves' expression was found to be computationally impractical and was reduced into a form suitable for numerical calculation by Estes (1965) in his unpublished work. This latter result is used to calculate the unconditional error rate for the SLDF in their paper. Also, the expected value of the conditional error rates for the SQDF and SEDF were derived in the form of non-closed integrals, and solved numerically to estimate the unconditional error rate for the classifiers. The unconditional error rate for the SLDF for independent measurements was studied by approximate formulae and by simulation. The expected values of the conditional error rates for the SLDF and SQDF were evaluated in the case of spherical normal populations with $\mathbf{\Sigma}_i = \mathbf{I}$, $i = 1, 2$.

The authors (Raudys and Pikelis (1980)) also performed a simulation study using four sets of data from various populations, and compared the error rate of each classifier. A major result of the simulation study was that the SEDF performed better than the SLDF when p is large relative to the training sample size, n. In

fact, over the whole study, the SEDF performed at least as well as the SLDF, even for some non-spherical covariance configurations.

2.2.2 Peck and Van Ness (1982)

Peck and Van Ness (1982), noted that one problem with using the SLDF is that the unbiased population parameter estimators in high dimensions are often of poor quality. This applies particularly to the estimates of the population covariance matrix. The problem is often evident even for Gaussian data. A shrinkage estimator for the covariance matrix in the SLDF was investigated to try to ascertain its effect in addressing this problem. A shrinkage estimator is usually a function of $\mathbf{S}_p{}^{-1}$, where $\mathbf{S}_p$ is the pooled sample covariance matrix. This function then replaces $\mathbf{S}_p{}^{-1}$ in the SLDF.

There are a number of shrinkage estimators, including the characteristic roots method (Stein (1975)), the correlation matrix methods (Lin, S. (1978)) and the empirical Bayes method (Haff (1979, 1980)). Lin, H. (1979) compared the three approaches via a Monte Carlo study and concluded that for many, but not all covariance structures, the characteristic roots method and correlation matrix method out-performed the classical estimator used in the standard SLDF. The empirical Bayes method improved upon the classical estimator for all covariance structures and because of this latter fact, Peck and Van Ness chose the empirical Bayes method, where the pooled sample estimate of the (assumed common) population covariance matrix is replaced by a function of it (called the Bayes estimator),

$$B = (1 - \mathcal{T}(U))(2n - p - 3)\mathbf{S}_p{}^{-1} + \left(\frac{\mathcal{T}(U)b}{\mathrm{tr}\{\mathbf{S}_p\}}\right)\mathbf{I}. \tag{2.1}$$

Here b is a positive constant, U is a measure of disparity among the sample (covariance) eigenvalues (it is the geometric mean of the eigenvalues divided by the arithmetic mean),

$$U = \frac{p|\mathbf{S}_p|^{1/p}}{\mathrm{tr}\{\mathbf{S}_p\}},$$

the function $\mathcal{T}(.)$ is a non-decreasing solution to

$$(2n - p - 1)\mathcal{T}^2 - 4\mathcal{T} + (4U/p)\mathcal{T}' < 0$$

and $0 \leq \mathcal{T}(U) \leq 1$. Here, it is assumed equal training sample sizes, i.e. $n_1 = n_2 = n$ in the two-group case.

Peck and van Ness made the assumption that the common group covariance matrix, Σ, was a diagonal matrix with all leading diagonal elements equal to some constant σ^2, so that

$$\frac{b}{\text{tr}\{S_p\}}\boldsymbol{I}$$

is a natural estimator of Σ^{-1}. The quantity b was chosen to be $p(2n-2)-2$, so as to yield an unbiased estimate of Σ^{-1}. Their simulation results showed that the discriminant function using shrinkage estimators performed better than the standard SLDF in most cases, but this improvement was highly dependent on the Mahalanobis distance between the two populations. A further conclusion was that if the Euclidean distance between the population means is small then the shrinkage estimator is of little use because the effect of poor estimation of the population means on the probability of correct classification is more detrimental than the effect of poor estimation of the covariance matrix.

2.2.3 Marco, Young and Turner (1987)

A Monte Carlo simulation study by Marco et al. (1987) compared the SLDF and SEDF under conditions derived so as to ensure that the two classifiers were (i) equivalent and (ii) non-equivalent. Here "equivalence" means that the classifiers have the same true error rates (see Section 2.2.4), and arises from a contrived arrangement of the population parameters, which are assumed known.

The results of their simulation study indicated that the SEDF out-performs the SLDF when the underlying parameter configurations are such that the SEDF is equivalent to the SLDF, assuming all parameters are known. The SEDF appears to do as well as the SLDF even for non-equivalent situations. Another feature of their results was that when the ratio of the Mahalanobis distance to Euclidean distance is small, the SEDF tends to perform better than the SLDF, whereas when the ratio is large, the converse is true. Overall, the studies by Marco et al. showed that the SEDF performs as well as or better than the SLDF especially when the Euclidean and Mahalanobis distances were similar, and when the dimension of the data is large relative to the training sample sizes, and the variables in the data are mildly or moderately correlated (positive). Simulations under conditions of medium to high correlation were not performed, but it was stated that the results suggested the SEDF would also perform very well in such conditions.

An obvious advantage of the SEDF over the SLDF is its computational simplicity, which may be important with high dimensional data in real-time pattern recognition systems. Marco et al. also stated that a further advantage of the SEDF is that the SEDF is invariant to correlated training observations whereas the SLDF is not. This claim is unsubstantiated, however, and it appears that there is confusion over the type of correlation meant. The comparison is made to the SLDF, which is indeed affected by such correlation probably because the assumption of independence is not satisfied within the training samples. (See Basu and Odell (1974), Tubbs (1980), Lawoko and McLachlan (1983) and Koolaard and Lawoko (1993). However, these papers deal with correlation between training observations specifically, rather than correlation between measurement variables such as when Σ is as in (2.15), which is the only type of correlation used in the paper by Marco et al.

Comparison of their results with those of Peck and Van Ness (1982) show that the SEDF performs as well as or better than the SLDF using a shrinkage estimator of Σ. Exceptions to this include the case when the underlying parameter configurations are such that the Mahalanobis distance is considerably larger than the Euclidean distance, and the latter is of a small magnitude.

2.2.4 Implications of results from Marco, Young and Turner (1987)

This paper derives conditions under which the LDF and the EDF are "equivalent" (i.e. have the same overall error rates for known population parameters). The authors also report results of simulation studies to compare the SLDF and SEDF not only under conditions of equivalence but also under certain situations of "non-equivalence".

To be specific, in the two-group case, assuming equal prior probabilities and costs of misclassification and common group covariance matrix Σ, it is well known that the "true" error rate (i.e. when all population parameters are known) for misallocating an object from group 1 to group 2 by the SLDF is the same as the error rate for misallocating an object from group 2 to group 1. That is,

$$P_{21}^{L} = P_{12}^{L} = \Phi\left(\frac{-\Delta}{2}\right), \tag{2.2}$$

where Δ is given in expression (1.8). The corresponding error rates for the SEDF

are

$$P_{21}^{E} = P_{12}^{E} = \Phi\left(-\frac{1}{2}\frac{(\mu_1 - \mu_2)'(\mu_1 - \mu_2)}{[(\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2)]^{1/2}}\right). \tag{2.3}$$

Thus the overall error rates for the SLDF and SEDF (to be denoted here by P^L and P^E) are equal to expressions (2.2) and (2.3) respectively.

Marco et al. proved the following related results.

(i) Let $\mathbf{V}$ be a $p \times p$ full rank matrix and $\mathbf{F}$ any $p \times 1$ matrix with pseudo inverse $\mathbf{F}^+$. If $\mathbf{F}\mathbf{F}^+$ and $\mathbf{V}^{-1}$ commute then

$$\mathbf{F}'\mathbf{V}^{-1}\mathbf{F} = (\mathbf{F}^+\mathbf{V}\ \mathbf{F}^{+\prime})^{-1}. \tag{2.4}$$

(ii) If we add the requirement that $\mathbf{V}$ be symmetric, then

$$[\mathbf{F}'\mathbf{V}^{-1}\mathbf{F}]^{1/2} = \frac{\mathbf{F}'\mathbf{F}}{[\mathbf{F}'\mathbf{V}\mathbf{F}]^{1/2}}. \tag{2.5}$$

(iii) If we set $\mathbf{F} = (\mu_1 - \mu_2)$ and $\mathbf{V} = \Sigma$ in result (ii) where $(\mu_1 - \mu_2)$ and Σ satisfy the requirements for results (i) and (ii), then $P^L = P^E$.

The authors argue that in view of result (iii) "... in many practical situations the SEDF might perform better than the SLDF since considerably fewer parameters must be estimated for the SEDF". Thus, since "... the performance of the SLDF deteriorates significantly as the dimension becomes large relative to the training sample sizes, the computationally simpler SEDF may be the preferred discrimination algorithm in this situation". In view of this last argument, the authors conjectured that "the SEDF may perform as well as the SLDF even for (some) 'non-equivalent' situations". The authors then performed a simulation experiment for a very special structure of Σ and concluded that there were indeed situations when the SEDF performed better than the SLDF. They found that the "improvement of the SEDF over the SLDF is highly dependent on the ratio of Mahalanobis distance to Euclidean distance". In particular, "whenever this ratio is small, the SEDF tends to out-perform the SLDF, (and) when the ratio is large the reverse is true".

One possible explanation for the observed relative behaviours of the two discriminant functions follows from Peck and Van Ness (1982), who conjectured that

all this is due to the relative effects of the errors in estimating Σ to that in estimating μ_1 and μ_2, and the relative seriousness of these effects depends on the sizes of the Mahalanobis and Euclidean distances; see the original article or Marco et al.(1987) for further details and illustrations. Of course, in "non-equivalent" situations when the SLDF has a lower (true) error rate than the SEDF, it would be anticipated that the SLDF would perform better than the SEDF.

On the matter of when the SEDF performs better than the SLDF, consider the proof of result (iii) in Marco et al. (1987), where, if the conditions for the result are satisfied, then

$$\frac{(\mu_1 - \mu_2)'(\mu_1 - \mu_2)}{[(\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2)]^{1/2}} = [(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)]^{1/2}. \tag{2.6}$$

By swapping Σ and Σ^{-1} in the above result, we get the equivalent result that

$$\frac{\Delta_E^4}{\Delta^2} = \frac{[(\mu_1 - \mu_2)'(\mu_1 - \mu_2)]^2}{[(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)]} = (\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2), \tag{2.7}$$

where Δ_E is the Euclidean distance between the two populations. Thus the size of the ratio between the distance measures (Euclidean and Mahalanobis) can be reduced to an explicit function of the elements of μ_1, μ_2 and Σ.

Without loss of generality, one can set $\mu_2 = (0, 0, \ldots, 0)'$. Marco et al. (1987) set the values of $\mu_1 = (m, m, \ldots, m)'$ under "equivalence" and $\mu_1 = (m^*, 0, 0, \ldots, 0)'$ under "non-equivalence", where m and m^* are appropriately chosen scalars so that the Mahalanobis distances can be set equal (under equivalence and non-equivalence) for purposes of comparison. This study concentrates on the "equivalence" situation since it provides fair comparison between the two discriminant functions (both being optimal Bayes procedures for known population parameters under "equivalence"). Since $\mu_2 = (0, 0, \ldots, 0)'$ and $\mu_1 = (m, m, \ldots, m)'$ in this situation, it follows that $\Delta_E^2 = pm^2$ and

$$\frac{\Delta_E^2}{\Delta^2} = \left(\frac{\Delta_E^4}{\Delta^2}\right) \times \frac{1}{\Delta_E^2} = (\mu_1'\Sigma\mu_1)/pm^2 = \sum_{i,j} \sigma_{ij}/p, \tag{2.8}$$

where $\Sigma = \{\sigma_{ij}\}$.

Thus the only factors which determine the size of the ratio of the two distance functions in this situation are the elements of the covariance matrix, Σ. If, as in Marco et al., standardisation is done and the covariance matrix is effectively a correlation matrix, then it follows that in general high positive correlations yield

large values of $\sum_{i,j} \sigma_{ij}$. If there is no standardisation of the observation vectors then large (small) variances and/or large positive (negative) correlations would result in large (small) values of $\sum_{i,j} \sigma_{ij}$.

Note that in their discussions Marco et al.(1987) refer to the size of the ratio of Δ^2 to Δ_E^2, which is the reciprocal of the ratio in expression (2.8). In terms of the ratio in (2.8) these authors' simulation experiments suggest that: SEDF performs better (worse) than SLDF when $\sum_{ij} \sigma_{i,j}$ is large (small). It can be concluded from the arguments above that it is the type and extent of correlations (or covariances) among the observations which determine this observed behaviour (see Koolaard and Lawoko (1996)).

2.2.5 Motivation for the present study

As mentioned earlier the expansions of the error rates that Raudys and Pikelis (1980) obtained involved numerical integration, which is often a complicated technique. Furthermore, they only considered the trivial case of equivalence of the SLDF and SEDF when $\mathbf{\Sigma} = \mathbf{I}$. Meanwhile, Marco et al.(1987) compared the performances of the SLDF and SEDF through simulations only. Thus it is of interest to broaden these comparisons by using expected values of the error rates of each of the discriminant classifiers. In this chapter, expectations of asymptotic expansions of the conditional (i.e. actual) error rates are obtained for each classifier, and comparisons of the performances are made under the same conditions as those used by Marco et al. It is of interest to determine if the results and deductions arrived at from the asymptotic expansions in this study are consistent with the simulation results of the aforementioned authors, which indicated that the simpler SEDF is often superior in performance to the SLDF.

Lim (1992) derived the asymptotic expansions for the conditional error rates of the SLDF and SEDF in the case of "nonequivalence" of the two classifiers. Lim also derived the asymptotic expansion for the conditional error rate of the SEDF under "equivalence", but was unable to derive the corresponding expansion for the SLDF under these conditions, so a proper comparison of the classifiers was not able to be achieved. The outstanding expansions are derived in this chapter, enabling the comparisons to be made.

2.3 ASYMPTOTIC EXPANSIONS

In this chapter, the asymptotic expected error rates were obtained using Taylor series expansions of the conditional error rates and taking expectations over the distributions of $\bar{x}_1$, $\bar{x}_2$ and S_p. In particular, if $\mathcal{H}(.)$ is a differentiable function of parameters $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s)$, where $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_s)$, are consistent estimators of $(\beta_1, \beta_2 \ldots, \beta_s)$, then the Taylor series expansion (up to order one) of $\mathrm{E}(\mathcal{H})$ ($\mathrm{E}(.)$ denoting expectation) about the point $(\beta_1, \beta_2 \ldots, \beta_s)$ can be expressed as

$$\begin{aligned} \mathrm{E}(\mathcal{H}) \approx \mathcal{H}(\beta_1, \beta_2, \ldots, \beta_s) &+ \sum_{j=1}^{s} \frac{\partial \mathcal{H}}{\partial \hat{\beta}_j} \mathrm{E}(\hat{\beta}_j - \beta_j) \\ &+ \frac{1}{2} \sum_{i,j} \frac{\partial^2 \mathcal{H}}{\partial \hat{\beta}_i \partial \hat{\beta}_j} \mathrm{E}\left[(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)'\right]. \end{aligned} \tag{2.9}$$

For our expansions $\mathcal{H} = \Phi(.)$ is the standard normal distribution function, and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s$ are the elements of $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, S_p. The expansions are evaluated at the point $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$.

The two asymptotic error rates considered here are (i) the expected error rate based on samples of n_1 from population 1 and n_2 from population 2 (This is often called the expected actual or unconditional error rate) and (ii) the expected plug-in error rate. For these error rates, the function $\mathcal{H}(.)$ takes the following forms (for misclassification of an object from population 1 to population 2), where the subscript 'A' and 'P' refer to the "actual" and "plug-in" error rates respectively:

$$P_{21(A)}^{SLDF} = \Phi\left(-\frac{[\boldsymbol{\mu}_1 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)]' S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} \Sigma S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right), \tag{2.10}$$

$$P_{21(A)}^{SEDF} = \Phi\left(-\frac{[\boldsymbol{\mu}_1 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)]' (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \Sigma (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right), \tag{2.11}$$

$$P_{21(P)}^{SLDF} = \Phi(-\tfrac{1}{2}[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}) \tag{2.12}$$

and

$$P_{21(P)}^{SEDF} = \Phi\left(-\frac{1}{2} \frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right). \tag{2.13}$$

Corresponding probabilities of misclassifying an object from population 2 to population 1 are:

$$P_{12(A)}^{SLDF} = \Phi\left(\frac{[\boldsymbol{\mu}_2 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)]' S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} \Sigma S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right),$$

$$P_{12(A)}^{SEDF} = \Phi\left(\frac{[\boldsymbol{\mu}_2 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)]' (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \Sigma (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right),$$

$$P_{12(P)}^{SLDF} = \Phi(-\tfrac{1}{2}[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2})$$

and $$P_{12(P)}^{SEDF} = \Phi\left(-\frac{1}{2}\frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\mathrm{S}_p(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right).$$

The following result by Okamoto (1963) was used to obtain the partial derivative terms in the expansion:

> If the covariance matrix $\boldsymbol{\Sigma}$ is symmetric and invertible, and we let $\boldsymbol{\Sigma}^{-1} = \{\sigma^{ij}\}$ (where $\{\sigma^{ij}\}$ is a function of σ_{rs} (the $(r,j)^{th}$ element of $\boldsymbol{\Sigma}$), then
>
> $$\frac{\partial(\sigma^{ij})}{\partial\sigma_{rs}} = -\frac{1}{1+\delta_{rs}}\left(\sigma^{ir}\sigma^{sj} + \sigma^{is}\sigma^{rj}\right) \qquad (r \leq s) \qquad (2.14)$$
>
> where δ_{rs} is the Kronecker delta.

In a series of papers, McLachlan (1972, 1973, 1974a, 1974b) obtained asymptotic expansions of error rates for the SLDF. No such results appear to have been obtained for the SEDF. This is partly due to the fact that for the SLDF the function $\mathcal{H}(.)$ can be reduced to a relatively simple function (usually referred to as "canonical form") through a linear transformation of the observation vector. This simplifies the algebra considerably, and makes the final result dependent on only a few parameters. Unfortunately, no similar technique can be used for the corresponding $\mathcal{H}(.)$ function for the SEDF. The canonical form that has been traditionally adopted (after the transformation) has been $\boldsymbol{\mu}_1 = (\Delta, 0, 0, \ldots, 0)'$, $\boldsymbol{\mu}_2 = (0, 0, 0, \ldots, 0)'$ and $\boldsymbol{\Sigma} = \mathbf{I}$, which would not allow us to investigate the distinction between SLDF and SEDF, since such a parametric configuration means there is no difference between the SLDF and the SEDF. Also, this represents a different parametric configuration to that of the "equivalence" situation of the error rates (see Section 2.2.4). Hence, such an investigation (as is being planned here) would require that a particular structure of $\boldsymbol{\Sigma}$ be assumed. Consequently each asymptotic expansion takes a different form, depending on (i) the assumed structure of $\boldsymbol{\Sigma}$, (ii) whether the expansion is obtained under "equivalence" or "non-equivalence", (iii) whether the expansion is for the SLDF or the SEDF, and also (iv) whether the expansion is for the "actual" or "plug-in" error rate.

The two distinct structures of $\boldsymbol{\Sigma}$ that will be considered are:

1. $\boldsymbol{\Sigma}$ an (intra-class) equi-correlation matrix (Denoted $\boldsymbol{\Sigma}_A$):

$$\boldsymbol{\Sigma}_A = (1-\rho)\mathbf{I} + \rho\mathbf{J}, \qquad \left(\frac{-1}{p-1} \leq \rho \leq 1\right)$$

where $\mathbf{I}$ is a $p \times p$ identity matrix and $\mathbf{J}$ is a $p \times p$ matrix of ones. That is,

$$\Sigma_A = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \vdots \\ \vdots & \vdots & & \ddots & \rho \\ \rho & \rho & \dots & \rho & 1 \end{bmatrix}. \tag{2.15}$$

2. Σ exhibiting the auto-correlation structure of an auto-regressive process of order 1 (Denoted Σ_B). That is,

$$\Sigma_B = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \dots & \vdots \\ \vdots & \vdots & & \ddots & \rho \\ \rho^{p-1} & \rho^{p-2} & \dots & \rho & 1 \end{bmatrix}. \tag{2.16}$$

Asymptotic expansions are obtained under the following conditions:

1. **Non-Equivalence of the LDF and EDF**

   Asymptotic expansions of the actual error rate and plug-in error rates of the SLDF and SEDF for the case of non-equivalence under the following conditions

   - $\mu_1 = (m^*, 0, 0, \dots, 0)'$, $\mu_2 = (0, 0, \dots, 0)'$ and $\Sigma = \Sigma_A$.
   - $\mu_1 = (m^*, 0, 0, \dots, 0)'$, $\mu_2 = (0, 0, \dots, 0)'$ and $\Sigma = \Sigma_B$.

   These expansions have been obtained and reported in Lim (1992).

2. **Equivalence of the LDF and EDF**

   Asymptotic expansions of the actual and plug-in error rates of the SLDF and SEDF for the case of equivalence under the following conditions

   - $\mu_1 = (m, m, \dots, m)'$, $\mu_2 = (0, 0, \dots, 0)'$ and $\Sigma = \Sigma_A$.
   - $\mu_1 = (m, m, \dots, m)'$, $\mu_2 = (0, 0, \dots, 0)'$ and $\Sigma = \Sigma_B$.

For these two cases the asymptotic expansions of the plug in error rates for both the SEDF and SLDF, as well as the asymptotic expansions of the expected actual error rate for the SEDF (but not for the SLDF), are given in Lim (1992).

The Taylor Series expansion (up to first order) of the actual error rate associated with the SLDF under "equivalence" conditions is of the form

$$\begin{aligned} P_A^{SLDF} \approx\ & \Phi\left(-\frac{m}{2}\sum_w\sum_v(\sum_u s^{uv})(\sum_u \sigma_{vu}s^{uw})^{-1/2}\{\sum_k\sum_l s^{lk}\}\right) \\ & + \frac{1}{2n_i}\sum_i\sum_j \frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}\sigma_{ij} + \frac{1}{2n_2}\sum_i\sum_j \frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}\sigma_{ij} \\ & + \frac{1}{2}\frac{(n_1+n_2)}{(n_1+n_2-2)^2}\sum_k\sum_l\sum_i\sum_j \frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}(\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}) \quad (2.17) \end{aligned}$$

where the quantities $\frac{\partial^2\Phi(.)}{\partial\theta_1\partial\theta_2}$ are obtained separately for each assumed structure of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ for any variables θ_1 and θ_2. Here, $\mathrm{S}_p{}^{-1} = \{s^{ij}\}$, and s_{rs} is the $(r,s)^{th}$ element of S_p. Details of the full algebraic expressions of the asymptotic expansions, which are the quantities

$$\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}, \frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}, \frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}},$$

are given in Appendix A for each structure of Σ. Evaluation of this expansion involves taking the expectations of the above expression, yielding the expected actual (i.e. unconditional) error rate associated with the SLDF.

2.4 NUMERICAL EVALUATIONS OF ASYMPTOTIC EXPANSIONS

In addition to comparing the performances of the SLDF and the SEDF via asymptotic expansions, simulations are performed to give a further assessment of the actual error rates of each classifier. Lim (1992) obtained the expected plug-in error rate associated with each classifier, and her results will be incorporated in the present discussion. As mentioned previously, the comparison is done under similar but more extensive conditions to Marco, Young and Turner (1987), in order to made direct comparison with their results. Values of m are chosen such that the Mahalanobis Distance, Δ, is the same in both cases of equivalence and

non-equivalence of the SLDF and SEDF. The values of the (squared) Mahalanobis distance used are $\Delta^2 = 0.5$, 1.0, 1.5, 2.0 and 2.5. The covariance matrices, whether they are of the form $\Sigma = \Sigma_A$ or $\Sigma = \Sigma_B$, are determined by the parameter ρ. The values of ρ used are 0.0, 0.2, 0.4 and 0.65. In some cases for $\Sigma = \Sigma_B$, additional negative values of ρ have been used to allow more extensive investigation. The negative values of ρ used are: ρ = -0.2, -0.4 and -0.65. It was decided not to use these negative values when $\Sigma = \Sigma_A$, since Σ will not be positive definite if ρ is less than $-1/(p-1)$. For the case of non-equivalence of the SLDF and SEDF (where $\boldsymbol{\mu}_1 = (m^*, 0, 0 \ldots 0)'$ and $\boldsymbol{\mu}_2 = (0, 0, \ldots, 0)'$) the value of m^* is given by

$$m^* = \sqrt{\{\Delta^2/\sigma^{11}\}}. \tag{2.18}$$

Meanwhile, for the case of equivalence of the two classifiers (where $\boldsymbol{\mu}_1 = (m, m, \ldots, m)'$ and $\boldsymbol{\mu}_2 = (0, 0, \ldots, 0)'$), the value of m is given by

$$m = \sqrt{\{\Delta^2/\sum_i \sum_j \sigma^{ij}\}}. \tag{2.19}$$

For most combinations of values of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ, comparison of the SLDF and SEDF was carried out using two values of the dimension, namely $p = 4$ and $p = 8$. The complexity of the expansions meant that numerical evaluation of them at larger dimensions was not feasible in practice (surprisingly!), owing to the prohibitive amount of computation time required. The sample sizes n_1 and n_2 from populations (or groups) 1 and 2 were taken to be equal at $n_1 = n_2 = 50 = n$.

The relative performances of the SLDF and SEDF are compared using three criteria:

1. comparisons of the expected actual error rates under conditions of equivalence and non-equivalence.

2. comparisons of the expected plug-in error rates, obtained by Lim (1992), under conditions of equivalence and non-equivalence.

3. comparisons of the estimated actual error rates, obtained via Monte Carlo simulation methods, under conditions of equivalence and non-equivalence.

Since comparison of the classifiers under conditions of equivalent error rates is the fair course of action, the discussion will primarily focus on results obtained under this scenario. For the purpose of this discussion we will refer to the error rate when all parameters are known as the true error rate. The asymptotic expectations of the actual error rate and the plug in error rate will be referred to as "the expected

actual error rate" and "the expected plug in error rate" respectively. The expected actual error rate is also called the unconditional error rate. These various error rates will be denoted as follows:

e^L_{true}, e^E_{true}	=	true error rates for the SLDF (superscript L) and SEDF (superscript E).
e^L, e^E	=	asymptotic expected actual (i.e. unconditional) error rates.
$\hat{e}^L, \hat{e}^E$	=	asymptotic expected plug-in error rates.
e^L_s, e^E_s	=	cross-validation error rates from simulation experiments.

The tables presented (Tables 2.3 to 2.5) give the order of magnitudes of various expected error rates, and error rates obtained through simulation. The simulation results in these tables will be discussed in Section 2.5. The results for the asymptotic expansions under the scenarios in Section 2.3 are now discussed separately.

Non-equivalence situation ($\Sigma = \Sigma_A$)

Lim (1992) derived asymptotic expansions of the actual and plug-in error rates - expressions (2.10), (2.11), (2.12) and (2.13) - under the previously defined conditions of non-equivalence of these error rates. A reduced results table for this case is presented (Table 2.1), and the main features of the results of the numerical evaluation of these expansions are now discussed.

The expansions appear to be affected by a combination of large p and high correlation between variables in this case of $\Sigma = \Sigma_A$. When $\rho = 0.65$ and $p = 8$ the value of e^E and $\hat{e}^E$ are substantially lower than expected, with $\hat{e}^E$ being particularly low. It would not be advisable to use these expansions in this rather extreme condition of high intra-class correlation coupled with high dimension.

It is observed that e^L and e^E increase as the dimension changes from $p = 4$ to $p = 8$, but that $\hat{e}^L$ and $\hat{e}^E$ decrease for the same change in dimension. It is well known that the plug-in error rate for the SLDF is usually too optimistic, and this bias appears to exhibit itself more strongly in the higher dimensional settings. The expected plug-in rates $\hat{e}^L$ and $\hat{e}^E$ usually give very poor estimates of e^L and e^E respectively, except for conditions of small p, large Δ^2, and ρ close to zero. The quantity e^L generally decreases as ρ increases from 0 to 0.65. This is consistent with Cochran (1962), but is a phenomenon which is not entirely predictable as it is likely to be produced by the combined effects of ρ, p and Δ^2, as well as the

Table 2.1: The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "non-equivalence" with $\Sigma = \Sigma_A$.

		$p = 4$			$p = 8$		
Δ^2	ρ	e^E_{true} $\mathbf{e^L_{true}}$	e^E $\mathbf{e^L}$	$\hat{e}^E$ $\mathbf{\hat{e}^L}$	e^E_{true} $\mathbf{e^L_{true}}$	e^E $\mathbf{e^L}$	$\hat{e}^E$ $\mathbf{\hat{e}^L}$
0.5	0.0	.3618 **.3618**	.3784 **.3857**	.3456 **.3373**	.3618 **.3618**	.3996 **.3954**	.3244 **.3051**
	0.2	.3677 **.3618**	.3823 **.3766**	.3491 **.3370**	.3706 **.3618**	.4075 **.3899**	.3293 **.3047**
	0.4	.3810 **.3618**	.3847 **.3685**	.3515 **.3360**	.3861 **.3618**	.3891 **.3673**	.3093 **.3036**
2.0	0.0	.2398 **.2398**	.2474 **.2277**	.2328 **.2191**	.2398 **.2398**	.2562 **.2319**	.2240 **.1920**
	0.2	.2495 **.2398**	.2566 **.2280**	.2414 **.2187**	.2544 **.2397**	.2714 **.2308**	.2367 **.1914**
	0.4	.2724 **.2398**	.2754 **.2252**	.2592 **.2172**	.2813 **.2397**	.2861 **.2155**	.2483 **.1897**

accuracy of the asymptotic expansions used in this study, and that of Lim (1992). In these conditions, the expected error rate e^E tends to be closer to the true error rate e^E_{true}, than e^L is to e^L_{true}.

Non-equivalence situation ($\Sigma = \Sigma_B$)

In this situation, Lim (1992) also derived the appropriate asymptotic expansions and the results of evaluating these expansions are now summarised. A table showing some of the results is presented (Table 2.2). The missing values in this table result from the evaluation of the asymptotic expansion for the LDF going out of bounds for the case where $\rho = -0.4$ (Lim). The expected actual error rates e^L, e^E, as well as the expected plug-in error rates $\hat{e}^L$ and $\hat{e}^E$ exhibit more consistent behaviour in relation to ρ, p and Δ^2 under these conditions. The expected rates e^L and e^E increase as p increases from $p = 4$ to $p = 8$, and also as ρ becomes closer to either 1 or -1. This trend was not apparent when $\Sigma = \Sigma_A$. A reason for this may be that for a given value of ρ, the correlations between variables in the data when $\Sigma = \Sigma_A$ are generally stronger than the corresponding correlations when $\Sigma = \Sigma_B$. The true error rate for the SLDF, e^L_{true}, remains constant with respect to p and ρ. This is due to the fact that it only depends on the Mahalanobis distance,Δ,

Table 2.2: The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "non-equivalence" with $\Sigma = \Sigma_B$.

		$p = 4$			$p = 8$		
		e^E_{true}	e^E	$\hat{e}^E$	e^E_{true}	e^E	$\hat{e}^E$
Δ^2	ρ	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$
0.5	-0.4	.3730	.3886	.3525	.3730	.4207	.3336
		.3618		**.3365**	**.3618**		**.3043**
	0.0	.3618	.3784	.3456	.3618	.3996	.3244
		.3618	**.3906**	**.3373**	**.3618**	**.4228**	**.3051**
	0.4	.3730	.3886	.3537	.3730	.4207	.3352
		.3618	**.4064**	**.3365**	**.3618**	**.4669**	**.3043**
2.0	-0.4	.2585	.2661	.2494	.2585	.2799	.2414
		.2398		**.2178**	**.2398**		**.1908**
	0.0	.2398	.2474	.2328	.2398	.2562	.2240
		.2398	**.2447**	**.2191**	**.2398**	**.2719**	**.1920**
	0.4	.2585	.2661	.2499	.2585	.2799	.2420
		.2398	**.2733**	**.2178**	**.2398**	**.3414**	**.1908**

which, under these conditions and for the present structure of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ, is not affected by p or ρ. Since e^L increases as both p and ρ increase, this leads to the result that e^L substantially overestimates e^L_{true} for large p and ρ. On the other hand, since both e^E and e^E_{true} increase as p and ρ increase, these two error rates are generally closer together in magnitude.

Once again the expected plug-in error rates $\hat{e}^E$, and particularly $\hat{e}^L$, are affected by bias, and underestimate e^L and e^E. This is a well known result and is especially true for larger dimension ($p = 8$) and ρ not close to 0 (either positive or negative). Note that $\hat{e}^L$ substantially decreases, and $\hat{e}^E$ slightly decreases as p increases from 4 to 8. Generally, $\hat{e}^L$ is a poor estimate of e^L, while $\hat{e}^E$ is a better estimate of e^E, although it too is an underestimate.

Equivalence situation ($\Sigma = \Sigma_A$)

The results for this case are presented in Table 2.3. The true error rate is equal for SEDF and SLDF for all combinations of parameters, and is only affected by Δ^2 and not p or ρ. Because e^E is higher than e^E_{true} for $\rho = 0$ and decreases as ρ increases, it is closest to the true value when $\rho = 0.65$. For the SLDF, e^L also decreases slightly as ρ increases, especially for larger values of Δ^2, but usually it is

as close to e^L_{true} as e^E is to e^E_{true}. The expected error rate for the SLDF (e^L) also increases slightly when the dimension increases from 4 to 8, except for those cases of small Δ^2 and high correlation.

Overall, there is little difference in the performance of the SLDF and SEDF under these conditions, although the asymptotic expansion of e^L is extremely large and complex and hence of doubtful practical use (see Appendix A).

Regarding the expected plug-in error rates, both $\hat{e}^E$ and $\hat{e}^L$ underestimate the expected actual error rates. The plug-in error rates decrease when the dimension increases from $p = 4$ to $p = 8$, while e^L and e^E increase. It is also evident from Table 2.3 that $\hat{e}^L$ more seriously underestimates the expected actual error rate than $\hat{e}^E$, particularly for larger values of p and Δ^2. For medium to large separation between the populations ($\Delta^2 > 1.0$), $\hat{e}^E$ gives a more accurate estimate of the expected actual error rate than $\hat{e}^L$, especially when $p > 4$. On the other hand, for small Δ^2, there is no difference in performance between $\hat{e}^E$ and $\hat{e}^L$.

Equivalence situation ($\Sigma = \Sigma_B$) *Positive correlation* ($\rho > 0$)

The results for this case are given in Table 2.4. The expected error rates e^L and e^E both decrease only slightly as the correlation strength increases from $\rho = 0$ to 0.65 under this covariance structure, although e^L shows a larger reduction for the highest value of ρ ($\rho = 0.65$). Both expected error rates estimate their corresponding true error rate reasonably well. In particular, e^L estimates e^L_{true} a little better than e^E estimates e^E_{true} when Δ^2 and ρ are small, whereas the reverse is true when Δ^2 is larger and ρ is moderate to high.

The behaviour of the expected plug-in error rates $\hat{e}^L$ and $\hat{e}^E$ in this case is similar to that under the previous covariance structure, where $\Sigma = \Sigma_A$.

Equivalence situation ($\Sigma = \Sigma_B$) *Negative correlation* ($\rho < 0$)

The results for this case are given in Table 2.5. The expected error rates e^L and e^E both increase in magnitude as the dimension p increases from 4 to 8 and as the correlation ρ becomes more strongly negative (ρ decreases from -0.2 to -0.65). In particular, e^E increases considerably as ρ decreases, while e^L remains relatively stable. Both e^L and e^E usually overestimate e^L_{true} and e^E_{true} respectively, but e^L is much closer to e^L_{true} than e^E is to e^E_{true}. In fact for $\rho = -.65$ and large p, e^E can be two or three times larger than e^E_{true}.

Table 2.3: The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "equivalence" with $\Sigma = \Sigma_A$.

		$p = 4$				$p = 8$			
		e^E_{true}	e^E	$\hat{e}^E$	e^E_s	e^E_{true}	e^E	$\hat{e}^E$	e^E_s
Δ^2	ρ	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_s$	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_s$
0.5	0.0	.3618	.3788	.3470	.3820(.062)	.3618	.4001	.3261	.4024(.056)
		.3618	**.3597**	**.3373**	**.3866(.066)**	**.3618**	**.3695**	**.3037**	**.4072(.067)**
	0.2	.3618	.3669	.3510	.3704(.049)	.3618	.3671	.3425	.3672(.068)
		.3618	**.3572**	**.3378**	**.3784(.061)**	**.3618**	**.3624**	**.3046**	**.3930(.059)**
	0.4	.3618	.3641	.3554	.3574(.056)	.3618	.3639	.3524	.3686(.063)
		.3618	**.3521**	**.3381**	**.3752(.063)**	**.3618**	**.3498**	**.3052**	**.3958(.076)**
	0.65	.3618	.3631	.3593	.3656(.054)	.3618	.3631	.3586	.3588(.056)
		.3618	**.3331**	**.3384**	**.3856(.063)**	**.3618**	**.3051**	**.3059**	**.4016(.071)**
1.0	0.0	.3085	.3205	.2294	.3236(.056)	.3085	.3347	.2857	.3284(.054)
		.3085	**.3110**	**.2867**	**.3240(.057)**	**.3085**	**.3245**	**.2562**	**.3274(.059)**
	0.2	.3085	.3125	.3020	.3084(.056)	.3085	.3128	.2967	.3266(.054)
		.3085	**.3092**	**.2873**	**.3180(.056)**	**.3085**	**.3194**	**.2574**	**.3340(.061)**
	0.4	.3085	.3107	.3050	.3104(.061)	.3085	.3107	.3032	.3132(.054)
		.3085	**.3057**	**.2877**	**.3252(.059)**	**.3085**	**.3109**	**.2580**	**.3360(.056)**
	0.65	.3085	.3101	.3076	.3128(.053)	.3085	.3102	.3074	.3102(.054)
		.3085	**.2930**	**.2882**	**.3212(.060)**	**.3085**	**.2812**	**.2590**	**.3394(.065)**
2.0	0.0	.2398	.2481	.2351	.2542(.053)	.2398	.2571	.2268	.2478(.054)
		.2398	**.2461**	**.2196**	**.2643(.056)**	**.2398**	**.2633**	**.1902**	**.2580(.053)**
	0.2	.2398	.2431	.2367	.2376(.046)	.2398	.2434	.2336	.2434(.052)
		.2398	**.2448**	**.2204**	**.2462(.052)**	**.2398**	**.2596**	**.1917**	**.2622(.056)**
	0.4	.2398	.2420	.2386	.2490(.051)	.2398	.2421	.2377	.2504(.060)
		.2398	**.2425**	**.2209**	**.2530(.047)**	**.2398**	**.2541**	**.1926**	**.2674(.051)**
	0.65	.2398	.2416	.2402	.2376(.053)	.2398	.2418	.2403	.2396(.053)
		.2398	**.2345**	**.2214**	**.2508(.057)**	**.2398**	**.2355**	**.1939**	**.2638(.065)**
2.5	0.0	.2146	.2219	.2112	.2172(.053)	.2146	.2296	.2043	.2236(.058)
		.2146	**.2220**	**.1951**	**.2278(.052)**	**.2146**	**.2400**	**.1661**	**.2340(.052)**
	0.2	.2146	.2178	.2126	.2130(.052)	.2146	.2181	.2101	.2142(.054)
		.2146	**.2208**	**.1959**	**.2200(.058)**	**.2146**	**.2367**	**.1677**	**.2328(.054)**
	0.4	.2146	.2168	.2141	.2164(.050)	.2146	.2170	.2135	.2174(.054)
		.2146	**.2188**	**.1964**	**.2274(.054)**	**.2146**	**.2321**	**.1686**	**.2346(.053)**
	0.65	.2146	.2165	.2155	.2132(.050)	.2146	.2167	.2156	.2134(.048)
		.2146	**.2121**	**.1970**	**.2174(.042)**	**.2146**	**.2164**	**.1699**	**.2320(.054)**

Table 2.4: The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "equivalence" with $\Sigma = \Sigma_B$ and positive ρ.

		$p = 4$				$p = 8$			
		e^E_{true}	e^E	$\hat{e}^E$	e^E_s	e^E_{true}	e^E	$\hat{e}^E$	e^E_s
Δ^2	ρ	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_s$	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_s$
0.5	0.0	.3618	.3788	.3470	.3828(.055)	.3618	.4001	.3261	.3858(.060)
		.3618	**.3597**	**.3373**	**.3852(.065)**	**.3618**	**.3695**	**.3037**	**.3936(.063)**
	0.2	.3624	.3704	.3495	.3676(.054)	.3623	.3806	.3316	.3874(.058)
		.3618	**.3583**	**.3376**	**.3788(.061)**	**.3618**	**.3671**	**.3042**	**.3954(.069)**
	0.4	.3634	.3673	.3541	.3696(.064)	.3636	.3725	.3417	.3616(.054)
		.3618	**.3548**	**.3378**	**.3812(.065)**	**.3618**	**.3615**	**.3044**	**.3832(.059)**
	0.65	.3642	.3659	.3597	.3646(.069)	.3660	.3692	.3549	.3750(.057)
		.3618	**.3400**	**.3378**	**.3786(.066)**	**.3618**	**.3361**	**.3046**	**.3984(.072)**
1.0	0.0	.3085	.3205	.2994	.3156(.059)	.3085	.3347	.2857	.3364(.052)
		.3085	**.3110**	**.2867**	**.3136(.057)**	**.3085**	**.3245**	**.2562**	**.3368(.061)**
	0.2	.3092	.3153	.3014	.3160(.059)	.3092	.3221	.2897	.3182(.064)
		.3085	**.3099**	**.2871**	**.3238(.058)**	**.3085**	**.3227**	**.2567**	**.3322(.063)**
	0.4	.3106	.3139	.3052	.3174(.058)	.3109	.3176	.2973	.3184(.051)
		.3085	**.3076**	**.2873**	**.3280(.061)**	**.3085**	**.3189**	**.2570**	**.3410(.059)**
	0.65	.3117	.3135	.3095	.3162(.054)	.3140	.3170	.3079	.3070(.061)
		.3085	**.2980**	**.2874**	**.3250(.053)**	**.3085**	**.3022**	**.2573**	**.3394(.057)**
2.0	0.0	.2398	.2481	.2351	.2532(.057)	.2397	.2571	.2268	.2530(.053)
		.2398	**.2461**	**.2196**	**.2532(.059)**	**.2397**	**.2633**	**.1902**	**.2640(.059)**
	0.2	.2406	.2453	.2368	.2488(.048)	.2405	.2496	.2296	.2528(.050)
		.2398	**.2453**	**.2200**	**.2532(.047)**	**.2397**	**.2619**	**.1908**	**.2640(.054)**
	0.4	.2424	.2452	.2400	.2434(.049)	.2428	.2479	.2355	.2476(.056)
		.2398	**.2439**	**.2203**	**.2472(.053)**	**.2398**	**.2595**	**.1913**	**.2684(.059)**
	0.65	.2437	.2457	.2433	.2438(.052)	.2467	.2495	.2439	.2350(.054)
		.2398	**.2381**	**.2204**	**.2548(.051)**	**.2398**	**.2491**	**.1917**	**.2618(.055)**
2.5	0.0	.2146	.2219	.2112	.2200(.048)	.2146	.2296	.2043	.2384(.054)
		.2146	**.2220**	**.1951**	**.2262(.052)**	**.2146**	**.2400**	**.1661**	**.2478(.055)**
	0.2	.2155	.2198	.2127	.2302(.052)	.2154	.2234	.2208	.2148(.052)
		.2146	**.2213**	**.1951**	**.2358(.058)**	**.2146**	**.2388**	**.1667**	**.2260(.049)**
	0.4	.2173	.2201	.2158	.2234(.050)	.2178	.2225	.2122	.2368(.050)
		.2146	**.2200**	**.1959**	**.2258(.045)**	**.2146**	**.2367**	**.1671**	**.2430(.055)**
	0.65	.2188	.2208	.2189	.2286(.050)	.2219	.2246	.2201	.2274(.049)
		.2146	**.2153**	**.1959**	**.2266(.052)**	**.2146**	**.2281**	**.1676**	**.2384(.059)**

Table 2.5: The true (e_{true}), expected actual (e), expected plug-in ($\hat{e}$) and mean simulated (e_s)(with standard deviation) error rates of the SEDF and SLDF under the case of "equivalence" with $\Sigma = \Sigma_B$ and with negative ρ.

		$p = 4$				$p = 8$			
		e^E_{true}	e^E	$\hat{e}^E$	e^E_s	e^E_{true}	e^E	$\hat{e}^E$	e^E_s
Δ^2	ρ	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_s$	$\mathbf{e}^L_{true}$	$\mathbf{e}^L$	$\hat{\mathbf{e}}^L$	$\mathbf{e}^L_s$
0.5	-0.2	.3626	.3994	.3522	.3878(.065)	.3624	.4505	.3400	.4030(.066)
		.3618	**.3605**	**.3367**	**.3770(.078)**	**.3618**	**.3709**	**.3031**	**.3900(.060)**
	-0.4	.3654	.4549	.3841	.4008(.052)	.3644	.6034	.4346	.4286(.057)
		.3618	**.3614**	**.3350**	**.3858(.064)**	**.3618**	**.3729**	**.3009**	**.3986(.067)**
	-0.65	.3721	.7889	.6551	.4370(.062)	.3706	**	**	.4594(.059)
		.3618	**.3643**	**.3250**	**.3766(.068)**	**.3618**	**.3815**	**.2875**	**.3932(.068)**
1.0	-0.2	.3096	.3347	.3034	.3324(.050)	.3092	.3686	.2953	.3440(.050)
		.3085	**.3118**	**.2859**	**.3312(.058)**	**.3085**	**.3258**	**.2552**	**.3318(.060)**
	-0.4	.3133	.3636	.3264	.3476(.059)	.3119	.4717	.3595	.3632(.060)
		.3085	**.3128**	**.2837**	**.3192(.057)**	**.3085**	**.3282**	**.2524**	**.3360(.060)**
	-0.65	.3222	.6022	.5125	.3948(.061)	.3203	**	.9497	.4292(.053)
		.3085	**.3167**	**.2790**	**.3288(.052)**	**.3085**	**.3395**	**.2350**	**.3266(.058)**
2.0	-0.2	.2411	.2576	.2382	.2516(.056)	.2407	.2787	.2332	.2760(.066)
		.2398	**.2469**	**.2186**	**.2476(.049)**	**.2398**	**.2646**	**.1891**	**.2700(.062)**
	-0.4	.2457	.2844	.2549	.2756(.054)	.2439	.3451	.2751	.3118(.058)
		.2398	**.2481**	**.2161**	**.2470(.055)**	**.2398**	**.2674**	**.1856**	**.2630(.057)**
	-0.65	.2570	.4359	.3789	.3254(.054)	.2546	.7905	.6548	.3746(.050)
		.2398	**.2529**	**.2011**	**.2484(.050)**	**.2398**	**.2814**	**.1646**	**.2614(.056)**
2.5	-0.2	.2160	.2302	.2141	.2264(.051)	.2155	.2479	.2099	.2402(.051)
		.2146	**.2227**	**.1941**	**.2156(.048)**	**.2146**	**.2414**	**.1648**	**.2312(.053)**
	-0.4	.2209	.2539	.2291	.2496(.053)	.2190	.3046	.2458	.2618(.057)
		.2146	**.2240**	**.1915**	**.2252(.055)**	**.2146**	**.2442**	**.1612**	**.2350(.059)**
	-0.65	.2328	.3848	.3365	.2952(.058)	.2302	.6843	.5695	.3564(.060)
		.2146	**.2290**	**.1762**	**.2226(.052)**	**.2146**	**.2590**	**.1394**	**.2334(.052)**

In general, when $\rho < 0$, the SLDF performs better than the SEDF in terms of expected error rate. This is consistent with the Marco et al.(1987) theory about the relative performance of the two classifiers depending on the ratio of Mahalanobis to Euclidean distance. It was noted in Section 2.2.4 that examination of this ratio suggests that negative values of ρ in Σ_A or Σ_B would result in small values of $\sum_{i,j} \sigma_{i,j}$. This value is inversely proportional to the ratio referred to in Marco et al., explaining why the SEDF performs more poorly than the SLDF.

The plug-in error rate $\hat{e}^E$ increases dramatically (similarly to e^E) as ρ decreases to $\rho = -0.65$, especially under conditions where the population separation (Δ^2) is small. Clearly the error rates e^E and $\hat{e}^E$ are badly affected by strong negative correlation. In comparison, the behaviour of e^L and $\hat{e}^L$ are much more stable as $|\rho|$ increases under these conditions, although the accuracy of $\hat{e}^L$ as an estimate of e^L also deteriorates for $|\rho|$ large ($\rho = -0.65$).

2.5 SIMULATION RESULTS

A Monte Carlo simulation study was performed to verify and compare the results of the asymptotic expansions from the previous section. The values of Δ^2, p, n_1, n_2, ρ and Σ were fixed to be the same as the values used for the evaluations of the expansions. The value of m was obtained using equation (2.19), and data was generated from two multivariate normal distributions with equal covariance matrices Σ (where $\Sigma = \Sigma_A$ or $\Sigma = \Sigma_B$), and with $\boldsymbol{\mu}_1 = (m, m, \ldots, m)'$ and $\boldsymbol{\mu}_2 = (0, 0, \ldots, 0)'$.

The random samples drawn from each population were of size 50 (i.e. $n = n_1 = n_2 = 50$), and 100 simulations were performed for each combination of parameters (Δ^2, p, ρ). The Fortran 77 computer language was used along with NAG (1983) libraries for the simulation experiments. Sample observations were allocated to one of two multivariate normal populations having various mean and variance combinations as described in Section 2.3. Allocation was made using each classifier and the error rates for each were assessed using all three estimating techniques (bootstrapping, cross-validation, resubstitution). Although these various estimates of the error rates were obtained from the simulation experiments, previous work (e.g. Ganeshanandam and Krzanowksi (1990)) suggest that the cross-validated error rate is a good and reliable estimate to use. Consequently, the discussions here

on simulated error rates (for the various scenarios in Section 2.3) will be based on cross-validated error rates only, denoted by e_s^E and e_s^L. Its values are compared with the asymptotic expected (unconditional) error rates.

Equivalence situations ($\Sigma = \Sigma_A$)

Increasing the correlation ρ has very little effect on the simulated error rates (e_s^L and e_s^E). This behaviour is slightly different from that of the evaluated expansions of the expected actual error rates for the SLDF, where increasing ρ decreased the error rate of the SLDF, e^L (for small Δ^2 and larger p in particular). The increase in simulated error rates e_s^L and e_s^E when the dimension p increases from 4 to 8 is slight but consistent over all Δ^2 and ρ, whereas the results from the asymptotic expansions showed the expected (unconditional) error rates were affected differently for different combinations of values of Δ^2, ρ and p. When ρ is zero, or very small ($|\rho| < 0.2$), the asymptotic expansions for SLDF yields similar error rates to the simulated values, but as ρ increases it appears to unduly affect the expansion evaluations for the SLDF which yield underestimates of the true error rate. The expansion for SEDF always yield similar error rates to those from the simulation experiments. In general, under these conditions, the agreement between simulated and expected values (from the asymptotic expansions) is better for the SEDF than for the SLDF.

Equivalence situation ($\Sigma = \Sigma_B$) *positive correlation* ($\rho > 0$)

Increasing the correlation ρ from 0 to 0.65 did not appear to have any effect on the simulated error rates whereas the expansion of the expected actual error rate of the SLDF indicated the error rate decreases especially for high ρ ($\rho = 0.65$). The simulation results confirmed the expansion evaluations of the SEDF error rate across all values of ρ, Δ^2 and p. However, the asymptotic expansion for the SLDF appears to be affected by high correlation, ρ, leading to values of e^L which are significantly lower than the true error rate for small Δ^2, whereas the values of e_s^L are consistently higher than e_{true}^L and e^L .

As a general conclusion, agreement between simulated and expected values from the asymptotic expansions is better for the SEDF than for the SLDF, although the difference is not great and only appears when the group separation is small ($\Delta^2 \leq 1$).

Equivalence situation ($\Sigma = \Sigma_B$) *negative correlation* ($\rho < 0$)

The SEDF appears to be unduly affected when the correlation ρ is negative, and particularly when ρ is high and negative. The asymptotic expansion value e^E yields higher values than the simulated results in general, particularly for $\Delta^2 \leq 2$ and $\rho < -0.2$. The simulation results confirmed that the error rate of the SEDF does indeed increase as ρ becomes more strongly negative, but that the error rate of the SLDF remained at a similar level for all (negative) values of ρ. On the other hand, e^L is much closer to the simulated results, and there is good agreement between them. Under these conditions of negative correlation it seems clear that the SLDF performs better than the SEDF.

In summary, each classifier (i.e. SLDF or SEDF) has particular conditions under which it performs better than the other, although the SEDF performed better than the SLDF under the majority of conditions in this section. The simplicity of use of the SEDF, and its overall performance relative to the SLDF make it preferable as a rule for discrimination for the kind of conditions studied in this work.

2.6 GRAPHICAL DISPLAYS

It is instructive to demonstate the relative performances of the SLDF and SEDF, through a graphical presentation of their error rates as they vary with the magnitude of the square of the Mahalanobis distance Δ^2 between populations. This enables some of the observations made in Sections 2.4 and 2.5 to be more easily seen. Define the differences between the estimated and true error rates as:

$\zeta^L = e^L - e^L_{true}$ = difference between the expected actual error rate and true error rate for the SLDF.

$\zeta^E = e^E - e^E_{true}$ = difference between the expected actual error rate and the true error rate for the SEDF.

$\zeta^L_s = e^L_s - e^L_{true}$ = difference between the simulated error rate and the true error rate for the SLDF.

$\zeta^E_s = e^E_s - e^E_{true}$ = difference between the simulated error rate and the true error rate for the SEDF.

Graphical displays of the *Absolute Difference from True Error Rate* (i.e. $|\zeta^L|$, $|\zeta^E|$, $|\zeta^L_s|$ and $|\zeta^E_s|$) versus the *Mahalanobis distance squared* (Δ^2), for various levels

of correlation (ρ (*rho*)) among the observations, are presented in Figures 2.1 to 2.6. Results for positive autocorrelation structures are presented in Figures 2.1 and 2.2 ($\Sigma = \Sigma_A$) and Figures 2.3 and 2.4 ($\Sigma = \Sigma_B$), while results for negative autocorrelation between neighbouring observations with $\Sigma = \Sigma_B$ are presented in Figures 2.5 and 2.6. Since Figures 2.1 to 2.6 show absolute differences between the error rates, they hide any bias that an estimator might tend to have. Consequently, six corresponding graphs have been provided to illustrate this bias issue: Figures 2.7 to 2.12, which display values of ζ^L, ζ^E, ζ_s^L and ζ_s^E.

It was hypothesised in Marco, Young and Turner (1987) that the SEDF performs better than the SLDF if the ratio Δ^2/Δ_E^2 (Mahalanobis to Euclidean Distance) is large. It was also established in Section 2.2.4 that this condition can be reduced to the size of $\sum_{ij} \sigma_{ij}$. Since large (positive) ρ means large $\sum_{ij} \sigma_{ij}$, a comparison of the plots showing $|\zeta^L|$ and $|\zeta^E|$ for a given value of ρ indicates that the expected error rates provide support for this conjecture. The plots of ζ^L and ζ^E in Figures 2.7 to 2.12 also support these results.

The plots in Figures 2.7 and 2.9 show that for positive correlation, e^L usually initially underestimates the true error rate (when Δ^2 is small) and this estimation improves as Δ^2 increases until it overestimates the true error rate for very large Δ^2.

The plots for the simulated error rates in Figures 2.2 and 2.4 suggest that for positive ρ, $|\zeta_s^E|$ tends to be smaller than $|\zeta_s^L|$, and Figure 2.6 suggests that for negative ρ, the reverse happens. Although it is the absolute values of the simulated error rates which are shown in these figures, the graphs of ζ_s^E and ζ_s^L are similar (Figures 2.8, 2.10 and 2.12), indicating that the simulated error rates show that the SEDF performs better than the SLDF for positive ρ . Also, the simulated error rates tend to be generally larger than the true error rates, which is to be expected.

From Figure 2.9, an interesting difference between ζ^L and ζ^E is exhibited. As ρ increases, ζ^E decreases from positive values towards zero, particularly as Δ^2 increases. Meanwhile, ζ^L increases from negative values, through zero, to positive values. Thus e^L generally underestimates the true error rate when Δ^2 is small, but overestimates it for large values of Δ^2.

When we compare the results for $\Sigma = \Sigma_A$ with those for $\Sigma = \Sigma_B$ we find that the corresponding values of $|\zeta^L|$, $|\zeta^E|$, $|\zeta_s^L|$ and $|\zeta_s^E|$ are quite similar. In fact, it can be seen from the orders of magnitude of these differences in error rates that

the estimation of the error rates provided by the asymptotic expansions are quite reasonable in both situations. This is confirmed by the simulated error rates being of similar order of magnitude. It appears however, that when ρ is negative and Δ^2 is large, the approximation provided by e^E is an overestimate. The problem is worsened as p increases. Note, however, the simulated error rates are also unusually large under this situation (see Figures 2.6 and 2.12).

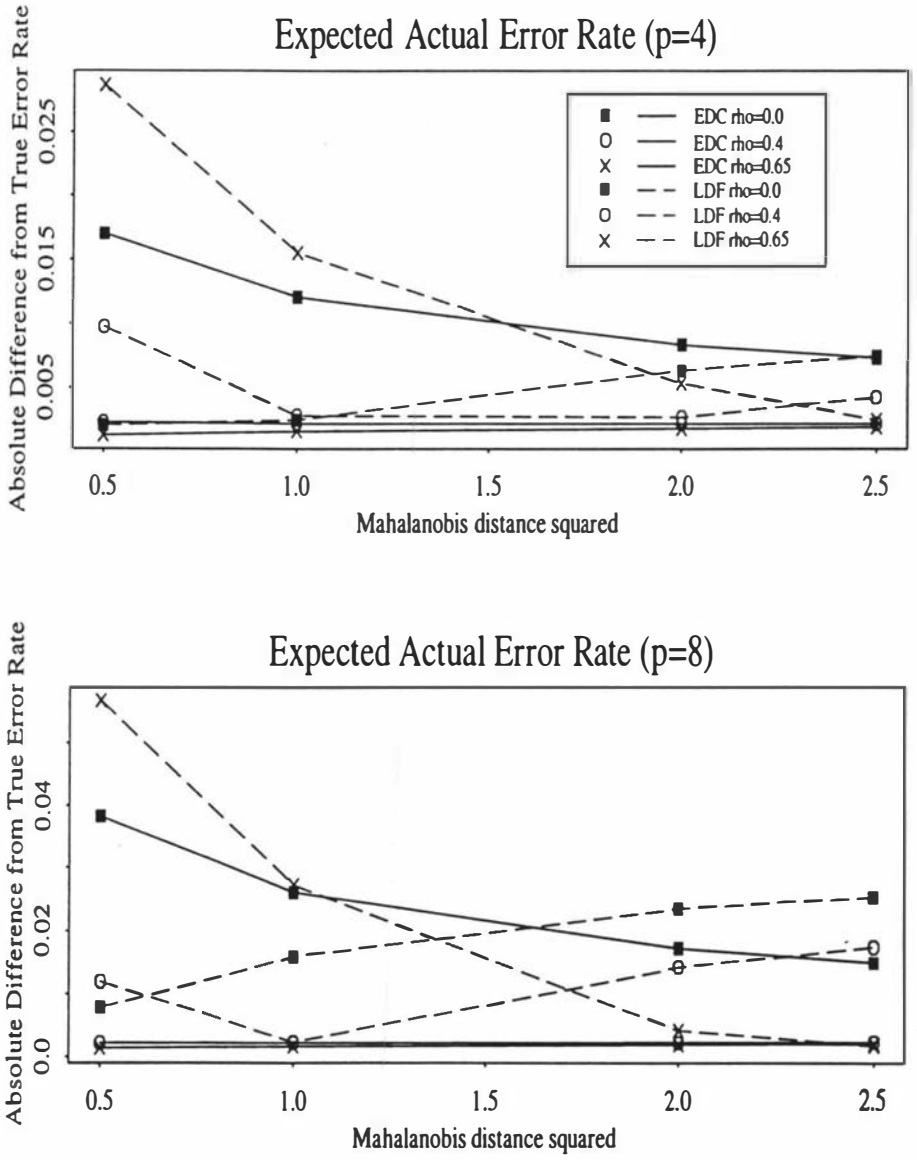


Figure 2.1: $|\zeta^L|$ and $|\zeta^E|$ for $\mathbf{\Sigma} = \mathbf{\Sigma}_A$, and various Δ^2 and ρ values ($\rho > 0$).

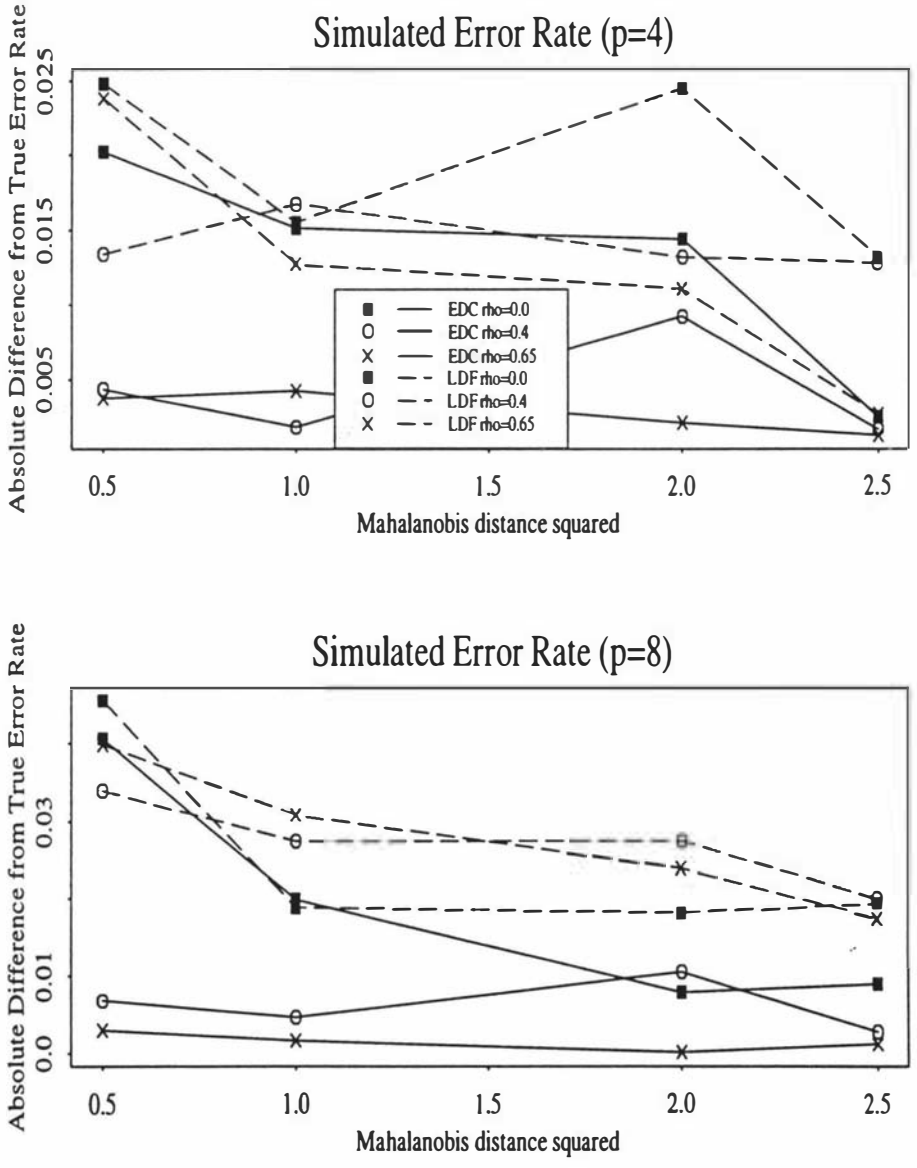


Figure 2.2: $|\zeta_s^L|$ and $|\zeta_s^E|$ for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_A$, and various Δ^2 and ρ values ($\rho > 0$).

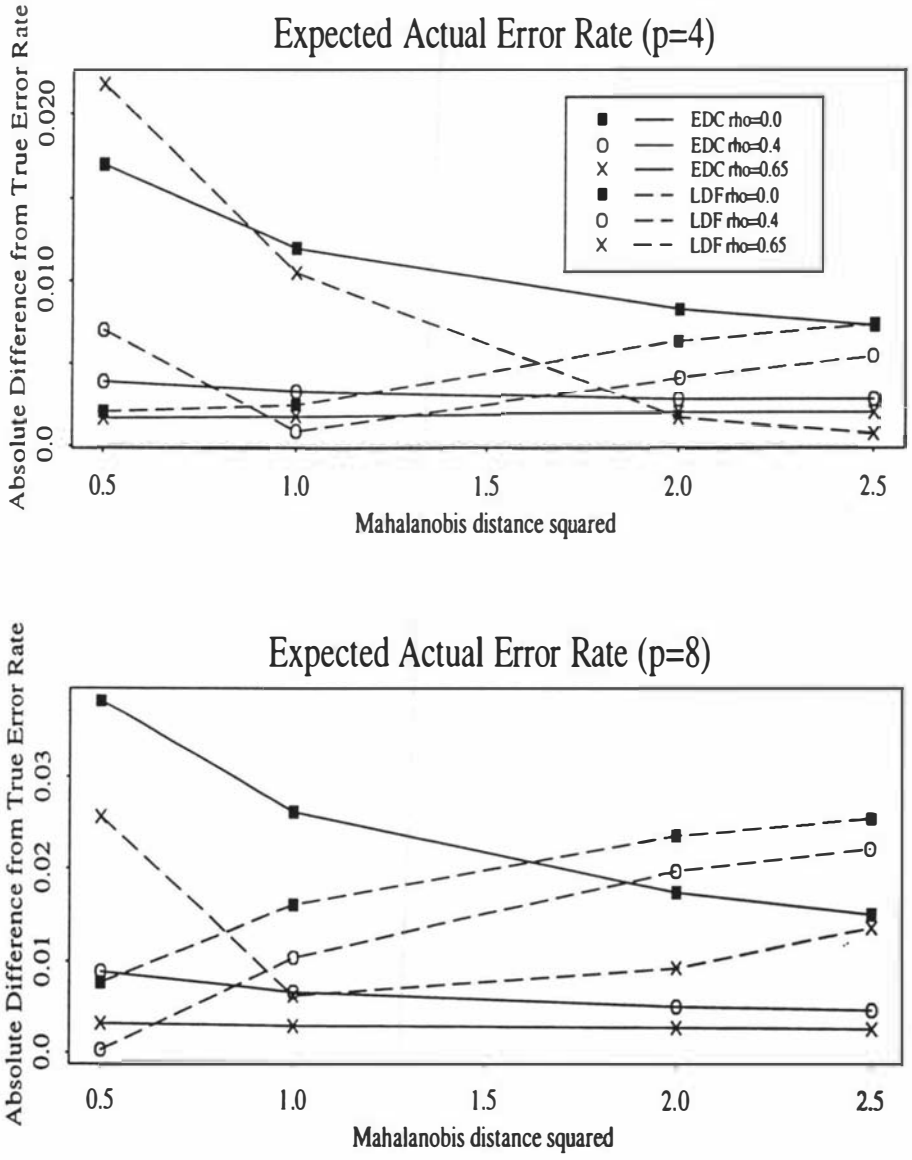


Figure 2.3: $|\zeta^L|$ and $|\zeta^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho > 0$).

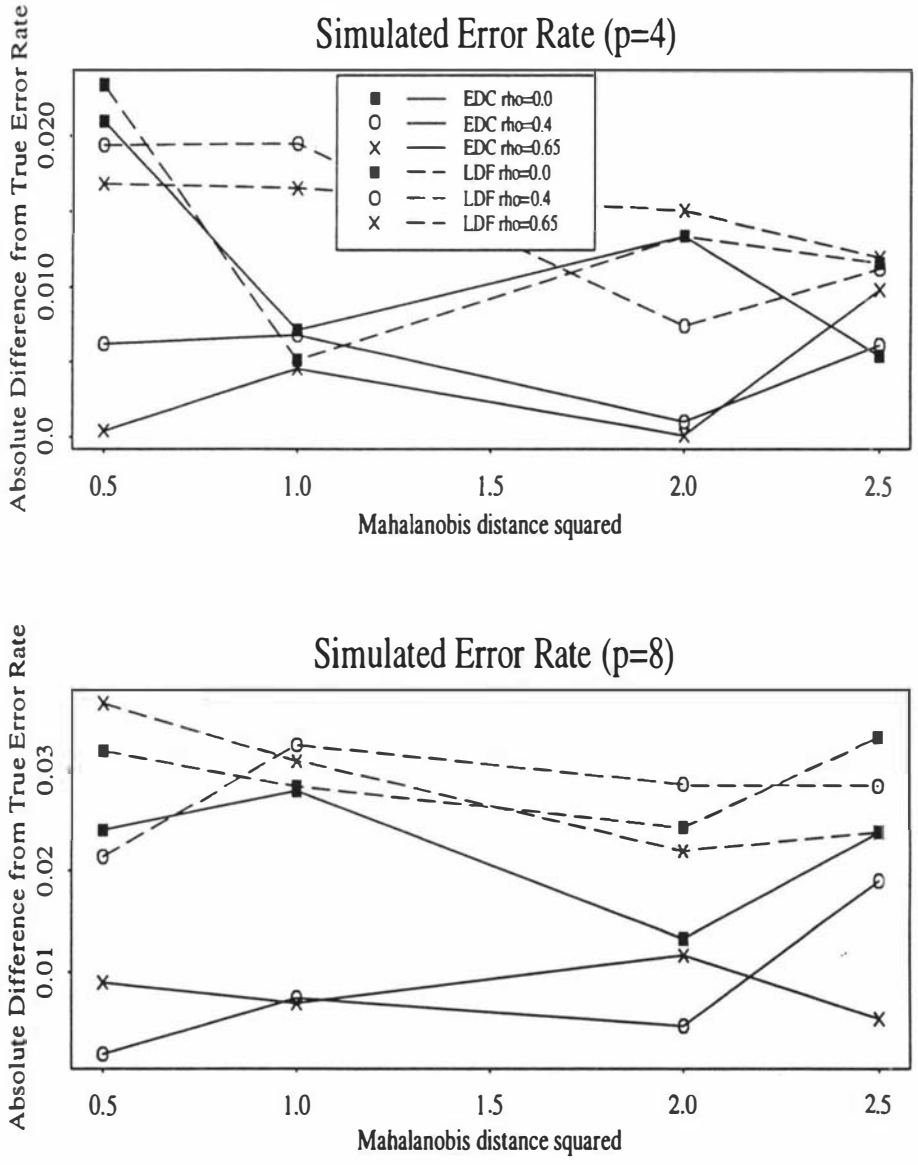


Figure 2.4: $|\zeta_s^L|$ and $|\zeta_s^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho > 0$).

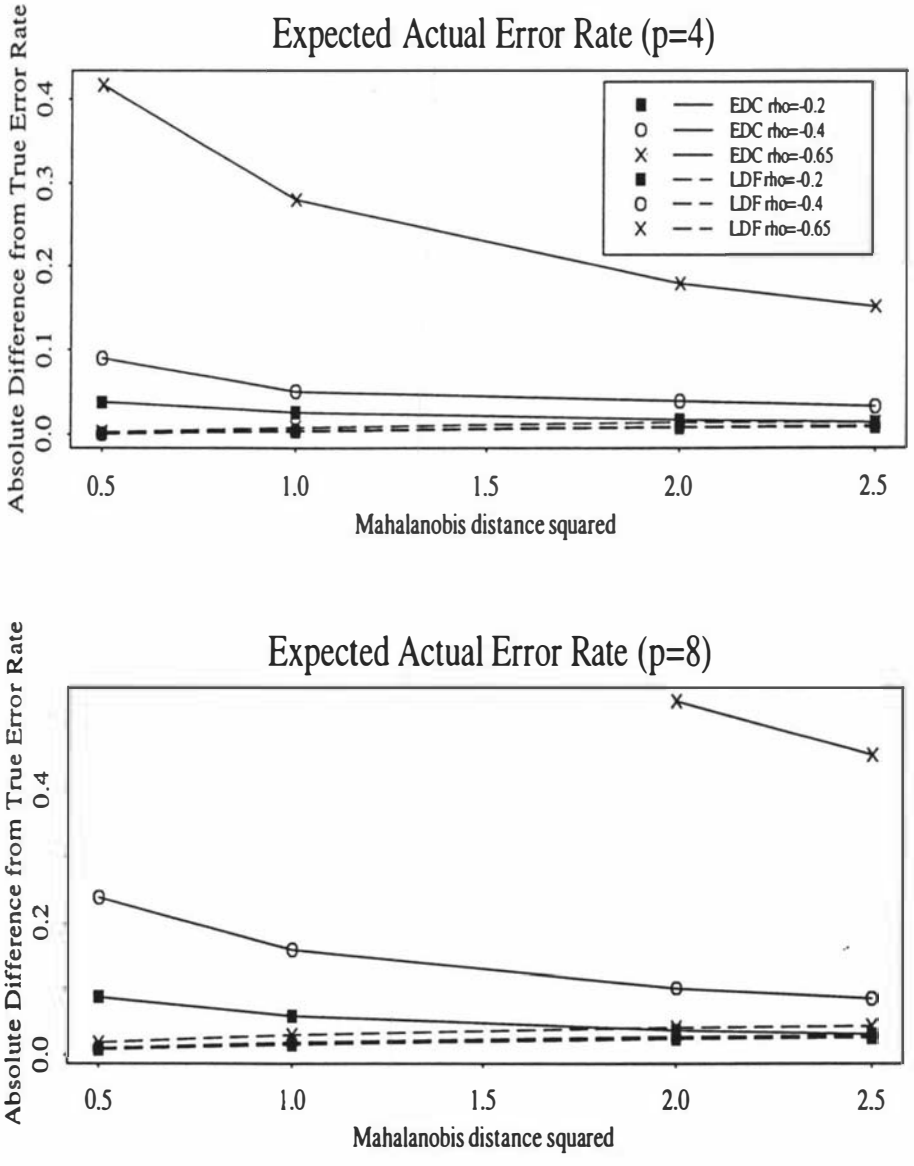


Figure 2.5: $|\zeta^L|$ and $|\zeta^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$).

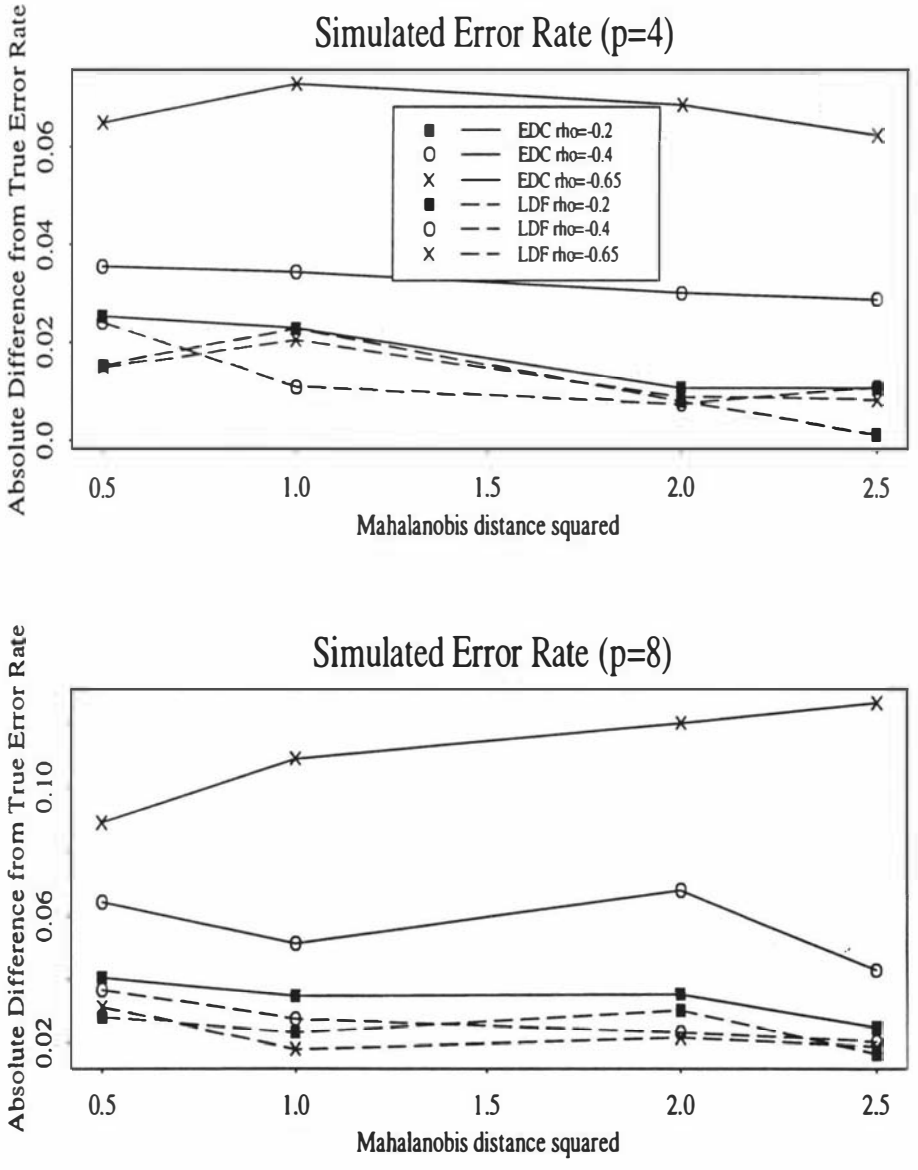


Figure 2.6: $|\zeta_s^L|$ and $|\zeta_s^E|$ for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$).

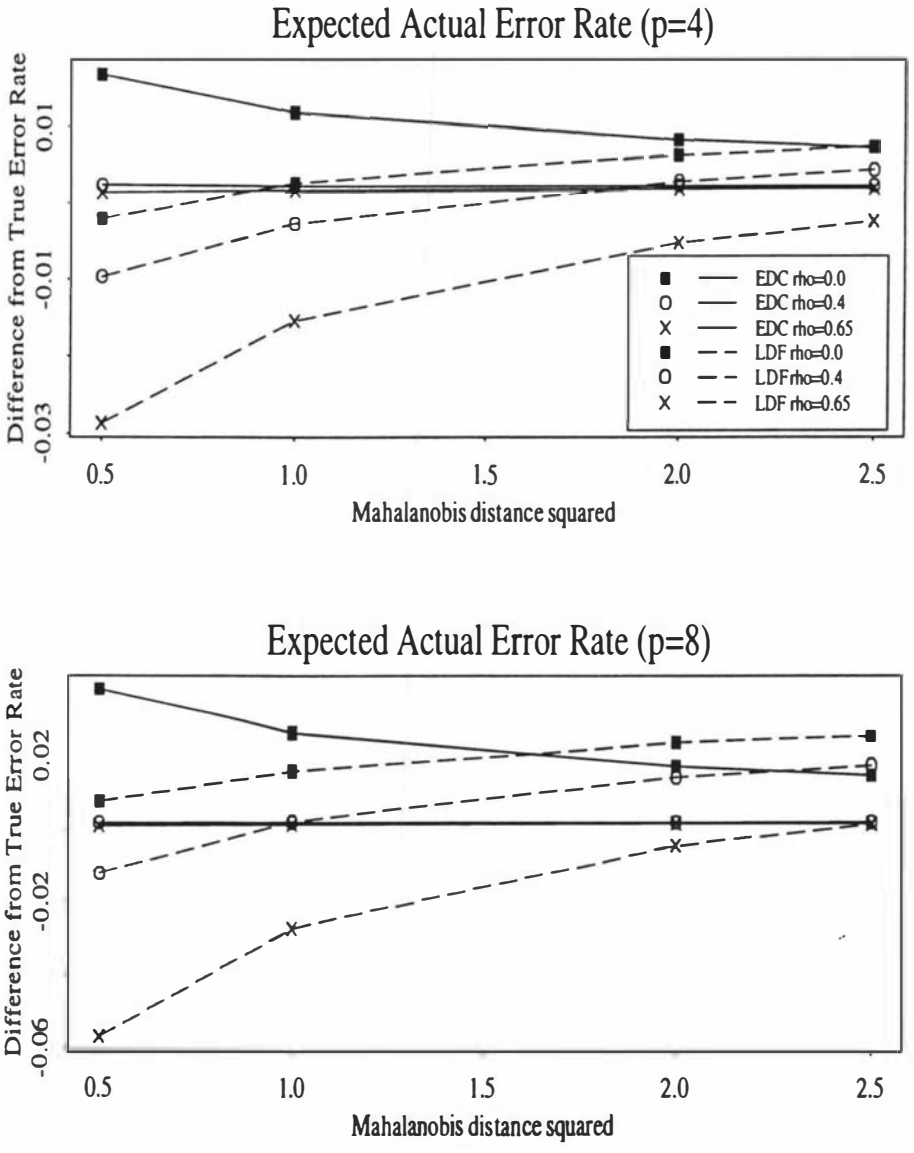


Figure 2.7: ζ^L and ζ^E for $\Sigma = \Sigma_A$, and various Δ^2 and ρ values ($\rho > 0$).

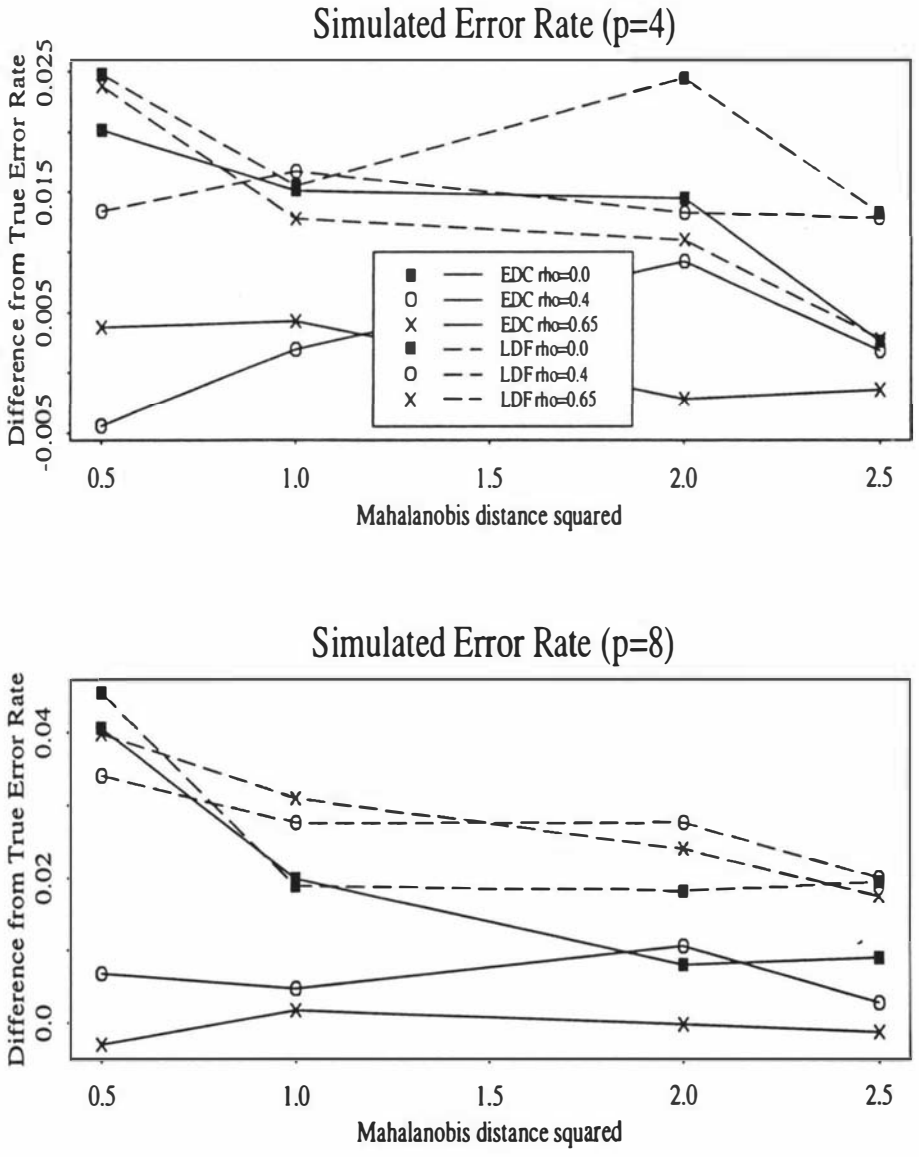


Figure 2.8: ζ_s^L and ζ_s^E for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_A$, and various Δ^2 and ρ values ($\rho > 0$).

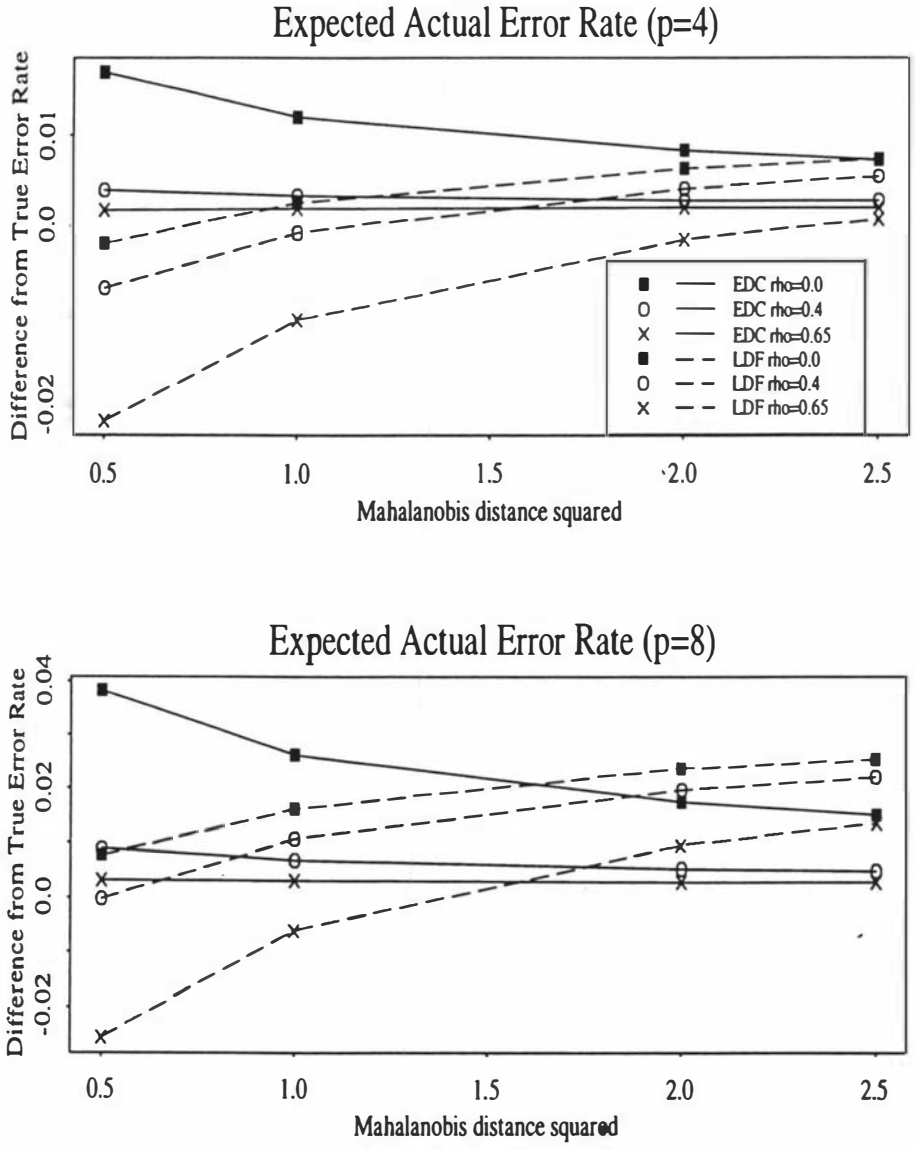


Figure 2.9: ζ^L and ζ^E for $\mathbf{\Sigma} = \mathbf{\Sigma}_B$, and various Δ^2 and ρ values ($\rho > 0$).

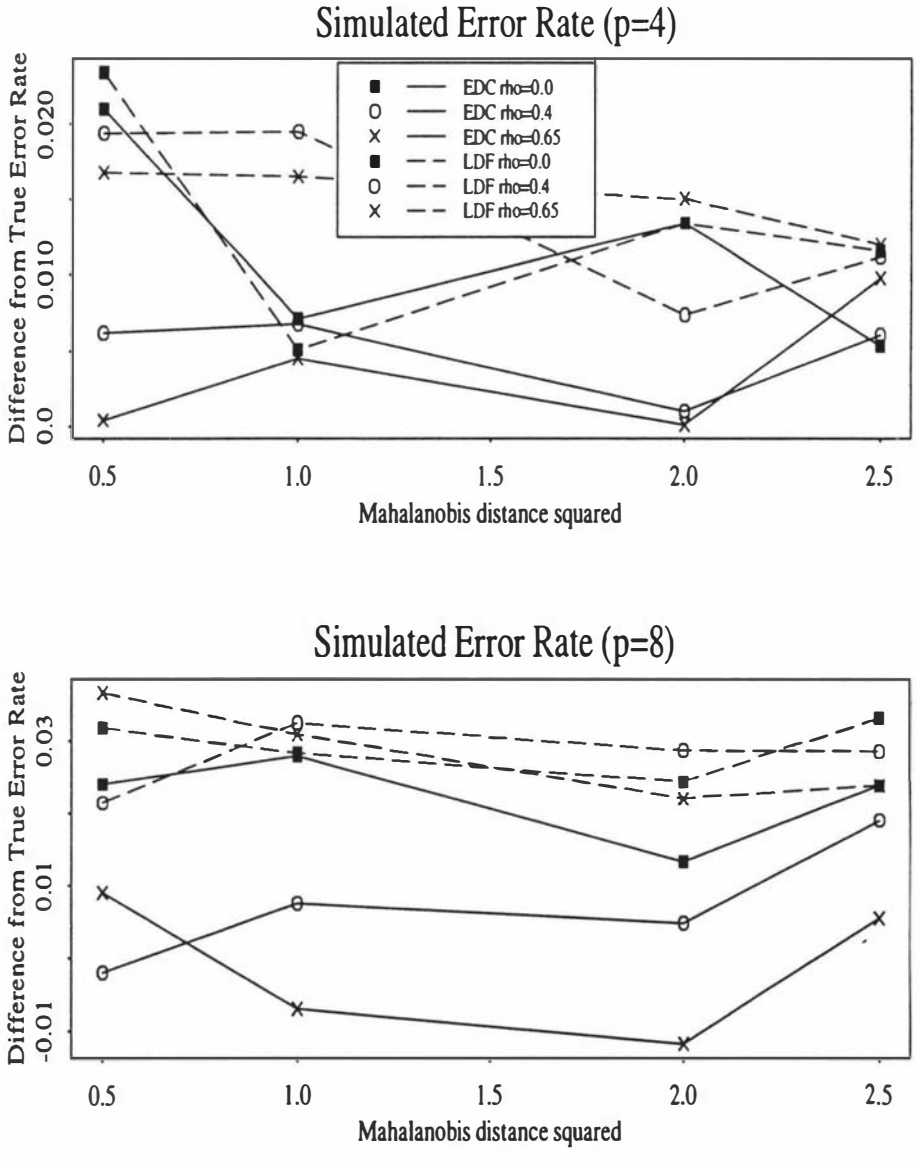


Figure 2.10: ζ_s^L and ζ_s^E for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B$, and various Δ^2 and ρ values ($\rho > 0$).

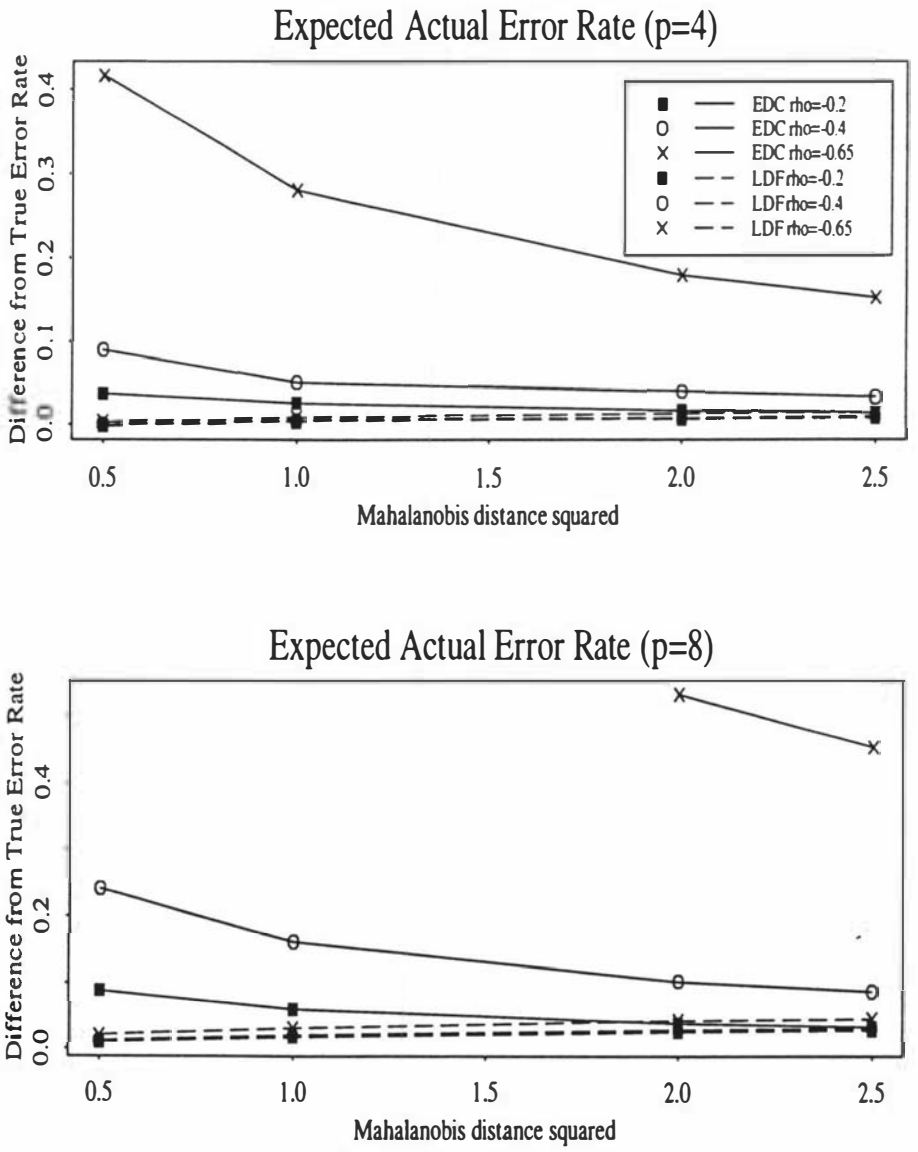


Figure 2.11: ζ^L and ζ^E for $\Sigma = \Sigma_B$, and various Δ^2 and ρ values ($\rho < 0$).

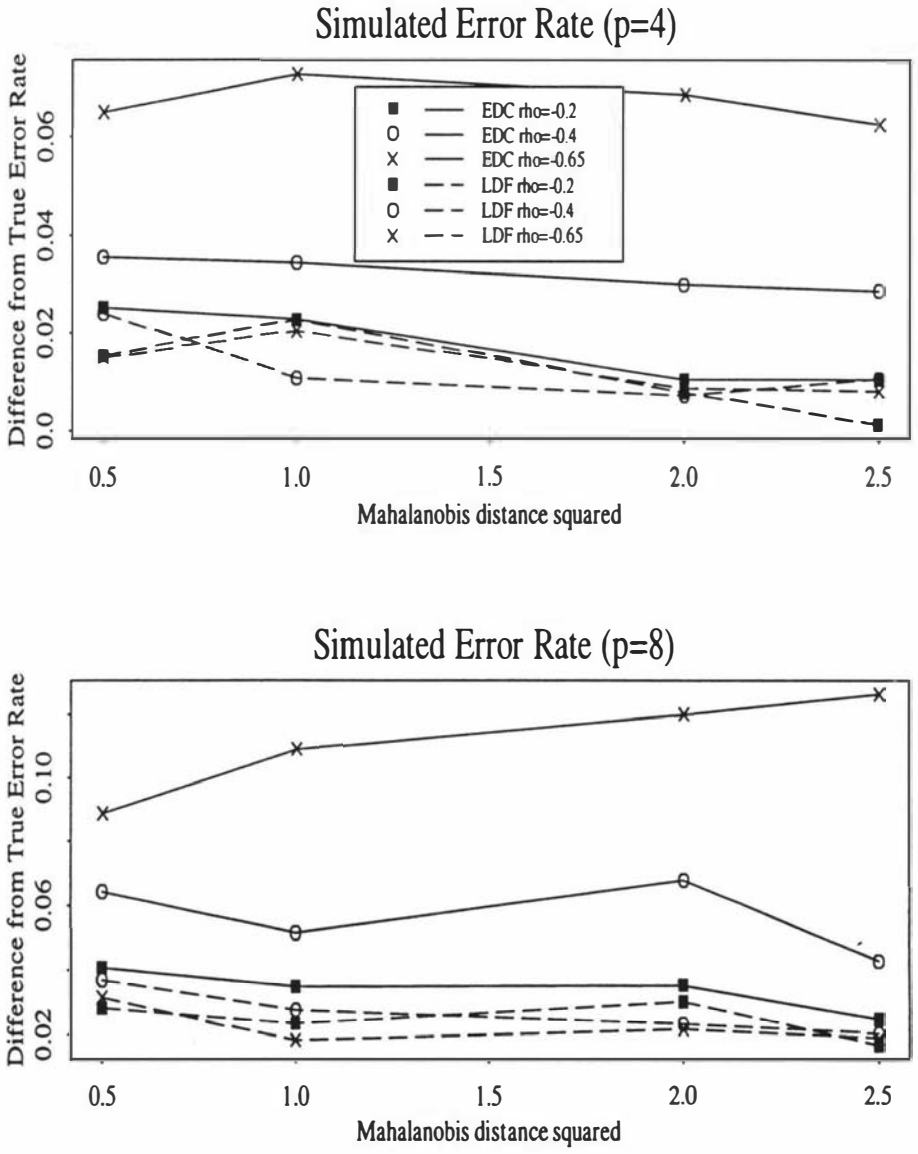


Figure 2.12: ζ_s^L and ζ_s^E for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B$, and various Δ^2 and ρ values ($\rho < 0$).

Chapter 3

REGULARISED DISCRIMINANT ANALYSIS

3.1 INTRODUCTION

Problems associated with estimating the K population (group) covariance matrices, $\boldsymbol{\Sigma}_k$ $(1 \leq k \leq K)$, were mentioned in Chapter 2. In the situation where the sample size is small in relation to the dimension, the usual discriminant rules (i.e. sample quadratic discriminant function SQDF, and sample linear discriminant function SLDF) are both affected by the quality of the sample based estimates of the population parameters, especially the covariance matrix. Friedman (1989) considered using alternative estimates of the covariance matrix, instead of the usual maximum likelihood ones. His regularisation technique (RDF) is described in detail in this chapter, since this will be helpful to the reader throughout this and subsequent chapters where modifications to the process will be examined and tested. The details of the RDF need to be identified clearly in order to do this.

The technique is also compared through simulation to the SQDF, SLDF and the sample Euclidean distance function (SEDF). The SEDF was compared to the SLDF under limited conditions in Chapter 2, and is included in this chapter since, under some circumstances, it can be a viable alternative discriminant rule to the commonly used SQDF and SLDF. A modification to Friedman's technique is explained and used to gain further understanding of the method.

3.2 PROBLEMS WITH ESTIMATING COVARIANCE MATRICES

The sample quadratic discriminant function (SQDF) requires approximately normal group conditional densities and reasonably large training sample sizes before it can be expected to perform well in discrimination. The sample linear discriminant function (SLDF) is more robust to non-normality, and requires less parameter estimation than the SQDF. However, it too can produce poor estimates of the pooled between-groups covariance matrix, particularly if the size of the training sample from group k, n_k, is small in relation to the dimension of the measurement space, p. The estimates of the covariance matrices can be highly variable in this situation, and Friedman (1989) showed the effect of this phenomenon on discriminant analysis by representing the group covariance matrices in terms of their spectral decompositions. That is, Σ_k can be represented as

$$\Sigma_k = \sum_{i=1}^{p} e_{ik}\eta_{ik}\eta_{ik}',$$

where e_{ik} is the ith eigenvalue of Σ_k, and η_{ik} is its corresponding eigenvector. The inverse may be written as

$$\Sigma_k^{-1} = \sum_{i=1}^{p} \frac{\eta_{ik}\eta_{ik}'}{e_{ik}}.$$

The discriminant score in expression (1.3) may then be written as

$$d_k(\boldsymbol{x}) = \sum_{i=1}^{p}\left(\frac{\left[\eta_{ik}'\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)\right]^2}{e_{ik}}\right) + \sum_{i=1}^{p}\ln\left\{e_{ik}\right\} - 2\ln\left\{\pi_k\right\}. \tag{3.1}$$

It is clear that small eigenvalues will have a large effect on this quantity. Sample-based estimates ($\mathbf{S}_k$) of Σ_k are known to produce biased estimates of the eigenvalues, especially when the size of the training sample used to obtain the estimate is small relative to the dimension. It is well known that the smallest eigenvalues are biased towards values that are too small, while the largest eigenvalues are biased towards values that are too large. This bias is even more pronounced when the eigenvalues of $\mathbf{S}_k$ are similar. When $n_k < p$, the sample covariance matrix is singular with rank $\leq n_k$. Thus the smallest $p - n_k + 1$ eigenvalues of $\mathbf{S}_k$ are zero. In such a case the sample discriminant score $\hat{d}_k(\boldsymbol{x})$, in expression (3.2), cannot be obtained since the first term of this equation involves division by the eigenvalue

estimates. The sample discriminant score may be written as

$$\hat{d}_k(\boldsymbol{x}) = \sum_{i=1}^{p} \left(\frac{\left[\hat{\boldsymbol{\eta}}_{ik}' (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)\right]^2}{\hat{e}_{ik}} \right) + \sum_{i=1}^{p} \ln \{\hat{e}_{ik}\} - 2 \ln \{\pi_k\}, \tag{3.2}$$

where $\hat{e}_{ik}$ is the i^{th} eigenvalue of $\mathbf{S}_k$, and $\hat{\boldsymbol{\eta}}_{ik}$ is its corresponding eigenvector.

If the sample covariance matrix is nearly singular, the smallest values of $\hat{e}_{ik}$, $(i = 1, \ldots, p$ and $1 \leq k \leq K)$ will be close to zero and will inflate the quantity $\hat{d}_k(\boldsymbol{x})$. The effect of this bias in discriminant analysis is to exaggerate the importance associated with the low-variance subspace which is spanned by those eigenvectors corresponding to the smallest eigenvalues. In fact, most of the variation in the sample discriminant score is associated with directions of low sample variance in the measurement space (Friedman (1989)).

3.3 REGULARISED ESTIMATES OF $\boldsymbol{\Sigma}_k$

One can reduce the variance associated with sample-based estimates of $\boldsymbol{\Sigma}_k$ by biasing the estimates away from the sample values and towards values that are more plausible in practice. Regularisation parameters may be introduced which control the amount of biasing, and the sample data can be used to estimate appropriate values for these parameters.

For example, consider the quadratic discriminant rule in expressions (1.4) and (1.5), where each $\mathbf{S}_k$ is replaced by the pooled sample covariance matrix $\mathbf{S}_p$. The resulting discriminant rule is the linear discriminant function. This is a more popular rule than the SQDF because of its greater robustness to (i) non-normality in the population distributions and (ii) poor estimates of the population parameters. The latter advantage of the SLDF over SQDF is enhanced by the decrease in variance associated with the estimation of the population covariance matrices.

A researcher who is applying normal-based classification procedures would normally test for homogeneity of variance between groups in the first instance. If the choice of rules is only between SLDF and SQDF, an initial test of $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ could be performed. If H_0 is rejected then the SQDF would be used, otherwise the SLDF may be used. An alternative approach is to introduce a regularisation parameter α, which regulates the shrinkage of the $\mathbf{S}_k$ to $\mathbf{S}_p$. Thus $\mathbf{S}_k$ in expressions (1.4) and (1.5) is replaced by

$$\hat{\boldsymbol{\Sigma}}_k(\alpha) = \alpha \mathbf{S}_k + (1 - \alpha)\mathbf{S}_p \quad (0 \leq \alpha \leq 1), \tag{3.3}$$

where α is determined from the data. Variations of this middle-of-the-road type of discriminant function were developed independently by Friedman (1989) and Greene and Rayens (1989). The results in these (and related) papers are now presented.

(i) **Greene and Rayens (1989)**

In their paper, these authors obtained empirical Bayes formulation for estimating the $\mathbf{\Sigma}_k$. That is, assuming that the training data from group k are independent observations from $\mathrm{N}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$, it follows that (conditioning on the $\mathbf{\Sigma}_k$)

$$(n_k - 1)\mathbf{S}_k \sim \mathrm{W}_p\left(\mathbf{\Sigma}_k, (n_k - 1)\right),$$

where $\mathrm{W}_p(.)$ denotes the central Wishart distribution with parameter matrix $\mathbf{\Sigma}_k$ and degrees of freedom $(n_k - 1)$. A conjugate prior distribution for $\mathbf{\Sigma}_k$ is assumed, which is the inverted Wishart distribution. That is, the $\mathbf{\Sigma}_k$ are assumed to be mutually independent with

$$\mathbf{\Sigma}_k \sim \mathrm{W}_p^{-1}\left((\omega - p - 1)\Psi, \omega\right),$$

where Ψ is the matrix of hyperparameters, and ω (where $\omega > p+1$) represents the degree of "concentration" of the $\mathbf{\Sigma}_k$ around Ψ. In particular, it can be established that

$$\mathrm{E}\left(\mathbf{\Sigma}_k\right) = \Psi,$$

and

$$\begin{aligned}\mathrm{cov}\left((\mathbf{\Sigma}_k)_{hj}, (\mathbf{\Sigma}_k)_{lm}\right) &= \frac{(\omega - p - 1)}{(\omega - p)(\omega - p - 3)}\left(\Psi_{hl}\Psi_{jm} + \Psi_{hm}\Psi_{lj}\right) \\ &\quad + \frac{2}{(\omega - p)(\omega - p - 3)}\Psi_{hj}\Psi_{lm}. \qquad (3.4)\end{aligned}$$

After some algebra and further results, it can be shown that the empirical Bayes estimate of $\mathbf{\Sigma}_k$ for a given ω is

$$\hat{\mathbf{\Sigma}}_k(\omega) = \frac{f_k}{f_k + \omega - p - 1}\mathbf{S}_k + \frac{\omega - p - 1}{f_k + \omega - p - 1}\mathbf{S}_p(\omega), \qquad (3.5)$$

where $f_k = n_k - 1$. The unknown parameter ω is estimated by either conditionally maximising the marginal likelihood of $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k$ over ω or using a method-of-moments type estimator. Details of this non-trivial computational task are given in the paper.

(ii) **Friedman (1989)**

Friedman's (1989) approach begins with the introduction of a regularisation parameter, λ, which controls the degree of shrinkage of the individual group covariance matrix estimates ($\mathbf{S}_k$) to the pooled estimate ($\mathbf{S}_p$). The following set of alternatives are obtained,

$$\hat{\boldsymbol{\Sigma}}_k(\lambda) = \frac{(1-\lambda)(n_k-1)\mathbf{S}_k + \lambda \mathbf{S}_p}{(1-\lambda)(n_k-1) + \lambda(N-K)}. \tag{3.6}$$

where

$$\sum_{k=1}^{K} n_k = N.$$

The parameter λ takes on values $0 \leq \lambda \leq 1$, and it is evident that if $\hat{\boldsymbol{\Sigma}}_k(\lambda)$ is used in expression (1.5) in place of $\mathbf{S}_k$, the scenario $\lambda = 0$ yields the SQDF, while the SLDF may be obtained by setting $\lambda = 1$ in expression (3.6).

Note that expression (3.6) yields discriminant rules where the only shrinkage is to the pooled estimate by varying degrees. This may not provide for sufficient regularisation, especially if the total sample size, N, is small in relation to the dimension p. In these cases, even for linear discriminant analysis, the number of parameters to be estimated is close to, or less than, the number of observations available. Also, biasing the group covariance estimates to the pooled covariance matrix may not be appropriate in some situations.

Friedman (1989), therefore, allowed for further regularisation of the sample covariance matrix. Thus $\boldsymbol{\Sigma}_k$ is estimated by

$$\hat{\boldsymbol{\Sigma}}_k(\lambda,\gamma) = (1-\gamma)\hat{\boldsymbol{\Sigma}}_k(\lambda) + \gamma\frac{\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_k(\lambda)\right\}}{p}\mathbf{I} \tag{3.7}$$

where $\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_k(\lambda)\right\}$ is the trace of the matrix $\hat{\boldsymbol{\Sigma}}_k(\lambda)$ in expression (3.6), $\mathbf{I}$ is a $p \times p$ identity matrix and γ is the additional parameter which regulates shrinkage towards a multiple of the identity matrix (the multiplier simply being the average eigenvalue of $\hat{\boldsymbol{\Sigma}}_k(\lambda)$). Shrinking in this way acts counter to the bias (described in Section 3.2) which is produced by sample estimation of the eigenvalues, by decreasing the larger eigenvalues of $\hat{\boldsymbol{\Sigma}}_k(\lambda)$ and increasing the smaller ones.

Friedman proposed that the regularised sample group covariance matrix $(\hat{\boldsymbol{\Sigma}}_k(\lambda,\gamma))$ replace $\mathbf{S}_k$ $(1 \leq k \leq K)$ in the sample quadratic discriminant rule

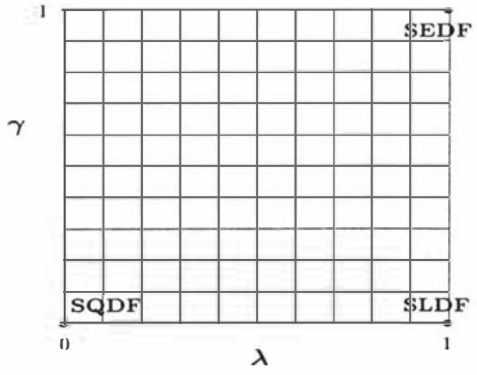


Figure 3.1: The extreme points on the (λ, γ) grid, and what each represents.

(expressions (1.4) and (1.5) for discriminant analysis. However, as $0 \leq \lambda, \gamma \leq 1$, a technique is required to select an appropriate (λ, γ) combination for use in the model. Friedman employed a technique which selects that combination which minimises an estimate of the future error rate (See Section 3.4 below). He termed this procedure regularised discriminant analysis (RDA).

RDA provides a rich class of regularisation alternatives. The possible (λ, γ) combinations may be thought of as lying on a plane with four corners (see Figure 3.1). The bottom left vertex $(\lambda = 0, \gamma = 0)$ corresponds to the SQDF, $(\lambda = 1, \gamma = 0)$ gives the SLDF, $(\lambda = 1, \gamma = 1)$ yields a discriminant rule based on the minimum Euclidean distance between groups, while $(\lambda = 0, \gamma = 1)$ yields a weighted minimum Euclidean distance rule where the group weights are inversely proportional to the average variance of the measurement variables in the group, i.e. $\mathrm{tr}\left\{\tilde{\Sigma}_k\right\}/p$. If γ is fixed at zero and λ varied, intermediate rules between the SQDF and SLDF are obtained. If λ is fixed at 1 and γ increased from 0, one obtains an analogy to ridge regression for linear discriminant analysis.

(iii) **Rayens and Greene (1991)**

As a consequence of the ideas in Friedman's article, Rayens and Greene (1991) modified their regularisation method to accommodate eigenvalue shrinkage

using the regularisation parameter γ as in Friedman's paper. They also proposed an alternative cross-validation approach for estimating the covariance mixing parameter λ, following a result which arises out of using the Kullback-Leibler distance measure for discrimination. Extensive use of cross-validation makes this also a computationally intensive option.

3.4 SELECTING REGULARISATION PARAMETER VALUES

Optimal values for the regularisation parameters λ and γ are not known in advance and Friedman (1989) suggested they be estimated from the training data. The selected (λ,γ) combination is that which gives rise to the minimum cross-validated estimate of the error rate associated with the sample regularised discriminant function (SRDF). A grid of points is chosen on the (λ,γ) plane ($0 \leq \lambda, \gamma \leq 1$) containing typically between 25 and 50 combinations of the regularisation parameters λ and γ. At each grid-point, the parameter values are used to create the classification rule. Cross-validation is used to estimate the misclassification risk of the rule for each combination of λ and γ for a given set of training data. The point $(\hat{\lambda}, \hat{\gamma})$ with the lowest estimated error rate is used as an estimate of the optimal values of λ and γ in a given situation.

This two-parameter optimisation problem would require excessive computation were it to be implemented directly. However, Friedman developed updating formulae for the computation of the regularised sample covariance matrix and its inverse, when a single and different observation is successively omitted from the sample (as occurs during cross-validation).

It should be noted that in his article, Friedman used robust versions of $\mathbf{S}_k$ and $\mathbf{S}_p$ in expressions (3.6) and (3.7) in place of the usual estimate in expression (1.10). We may write expression (3.6) as

$$\tilde{\Sigma}_k(\lambda) = \frac{\left[(1-\lambda)\tilde{\Sigma}_k + \lambda\tilde{\Sigma}_p\right]}{(1-\lambda)W_k + \lambda W} \tag{3.8}$$

where

$$\tilde{\Sigma}_k = \sum_{C(\nu)=k} w_\nu (\boldsymbol{x}_\nu - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_\nu - \bar{\boldsymbol{x}}_k)'$$

$$\tilde{\Sigma}_p = \sum_{k=1}^{K} \tilde{\Sigma}_k$$

$$W_k = \sum_{c(\nu)=k} w_\nu$$

$$W = \sum_{k=1}^{K} W_k,$$

and $c(\nu) = k$ is the group to which the ν^{th} observation ($\boldsymbol{x}_\nu$) ($1 \leq \nu \leq N$) belongs. Also, w_ν is the weight ($0 \leq w_\nu \leq 1$) assigned to the ν^{th} observation, and if all observations are given equal weight, then W_k is the size of the sample from group k.

The updating formula constructed by Friedman (1989) applied to the use of the robust estimator $\tilde{\Sigma}_k(\lambda, \gamma)$ which is defined by using $\tilde{\Sigma}_k(\lambda)$ instead of $\tilde{\Sigma}_k(\lambda)$ in expression (3.7). It was shown that if an observation is removed from the kth training sample, then $\tilde{\Sigma}_{k/\nu}(\lambda, \gamma)$ is obtained from $\tilde{\Sigma}_k(\lambda, \gamma)$ by subtracting a rank one matrix and a multiple of the identity matrix. Here "$/\nu$" indicates that the ν^{th} observation has been removed from calculations. The inverse of $\tilde{\Sigma}_{k/\nu}(\lambda, \gamma)$ is obtained in a similar way, making use of its spectral decomposition. Despite the updating formulae, this is still a computationally intensive process.

It should be noted that selection of appropriate parameter values is not as straight-forward as it may appear. Rayens and Greene (1991) noted from their simulation trials involving the SRDF that the minimum cross-validated estimate of the misclassification risk is often constant for a wide range of (λ, γ) combinations. This may be due to the fact that the error surface is fairly flat over a range of values of λ and γ. Hence the optimal choice $(\hat{\lambda}, \hat{\gamma})$ for the model will often not be uniquely determined. This was found to be commonly the case in a simulation study done in this project, which is described in Section 3.6.

Friedman did not address the issue of breaking of ties in the situation of multiple minima. However, Rayens and Greene demonstrated it as an issue which needs attention since it gives rise to a related phenomenon of concern. That is, some situations occur where only a very small proportion of the sample data influences the optimal choice of $(\hat{\lambda}, \hat{\gamma})$. In such a situation, while most of the observations are correctly classified for all (or almost all) points on the (λ,γ) plane, the remaining few observations are incorrectly classified for some values of λ or γ, and hence they exclusively determine the minimum cross-validated error rate. In effect, the choice of values for the regularisation parameters often depends on only a subset

of the available data, and the remainder of the data has no influence, and thus is effectively ignored in the model selection process. This occurs especially when the groups are fairly well separated.

Friedman employed a strategy of maximum regularisation in the case of ties, where, for all points yielding the minimum error rate on the (λ, γ) grid, that point $(\hat{\lambda}, \hat{\gamma})$ is selected which gives rise to the largest value of γ for the largest value of λ . This may not always be the ideal course of action. For example, in Section 4.3 the effect of using an alternative rule for the selection of $(\hat{\lambda}, \hat{\gamma})$ in situations where there are ties of this nature is discussed.

3.5 ASSESSMENT OF THE SRDF

3.5.1 Comparison of SRDF with other classifiers

It should be noted at this stage that in all subsequent work in this thesis, we concentrate on the regularised discriminant function as defined by Friedman (denoted here as SRDF), or some variants of it. Thus, unless stated otherwise, we do not consider any further the work of Greene and Rayens (1989) and Rayens and Greene (1991).

Friedman (1989) performed a simulation study to compare the regularised discriminant rule with the linear and quadratic discriminant functions in terms of their simulated overall error rates. The simulation conditions represented a wide range of situations in terms of the general structures of the group means and covariance matrices. Some of these conditions were chosen because they were expected to be unfavourable to the SRDF in that any regularisation away from the SQDF or SLDF would be detrimental to the discrimination process. On the other hand, some conditions were chosen because they were expected to be favourable to regularisation. Friedman (1989) considered six conditions for simulation and these are listed below.

In each example the training samples (of size 40) comprised observations randomly generated in equal proportions from three p–dimensional normal populations where $p = 6, 10, 20$ and 40. The optimisation grid over the (λ, γ) unit square consisted of 25 points. Each simulation trial involved the formation of the linear, quadratic and regularised discriminant rules from the training data. These rules were then applied to a test sample of observations (of size 100) which were

generated from the same populations as the training samples. The probability of misclassification (error rate) for each rule could therefore be estimated from the test sample. The average of 100 replications of this simulation trial was obtained.

The six conditions, defined in terms of the population covariance matrices and means, which are also employed extensively in this thesis for purposes of comparison, are:

1. Equal spherical population covariance matrices.
   A spherical matrix may be thought of as one where all the eigenvalues are similar in magnitude.
2. Unequal, spherical population covariance matrices.
3. Equal, highly ellipsoidal population covariance matrices with group mean differences in the low variance subspace.
   By ellipsoidal it is meant that there is a large difference in magnitude between the smallest and largest eigenvalues. This was achieved by making the leading diagonal elements of Σ_k highly disparate.
4. Equal, highly ellipsoidal population covariance matrices with group mean differences in the high variance subspace.
5. Unequal, highly ellipsoidal population covariance matrices with zero mean differences.
6. Unequal, highly ellipsoidal population covariance matrices with non-zero mean differences.

The following is a summary and discussion of the results of the simulation study by Friedman (1989) comparing the three discriminant functions: SRDF, SQDF and SLDF. The reason for repeating many of these results is to establish patterns that occur in the behaviour of the SRDF over these varying sets of conditions for comparison purposes (later), and to highlight its superiority over the commonly used discriminant rules under these circumstances.

In all the above simulation conditions, the SRDF-assessed optimum regularisation parameter values λ and γ were concentrated near to what would be expected in order to obtain a near-optimum classification rule. Hence the overall conclusion of the study was that the SRDF performs much better than the SLDF or SQDF in conditions that favoured regularisation of the types available. Further, the SRDF

does not lose much in performance to the SLDF or SQDF in conditions where either of these latter rules were optimal. The superior performance of the SRDF in reducing the error rate in situations where the sample size n_k $(1 \leq k \leq K)$ is small relative to the dimensionality (p) was the highlight of the results of the study.

The minimum cross-validated estimate of the error rate assessed from the training sample underestimates the (actual) error rate estimate obtained from the test sample by about 20%. Such a result is not unexpected since an estimate of the error rate obtained from the same training data as is used to construct the discriminant rule will always be optimistic. Friedman was surprised to find that there was only low correlation between these two error rate estimates, however. The implication was that the minimised cross-validated error rate provides an assessment of the unconditional error rate of the SRDF rather than its conditional error rate for a given set of training data.

In all simulation conditions where the total training sample size, N, was equal to the dimension p, the SRDF proved far superior to the SLDF or SQDF. The average assessed value of the regularisation parameter γ ranged from 0.45 to 1.0. This indicates that some shrinkage of the eigenvalues of $\mathbf{S}_k$ towards equality enhances discrimination, even under conditions where shrinkage of this sort would be thought to be counterproductive. This is because when the ratio of n_k to p becomes small, the effect of this shrinkage is to stabilise the extreme (both small and large) eigenvalues in the covariance estimates.

The case of spherical group covariance matrices (either equal or unequal) suited the SRDF. In particular, shrinkage of the eigenvalues towards equality is desirable in these situations, and indeed the average (over 100 replications) of the selected regularisation parameter values $\hat{\gamma}$, (i.e. $\bar{\hat{\gamma}}$), was close to 1. The SRDF was superior to the other rules under these conditions, especially for large p.

The case of equal but highly ellipsoidal group covariance matrices (with group mean differences concentrated in the low variance subspace) ought to have favoured the SLDF since any shrinkage away from the point $(\lambda = 1, \gamma = 0)$ would be counterproductive. This is because any use of γ would tend to obscure the differences in group means since increasing the smaller eigenvalues would increase the variance in the low variance subspace. The regularisation parameters for the SRDF have average values near this point. When p is very large, $\bar{\hat{\gamma}}$ increases in magnitude, and the resulting reduction in variance in the high variance subspace enables the

SRDF to out-perform the SLDF, even though such shrinkage introduces bias.

When the group covariances are unequal and highly ellipsoidal, very little regularisation of either type is desirable since, in the absence of substantial differences between group means, the differences in the covariances are heavily relied upon to separate the observations into their correct groups. The averages $\bar{\hat{\lambda}}$ and $\bar{\hat{\gamma}}$ are close to what is expected (i.e. $\lambda = 0, \gamma = 0$), although the $\hat{\gamma}$ values again tend to increase for larger p. This enables the SRDF to perform better than the SQDF in the larger dimensional settings, and comparable to it when p is small.

A related problem noted earlier, is that the optimal values ($\hat{\lambda}$ and $\hat{\gamma}$) are often not unique. The extent of the implications associated with this feature of the SRDF will be addressed as they occur in the discussion of the results in the following sections. The problems are addressed in Chapter 4, when a modification to the model selection procedure of the SRDF is implemented. The following sections describe various simulation studies which are aimed at further evaluating the SRDF, and investigating modifications to the technique. For this purpose it was necessary to develop software for its implementation. The software was written in a series of subroutines using MATLAB™ (1992), to implement the technique as developed by Friedman (1989).

For the studies in this chapter, however, one procedure employed by Friedman relating to the practical application of the SRDF was not implemented. The situation may arise where the estimate $\mathbf{S}_k$ or $\mathbf{S}_p$, of a group covariance matrix is singular, usually due to the sample size being less than the dimension. To enable the inverse of a singular sample covariance matrix to be obtained, Friedman (1989) advocated replacing the zero eigenvalues with a small number of sufficient magnitude to enable numerically stable inversion. The effect of this would be to produce a classification rule based on Euclidean distance in the zero-variance subspace. In other words, the variance of the subspace spanned by the eigenvectors corresponding to the zero eigenvalues is effectively ignored in the classification rule. In the present study, this manipulation of the eigenvalues was omitted, and samples of sufficient size ensured the problem of singularity did not arise.

The first step in the study was to perform simulations under the same six conditions as Friedman (1989) in order to verify that the implementation of Friedman's technique was correct and to establish a correspondence of results. The training samples varied in size according to the dimension of the population or group they

were sampled from. For $p = 6$, samples of size 14 were drawn from each population ($n_k = 14$ $(1 \leq k \leq K)$); for $p = 10$, $n_k = 16$, and for $p = 20$, $n_k = 28$. These sample sizes are sufficient to avoid singularity yet not so large that all classifiers perform well because of reduced problems with parameter estimation. In all cases there were $K = 3$ populations, and the optimisation grid of twenty-five (λ, γ) values was defined by the outer product of $\lambda = (0, .125, .354, .65, 1.0)$ and $\gamma = (0, .25, .5, .75, 1.0)$. These were the same values used by Friedman.

The training sample data was used to construct the various discriminant rules (SQDF, SLDF, SEDF and SRDF). An additional test data set of size 100 was generated from the same three populations, in equal proportion, and classified using the discriminant rules derived from the training data. The test data was used to estimate the misclassification or error rate for each rule, with each classification error assigned equal loss, irrespective of its type. One hundred replications of the above procedure were made.

The average error rate of each classifier, with its standard deviation, are given in Tables 3.1 through 3.6. Note that e_{cv}^{SRDF} is the minimum cross-validated error rate for the SRDF. The average regularisation parameter values are denoted $\bar{\hat{\lambda}}$ and $\bar{\hat{\gamma}}$.

The results in Tables 3.1 through 3.6 are generally comparable to those of Friedman (1989) although the performance of the SLDF and SQDF relative to the SRDF was often better because of the larger sample sizes used to enable full parameter estimation, which is particularly important for the SQDF. The minimum cross-validated error rate estimate underestimated the actual error rate by about 20% on average over the range of conditions, and by more in the situations where the group covariance matrices were unequal.

The sample Euclidean distance function (SEDF) was included amongst the discriminant rules which are being compared owing to its simplicity and good performance over against the SLDF in previous studies as discussed in Chapter 2. The SEDF used in this section is that obtained by using the SRDF model and setting the regularisation parameter values λ and γ both to one. This means that

$$\hat{\mathbf{\Sigma}}_k(\lambda, \gamma) = \frac{\mathrm{tr}\{\mathbf{S}_p\}}{p}\mathbf{I}, \text{ for all } k \ (1 \leq k \leq K)$$

replaces $\mathbf{S}_k$ in expression (1.5). Although the usual SEDF uses $\mathbf{S}_k = \mathbf{I}$, for all k, the allocation of a given observation to one group will not be affected. The performance of the SEDF in relation to the other discriminant rules is now discussed.

Table 3.1: **Equal Spherical Covariance Matrices.** Average error rate (with standard deviation) for several discriminant functions.

	$p = 6$	$p = 10$	$p = 20$
SRDF	.11 (.04)	.12 (.04)	.12 (.04)
SLDF	.13 (.04)	.14 (.04)	.15 (.04)
SQDF	.24 (.06)	.32 (.07)	.41 (.07)
SEDF	.11 (.04)	.11 (.03)	.11 (.03)
e_{cv}^{SRDF}	.09 (.05)	.10 (.04)	.10 (.04)
$\bar{\hat{\lambda}}^{SRDF}$	.87 (.29)	.85 (.30)	.80 (.34)
$\bar{\hat{\gamma}}^{SRDF}$	.78 (.34)	.81 (.26)	.81 (.24)

Table 3.2: **Unequal Spherical Covariance Matrices.** Average error rate (with standard deviation) for several discriminant functions.

	$p = 6$	$p = 10$	$p = 20$
SRDF	.14 (.04)	.18 (.05)	.11 (.04)
SLDF	.18 (.05)	.27 (.05)	.26 (.05)
SQDF	.25 (.06)	.48 (.07)	.48 (.05)
SEDF	.16 (.04)	.23 (.04)	.21 (.04)
e_{cv}^{SRDF}	.10 (.04)	.14 (.06)	.10 (.03)
$\bar{\hat{\lambda}}^{SRDF}$	.37 (.38)	.25 (.28)	.09 (.10)
$\bar{\hat{\gamma}}^{SRDF}$	.78 (.31)	.86 (.21)	.90 (.19)

Table 3.3: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace.** Average error rate (with standard deviation) for several discriminant functions.

	$p = 6$	$p = 10$	$p = 20$
SRDF	.07 (.05)	.12 (.04)	.15 (.04)
SLDF	.06 (.03)	.11 (.04)	.14 (.04)
SQDF	.14 (.05)	.29 (.06)	.39 (.06)
SEDF	.24 (.06)	.29 (.06)	.32 (.05)
e_{cv}^{SRDF}	.06 (.04)	.11 (.04)	.13 (.04)
$\tilde{\lambda}^{SRDF}$	.87 (.24)	.89 (.23)	.87 (.19)
$\tilde{\gamma}^{SRDF}$	.05 (.14)	.04 (.11)	.04 (.09)

The SEDF gave the lowest average error rate (with smallest standard deviation) under conditions of equal and spherical group covariance matrices (Table 3.1), but was similar to the SDRF. This is not surprising since in these conditions the optimal value for λ and γ is $(1, 1)$ since such regularisation would bias the covariance estimates towards exactly the correct value. Even when the group covariance matrices are unequal, but spherical (Table 3.2), the SEDF gives a comparable error rate to the SRDF when the dimensionality is small. However as p becomes large, the error rate of the SEDF becomes much larger than that of the SRDF. Under these simulation conditions the SEDF gives a lower error rate than either the SQDF or SLDF. This is consistent with the findings in Chapter 2 regarding the relative performance of the SLDF and SEDF for various scenarios involving the ratio of the Euclidean to Mahalanobis distance. In these conditions of spherical group covariances the ratio is not small, whereas in those conditions involving highly ellipsoidal covariance matrices, the Mahalanobis distance is much larger than the Euclidean distance and the SEDF performs worse than the SLDF.

For the case of equal, highly ellipsoidal group covariance matrices (with group mean differences concentrated in the low variance subspace) (Table 3.3), the SEDF performs poorly compared to the SRDF and SLDF. A high degree of shrinkage of the covariance matrix eigenvalues to equality is clearly not helpful to the classification process here since the mean differences may become obscured. When the mean differences are concentrated in the high variance subspace (Table 3.4), the three methods, SRDF, SLDF and SEDF perform equally well for p not large.

Table 3.4: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace.** Average error rate (with standard deviation) for several discriminant functions.

	$p = 6$	$p = 10$	$p = 20$
SRDF	.06 (.03)	.10 (.03)	.11 (.03)
SLDF	.07 (.03)	.12 (.04)	.14 (.04)
SQDF	.16 (.06)	.30 (.08)	.42 (.06)
SEDF	.06 (.03)	.10 (.03)	.11 (.03)
e_{cv}^{SRDF}	.04 (.03)	.07 (.04)	.10 (.03)
$\hat{\lambda}^{SRDF}$	.85 (.31)	.86 (.29)	.79 (.33)
$\hat{\gamma}^{SRDF}$	.58 (.37)	.62 (.33)	.67 (.27)

Table 3.5: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences.** Average error rate (with standard deviation) for several discriminant functions.

	$p = 6$	$p = 10$	$p = 20$
SRDF	.20 (.06)	.12 (.05)	.03 (.02)
SLDF	.60 (.06)	.59 (.06)	.58 (.05)
SQDF	.17 (.05)	.14 (.06)	.14 (.04)
SEDF	.60 (.06)	.59 (.06)	.58 (.05)
e_{cv}^{SRDF}	.17 (.06)	.11 (.04)	.02 (.02)
$\hat{\lambda}^{SRDF}$	.04 (.07)	.04 (.06)	.04 (.06)
$\hat{\gamma}^{SRDF}$	.12 (.15)	.25 (.16)	.35 (.18)

This suggests that the error rate surface over the (λ, γ) plane is very "flat" over a wide range of values of λ and γ. For large dimensional settings, a high degree of regularisation with γ results in an overall reduction in variance so that the mean differences become more apparent. In these conditions, therefore, the SEDF and SRDF prove to be the superior methods, especially the SEDF with its maximal eigenvalue shrinkage.

The SEDF does not perform well under conditions of unequal, ellipsoidal group covariance matrices (Tables 3.5 and 3.6) since very little of either type of regularisation is appropriate in this case.

Table 3.6: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences.** Average error rate (with standard deviation) for several discriminant functions.

	$p = 6$	$p = 10$	$p = 20$
SRDF	.06 (.04)	.06 (.04)	.02 (.02)
SLDF	.17 (.05)	.18 (.04)	.21 (.04)
SQDF	.04 (.03)	.05 (.04)	.06 (.04)
SEDF	.16 (.04)	.17 (.04)	.17 (.04)
e_{cv}^{SRDF}	.04 (.03)	.03 (.03)	.01 (.01)
$\bar{\hat{\lambda}}^{SRDF}$	.10 (.20)	.10 (.14)	.07 (.06)
$\bar{\hat{\gamma}}^{SRDF}$	.19 (.27)	.29 (.22)	.35 (.19)

3.5.2 Simulations for groups with small mean differences

Most of the simulations performed by Friedman involved parameter settings where the mean differences between groups were quite large. It is of interest to examine the behaviour of the model selection process of the SRDF (i.e. the process which selects the regularisation parameters λ and γ) when the differences between group means is much smaller than before. This may indicate whether a greater or lesser degree of regularisation is generally required when the mean differences between groups decrease, and the conditions for discrimination become more difficult. In this section, a simulation study is performed under the same group covariance structures as in Subsection 3.5.1, but the Euclidean distance between each pair of population means is reduced by approximately 75 %. The average Mahalanobis distance between pairs of populations is reduced by a similar amount for most of the simulation conditions. For condition 6 (Table 3.11), the reduction in average Mahalanobis distance between pairs of populations is nearly 90 % through the smaller population mean differences used. Average parameter values (with standard deviations) for all conditions are given in Tables 3.7 to 3.11 for the case of smaller group mean differences. In the following discussion, a comparison is made between the simulation results in Subsection 3.5.1, and the results obtained under the same conditions except for smaller group mean differences, with emphasis on the average values of $\hat{\lambda}$ and $\hat{\gamma}$ (i.e. $\bar{\hat{\lambda}}$ and $\bar{\hat{\gamma}}$).

Overall, the relative performance of the various classification rules (in terms of their error rate estimates) is not changed by closer group means, but obviously the

Table 3.7: **Equal Spherical Covariance Matrices.** Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.1.

	$p = 6$	$p = 10$	$p = 20$
$\bar{\hat{\lambda}}^{SRDF}$	.69 (.39)	.73 (.35)	.70 (.37)
$\bar{\hat{\gamma}}^{SRDF}$	.69 (.35)	.65 (.37)	.67 (.33)

Table 3.8: **Unequal Spherical Covariance Matrices.** Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.2.

	$p = 6$	$p = 10$	$p = 20$
$\bar{\hat{\lambda}}^{SRDF}$	.26 (.30)	.12 (.14)	.05 (.08)
$\bar{\hat{\gamma}}^{SRDF}$	.73 (.32)	.84 (.22)	.89 (.16)

error rates of the rules increase substantially and to a different extent depending on the parameter settings. Higher average error rates are coupled with increases in the variance of the error rate estimates (approximately 20% higher with closer group means).

In the situation where the group covariance matrices are equal and spherical, the average selected $\hat{\lambda}$ and $\hat{\gamma}$ values are slightly lower since in general the information from the covariance estimates is more necessary for discrimination purposes than when the group means are well separated, and hence less regularisation is appropriate.

Under conditions of unequal, spherical group covariance matrices the average

Table 3.9: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace.** Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.3.

	$p = 6$	$p = 10$	$p = 20$
$\bar{\hat{\lambda}}^{SRDF}$	.75 (.32)	.81 (.26)	.73 (.29)
$\bar{\hat{\gamma}}^{SRDF}$	.01 (.04)	.06 (.16)	.09 (.22)

Table 3.10: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace.** Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.4.

	$p = 6$	$p = 10$	$p = 20$
$\bar{\hat{\lambda}}^{SRDF}$	.75 (.36)	.78 (.32)	.55 (.35)
$\bar{\hat{\gamma}}^{SRDF}$	.56 (.34)	.61 (.33)	.65 (.29)

Table 3.11: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences.** Average regularisation parameter values (with standard deviation) in the case of smaller mean differences than in Table 3.6.

	$p = 6$	$p = 10$	$p = 20$
$\bar{\hat{\lambda}}^{SRDF}$	.03 (.07)	.05 (.07)	.05 (.06)
$\bar{\hat{\gamma}}^{SRDF}$	.13 (.18)	.28 (.17)	.31 (.19)

selected $\hat{\lambda}$ value is reduced with higher dimensionality and small separation between groups. This is to be expected since regularisation of the covariance estimates to commonality is likely to be more detrimental to the classification process if group mean differences are small. The value of $\bar{\hat{\gamma}}$ under those conditions remains very high despite the separation between the groups becoming small. This indicates that those conditions are ideal for eigenvalue shrinkage.

The best two classifiers under conditions of equal, highly ellipsoidal group covariance matrices (with group mean differences concentrated in the low variance subspace) are the SLDF and SRDF. When the groups are close together, the average $\hat{\lambda}$ value for the SRDF is still close to one which is the optimal value, but is again slightly reduced because of the need to retain covariance information for the classification process. The value of $\bar{\hat{\gamma}}$ is not affected by smaller mean differences and remains very close to zero.

In summary, the SRDF is still generally superior to the other techniques in terms of assessed error rate even for very small separation between group means. The values of $\bar{\hat{\lambda}}$ and $\bar{\hat{\gamma}}$ are not greatly affected by closer group means, but any effect that is present is towards less regularisation. This is so that more information may be retained from the group covariance estimates in order to enhance classification

under difficult discrimination conditions.

3.6 FURTHER MODEL SELECTION CONSIDERATIONS FOR THE SRDF: BREAKING OF TIES

In Section 3.4 it was pointed out that often the choice of $(\hat{\lambda}, \hat{\gamma})$ for the sample regularised discriminant function (SRDF) is not uniquely determined. If the cross-validated error rate of the SRDF over the (λ, γ) plane is thought of as a response surface, then it is often the case that the surface is very flat in the neighbourhood of its minimum. Thus, there is a range of (λ, γ) combinations that result in the same or very similar minimum cross-validated error rate being obtained. Consequently, a decision must be made as to how to break the tie and choose a particular $(\hat{\lambda}, \hat{\gamma})$ combination to use in the model.

It is of interest to study the effect of a different procedure than that employed by Friedman (1989) for selecting $\hat{\lambda}$ and $\hat{\gamma}$ in these tied situations. Friedman's approach was one of maximum regularisation: choosing the largest $\hat{\gamma}$ value for the largest $\hat{\lambda}$ among those combinations with the minimum cross-validated error rate. Rayens and Greene (1989) showed the importance of any specific procedure used to break ties by giving an example where the minimum cross-validated error rate occurs at more than one-third of the points on the (λ, γ) grid. They noted that in this case if the ties were broken by taking the largest $\hat{\lambda}$ value for the largest $\hat{\gamma}$ value, a completely different grid point would have been selected. The results of a simulation study are reported in this section, which (again) involved performing 100 replications in identical settings to those described in Section 3.5, and under the same sets of conditions. The difference here is that a policy of minimum regularisation for the SRDF is employed in those cases where the minimum cross-validated error rate is not uniquely determined. Thus, if there are more than one $(\hat{\lambda}, \hat{\gamma})$ combinations associated with the minimum cross-validated training sample error rate, the point chosen is that which has the smallest values of $\hat{\gamma}$ for the smallest $\hat{\lambda}$. The classification rule which employs this tie-breaking procedure will be denoted as SRDF1, which represents "minimum regularisation" as opposed to Friedman's "maximum regularisation" option. Results are given in Tables 3.12 to 3.17, where the average error rates and average regularisation parameter values are given for various dimensions.

Table 3.12: **Equal Spherical Covariance Matrices.** Comparison of SRDF and SRDF1 error rates and regularisation parameter values.

	$p = 6$	$p = 10$	$p = 20$
SRDF1	.12 (.03)	.14 (.04)	.12(.03)
SRDF	.11 (.04)	.12 (.04)	.12 (.04)
e_{cv}^{RDF1}	.09 (.05)	.10 (.05)	.10(.04)
$\bar{\hat{\lambda}}^{SRDF1}$	.15 (.26)	.20 (.33)	.24(.33)
$\bar{\hat{\gamma}}^{SRDF1}$	.67 (.32)	.69 (.30)	.80(.25)
e_{cv}^{SRDF}	.09 (.05)	.10 (.04)	.10 (.04)
$\bar{\hat{\lambda}}^{SRDF}$	.87 (.29)	.85 (.30)	.80 (.34)
$\bar{\hat{\gamma}}^{SRDF}$	.78 (.34)	.81 (.26)	.81 (.24)

The first and major finding from this study which compares SRDF1 with the SRDF (see Tables 3.12 to 3.17), is that whether minimum or maximum regularisation is used to break ties does not matter greatly in most of the parameter settings considered, even though the assessed values $\hat{\lambda}$ and, to a lesser extent $\hat{\gamma}$, are quite different. It also indicates the degree of homogeneity in the error rate response surface over the $(\hat{\lambda}, \hat{\gamma})$ plane. This homogeneity is greater with respect to the covariance mixing parameter λ, while the error rate surface is clearly more sensitive to the parameter γ.

Examining Table 3.12, for example, it can be seen that when the group covariance matrices are all equal and spherical, the error rate for SRDF1 is only slightly greater than that for the SRDF, even though $\bar{\hat{\lambda}}$ is much smaller. For SRDF1, $\bar{\hat{\lambda}}$ is close to zero while for SRDF it is close to one, even though the optimal parameter configuration is $\lambda = \gamma = 1$. The value of $\bar{\hat{\gamma}}$ is only slightly lower for SRDF1, indicating that substantial shrinkage of the eigenvalues is important for classifying under these conditions. The standard deviation of the sample based $\hat{\lambda}$ and $\hat{\gamma}$ estimates are similar for both rules, and its large size (in general) is further evidence that the cross-validated error rate response surface is quite flat at its minimum.

SRDF1 might be expected to give a better error rate than the SRDF if the group covariances are unequal but spherical (Table 3.13), since a low value of $\hat{\lambda}$ is desirable. In fact the difference is only slight and only occurs in the high dimensional settings. For SRDF1, $\bar{\hat{\lambda}}$ is very close to the optimal value of zero, and

Table 3.13: **Unequal Spherical Covariance Matrices.** Comparison of SRDF and SRDF1 error rates and regularisation parameter values.

	$p = 6$	$p = 10$	$p = 20$
SRDF1	.18 (.05)	.16 (.04)	.11 (.03)
SRDF	.14 (.04)	.18 (.05)	.11 (.04)
e_{cv}^{RDF1}	.15 (.06)	.14 (.05)	.10 (.03)
$\bar{\hat{\lambda}}^{SRDF1}$	.10 (.18)	.06 (.09)	.03 (.06)
$\bar{\hat{\gamma}}^{SRDF1}$	.71 (.30)	.84 (.22)	.90 (.14)
e_{cv}^{SRDF}	.10 (.04)	.14 (.06)	.10 (.03)
$\bar{\hat{\lambda}}^{SRDF}$	.37 (.38)	.25 (.28)	.09 (.10)
$\bar{\hat{\gamma}}^{SRDF}$	.78 (.31)	.86 (.21)	.90 (.19)

Table 3.14: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace.** Comparison of SRDF and SRDF1 error rates and regularisation parameter values.

	$p = 6$	$p = 10$	$p = 20$
SRDF1	.08 (.04)	.13 (.05)	.16 (.04)
SRDF	.07 (.05)	.12 (.04)	.15 (.04)
e_{cv}^{RDF1}	.05 (.03)	.11 (.05)	.14 (.04)
$\bar{\hat{\lambda}}^{SRDF1}$	.41 (.28)	.56 (.30)	.73 (.27)
$\bar{\hat{\gamma}}^{SRDF1}$	.02 (.07)	.03 (.11)	.02 (.07)
e_{cv}^{SRDF}	.06 (.04)	.11 (.04)	.13 (.04)
$\bar{\hat{\lambda}}^{SRDF}$	.87 (.24)	.89 (.23)	.87 (.19)
$\bar{\hat{\gamma}}^{SRDF}$	.05 (.14)	.04 (.11)	.04 (.09)

Table 3.15: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace.** Comparison of SRDF and SRDF1 error rates and regularisation parameter values.

	$p = 6$	$p = 10$	$p = 20$
SRDF1	.07 (.03)	.10 (.03)	.11 (.03)
SRDF	.06 (.03)	.10 (.03)	.11 (.03)
e_{cv}^{RDF1}	.04 (03)	.08 (.04)	.09 (.03)
$\bar{\hat{\lambda}}^{SRDF1}$	.15 (.25)	.26 (.32)	.32 (.34)
$\bar{\hat{\gamma}}^{SRDF1}$	.50 (.35)	.55 (.26)	.67 (.27)
e_{cv}^{SRDF}	.04 (.03)	.07 (.04)	.10 (.03)
$\bar{\hat{\lambda}}^{SRDF}$	.85 (.31)	.86 (.29)	.79 (.33)
$\bar{\hat{\gamma}}^{SRDF}$	.58 (.37)	.62 (.33)	.67 (.27)

Table 3.16: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences.** Comparison of SRDF and SRDF1 error rates and regularisation parameter values.

	$p = 6$	$p = 10$	$p = 20$
SRDF1	.18 (.06)	.11 (.04)	.03 (.02)
SRDF	.20 (.06)	.12 (.05)	.03 (.02)
e_{cv}^{RDF1}	.18 (.06)	.09 (.04)	.02 (.01)
$\bar{\hat{\lambda}}^{SRDF1}$	.01 (.04)	.01 (.04)	.02 (.05)
$\bar{\hat{\gamma}}^{SRDF1}$	.10 (.14)	.26 (.15)	.26 (.15)
e_{cv}^{SRDF}	.17 (.06)	.11 (.04)	.02 (.02)
$\bar{\hat{\lambda}}^{SRDF}$	.04 (.07)	.04 (.06)	.04 (.06)
$\bar{\hat{\gamma}}^{SRDF}$	.12 (.15)	.25 (.16)	.35 (.18)

Table 3.17: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences.** Comparison of SRDF and SRDF1 error rates and regularisation parameter values.

	$p = 6$	$p = 10$	$p = 20$
SRDF1	.05 (.02)	.05 (.04)	.01 (.01)
SRDF	.06 (.04)	.06 (.04)	.02 (.02)
e_{cv}^{RDF1}	.04 (.03)	.03 (.02)	.01 (.01)
$\bar{\hat{\lambda}}^{SRDF1}$	.01 (.03)	.02 (.04)	.00 (.00)
$\bar{\hat{\gamma}}^{SRDF1}$	.10 (.13)	.22 (.15)	.27 (.09)
e_{cv}^{SRDF}	.04 (.03)	.03 (.03)	.01 (.01)
$\bar{\hat{\lambda}}^{SRDF}$	.10 (.20)	.10 (.14)	.07 (.06)
$\bar{\hat{\gamma}}^{SRDF}$	.19 (.27)	.29 (.22)	.35 (.19)

its standard deviation is very small. The values of $\bar{\hat{\gamma}}$ for both SRDF1 and the SRDF are very similar in magnitude, indicating that a certain level of $\hat{\gamma}$ is necessary in this case.

In the cases of equal but highly ellipsoidal group covariances (Tables 3.14 and 3.15), altering the procedure for the breaking of ties has little effect in terms of error rates. The minimum level of the cross-validated error rate response surface on the (λ, γ) plane occurs over a wide range of $\hat{\lambda}$, but a much narrower range of $\hat{\gamma}$ values, again indicating that eigenvalue shrinkage is more critical for classification purposes under these conditions, since the error rate is very sensitive to the parameter γ.

Error rates are slightly lower for the SRDF1 than for the SRDF under simulation conditions where the group covariance matrices are highly ellipsoidal and dissimilar. In the situation of equal group means (Table 3.16), both $\bar{\hat{\lambda}}$ and $\bar{\hat{\gamma}}$ are very close to zero for both the SRDF and SRDF1, although $\bar{\hat{\gamma}}$ increases for larger p. This shows that the minimum error rate would usually occur in a very small region of the (λ, γ) plane, situated near the vertex $\lambda = \gamma = 0$, which is the optimum combination for this parameter configuration.

Conditions for discrimination are usually improved by having non-zero differences between group means (Table 3.17). The SRDF1 performs better than the SRDF here because the distribution of $\hat{\lambda}$ for SRDF1 has a mean ($\bar{\hat{\lambda}}$) closer to the desirable value of zero. Its standard deviation is also smaller, indicating a greater consistency of low $\hat{\lambda}$ values selected. The distribution of $\hat{\gamma}$ also has a lower standard

deviation for SRDF1.

In conclusion therefore, modifications to the way in which the SRDF breaks ties in the selection of regularisation parameter values $\hat{\lambda}$ and $\hat{\gamma}$ cannot be said to significantly change its error rate. Under most of the conditions looked at, the error rates with and without the modification were quite similar, although some parameter configuration settings favoured the lesser degree of regularisation offered by SRDF1, while others favoured more regularisation. Thus the issue of how the SRDF should break ties in the minimum cross-validated error rate is not a crucial one in terms of affecting the error rate of the discriminant rule for most of the conditions trialled. Also, of the two regularisation parameters, the choice of $\hat{\lambda}$ is less crucial and less precise, than the choice of $\hat{\gamma}$.

Chapter 4

INVARIANCE AND SAMPLE SIZE CONSIDERATIONS FOR THE SAMPLE REGULARISED DISCRIMINANT FUNCTION

4.1 INTRODUCTION

Friedman (1989) noted that the regularised discriminant function is not generally scale invariant. The reason for this relates to the presence of the eigenvalue shrinkage parameter γ. Changing the relative scales of the measurement variables, or their linear combinations, will usually alter the eigenvalues of the sample covariance matrix and change the classification rule and results. In particular, if $\gamma = 0$, the SRDF is scale invariant. Since scale invariance is often regarded as an important characteristic of discriminant functions, it is of considerable interest to investigate whether a similar level of discriminatory success can be achieved with a modification of the regularisation rule. A modification to the SRDF is introduced in Sections 4.2 and 4.3, and is compared with the original SRDF.

A further study is implemented in this chapter to investigate the effect of sample size on the various classifiers. From the studies in the previous chapter, plus other published results (see, for example, Aeberhard et al. (1994)), it was found that the SRDF is at least equal to but usually superior to the other classification rules under a fairly wide range of situations. The conditions under which these studies have been implemented involved reasonably small sample sizes compared to dimension, p. This simulation study is undertaken using a range of larger sample sizes (in relation to p) in an attempt to determine if regularisation – and in particular the

eigenvalue shrinkage feature of the SRDF – loses its advantage over the common discriminant rules once the sample size becomes sufficiently larger than p.

4.2 INVARIANCE

In order to achieve a classification rule possessing scale invariance, the effect of removing the eigenvalue shrinkage parameter γ from the model is examined. However, if one simply removed γ from the SRDF model, the resulting discriminant rule would allow for a reduced set of regularised models between the SQDF and the SLDF only, as defined in expression (3.6). It was mentioned there that this set of alternatives is rather restrictive. Further, the resulting model may not provide appropriate regularisation if the group covariance matrices are of quite a different nature. In such a situation, it is plausible that some improvement could be made if *each* covariance matrix were independently regularised to the pooled estimate by an appropriate degree, which would be estimated from the training data. Using such shrinkage could overcome, to some extent, the problem of inappropriate regularisation, as the model would be more sensitive to variations in the "shape" among the various populations.

In the single parameter regularisation model of equation (3.6), it may occur that in the selection of λ, a large proportion of the training observations misclassified by cross-validation come from one group. This may be in part due to the shrinkage employed being inappropriate for that group but appropriate for the others. The following model is proposed to obtain separate regularised group covariance estimates:

$$\hat{\Sigma}_k(\lambda_k) = \frac{(1-\lambda_k)(n_k-1)\tilde{\mathbf{S}}_k + \lambda_k \tilde{\mathbf{S}}_p}{(1-\lambda_k)(n_k-1) + \lambda_k (N-K)} \tag{4.1}$$

where $k = 1, \ldots, K$ groups, and $\tilde{\mathbf{S}}_p$ is the pooled covariance matrix.

The K regularisation parameters λ_k control the degree of shrinkage of the individual group covariance matrix estimates towards the pooled estimate. The value $\lambda_k = 0$ gives $\hat{\Sigma}_k(\lambda_k) = \tilde{\mathbf{S}}_k$ and $\lambda_k = 1$ yields $\hat{\Sigma}_k(\lambda_k) = \tilde{\mathbf{S}}_p$. Each λ_k is obtained by minimising the group conditional cross-validated error rate over the range $0 \leq \lambda_k \leq 1$, $k = 1, \ldots, K$. Each $\mathbf{S}_k$ in expression (1.5) is replaced by $\hat{\Sigma}_k(\lambda_k)$ for discriminant analysis. This approach will be denoted SRDF-Modified (or SRDF-M). To demonstrate that the SRDF-M rule is invariant under a linear

scale transformation, let

$$d_1(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_1)'\hat{\boldsymbol{\Sigma}}_1^{-1}(\lambda_1)(\boldsymbol{x} - \bar{\boldsymbol{x}}_1) + \ln\left|\hat{\boldsymbol{\Sigma}}_1^{-1}(\lambda_1)\right|$$

be the discriminant score for observation $\boldsymbol{x}$ in group 1 ($0 \leq \lambda_1 \leq 1$). Similarly, let

$$d_2(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_2)'\hat{\boldsymbol{\Sigma}}_2^{-1}(\lambda_2)(\boldsymbol{x} - \bar{\boldsymbol{x}}_2) + \ln\left|\hat{\boldsymbol{\Sigma}}_2^{-1}(\lambda_2)\right|$$

be the score for observation $\boldsymbol{x}$ in group 2, for λ_2 ($0 \leq \lambda_2 \leq 1$) not necessarily equal to λ_1. Given a symmetric non-singular matrix $\mathbf{A}$, one can form a linear transformation $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$ and it can be shown that

$$d_1(\boldsymbol{y}) = d_1(\boldsymbol{x}) + 2\ln\left|\mathbf{A}^{-1}\right|$$

and

$$d_2(\boldsymbol{y}) = d_2(\boldsymbol{x}) + 2\ln\left|\mathbf{A}^{-1}\right|.$$

Thus the transformed discriminant scores for all the K groups only differ from the untransformed scores by the addition of a constant ($2\ln|\mathbf{A}|$), and hence the discriminant rule is not affected by the transformation. The following section will report on a simulation study to investigate the relative performance of the scale invariant SRDF-M compared with the SRDF and other classification rules.

4.3 ASSESSING THE PERFORMANCE OF THE MODIFIED REGULARISED DISCRIMINANT FUNCTION (SRDF-M)

4.3.1 The performance of SRDF-M when the population shapes are similar

A Monte Carlo simulation study was performed under the same conditions (defined by the various population parameter configurations) and sample sizes as in the previous chapter (See Sections 3.5 and 3.6). Results for the SRDF-M rule are presented in Tables 4.1 to 4.6, along with the error rates of the other discriminant rules. These are repeated from tables given in Chapter 3, thus allowing comparison with the other classification rules.

Table 4.1: **Equal Spherical Covariance Matrices.** Average error rate (with standard deviation) and parameter values for several discriminant functions.

	Dimension p		
	6	**10**	**20**
SRDF	.11 (.04)	.12 (.04)	.12 (.04)
SRDF-M	.14 (.04)	.17 (.05)	.16 (.04)
SLDF	.13 (.04)	.16 (.05)	.15 (.04)
SQDF	.23 (.05)	.39 (.07)	.42 (.05)
SEDF	.11 (.03)	.12 (.04)	.12 (.03)
e_{cv}^{SRDF}	.09 (.05)	.10 (.04)	.10 (.04)
$\hat{\bar{\lambda}}^{SRDF}$	.87 (.29)	.85 (.30)	.80 (.34)
$\hat{\bar{\gamma}}^{SRDF}$	.78 (.34)	.81 (.26)	.81 (.24)
$e_{cv(1)}^{SRDF-M}$	.17 (.08)	.17 (.09)	.21 (.09)
$e_{cv(2)}^{SRDF-M}$	.09 (.07)	.12 (.07)	.13 (.07)
$e_{cv(3)}^{SRDF-M}$	.09 (.07)	.10 (.07)	.12 (.06)
$\hat{\bar{\lambda}}_1^{SRDF-M}$	.79 (.35)	.81 (.28)	.84 (.26)
$\hat{\bar{\lambda}}_2^{SRDF-M}$	.91 (.25)	.93 (.19)	.90 (.21)
$\hat{\bar{\lambda}}_3^{SRDF-M}$	.92 (.21)	.87 (.25)	.83 (.25)

It is evident from the results in these tables that having the option to use the regularisation parameter γ to shrink the covariance matrix eigenvalues to equality undoubtedly enhances discrimination in many situations, and not only when the populations are spherical. This type of shrinkage reduces the variance, which, despite the introduced bias, is beneficial for discrimination purposes especially in the high dimensional setting. This extra variance-reduction factor probably explains why the minimum cross-validated error rate for SRDF-M sometimes underestimates the actual error rate by a greater degree than for the SRDF, especially for large dimensions (p). The magnitude of the minimum cross-validated error rate over the whole training sample for SRDF-M is at a comparable level to those for the SRDF, and it is the *actual* error rate which is usually higher for SRDF-M.

When the group covariances are spherical and set to be equal, SRDF-M yielded error rate estimates of between 30% and 40% higher than the SRDF (Table 4.1). Under these conditions, eigenvalue shrinkage (to equality) clearly enhances discrimination, as evidenced by the fact that the SEDF performs well. The mean

Table 4.2: **Unequal Spherical Covariance Matrices.** Average error rate (with standard deviation) and parameter values for several discriminant functions.

	Dimension p		
	6	**10**	**20**
SRDF	.14 (.04)	.18 (.05)	.11 (.04)
SRDF-M	.24 (.07)	.28 (.08)	.28 (.08)
SLDF	.23 (.06)	.26 (.05)	.26 (.05)
SQDF	.32 (.06)	.44 (.07)	.48 (.05)
SEDF	.20 (.04)	.22 (.05)	.21 (.04)
e_{cv}^{SRDF}	.10 (.04)	.14 (.06)	.10 (.03)
$\hat{\bar{\lambda}}^{SRDF}$	.37 (.38)	.25 (.28)	.09 (.10)
$\hat{\bar{\gamma}}^{SRDF}$	.78 (.31)	.86 (.21)	.90 (.19)
$e_{cv(1)}^{SRDF-M}$	.14 (.09)	.14 (.09)	.11 (.05)
$e_{cv(2)}^{SRDF-M}$	.19 (.10)	.19 (.09)	.24 (.08)
$e_{cv(3)}^{SRDF-M}$	.21 (.09)	.25 (.10)	.25 (.09)
$\hat{\bar{\lambda}}_1^{SRDF-M}$	.70 (.35)	.73 (.34)	.60 (.27)
$\hat{\bar{\lambda}}_2^{SRDF-M}$	.77 (.34)	.77 (.33)	.75 (.28)
$\hat{\bar{\lambda}}_3^{SRDF-M}$	.43 (.39)	.60 (.40)	.43 (.36)

Table 4.3: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace.** Average error rate (with standard deviation) and parameter values for several discriminant functions.

	Dimension p		
	6	10	20
SRDF	.07 (.05)	.12 (.04)	.15 (.04)
SRDF-M	.06 (.03)	.14 (.05)	.16 (.04)
SLDF	.06 (.03)	.13 (.05)	.16 (.04)
SQDF	.13 (.06)	.36 (.08)	.39 (.06)
SEDF	.24 (.06)	.32 (.06)	.34 (.05)
e_{cv}^{SRDF}	.06 (.04)	.11 (.04)	.13 (.04)
$\tilde{\lambda}^{SRDF}$	.87 (.24)	.89 (.23)	.87 (.19)
$\tilde{\gamma}^{SRDF}$	.05 (.14)	.04 (.11)	.04 (.09)
$e_{cv(1)}^{SRDF-M}$	.10 (.08)	.17 (.10)	.19 (.07)
$e_{cv(2)}^{SRDF-M}$	.04 (.05)	.10 (.09)	.12 (.05)
$e_{cv(3)}^{SRDF-M}$	.04 (.06)	.09 (.07)	.10 (.06)
$\tilde{\lambda}_1^{SRDF-M}$	.91 (.26)	.79 (.33)	.83 (.27)
$\tilde{\lambda}_2^{SRDF-M}$	.99 (.05)	.95 (.17)	.86 (.29)
$\tilde{\lambda}_3^{SRDF-M}$	.96 (.19)	.87 (.27)	.91 (.19)

minimizing cross-validated error rate over all groups underestimated the actual error rate by around 20% for $p \leq 10$, but by only about 5% for $p = 20$. The means of the group conditional minimizing cross-validated error rates differed significantly, with substantial variation.

If the group covariances are spherical but unequal (Table 4.2), SRDF-M gives error rate estimates around 70% higher than for SRDF, and worse for larger dimensions. It is clear that under such conditions, eigenvalue shrinkage is very desirable in order to reduce variation in the higher dimensions. The mean minimum cross-validated error rate over all groups underestimated the actual misclassification risk by 25% – 30%, although observations from the higher variance groups were more frequently misclassified.

The performance of the SRDF-M rule is comparable to that of the SRDF under conditions of equal but highly ellipsoidal group covariances (Tables 4.3 and 4.4). This is not surprising since eigenvalue shrinkage is expected to be counterproductive

Table 4.4: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace.** Average error rate (with standard deviation) and parameter values for several discriminant functions.

	Dimension p		
	6	**10**	**20**
SRDF	.06 (.03)	.10 (.03)	.11 (.03)
SRDF-M	.08 (.03)	.14 (.04)	.15 (.05)
SLDF	.07 (.03)	.13 (.04)	.14 (.04)
SQDF	.16 (.05)	.36 (.08)	.38 (.06)
SEDF	.07 (.03)	.11 (.03)	.11 (.03)
e_{cv}^{SRDF}	.04 (.03)	.07 (.04)	.10 (.03)
$\bar{\hat{\lambda}}^{SRDF}$	.85 (.31)	.86 (.29)	.79 (.33)
$\bar{\hat{\gamma}}^{SRDF}$	.58 (.37)	.62 (.33)	.67 (.27)
$e_{cv(1)}^{SRDF-M}$	.07 (.05)	.15 (.08)	.16 (.07)
$e_{cv(2)}^{SRDF-M}$	.06 (.06)	.09 (.08)	.09 (.05)
$e_{cv(3)}^{SRDF-M}$	.06 (.06)	.09 (.07)	.11 (.06)
$\bar{\hat{\lambda}}_1^{SRDF-M}$	.86 (.31)	.80 (.32)	.80 (.28)
$\bar{\hat{\lambda}}_2^{SRDF-M}$	.88 (.31)	.88 (.29)	.87 (.24)
$\bar{\hat{\lambda}}_3^{SRDF-M}$	.86 (.30)	.87 (.25)	.88 (.23)

in this situation. In the case where the group mean differences are concentrated in the low variance subspace (Table 4.3), and therefore more pronounced, the $\bar{\hat{\lambda}}_k$ $(k = 1, \ldots, K)$ values are very close to one, and the performance of SRDF-M approaches that of the SLDF, which is the optimal rule in these conditions. However, when the group means are concentrated in the high variance subspace (Table 4.4), SRDF-M is less successful compared to the SRDF. The high degree of covariance shrinkage towards the identity matrix enhances discrimination, because of the reduction in variance achieved. This is why the SEDF performs as well as the SRDF under these conditions, and yields a lower error rate than SRDF-M by about 40%. The mean minimizing cross-validated error rate for SRDF-M underestimates the actual rate by between zero and 15% when the group mean differences are more distinguishable in the low variance subspace, and around 20% when the means are obscured by high variance.

The final sets of simulation conditions represent the situation where the group covariances are unequal and highly ellipsoidal (Tables 4.5 and 4.6). The SRDF-M does not perform well here. Its average misclassification risk is 50% to 100% larger than for the SRDF, and much more in the higher dimensions, when the SRDF error rate decreases on account of increased use of γ. The standard deviation of the misclassification error was also very large for SRDF-M compared with that of the other rules.

These conditions are ideal for the SQDF, hence one might expect SRDF-M to perform comparably well if the model selection procedure chooses small values of λ. However, it performs considerably worse, as $\bar{\hat{\lambda}}$ shows that values of λ are being chosen which are too high. The minimizing cross-validated error rate based on the training sample is observed to be similar to that for the SRDF, although Friedman noted that this did not appear to be related to the actual error rate estimate obtained from the test sample. Despite this observation, it can be seen from Tables 4.5 and 4.6 that in fact the minimising cross-validated error rate severely underestimates the actual rate for SRDF-M, especially for the high dimensional settings. This is a curious phenomenon which exhibits itself strongly only in these simulation conditions where the groups have high and unequal variance. The reduction in variance obtained by eigenvalue shrinkage is not the complete explanation for, otherwise, SRDF-M should perform comparably to the SQDF, but it does not. It should be noted that the error rate estimates for SRDF-M also have unusually

Table 4.5: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences.** Average error rate (with standard deviation) and parameter values for several discriminant functions.

	Dimension p		
	6	**10**	**20**
SRDF	.20 (.06)	.12 (.05)	.03 (.02)
SRDF-M	.29 (.08)	.39 (.11)	.28 (.16)
SLDF	.60 (.06)	.60 (.06)	.59 (.06)
SQDF	.16 (.04)	.19 (.06)	.11 (.05)
SEDF	.60 (.07)	.59 (.06)	.58 (.06)
e_{cv}^{SRDF}	.17 (.06)	.11 (.04)	.02 (.02)
$\hat{\bar{\lambda}}^{SRDF}$	.04 (.07)	.04 (.06)	.04 (.06)
$\hat{\bar{\gamma}}^{SRDF}$	.12 (.15)	.25 (.16)	.35 (.18)
$e_{cv(1)}^{SRDF-M}$	.14 (.11)	.15 (.10)	.02 (.02)
$e_{cv(2)}^{SRDF-M}$	.07 (.08)	.10 (.07)	.01 (.02)
$e_{cv(3)}^{SRDF-M}$	.15 (.08)	.08 (.08)	.01 (.02)
$\hat{\bar{\lambda}}_1^{SRDF-M}$	.01 (.03)	.03 (.08)	.06 (.08)
$\hat{\bar{\lambda}}_2^{SRDF-M}$	.05 (.08)	.07 (.09)	.06 (.06)
$\hat{\bar{\lambda}}_3^{SRDF-M}$	.25 (.20)	.30 (.15)	.36 (.15)

Table 4.6: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences.** Average error rate (with standard deviation) and parameter values for several discriminant functions.

	Dimension p		
	6	**10**	**20**
SRDF	.06 (.04)	.06 (.04)	.02 (.02)
SRDF-M	.13 (.07)	.21 (.09)	.22 (.13)
SLDF	.20 (.05)	.21 (.04)	.20 (.05)
SQDF	.06 (.04)	.10 (.06)	.06 (.03)
SEDF	.20 (.05)	.20 (.04)	.17 (.04)
e_{cv}^{SRDF}	.04 (.03)	.03 (.03)	.01 (.01)
$\bar{\hat{\lambda}}^{SRDF}$	.10 (.20)	.10 (.14)	.07 (.06)
$\bar{\hat{\gamma}}^{SRDF}$	.19 (.27)	.29 (.22)	.35 (.19)
$e_{cv(1)}^{SRDF-M}$	.05 (.07)	.09 (.06)	.02 (.02)
$e_{cv(2)}^{SRDF-M}$	.04 (.05)	.06 (.07)	.01 (.01)
$e_{cv(3)}^{SRDF-M}$	.01 (.04)	.01 (.02)	.00 (.00)
$\bar{\hat{\lambda}}_1^{SRDF-M}$	.11 (.21)	.11 (.18)	.07 (.09)
$\bar{\hat{\lambda}}_2^{SRDF-M}$	.14 (.21)	.18 (.24)	.13 (.14)
$\bar{\hat{\lambda}}_3^{SRDF-M}$	.88 (.30)	.85 (.29)	.89 (.23)

high variance under these conditions. A possible explanation is that under these conditions the best rules are those where $\bar{\hat{\lambda}}$ is close to zero with low variability. Now, the values of $\bar{\hat{\lambda}}_k$ are not always close to zero for SRDF-M, and since each λ_k is obtained from such a small number of data points, it variability is high.

A feature of the performance of SRDF-M under these conditions is that $\bar{\hat{\lambda}}_3$ is much higher than $\bar{\hat{\lambda}}_1$ or $\bar{\hat{\lambda}}_2$. It happens that group 3 does not have quite the same extreme ellipsoidal nature of the other two groups. Significant shrinkage of the group 3 covariance matrix to the pooled covariance appears to lead to observations from that group becoming indistinguishable (to the classification rule) from those of the other high variance groups, and the error rate for that group becomes quite high.

It is noted that if a policy of minimum regularisation is used to break ties (similar to that employed by SRDF1 in Chapter 3), the performance of SRDF-M is enhanced because smaller values of λ_k are selected. The two tables (Tables 4.7 and 4.8) below compare the performance of SRDF-M with SRDF-M1. The difference between rules SRDF-M and SRDF-M1 lies only in the policy used to break ties when there is no unique value of λ_k which minimizes the cross-validated error rate for group k. That is, if the error rate is the same for several values of λ, SRDF-M selects the largest λ of those values, while SRDF-M1 selects the smallest. From Tables (4.7 and 4.8), it can be seen that the average values of λ for the SRDF-M rule have much higher variation than those for SRDF-M1. This is to be expected since the $\bar{\hat{\lambda}}$ values are generally much closer to 0 for the SRDF-M1 rule. The minimum cross-validated error rates for SRDF-M1 are, for these simulation cases, higher than those for SRDF-M, yet their average is closer to the actual error rate obtained from the test samples. This is because the minimum cross-validated error rates for SRDF-M1 do not underestimate the actual error rate as severely as those for SRDF-M. The minimum cross-validated error rates for both SRDF-M and SRDF-M1 are quite variable, as are the actual error rates achieved by each rule.

In conclusion, the proposed regularisation model SRDF-M was not as successful as the SRDF. This clearly shows the value of eigenvalue shrinkage, especially when p is large. The attempt to make SRDF-M more sensitive by employing a separate λ for each group caused other problems in certain circumstances as described above. If the problem of lack of scale invariance is to be avoided, other techniques

Table 4.7: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences.** Comparison of SRDF-M and SRDF-M1 classifiers.

	Dimension p		
	6	**10**	**20**
SRDF-M	.29 (.08)	.39 (.11)	.28 (.16)
SRDF-M1	.27 (.06)	.30 (.10)	.23 (.11)
$e_{cv(1)}^{SRDF-M}$	.14 (.11)	.15 (.10)	.02 (.02)
$e_{cv(2)}^{SRDF-M}$	.07 (.08)	.10 (.07)	.01 (.02)
$e_{cv(3)}^{SRDF-M}$	.15 (.08)	.08 (.08)	.01 (.02)
$\hat{\bar{\lambda}}_1^{SRDF-M}$	.01 (.03)	.03 (.08)	.06 (.08)
$\hat{\bar{\lambda}}_2^{SRDF-M}$	.05 (.08)	.07 (.09)	.06 (.06)
$\hat{\bar{\lambda}}_3^{SRDF-M}$	.25 (.20)	.30 (.15)	.36 (.15)
$e_{cv(1)}^{SRDF-M1}$	.21 (.11)	.18 (.09)	.07 (.05)
$e_{cv(2)}^{SRDF-M1}$	.09 (.08)	.10 (.07)	.04 (.04)
$e_{cv(3)}^{SRDF-M1}$	.14 (.08)	.07 (.06)	.01 (.02)
$\hat{\bar{\lambda}}_1^{SRDF-M1}$	.03 (.14)	.04 (.07)	.06 (.06)
$\hat{\bar{\lambda}}_2^{SRDF-M1}$	.00 (.02)	.02 (.05)	.03 (.05)
$\hat{\bar{\lambda}}_3^{SRDF-M1}$	.14 (.11)	.15 (.08)	.13 (.02)

Table 4.8: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences.** Comparison of SRDF-M and SRDF-M1 classifiers.

	Dimension p		
	6	**10**	**20**
SRDF-M	.13 (.07)	.21 (.09)	.22 (.13)
SRDF-M1	.07 (.04)	.12 (.06)	.12 (.08)
$e_{cv(1)}^{SRDF-M}$	.05 (.07)	.09 (.06)	.02 (.02)
$e_{cv(2)}^{SRDF-M}$	.04 (.05)	.06 (.07)	.01 (.01)
$e_{cv(3)}^{SRDF-M}$	.01 (.04)	.01 (.02)	.00 (.00)
$\tilde{\bar{\lambda}}_1^{SRDF-M}$	.11 (.21)	.11 (.18)	.07 (.09)
$\tilde{\bar{\lambda}}_2^{SRDF-M}$	.14 (.21)	.18 (.24)	.13 (.14)
$\tilde{\bar{\lambda}}_3^{SRDF-M}$	.88 (.30)	.85 (.29)	.89 (.23)
$e_{cv(1)}^{SRDF-M1}$	.08 (.05)	.07 (.06)	.04 (.03)
$e_{cv(2)}^{SRDF-M1}$	.04 (.05)	.04 (.04)	.03 (.03)
$e_{cv(3)}^{SRDF-M1}$	.00 (.02)	.00 (.01)	.00 (.01)
$\tilde{\bar{\lambda}}_1^{SRDF-M1}$	.05 (.12)	.06 (.08)	.09 (.07)
$\tilde{\bar{\lambda}}_2^{SRDF-M1}$	.01 (.05)	.05 (.10)	.04 (.07)
$\tilde{\bar{\lambda}}_3^{SRDF-M1}$	.11 (.15)	.12 (.07)	.13 (.02)

need to be devised to replace eigenvalue shrinkage, while ensuring the accuracy of classification attained by the SRDF is not compromised.

4.3.2 The performance of SRDF-M when the population shapes are very different

The proposal of the technique of SRDF-M, where a separate covariance mixing parameter λ is determined for each group, envisaged the situation where the groups had quite different covariance structures. Allowing for a different degree of shrinkage (to the pooled estimate), as appropriate for each group, would be expected to lead to a more sensitive model than one which employs only a single regularisation parameter, λ. The simulation conditions of Section 3.5 (used in Subsection 4.3.1) all involved group parameter settings where the covariance matrices for the three groups were all of the same type of structure: either all spherical or all ellipsoidal. Hence it is of interest to investigate the usefulness of SRDF-M in situations where the group covariances are not all of the same type, but are a mixture of spherical and ellipsoidal structures. It may be expected that the potential of SRDF-M to shrink each covariance to the average by a different and appropriate amount, would be one advantage it affords over the other classification rules (especially the standard SRDF).

A further simulation study was conducted to compare the performances (in terms of their error rates) of the following discriminant rules: SRDF, SRDF-M1, SQDF, SLDF and SEDF. The reason why SRDF-M1 was chosen instead of SRDF-M is because if the group covariances are dissimilar, as they are for this study, a lower degree of covariance mixing is usually appropriate. The number of groups in each case is three.

For this study, the following four sets of parameter configurations were used.

1. Two equal and highly ellipsoidal population covariance matrices, and one spherical covariance matrix (identity matrix). The population mean differences concentrated mainly in the low variance subspace of the two ellipsoidal populations.

2. One highly ellipsoidal population covariance matrix, one moderately ellipsoidal population covariance matrix, and one spherical covariance matrix (identity matrix). The population mean differences are spread

evenly across all dimensions.

3. Two unequal and highly ellipsoidal population covariance matrices, and one spherical covariance matrix (identity matrix). Zero population mean differences.
4. One highly ellipsoidal population covariance matrix, and two unequal spherical covariance matrices. The population mean differences are spread evenly across all dimensions.

The ellipsoidal covariance matrices were very similar to those used in Friedman (1989) and in the previous simulations in this thesis. The procedure employed for this simulation study was the same as that in Section 3.5, and once again 100 replications were performed and the results are presented in Tables 4.9 to 4.12.

Table 4.9: **Two equal and highly ellipsoidal covariances, one spherical covariance matrices.** Mean differences in the low variance subspace: Average error rates with standard deviations.

	Dimension p		
	6	**10**	**20**
SRDF-M1	.02 (.02)	.06 (.05)	.06 (.05)
SRDF	.04 (.03)	.07 (.04)	.07 (.03)
SLDF	.04 (.02)	.08 (.03)	.10 (.04)
SQDF	.03 (.03)	.15 (.07)	.24 (.08)
SEDF	.13 (.04)	.18 (.04)	.18 (.04)
e_{cv}^{SRDF}	.01 (.01)	.03 (.02)	.04 (.02)
$\bar{\hat{\lambda}}^{SRDF}$	.47 (.40)	.28 (.25)	.29 (.20)
$\bar{\hat{\gamma}}^{SRDF}$	.15 (.28)	.21 (.29)	.23 (.32)
$e_{cv(1)}^{SRDF-M1}$	.00 (.01)	.00 (.01)	.00 (.00)
$e_{cv(2)}^{SRDF-M1}$	.01 (.03)	.05 (.05)	.08 (.05)
$e_{cv(3)}^{SRDF-M1}$	.01 (.02)	.04 (.06)	.07 (.04)
$\bar{\hat{\lambda}}_1$	.09 (.06)	.13 (.06)	.13 (.00)
$\bar{\hat{\lambda}}_2$	.03 (.07)	.13 (.17)	.24 (.19)
$\bar{\hat{\lambda}}_3$	.04 (.13)	.16 (.20)	.23 (.18)

Two equal, highly ellipsoidal covariances; one (low variance) spherical covariance matrix.

In this situation, two of the groups have equal and highly ellipsoidal covariance matrices as in Friedman (1989), Section (6.3). The other has a covariance matrix equal to the identity. The group mean differences are concentrated in the low variance subspace of the first two groups. Table 4.9 shows the misclassification error rates for each discriminant rule. As mentioned earlier, these conditions are fairly well suited to the SLDF, although the spherical group would become almost indistinguishable from the other two groups if a high degree of covariance mixing were employed. The SRDF-M1 rule performs well – slightly better than both SLDF and SRDF, for all dimensions used, although the error rates for all three classifiers are quite similar in magnitude. The SRDF employs mild eigenvalue shrinkage. However this does not lead to lower error rates than SRDF-M1. The minimum cross-validated error rate for SRDF-M1 underestimates the actual error rate (as assessed by the test sample) by a similar margin to that for the SRDF (i.e. 40%

Table 4.10: **Three unequal covariance matrices - one highly ellipsoidal, one moderately ellipsoidal, one spherical.** Group mean differences spread equally over all subspaces: Average error rates with standard deviations.

	Dimension p		
	6	**10**	**20**
SRDF-M1	.03 (.02)	.08 (.05)	.09 (.06)
SRDF	.05 (.03)	.06 (.03)	.03 (.02)
SLDF	.13 (.03)	.16 (.04)	.19 (.04)
SQDF	.03 (.02)	.06 (.04)	.07 (.05)
SEDF	.12 (.03)	.14 (.04)	.15 (.04)
e_{cv}^{SRDF}	.02 (.02)	.03 (.02)	.01 (.01)
$\bar{\hat{\lambda}}^{SRDF}$	.14 (.16)	.17 (.19)	.12 (.10)
$\bar{\hat{\gamma}}^{SRDF}$	.27 (.30)	.40 (.29)	.29 (.24)
$e_{cv(1)}^{SRDF-M1}$	.03 (.04)	.05 (.05)	.02 (.02)
$e_{cv(2)}^{SRDF-M1}$	.03 (.04)	.05 (.05)	.03 (.03)
$e_{cv(3)}^{SRDF-M1}$	.00 (.00)	.00 (.00)	.00 (.00)
$\bar{\hat{\lambda}}_1$	.01 (.03)	.05 (.08)	.06 (.09)
$\bar{\hat{\lambda}}_2$	.05 (.12)	.05 (.07)	.06 (.08)
$\bar{\hat{\lambda}}_3$	.01 (.04)	.09 (.06)	.12 (.02)

to 50%).

Three unequal group covariances, one highly ellipsoidal, one moderately ellipsoidal and one spherical (low variance).

Here, the situation is considered where all three group covariance matrices are of a different nature: one highly ellipsoidal, as in the previous case; one with a less extreme ellipsoidal structure to the first, with the ratio between the largest and smallest eigenvalues halved; and one equal to the identity matrix. The differences in group means is spread equally over all subspaces. Table 4.10 presents the results.

The SQDF, SRDF and SRDF-M1 rules perform equally well in the small dimensional settings. However, for larger p, the SRDF once again emerges superior, employing moderate eigenvalue shrinkage. It should be noted, however, that the error rates concerned are all of a small magnitude. As with the previous conditions, the spherical group is the most correctly classified group by all classification rules

Table 4.11: **Three unequal covariance matrices: Two highly ellipsoidal, one spherical.** Zero group mean differences: Average error rates with standard deviations.

	Dimension p		
	6	**10**	**20**
SRDF-M1	.14 (.04)	.12 (.05)	.15 (.07)
SRDF	.12 (.05)	.10 (.04)	.03 (.02)
SLDF	.49 (.06)	.47 (.05)	.46 (.05)
SQDF	.12 (.04)	.22 (.07)	.21 (.05)
SEDF	.47 (.06)	.46 (.05)	.45 (.04)
e_{cv}^{SRDF}	.09 (.04)	.08 (.04)	.02 (.02)
$\bar{\hat{\lambda}}^{SRDF}$	.04 (.06)	.06 (.07)	.07 (.07)
$\bar{\hat{\gamma}}^{SRDF}$	.24 (.20)	.23 (.17)	.25 (.20)
$e_{cv(1)}^{SRDF-M1}$	.11 (.07)	.12 (.08)	.04 (.04)
$e_{cv(2)}^{SRDF-M1}$	.07 (.07)	.08 (.07)	.03 (.03)
$e_{cv(3)}^{SRDF-M1}$	.00 (.00)	.00 (.00)	.00 (.00)
$\bar{\hat{\lambda}}_1$	.02 (.04)	.06 (.06)	.06 (.07)
$\bar{\hat{\lambda}}_2$	.01 (.03)	.02 (.04)	.03 (.06)
$\bar{\hat{\lambda}}_3$	.12 (.01)	.13 (.00)	.13 (.00)

because of its low variance and the presence of group mean differences across all subspaces.

Three unequal group covariance matrices, two highly ellipsoidal and one spherical.

This example considers the situation where two of the group covariance matrices are highly ellipsoidal and very unequal (similar to those in Section 6.4, Friedman (1989)). The other is equal to the identity matrix. The group means are all located at the origin. Results are presented in Table 4.11.

SRDF-M1 again performs well relative to the SRDF, especially for $p = 6, 10$. The model selection procedure of SRDF-M1 again appears to behave appropriately, employing low covariance mixing in this case where shrinkage of this sort would generally be strongly counter-productive. Although the SRDF also shrinks the covariance matrices slightly under these conditions, it is the eigenvalue shrinkage

which is significant. The resulting decrease in variance enhances the discrimination process, especially for large dimension, and again makes the SRDF the superior classification rule. The SRDF-M1 rule performs better than the SQDF for the larger dimensional settings. This must be a consequence of the mild use of the covariance mixing parameter λ by SRDF-M1 to reduce variance in the higher variance subspaces.

One interesting feature of the behaviour of SRDF-M1 in these conditions is the large discrepancy between the minimizing cross-validated error rate based on the training sample, and the assessed actual error rate from the test sample. Despite the fact that the former is always an underestimate of the latter, and also that the methods of assessing the error rates are different, the large magnitude of the underestimation warrants closer examination. The average minimizing cross-validated error rate for SRDF-M1 is comparable to the corresponding quantity for the SRDF. However, the average actual error rate for SRDF-M1 is five times higher than that for the SRDF when $p = 20$, with a correspondingly large standard deviation which is greater than that for all the classification rules. When the actual error rate is examined by group, it is possible to determine how large the variation in error rate is between groups and also among the different (sampling) replications. For the larger dimensions ($p > 6$), the two highly ellipsoidal groups have a higher error rate than the spherical group on average, while for $p = 6$ there is little difference.

If large training samples are used (yielding better parameter estimates), the discrepancy between training sample and test sample error rate for the SRDF-M1 is still evident even though the error rates are smaller. Furthermore, experimental simulations were performed where the training sample was identical to the test sample, and it was found that the test sample error rate was still noticeably underestimated by the training sample minimizing cross-validated error rate. While the author of this thesis has not been able to ascertain fully why this phenomenon occurs, it is concluded that for conditions difficult for discrimination, such as these, the variation in the data is such that eigenvalue shrinkage is necessary in reducing variance as the dimension becomes large, and this leads to reduced error rates.

Two spherical and one ellipsoidal covariance matrix.

The final example considers the case where one covariance matrix is equal to the identity matrix, one is highly ellipsoidal, and the other is a multiple of the identity

Table 4.12: **Three unequal covariance matrices: one highly ellipsoidal, two spherical.** Group mean differences spread evenly over all subspaces: Average error rates with standard deviations.

	Dimension p		
	6	**10**	**20**
SRDF-M1	.04 (.03)	.04 (.04)	.03 (.04)
SRDF	.04 (.03)	.02 (.02)	.01 (.02)
SLDF	.13 (.04)	.11 (.04)	.06 (.03)
SQDF	.06 (.04)	.11 (.07)	.12 (.07)
SEDF	.18 (.03)	.15 (.03)	.11 (.03)
e_{cv}^{SRDF}	.01 (.01)	.00 (.01)	.00 (.00)
$\bar{\hat{\lambda}}^{SRDF}$	.14 (.14)	.23 (.16)	.44 (.18)
$\bar{\hat{\gamma}}^{SRDF}$	.68 (.38)	.74 (.34)	.70 (.33)
$e_{cv(1)}^{SRDF-M1}$	.00 (.01)	.00 (.00)	.00 (.00)
$e_{cv(2)}^{SRDF-M1}$	.04 (.04)	.01 (.03)	.00 (.01)
$e_{cv(3)}^{SRDF-M1}$	.01 (.03)	.01 (.03)	.00 (.01)
$\bar{\hat{\lambda}}_1$	.09 (.06)	.12 (.03)	.12 (.02)
$\bar{\hat{\lambda}}_2$	.09 (.11)	.12 (.06)	.13 (.03)
$\bar{\hat{\lambda}}_3$	.00 (.01)	.02 (.06)	.00 (.01)

matrix, where the multiplier is a scalar of moderately low magnitude. The non-zero group mean differences are spread evenly across all subspaces. Simulation results are shown in Table 4.12.

SRDF-M1 again performs well and is comparable to the SRDF for all dimensions. This is somewhat surprising since SRDF employs a high degree of eigenvalue shrinkage which does not result in a significantly lower error rate. It is clear that some form of regularisation is beneficial in that it reduces variance in the high dimensional settings. It is interesting to note that for all methods apart from the SQDF, the error rates reduce slightly as the dimension increases. Note that the SQDF is the only classifier which does not use any form of regularisation. The minimum cross-validated error rate for SRDF-M1 again underestimates the actual misclassification error assessed from the test sample, but all the error rates involved are very small.

In conclusion, from this supplementary study, we have compared the performance of SRDF against SRDF-M1. The conditions under which this comparison was made were designed to best use the flexibility that SRDF-M1 has, which is the potential to regularise each group-conditional covariance estimate separately. The SRDF-M1 classifier performed well in all situations, and was generally at least as good as the two established discriminant rules, SQDF and SLDF. If the model selection process that selects λ for a given sample of data is working well for SRDF-M1, then it is expected that the classifier should perform at least as well as either SQDF and SLDF. This indeed appears to be the case, assisted by the SRDF-M1 policy of minimum regularisation to break ties in the selection of λ. However, despite the good performance of SRDF-M1, on the whole it did not perform quite as well as the SRDF, particularly for large dimension, p. This again shows the benefit of permitting the use of eigenvalue shrinkage as in the SRDF classifier.

It could be expected that a classifier similar to SRDF-M1, but which also includes the eigenvalue shrinkage parameter γ, would perform slightly better than the SRDF. This would be consistent with á conjecture that if the number of regularisation parameters in a model is increased, the model will usually do better.

Computational considerations

The approximate computation times in CPU seconds for 100 repetitions of the sampling experiment described in this section are given in Table 4.13 for various p. These are the times required by the SRDF and SRDF-M1 rules to perform the simulations using MATLAB™ on a SUN Sparcstation ELC. Also given is the ratio (SRDF-M to SRDF) of CPU time needed to complete 100 simulations for those two regularised discriminant rules.

Table 4.13: Comparison of computation times between SRDF-M1 and SRDF.

	Dimension p		
	6	**10**	**20**
CPU time in seconds (SRDF)	1699	2864	12546
CPU time in seconds (SRDF-M1)	446	685	2586
SRDF-M1/SRDF	0.26	0.24	0.21

4.4 PERFORMANCE OF THE REGULARISED DISCRIMINANT FUNCTION IN TERMS OF THE SAMPLE SIZE TO DIMENSION RATIO.

4.4.1 Simulation study

From the study by Friedman (1989), as well as those in the previous sections, it is clear that the SRDF has proved itself at least equal to but usually superior to the other classification rules under a fairly wide range of situations. This superiority is greatest in the higher dimensional settings ($p > 10$). The comparisons with the SQDF, SLDF and in particular SRDF-M1 (in Section 4.3) indicate that the advantage the SRDF has over the other classification rules is a result of allowing for (γ) regularisation (or shrinkage) of the covariance matrix eigenvalues to equality.

The ratio of training sample size from each population, n_k $(1 \leq k \leq K)$, to the dimensionality p in the previous studies (Sections 3.5 through 4.3) was between 1.4 (for large p) and 2.2 (for smaller p). It is of interest to investigate the performance of the SRDF relative to the other classification rules over a wider range of n_k/p ratios. As mentioned earlier, the motivation for this is that presumably regularisation of the covariance matrix eigenvalues would no longer be advantageous for discrimination once the training sample size increases past some point sufficiently larger than p (see also Lawoko and Koolaard (1996) and Koolaard, Ganesalingam and Lawoko (1996)). The question addressed in this section is: to what extent do the benefits of covariance matrix regularisation (in particular eigenvalue shrinkage) diminish as the sample size to dimensionality ratio increases?

A further simulation study was implemented in the manner of the previous sections, and using the same six simulation conditions determined by assigning various settings of the population means and covariances to certain values (See Section 3.5). The discriminant rules compared with SRDF were SQDF, SLDF, SRDF-M1 and SEDF. The samples from each population are taken to be of equal size, so let $n = n_k (k = 1, \ldots, K)$ The various n/p ratios employed are 1.2, 1.5, 2, 3, 5, 10 for dimensions $p = 6, 10$ and 20. Again, the inordinate amount of computation time required precluded implementation of simulations for $p > 20$. In all cases there are three populations or groups involved. The (λ, γ) grid of values for use in the model selection procedure of the SRDF is defined by the outer product

of $\lambda = (0, .25, .5, .75, 1)$ and $\gamma = (0, .25, .5, .75, 1)$. The entire training sample is $3n$ in each case, since in all cases there are three groups. The test sample is 200. Fifty replications of each experiment were performed. Average error rate (with standard deviation in brackets) are given for each classification rule. The results are given in Tables 4.14 to 4.19. Graphical displays of the various classifier error rates for increasing n/p ratio are given in Figures 4.1 to 4.6.

The object of examination in this study is the eigenvalue regularisation technique as employed by the SRDF, hence the main interest is in comparing the SRDF with the methods which do not use this technique. These are the SLDF, SQDF and in particular SRDF-M1. Thus while the SEDF is included in the results, it involves maximum γ-regularisation and hence comparing its performance to that of the SRDF is less relevant to the issue being investigated here.

4.4.2 Simulation results

Equal and spherical group covariances (Table 4.14 and Figure 4.1)

The use of the γ parameter appears to enhance the classification process under conditions of equal, spherical group covariances only for small sample size to dimension ratios ($n/p < 3$). For larger ratios the advantage that the SRDF commands over the SLDF and SRDF-M1 diminishes to nothing for all dimensions. It is observed that for smaller dimensions a high degree of eigenvalue regularisation to equality is maintained in the regularisation process for all n/p ratios, as evidenced by the high value of $\bar{\hat{\gamma}}$. This is to be expected since the optimum value of γ in these conditions is one, as for the SEDF. For the higher dimensional settings, the $\bar{\hat{\gamma}}$ value is somewhat lower, yet it does not change with the ratio n/p.

All the discriminant rules give decreased error rates as the n/p ratio increases, with most significant change occuring for the SQDF, as expected. The SQDF is most sensitive to poor parameter estimates, and as the sample size increases its performance tends to improve quickly due to better parameter estimates. The performance of the SRDF improves slightly as the n/p ratio increases.

Unequal spherical covariances (Table 4.15 and Figure 4.2)

In conditions where the group covariances are not equal but are of spherical structure, the SRDF is the superior discriminant rule at all n/p ratios and dimensions

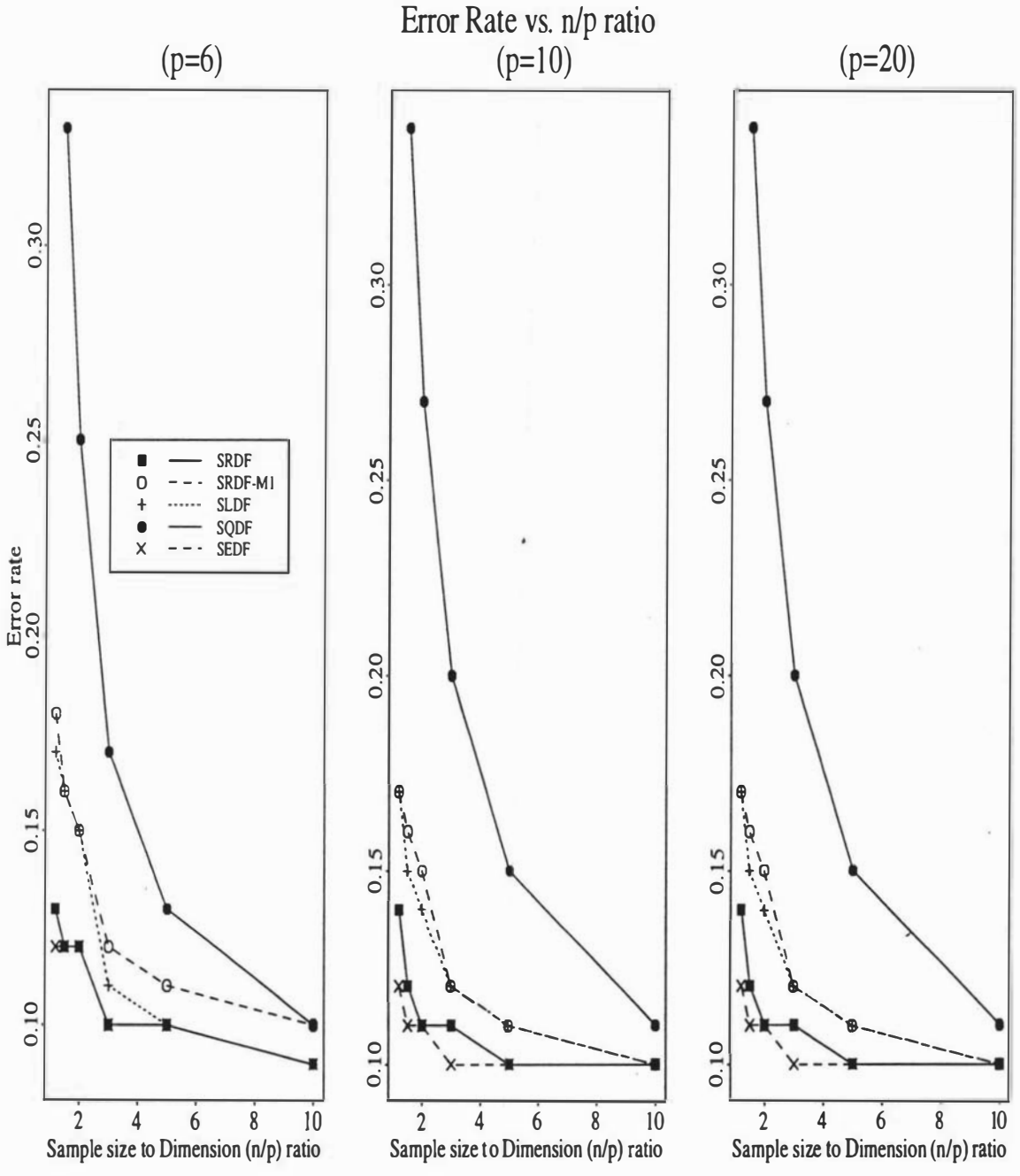


Figure 4.1: **Equal spherical population covariance matrices.** Classifier Error Rate vs. n/p ratio.

Table 4.14: **Equal Spherical Covariance Matrices.** Average error rates with standard deviations over a range of n/p ratios.

	n to p ratio					
$p = 6$	*1.2:1*	*1.5:1*	*2:1*	*3:1*	*5:1*	*10:1*
SRDF	.13 (.03)	.12 (.02)	.12 (.03)	.10 (.02)	.10 (.02)	.09 (.02)
SRDF-M1	.18 (.07)	.16 (.05)	.15 (.03)	.12 (.03)	.11 (.02)	.10 (.02)
SLDF	.17 (.05)	.16 (.04)	.15 (.03)	.11 (.03)	.10 (.02)	.09 (.02)
SQDF	.47 (.09)	.33 (.07)	.25 (.05)	.17 (.03)	.13 (.02)	.10 (.02)
SEDF	.12 (.02)	.12 (.02)	.12 (.02)	.10 (.02)	.10 (.02)	.09 (.02)
$p = 10$						
SRDF	.14 (.03)	.12 (.03)	.11(.02)	.11 (.02)	.10 (.02)	.10 (.02)
SRDF-M1	.17 (.05)	.16 (.04)	.15 (.04)	.12 (.04)	.11 (.03)	.10 (.02)
SLDF	.17 (.04)	.15 (.04)	.14 (.03)	.12 (.03)	.11 (.03)	.10 (.02)
SQDF	.46 (.07)	.34 (.07)	.27 (.05)	.20 (.05)	.15 (.03)	.11 (.02)
SEDF	.12 (.03)	.11 (.03)	.11 (.02)	.10 (.03)	.10 (.02)	.10 (.02)
$p = 20$						
SRDF	.13 (.03)	.12 (.02)	.11 (.02)	.11 (.02)	.18 (.04)	.13 (.02)
SRDF-M1	.18 (.04)	.15 (.03)	.14 (.02)	.12 (.02)	.11 (.03)	.10 (.02)
SLDF	.17 (.03)	.15 (.03)	.14 (.03)	.12 (.02)	.11 (.02)	.10 (.02)
SQDF	.49 (.06)	.39 (.06)	.32 (.04)	.24 (.03)	.17 (.03)	.13 (.02)
SEDF	.12 (.02)	.11 (.02)	.11 (.02)	.10 (.02)	.10 (.02)	.10 (.02)

used. Once again the average $\hat{\gamma}$ value used in the SRDF is high for most n/p ratios, however there are one or two aberrations. The effect of γ regularisation in these conditions is evident for all n/p ratios considered, but appears to begin to abate at $n/p = 10$. However, at this point the performance of SRDF-M1 only approaches that of the SRDF.

Equal, highly ellipsoidal covariances (Tables 4.16, 4.17 and Figures 4.3, 4.4)

As concluded earlier, there appears to be little advantage in eigenvalue shrinkage under these conditions where the group covariances are equal and of a highly ellipsoidal nature, and the group mean differences are concentrated in the low variance subspace. The performances of the SLDF and SRDF-M1 are comparable to that of the SRDF for all n/p ratios and for all dimensions studied. As eigenvalue shrinkage (towards equality) would be strongly counterproductive in these circumstances (since it would increase the variance in the low variance subspace), it is not surprising that the SRDF generally selects very low $\hat{\gamma}$ values (close to zero) for all p, and especially as the n/p ratio increases.

Table 4.15: **Unequal Spherical Covariance Matrices.** Average error rates with standard deviations over a range of n/p ratios.

	n to p ratio					
$p = 6$	*1.2:1*	*1.5:1*	*2:1*	*3:1*	*5:1*	*10:1*
SRDF	.22 (.04)	.20 (.04)	.20 (.03)	.17 (.03)	.17 (.03)	.16 (.03)
SRDF-M1	.31 (.07)	.27 (.07)	.26 (.05)	.22 (.04)	.19 (.03)	.17 (.02)
SLDF	.30 (.06)	.26 (.06)	.25 (.02)	.21 (.03)	.20 (.03)	.18 (.02)
SQDF	.53 (.07)	.43 (.07)	.34 (.06)	.25 (.04)	.19 (.04)	.17 (.02)
SEDF	.23 (.04)	.22 (.04)	.22 (.03)	.20 (.03)	.19 (.03)	.18 (.02)
$p = 10$						
SRDF	.20 (.05)	.17 (.04)	.15 (.03)	.15 (.03)	.13 (.03)	.10 (.03)
SRDF-M1	.28 (.05)	.28 (.05)	.27 (.06)	.20 (.04)	.18 (.03)	.14 (.03)
SLDF	.28 (.05)	.26 (.05)	.26 (.04)	.22 (.04)	.21 (.03)	.18 (.03)
SQDF	.52 (.07)	.43 (.06)	.35 (.05)	.25 (.05)	.19 (.03)	.14 (.03)
SEDF	.24 (.04)	.22 (.03)	.21 (.03)	.20 (.03)	.20 (.03)	.18 (.03)
$p = 20$						
SRDF	.13 (.03)	.12 (.02)	.10 (.02)	.19 (.03)	.12 (.02)	.09 (.02)
SRDF-M1	.29 (.07)	.30 (.08)	.26 (.06)	.18 (.04)	.14 (.03)	.13 (.02)
SLDF	.28 (.03)	.26 (.04)	.24 (.03)	.21 (.03)	.20 (.03)	.19 (.02)
SQDF	.55 (.04)	.47 (.06)	.37 (.04)	.27 (.04)	.18 (.03)	.12 (.02)
SEDF	.23 (.03)	.22 (.03)	.21 (.03)	.20 (.02)	.19 (.03)	.18 (.02)

When the group mean differences are concentrated in the high variance subspaces the SRDF is superior to the other discriminant rules (SEDF excepted) for $n/p < 3$. Beyond $n/p = 3$, SRDF-M1 and the SLDF discriminate as well as the SRDF. The benefits of employing the γ parameter disappear at $n/p = 3$ or more as the other rules improve in their performance at a faster rate than the SRDF, as the sample size increases.

The average $\hat{\gamma}$ value of the SRDF decreases as the sample size increases in relation to p. (From $\bar{\hat{\gamma}} \approx 0.75$ for $n/p = 1.2$ to $\bar{\hat{\gamma}} \approx 0.25$ for $n/p = 10$). Also, a lesser degree eigenvalue shrinkage is employed by the SRDF for larger values of p.

Unequal, highly ellipsoidal group covariances (Tables 4.18,4.19 and Figures 4.5, 4.6)

In the situation where the group means are equal, the SQDF, which represents no covariance regularisation, performs generally well as would be expected. For small n/p ratios ($n/p < 2$) and larger dimensions $p \geq 10$, the SRDF's performance is superior to that of the SQDF, suggesting that eigenvalue shrinkage is not advantageous once the sample size becomes twice as large as p. The other discriminant

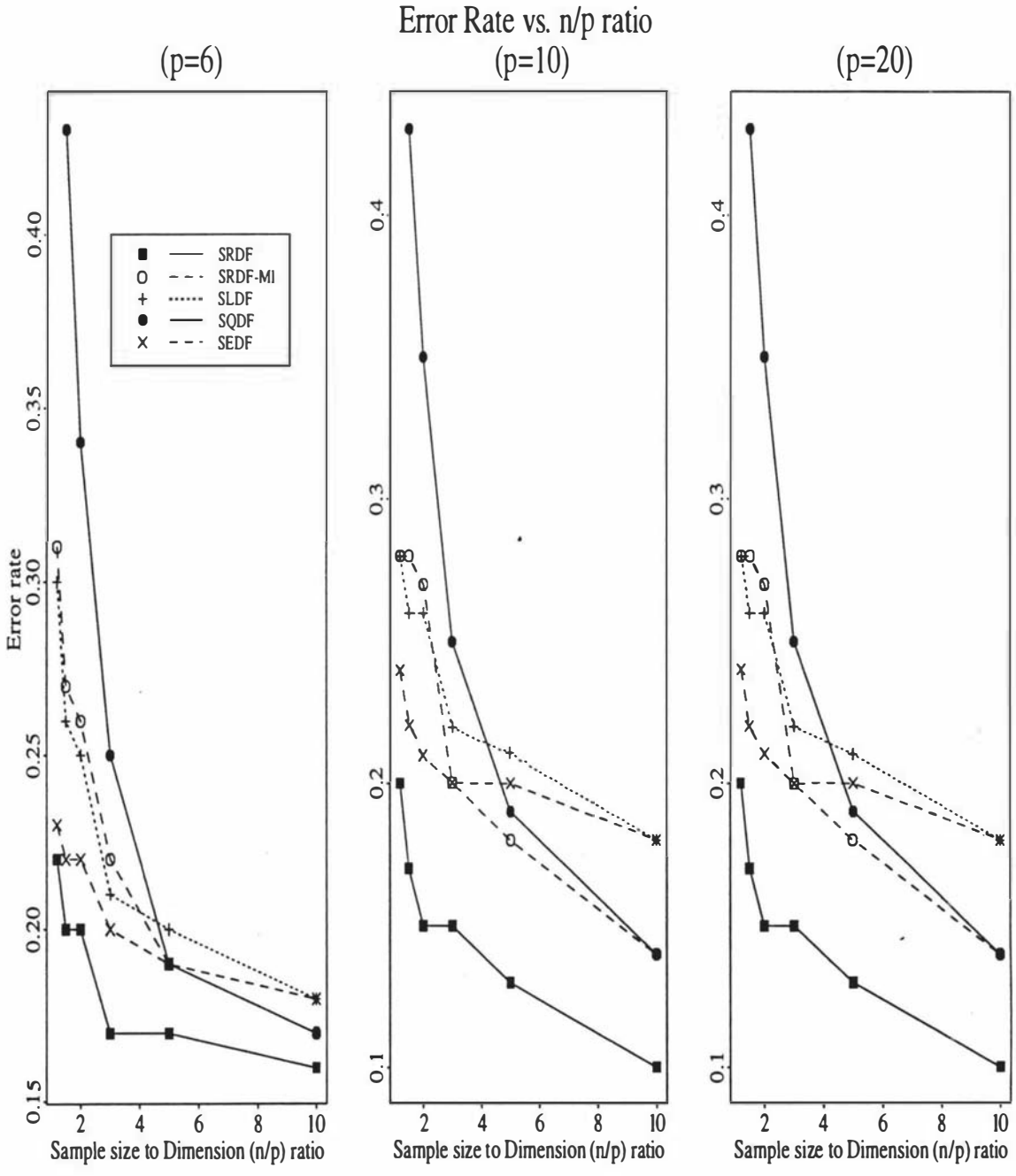


Figure 4.2: **Unequal spherical population covariance matrices.** Classifier Error Rate vs. n/p ratio.

Table 4.16: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in Low Variance Subspace.** Average error rates with standard deviations over a range of n/p ratios.

	n to p ratio					
$p = 6$	*1.2:1*	*1.5:1*	*2:1*	*3:1*	*5:1*	*10:1*
SRDF	.12 (.06)	.09 (.04)	.08 (.04)	.06 (.03)	.06 (.01)	.05 (.02)
SRDF-M1	.10 (.03)	.08 (.03)	.08 (.04)	.05 (.02)	.05 (.02)	.04 (.01)
SLDF	.10 (.03)	.08 (.02)	.07 (.02)	.05 (.02)	.05 (.02)	.04 (.01)
SQDF	.41 (.09)	.26 (.07)	.15 (.05)	.09 (.03)	.07 (.02)	.05 (.02)
SEDF	.28 (.05)	.27 (.06)	.26 (.05)	.22 (.04)	.21 (.04)	.20 (.03)
$p = 10$						
SRDF	.16 (.04)	.14 (.04)	.12 (.04)	.10 (.03)	.09 (.02)	.10 (.03)
SRDF-M1	.14 (.03)	.13 (.03)	.12 (.03)	.10 (.03)	.09 (.02)	.08 (.02)
SLDF	.14 (.03)	.12 (.03)	.11 (.02)	.09 (.03)	.09 (.02)	.08 (.02)
SQDF	.44 (.09)	.31 (.06)	.24 (.05)	.17 (.04)	.12 (.02)	.09 (.02)
SEDF	.32 (.05)	.30 (.05)	.28 (.04)	.26 (.04)	.24 (.03)	.23 (.03)
$p = 20$						
SRDF	.18 (.04)	.16 (.03)	.14 (.03)	.13 (.02)	.11 (.02)	.11 (.02)
SRDF-M1	.17 (.03)	.17 (.03)	.15 (.02)	.13 (.02)	.11 (.02)	.11 (.02)
SLDF	.17 (.03)	.16 (.02)	.14 (.02)	.12 (.02)	.11 (.02)	.11 (.02)
SQDF	.49 (.06)	.39 (.04)	.32 (.04)	.24 (.04)	.18 (.03)	.14 (.02)
SEDF	.33 (.04)	.32 (.04)	.30 (.04)	.27 (.04)	.26 (.04)	.24 (.03)

Table 4.17: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences in High Variance Subspace.** Average error rates with standard deviations over a range of n/p ratios.

	n to p ratio					
$p = 6$	*1.2:1*	*1.5:1*	*2:1*	*3:1*	*5:1*	*10:1*
SRDF	.08 (.03)	.07 (.02)	.07 (.02)	.07 (.02)	.06 (.02)	.06 (.02)
SRDF-M1	.13 (.05)	.10 (.04)	.09 (.03)	.07 (.03)	.06 (.02)	.05 (.01)
SLDF	.12 (.03)	.10 (.03)	.09 (.03)	.07 (.02)	.06 (.02)	.05 (.01)
SQDF	.43 (.10)	.29 (.09)	.18 (.05)	.11 (.03)	.07 (.02)	.06 (.01)
SEDF	.07 (.02)	.07 (.02)	.07 (.02)	.07 (.02)	.06 (.02)	.06 (.02)
$p = 10$						
SRDF	.10 (.03)	.10 (.03)	.10 (.02)	.10 (.02)	.08 (.02)	.11 (.04)
SRDF-M1	.15 (.03)	.13 (.03)	.12 (.04)	.10 (.02)	.09 (.02)	.08 (.02)
SLDF	.14 (.03)	.13 (.03)	.11 (.03)	.10 (.02)	.08 (.02)	.08 (.02)
SQDF	.45 (.07)	.32 (.06)	.23 (.05)	.16 (.04)	.12 (.03)	.09 (.02)
SEDF	.10 (.02)	.10 (.02)	.09 (.03)	.10 (.02)	.10 (.02)	.09 (.02)
$p = 20$						
SRDF	.12 (.03)	.12 (.02)	.10 (.02)	11 (.02)	.09 (.02)	.09 (.02)
SRDF-M1	.16 (.03)	.16 (.04)	.13 (.03)	.12 (.03)	.10 (.02)	.10 (.03)
SLDF	.16 (.03)	.15 (.04)	.13 (.02)	.12 (.03)	.10 (.02)	.13 (.03)
SQDF	.48 (.05)	.39 (.04)	.30 (.04)	.22 (.03)	.16 (.02)	.10 (.03)
SEDF	.12 (.02)	.12 (.03)	.11 (.03)	.11 (.03)	.10 (.02)	.11 (.03)

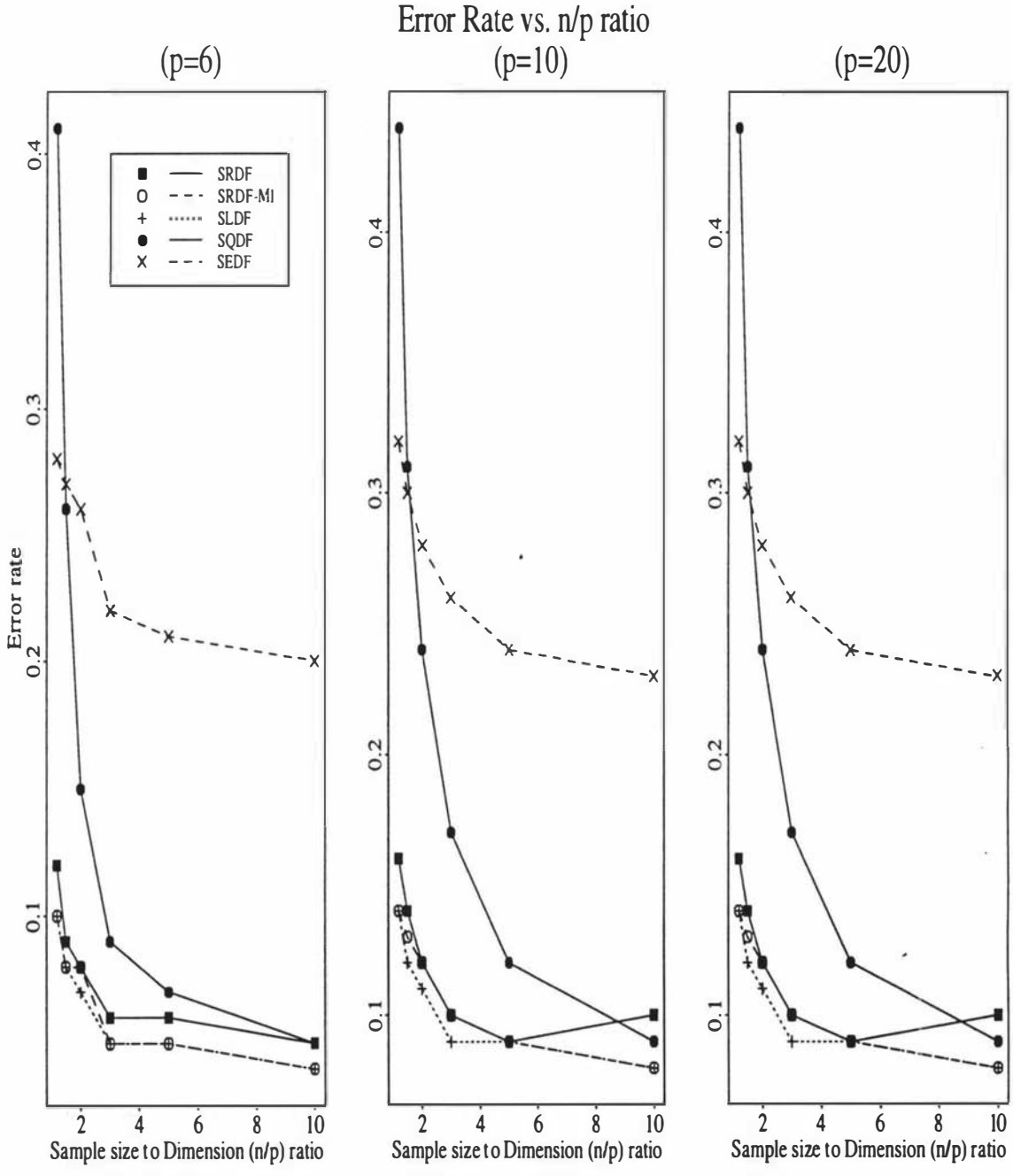


Figure 4.3: **Equal, highly ellipsoidal population covariance matrices.** Population mean differences concentrated in the low variance subspace. Classifier Error Rate vs. n/p ratio.

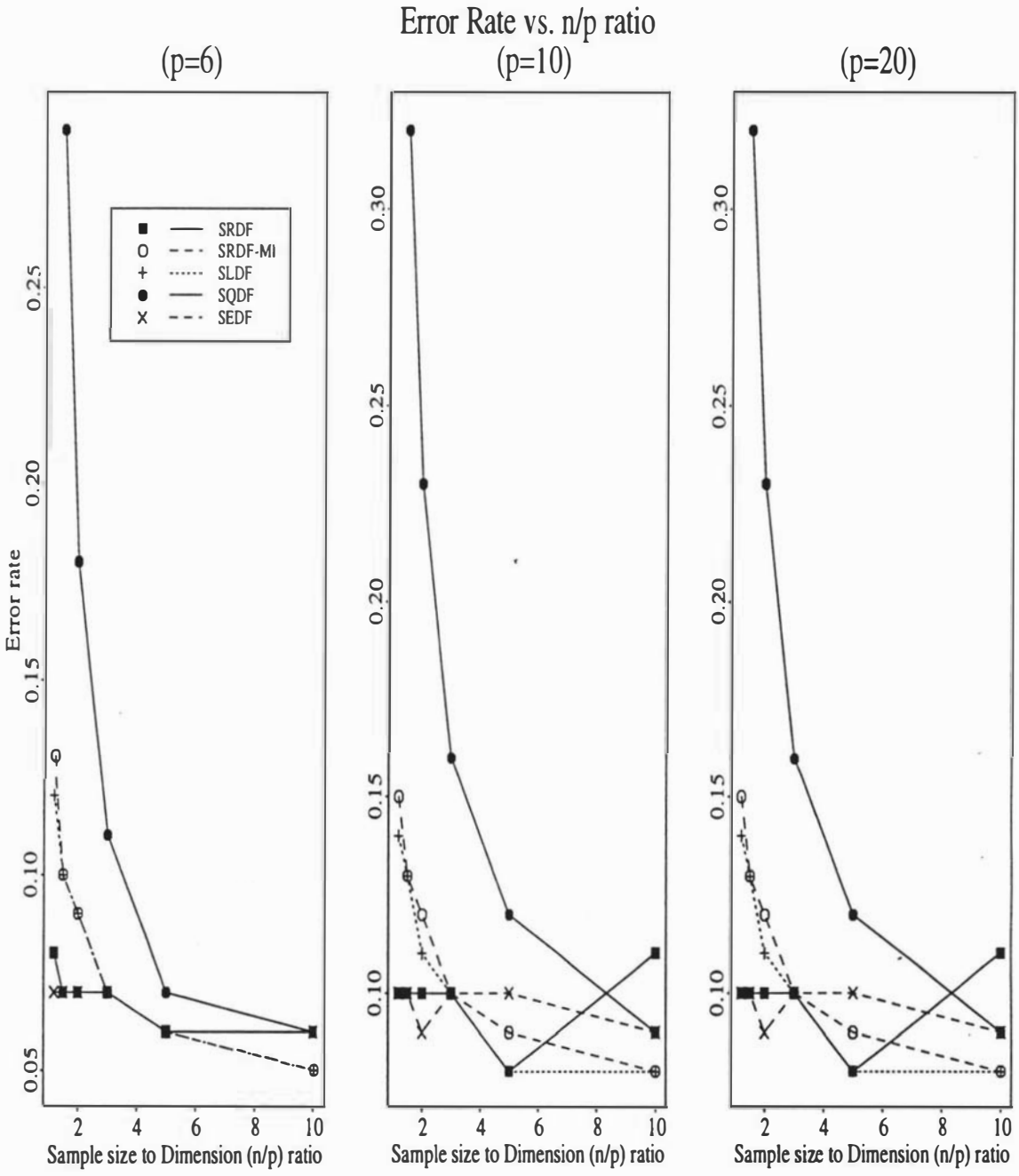


Figure 4.4: **Equal, highly ellipsoidal population covariance matrices.** Population mean differences concentrated in the high variance subspace. Classifier Error Rate vs. n/p ratio.

rules perform poorly in relation to these two, since any regularisation to the pooled covariance is strongly counterproductive.

Under these conditions, the n/p ratio at which the performance of SRDF-M1 approaches that of the SRDF is approximately $n/p = 2$: slightly smaller for small p and slightly larger for large p. The average $\hat{\gamma}$ value used in the SRDF is usually small, but there is substantial variation. This indicates that under these difficult discrimination conditions and substantial variance in the data, selection of $\hat{\gamma}$ is very sensitive to the particular training sample data at hand.

Once again, the SRDF error rate improves as the sample size to dimension ratio increases, although for $p \geq 10$ the reduction in error rate is not significant for $n/p > 3$. Even at $n/p = 10$, the performance of SRDF-M1 does not compare with that of the SRDF, indicating that regularisation to the pooled covariance alone, does not help the classification process under these conditions .

For the case where the group means differences are non-zero (but still unequal, highly ellipsoidal group covariances), the relative performance of the various rules remain the same as in the situation of equal group means above. The rules all yield lower error rates, since the groups now differ in location. For the SRDF, once n/p is greater than 3, there is no real reduction in error rate. On the other hand, the SRDF-M1 error rate decreases with increasing n/p until at $n/p = 10$, the two error rates are nearly equal. The SRDF is superior to the SQDF in these conditions only at the smallest sample size to dimension ratio, $n/p = 1.2$. Thus this is the n/p ratio beyond (that is, larger than) which regularising the covariance matrix eigenvalues towards equality no longer appears to be beneficial. The average γ value, $\bar{\hat{\gamma}}$, for the SRDF decreases as n/p increases. At $n/p = 10$, $\bar{\hat{\gamma}}$ is close to zero, which is the appropriate level given that the parameter estimates are good.

In conclusion, this simulation study underlines the usefulness of the eigenvalue shrinkage technique as employed in regularised discriminant analysis. The advantage that it commands over the other classification rules is strongest when the training sample size from each group is small in relation to the dimensionality, p. Furthermore, often that advantage remains even when the sample size increases to several times that of the dimensionality.

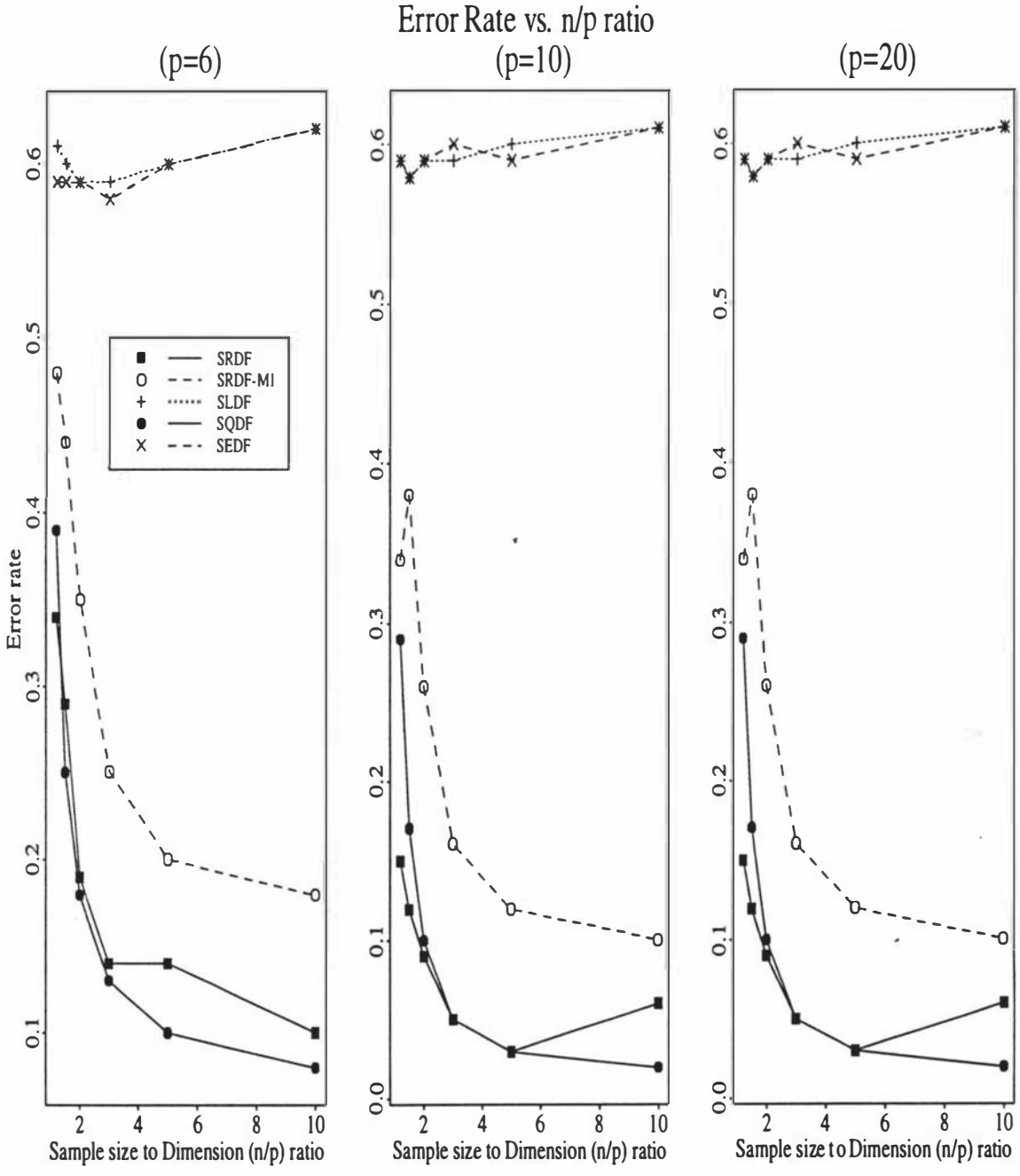


Figure 4.5: **Unequal, highly ellipsoidal population covariance matrices.** Population means equal. Classifier Error Rate vs. n/p ratio.

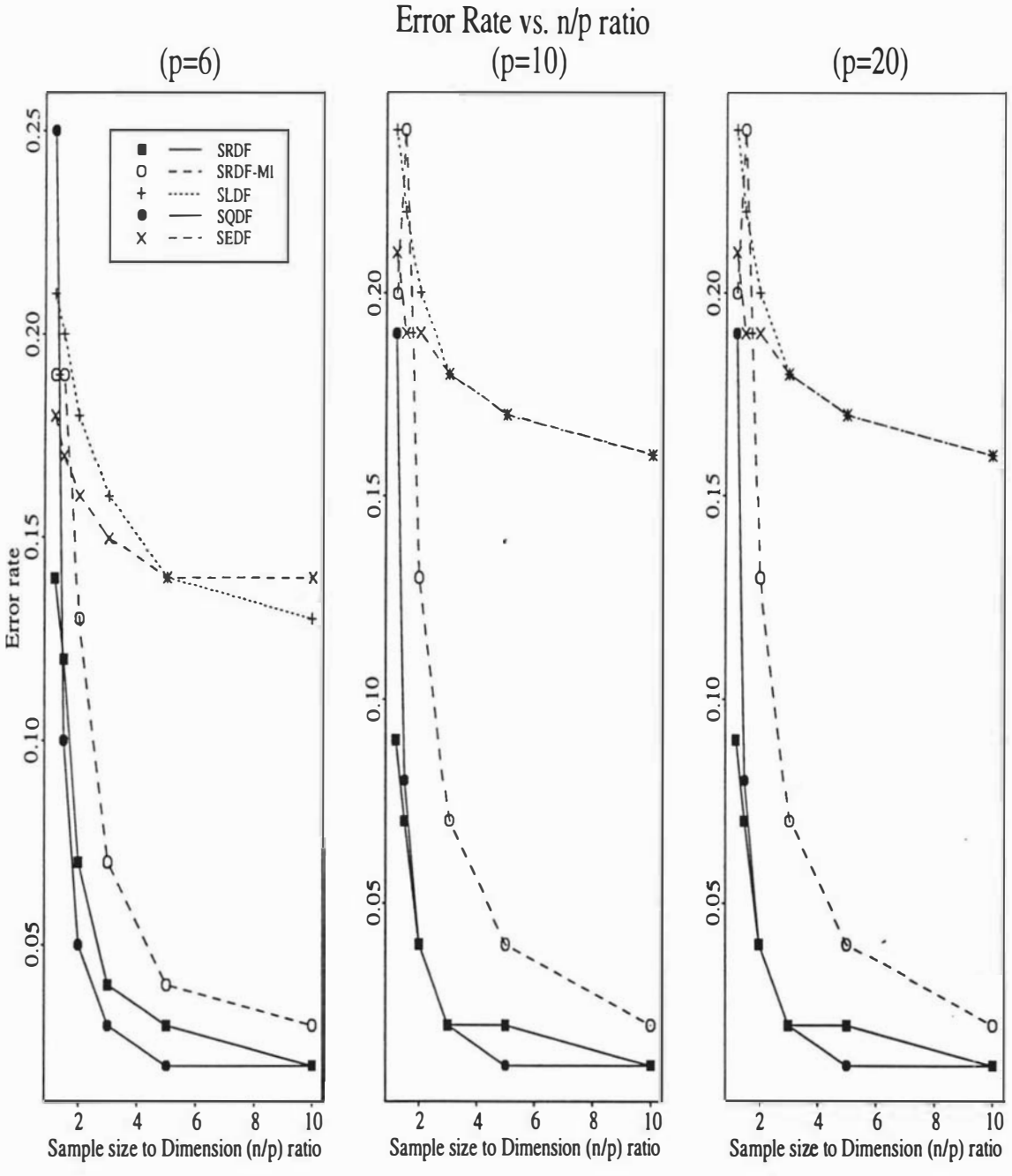


Figure 4.6: **Unequal, highly ellipsoidal population covariance matrices.** Population means unequal. Classifier Error Rate vs. n/p ratio.

Table 4.18: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences.** Average error rates with standard deviations over a range of n/p ratios.

	n to p ratio					
$p = 6$	*1.2:1*	*1.5:1*	*2:1*	*3:1*	*5:1*	*10:1*
SRDF	.34 (.11)	.29 (.07)	.19 (.05)	.14 (.04)	.14 (.07)	.10 (.04)
SRDF-M1	.48 (.09)	.44 (.08)	.35 (.09)	.25 (.05)	.20 (.07)	.18 (.06)
SLDF	.61 (.05)	.60 (.06)	.59 (.05)	.59 (.05)	.60 (.05)	.62 (.04)
SQDF	.39 (.09)	.25 (.06)	.18 (.04)	.13 (.03)	.10 (.02)	.08 (.02)
SEDF	.59 (.04)	.59 (.05)	.59 (.06)	.58 (.05)	.60 (.05)	.62 (.05)
$p = 10$						
SRDF	.15 (.06)	.12 (.04)	.09 (.03)	.05 (.02)	.03 (.02)	.06 (.03)
SRDF-M1	.34 (.10)	.38 (.09)	.26 (.08)	.16 (.05)	.12 (.05)	.10 (.04)
SLDF	.59 (.04)	.58 (.04)	.59 (.04)	.59 (.04)	.60 (.04)	.61 (.04)
SQDF	.29 (.09)	.17 (.06)	.10 (.03)	.05 (.02)	.03 (.01)	.02 (.01)
SEDF	.59 (.04)	.58 (.04)	.59 (.04)	.60 (.04)	.59 (.04)	.61 (.04)
$p = 20$						
SRDF	.03 (.02)	.02 (.02)	.02 (.02)	.01 (.01)	.00 (.01)	.00 (.00)
SRDF-M1	.40 (.18)	.36 (.11)	.20 (.07)	.11 (.03)	.06 (.03)	.05 (.02)
SLDF	.58 (.04)	.57 (.04)	.59 (.05)	.61 (.03)	.61 (.03)	.62 (.04)
SQDF	.20 (.07)	.10 (.03)	.04 (.02)	.01 (.01)	.00 (.00)	.00 (.00)
SEDF	.57 (.03)	.59 (.04)	.59 (.04)	.60 (.04)	.61 (.03)	.61 (.04)

Table 4.19: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences.** Average error rates with standard deviations over a range of n/p ratios.

	n to p ratio					
$p = 6$	*1.2:1*	*1.5:1*	*2:1*	*3:1*	*5:1*	*10:1*
SRDF	.14 (.04)	.12 (.04)	.07 (.03)	.04 (.03)	.03 (.02)	.02 (.01)
SRDF-M1	.19 (.06)	.19 (.07)	.13 (.05)	.07 (.03)	.04 (.02)	.03 (.01)
SLDF	.21 (.05)	.20 (.05)	.18 (.04)	.16 (.03)	.14 (.03)	.13 (.03)
SQDF	.25 (.12)	.10 (.06)	.05 (.02)	.03 (.01)	.02 (.01)	.02 (.01)
SEDF	.18 (.04)	.17 (.04)	.16 (.03)	.15 (.03)	.14 (.03)	.14 (.03)
$p = 10$						
SRDF	.09 (.05)	.07 (.03)	.04 (.03)	.02 (.01)	.02 (.01)	.01 (.01)
SRDF-M1	.20 (.07)	.24 (.09)	.13 (.05)	.07 (.03)	.04 (.02)	.02 (.01)
SLDF	.24 (.04)	.22 (.04)	.20 (.04)	.18 (.01)	.17 (.03)	.16 (.03)
SQDF	.19 (.10)	.08 (.05)	.04 (.02)	.02 (.01)	.01 (.01)	.01 (.01)
SEDF	.21 (.04)	.19 (.03)	.19 (.03)	.18 (.02)	.17 (.03)	.16 (.03)
$p = 20$						
SRDF	.03(.02)	.02 (.02)	.01 (.01)	.00 (.00)	.00 (.01)	.00 (.00)
SRDF-M1	.29 (.17)	.25 (.11)	.11 (.04)	.04 (.02)	.02 (.01)	.01 (.01)
SLDF	.22 (.04)	.20 (.03)	.18 (.03)	.17 (.03)	.15 (.02)	.15 (.03)
SQDF	.14 (.06)	.05 (.03)	.01 (.01)	.00 (.00)	.00 (.00)	.00 (.00)
SEDF	.18 (.03)	.17 (.03)	.17 (.03)	.16 (.02)	.15 (.02)	.14 (.03)

Chapter 5

MODEL SELECTION OF REGULARISATION PARAMETERS USING BHATTACHARYYA DISTANCE

5.1 INTRODUCTION

In the previous chapters, the advantage of the regularised discriminant model of Friedman has been demonstrated, particularly if the sample size is small in relation to the size of the dimension. The major reason for its success in many conditions that are difficult for discrimination stems from the rule's flexibility in allowing for eigenvalue regularisation towards equality in the sample covariance matrices. Since the importance of the parameter γ has been established, it will be maintained in subsequent models for discrimination that appear in this thesis.

Several potential weaknesses in the model selection procedure of the SRDF as developed by Friedman (1989) were noted by Rayens and Greene (1991). These included (i) the fact that the regularisation parameters were often determined by a small fraction of the data points available, and (ii) that in many instances (especially with smaller sample sizes) there will not be a unique choice of the parameters (λ, γ) for the model. These problems were discussed and studied in Chapters 3 and 4. Furthermore, despite the development of computationally efficient algorithms to enhance the attractiveness of what is inherently a computationally intensive model, the computation time is still rather high from the author's experience using MATLAB™ on a SUN Sparcstation ELC computer. Therefore it is of interest to explore other ways of arriving at appropriate regularisation parameter values

in place of minimising the cross-validated error rate at a range of points over the (λ, γ) grid. Because of the computational burden inherent in SRDF, and with regard to criticisms of the technique by Rayens and Greene (1991), it is investigated here whether information about appropriate values for the two regularisation parameters could be obtained by examining the behaviour of the Bhattacharyya distance (Bhattacharyya (1946)) between the various populations. Note that any determination of the optimal values of λ and γ (i.e. $\hat{\lambda}$ and $\hat{\gamma}$) from the data using the Bhattacharyya distance involves use of *all* the data points. A classification rule which uses regularisation parameters obtained from the Bhattacharyya distance is presented for the case of two populations or groups, and is compared via simulation with the original SRDF. An extension to the three group case is presented in Subsection 5.2.5, and its performance is also examined against the other rules. If this rule is to perform comparably to the SRDF in terms of its error rate, its model selection procedure must perform correctly in terms of selecting an appropriate degree of regularisation for a given situation. For example, if the populations are of similar shape and size (in terms of the magnitude of their variances), the covariance mixing parameter λ should be set to a reasonably high value. The Bhattacharyya distance is found to give information which leads to appropriate values of λ and γ being selected in general. The rule presented is also computationally much faster than Friedman's SRDF since it avoids re-sampling methods.

In Section 5.4, the various rules are compared in terms of their performances in correctly classifying observations from several real data sets. The results of the simulation studies and the case studies with real data sets show that the rule employing the Bhattacharyya distance in the model selection procedure generally performs as well as Friedman's SRDF.

5.2 CONSTRUCTION OF A MODEL SELECTION PROCEDURE BASED ON THE BHATTACHARYYA DISTANCE

5.2.1 Distance measures and their applications in discrimination

Distance measures have often been considered as alternatives to error rates in certain aspects of discriminant analysis. For example, Jain (1976) investigated the behaviour of an estimate of the Bhattacharyya distance when used as a criterion in variable selection. It was shown that the bias and variance of the estimate is related to the number of training samples and parameter values of the distribution. Kailath (1967) addressed the problem that minimising the error rate to determine optimum classification can be difficult to accomplish in practice. He investigated the idea of using simpler, albeit sub-optimal performance measures instead of the error rate, and compared the Bhattacharyya distance with an often-used measure, the divergence, which is closely related to Shannon's logarithmic measure of information. Not only is the Bhattacharyya distance easier to evaluate than the divergence, but in some examples in the study it was found to perform at least as well as the divergence in minimising the probability of misclassification. Kailath obtained an upper bound on the probability of misclassification in terms of the Bhattacharyya distance in the case of equal prior probabilities of the distributions. Note that Kailath only treated the case of two groups. Also, all his work assumed knowledge of the parameters, whereas, as we shall see later, if one has to use sample estimates of the parameters, the link between Bhattacharyya distance and error rate is much less clear. Also, Fukunaga and Hayes (1989) obtained an upper bound, in terms of the Bhattacharyya distance, on the Bayes error for classifying between two Gaussian distributions .

5.2.2 The Bhattacharyya distance

The Bhattacharyya distance between two multivariate normal density functions with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ is

$$B = B1 + B2 \tag{5.1}$$

where

$$B1 = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{5.2}$$

and

$$B2 = \frac{1}{2} \ln \left(\frac{\left| \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right|}{|\boldsymbol{\Sigma}_1|^{1/2} |\boldsymbol{\Sigma}_2|^{1/2}} \right). \tag{5.3}$$

The first term of the expression, $B1$, is similar to the well-known Mahalanobis distance between the densities. It measures the distance between the two distributions caused by the mean shift. The second term $B2$ utilises the determinants of the two covariance matrices as well as that of the average group covariance matrix. It gives a measure of the difference between the two distributions due to the covariance shift.

Fukunaga and Hayes (1989), in an extensive mathematical development, derived asymptotic expressions for the expected bias and variance of the sample estimates ($\widehat{B1}$ and $\widehat{B2}$) of terms $B1$ and $B2$, and showed that the bias of $\widehat{B1}$ is proportional to p/n (for $n_i = n$, $i = 1, 2, \ldots$), where n_i is the size of the sample taken from group i. They also showed that the bias of $\widehat{B2}$ is proportional to $(p+1)p/n$. In other words, estimates of the Bhattacharyya distance measure become increasingly biased as the ratio p/n increases, with $\widehat{B2}$ more seriously affected than $\widehat{B1}$. Thus in high dimensional space the bias present in the Bhattacharyya distance estimate is dominated by the bias inherent in estimation of term $B2$. They also showed that as the dimensionality increases, an increasingly large ratio of n/p is needed to maintain a constant expected value of B.

With the above knowledge of the Bhattacharyya distance function between two Gaussian distributions, it is plausible to expect that some degree of regularisation of the covariance, such as is provided for by the two-parameter model in expression (3.7), would improve the estimation of the Bhattacharyya distance. The reason for this stems from the accepted knowledge that covariance estimates based on expression (1.10) yield eigenvalue estimates which are biased. The largest ones are biased towards values which are too high, and the smallest ones are biased towards values which are too low. This bias will be worse in the situation where the true population eigenvalues are approximately equal, but in all cases this bias becomes more pronounced as the ratio of sample size to dimension decreases. The term $B2$ of the Bhattacharyya distance is most vulnerable to such bias occurring, being a ratio of determinants of sample covariance estimates, and regularisation of the

eigenvalues towards equality ought to prove useful in counteracting bias-induced anomalies in estimates of $B2$, particularly as p becomes large.

5.2.3 Behaviour of Bhattacharyya distance with regularised covariances

Kailath (1967) admitted that it would be hoping for too much, to expect a strong relationship between distance measures and error rates. Nevertheless, the author was able to obtain several useful theoretical results linking the two, assuming known population parameters. In the present covariance regularisation context with two parameters controlling the degree of shrinkage, as in expression (3.7), it would be too optimistic to expect that the $(\hat{\lambda}, \hat{\gamma})$ combination which maximises the Bhattacharyya distance for a given set of data would also yield a classification rule which minimises the future error rate. Instead, from the example (Table 5.1) below, we can often detect no such relationship between the sample Bhattacharyya distance and minimum error rate. Table 5.1 shows the values of the components $(\widehat{B1}, \widehat{B2})$ of the sample Bhattacharyya distance at a range of points over the (λ, γ) grid. The cross-validated error rate (e_{cv}) at each point is also stated to give an indication of the range within which the minimum actual error rate lies. The data set consisted of samples of size 13 from each of two normal populations ($p = 6$) with equal, highly ellipsoidal covariance matrices and mean differences in the high variance subspace (Condition 4 - see Chapter 3, Section 3.5).

Table 5.1: Example of (λ, γ) grid of Bhattacharyya distance values ($e_{cv}(B1, B2)$)

$\gamma = 1$	*.08* (3.84,0.05)	*.08* (3.84,0.00)	*.08* (3.84,0.00)
$\gamma = 0.5$	*.04* (2.93,0.10)	*.04* (2.93,0.01)	*.04* (2.93,0.00)
$\gamma = 0$	*.15* (2.73,0.59)	*.08* (2.73,0.05)	*.08* (2.73,0.00)
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$

It is evident from Table 5.1 (and other data sets) that the largest value of $B = B1 + B2$ for any given data set will always occur on the axis where $\lambda = 0$ on the (λ, γ) grid; i.e. suggesting no regularisation of the individual covariance matrices towards the average covariance. This is the case for samples from any two normal distributions. There are several reasons for this:

1. The value of $B1$ is not affected by λ when the regularised covariances are used. This is because the central component of $B1$ is

$$\left(\hat{\Sigma}_1(\lambda,\gamma)+\hat{\Sigma}_2(\lambda,\gamma)\right)/2,$$

and this is not affected by λ for a fixed value of γ.

2. The value of $B2$ decreases monotonically as λ increases, for fixed γ, since as λ approaches 1, the regularised covariances approach equality. When $\lambda = 1$, $\hat{\Sigma}_1(\lambda,\gamma)$ and $\hat{\Sigma}_2(\lambda,\gamma)$ are both equal to $(1-\gamma)\mathrm{S}_p + \gamma\left(\mathrm{tr}\left\{\mathrm{S}_p\right\}/p\right)\mathbf{I}$. In this case the numerator and denominator in the parenthesis in expression (5.3) are equal, and the term $B2$ becomes zero.

3. Term $B2$ is always non-negative since for two p-dimensional positive definite matrices, A and B,

$$|A|^{1/2}\,|B|^{1/2} < \left|\frac{A+B}{2}\right|.$$

4. The value of $B2$ decreases monotonically as γ increases from 0 to 1, for fixed λ. Since $B2$ is fundamentally a measure of the covariance shift between the two distributions and as the eigenvalues of the separate covariances are increasingly biased towards equality, the distributions become more similar in shape.

5.2.4 Model selection

As mentioned earlier, the simulations performed with Friedman's SRDF (and various modifications) in Chapters 3 and 4 have enabled us to observe that for a number of different simulation conditions, there is no unique combination of $\hat{\lambda}$ and $\hat{\gamma}$, using the criteria of minimum cross-validated error rate. Indeed, altering the rule for the breaking of such ties (Section 3.6) had little effect on the overall performance of the procedure. Thus it appears that the *degree* of regularisation (either covariance mixing or eigenvalue shrinkage, or both) is often not as important as its *presence* in any (roughly appropriate) form. It can therefore be conjectured that complex methods to obtain a precise selection of $\hat{\lambda}$ and $\hat{\gamma}$ are not warranted. A goal of the proposed model selection procedure using the Bhattacharyya distance is to provide a much faster algorithm to that proposed by Friedman using cross-validation. Also, the procedure should choose appropriate levels of the λ and γ parameters so that the classification rule obtained is comparable in performance to Friedman's SRDF.

Consequently, a relatively simple heuristic algorithm for selecting the values of λ and γ has been developed based on empirical data obtained from a number of repeated simulation experiments involving calculations of the quantities $B1$ and $B2$ over the (λ, γ) grid for a variety of conditions. A complex model selection procedure is not imperative since evidence was presented in Chapter 3, Section 3.6 that in most situations only very approximate values of λ and γ are required. Note, however, that γ is usually required to be estimated more accurately than λ. A dual optimisation (of λ and γ) approach is not possible here because of the behaviour of terms $B1$ and $B2$ outlined in points 1 - 4 in the previous section. Instead, the approach adopted is to first select one parameter, and then the other. Since both terms $B1$ and $B2$ exhibit similar behaviour in relation to λ for all values of γ, it is sensible to first choose a value for γ so as to narrow down the search area for λ on the (λ, γ) grid.

One conclusion from previous simulation studies (Chapter 4) is that as the sample size to dimension ratio decreases, an increasing degree of eigenvalue regularisation using γ (i.e. $\gamma > 0$) becomes necessary to counteract the bias in the estimated eigenvalues of the sample covariances. Also, an increasing amount of regularisation away from $\gamma = 0$ is required as p increases, even for those conditions where any shrinkage of the eigenvalues to equality would appear to be strongly counter-productive. See, for example, Chapter 4, Table 4.5 where the average $\hat{\gamma}$ value increases with dimension to substantial levels, even though no regularisation, or SQDF, would seem to be the best option in these conditions. The benefits of a decrease in variance from such regularisation has been shown to outweigh any introduced bias (see also Koolaard, Lawoko and Ganesalingam (1996)). The proposed method of selecting γ from the Bhattacharyya distance therefore only considers values of γ in the range $\theta \leq \gamma \leq 1$, where $\theta > 0$, but usually fairly close to zero, and where θ depends on both the magnitude of p and the sample size to dimensionality ratio.

Selection of the parameter γ

Increasing the value of the eigenvalue regularisation parameter γ typically decreases the term $B1$, but not always, and the trend is not always monotonic. However from point 4 above we see that $B2$ exhibits only monotonic behaviour in relation to γ. So it seems sensible to first look at the behaviour of $B1$ for a range of γ.

The bias inherent in the estimate of $B1$ would be expected to be less than that in estimates of $B2$, so that the principle upon which selection of the regularisation parameter γ is made involves giving $B1$ greater importance than $B2$. Thus, in general the aim is to choose that γ which gives a large or maximal value of $B1$ or $B1/B2$. From the behaviour of primarily $B1$, and secondarily $B2$, calculated for various γ over $\theta \leq \gamma \leq 1$, the following decision paths are proposed for the selection of an appropriate γ.

From empirical data we can identify three scenarios relating to $B1$. Note that all details, which define relative terms used here such as 'small' and 'large', are given in the algorithm in Appendix B:

I. *Magnitude of $B1$ small, and not greatly affected by the value of γ changing between θ and* 1. Under this scenario, $B1$ is not providing much information as to an appropriate value of γ, so look at the effect of γ on $B2$. If it is large, choose that γ which gives a minimal value of $B1/B2$, since in this case a dominant covariance shift over mean shift would seem to be important for enhancing classification. If γ also has little effect on $B2$, choose that γ which leads to a maximal value of $B1/B2$.

II. *Magnitude of $B1$ large and not greatly affected by the value of γ changing between θ and* 1. This indicates good conditions for classification due to the large Mahalanobis distance measure ($B1$) for all values of γ. Some average, approximate value of γ will suffice.

III. *$B1$ changes substantially as γ changes between θ and* 1. Under this scenario, if γ has little effect on $B2$, it is clearly desirable to select that γ yielding a large value of $B1$. However if $B2$ is greatly affected by γ also, some greater degree of reduction in the variance of the system (by increasing γ a little) is desirable for classification purposes, whilst still maintaining a sizeable Mahalanobis distance ($B1$) between the groups.

The above guide-lines lead to a simple algorithm for the selection of γ to use in expression (3.7) based on the three scenarios above and followed by the selection of λ depending on a crude estimate of the similarity of the group covariances. This algorithm is given in Appendix B. The critical values at each decision stage have been arrived at empirically through a heuristic procedure which involved observing

the values of $B1$ and $B2$ for various random samples from an extensive variety of normal population combinations.

Selection of the parameter λ

For the selection of the regularisation parameter λ, only the term $B2$ can be employed since $B1$ is constant over all values of λ for a given value of γ. Since the term $B2$ gives a measure of the difference between the two distributions due to the covariance shift, the value e^{-B2} at the point $\lambda = 0, \gamma = \theta$ (i.e. with minimal permitted eigenvalue shrinkage - see earlier) gives an indication of the similarity in the group covariance matrices, and this is used as the initial estimate of λ, denoted $\hat{\lambda}$. However, since it is known that $B2$ can be badly affected by bias if the sample size is small relative to p, a refinement to this estimate is proposed.

The magnitude of the term $B2$ when $\lambda = 0$ and $\gamma = 1$ gives further indication as to the similarity or dissimilarity of the group covariance estimates, and so can be used to obtain an appropriately adjusted value of $\hat{\lambda}$. Under this situation ($\gamma = 1$) of maximal eigenvalue shrinkage the determinants of the group covariances are reduced to their average eigenvalue raised to the power of the dimension, p. If the group covariances are similar, the average of their eigenvalues will be similar in magnitude and the term within the brackets in the expression for $B2$ will be close to one, resulting in the value of $B2$ itself being close to zero (see expression 5.3). Since it is not guaranteed that a value of $B2$ close to zero means that the two p–dimensional group covariances $\mathbf{S}_1$ and $\mathbf{S}_2$ are similar, a second quantity is used as a further check to determine the degree of similarity in the covariances in such a situation. Consider

$$z_{12} = \left(\frac{1}{p} \sum_{i=1}^{p} |\hat{e}_{1i} - \hat{e}_{2i}| \right)$$

where $\hat{e}_{1i}$ is the i^{th} eigenvalue of $\mathbf{S}_1$. Note that the $\mathbf{S}_1$ and $\mathbf{S}_2$ may have been minimally regularised (as explained earlier) using $\lambda = 0, \gamma = \theta$ to stabilize excessive variation in the original covariance estimates. This quantity is the average (absolute) difference between corresponding eigenvalues of $\mathbf{S}_1$ and $\mathbf{S}_2$ relative to a measure of the overall variance in the groups. It may occur that $B2(0, 1)$ (where $B2(a, b)$ denotes the value of $B2$ when $\lambda = a$ and $\gamma = b$) is close to zero, while dissimilarity between $\mathbf{S}_1$ and $\mathbf{S}_2$ is indicated by a large value of z_{12}. In such a situation the second quantity z_{12} serves to compliment $B2(0, 1)$ by detecting a phenomenon which the latter is incapable of detecting.

Taking into account the above procedures, the adjusted estimate of λ is $\hat{\lambda}'$, where

$$\hat{\lambda}' = \hat{\lambda}^{1/w}$$

where w is proportional to $1/B2(0,1)$. If w is large (suggesting similarity of the covariances), and z_{12} is large (suggesting dissimilarity), an adjustment of $\hat{\lambda}$ towards zero is made, if appropriate. Details of the heuristic algorithm, derived from purely empirical/simulation results are given in Appendix B.

Thus model selection using the Bhattacharyya distance consists of the following steps:

i. Evaluate $\widehat{B1}$ and $\widehat{B2}$ from the available data for varying degrees of covariance eigenvalue shrinkage (a range of γ), but using no covariance mixing ($\lambda = 0$).

ii. Select $\hat{\gamma}$ using decision algorithm in Appendix B that implements the guidelines given in this section .

iii. Using the amount of eigenvalue shrinkage determined by the selected parameter value $\hat{\gamma}$, estimate $\hat{\lambda}$ using $B2$ and confirm or adjust this estimate using the two checks of covariance similarity, the values w and z_{12}.

The regularised classification rule which uses the above model selection procedure will be denoted SRDF-B.

Re-sampling techniques are avoided in this procedure. This contrasts with Friedman's SRDF where a sample-reuse method (cross-validation) is performed at each of a whole grid of typically between 25 and 50 points. The result is a classification rule with a greatly reduced computational burden. Furthermore, the rule is one which avoids having to arbitrarily choose between apparently equally good (λ, γ) combinations, such as occurs when there is a non-unique minimum cross-validated error rate.

5.2.5 Model selection when there are more than two groups

The technique outlined above selects appropriate values of λ and γ using the Bhattacharyya distance in situations when there are only two groups. This is because the measure B as stated in expressions (5.1), (5.2) and (5.3) is written for the two-group case. If there are more than two groups, the above procedure must be

followed for each pair of groups, leading to estimates of the regularisation parameters being obtained for each pair. The final values of $\hat{\lambda}$ and $\hat{\gamma}$ are then calculated by simply taking the median of the various parameter values obtained from the different pairs. The median is used since there may be a small number of pairs of groups for which the model selection procedure leads to regularisation parameter values which are dissimilar to those obtained from the majority of pairs.

Since the model selection procedure is to be repeated for each pair of groups, the computation time required increases as the number of groups increases. However, since the model selection procedure proposed in this chapter is so much faster (in terms of computation time) than the cross-validation method employed by Friedman (see Table 5.14), the number of groups would have to be very large before the computation times of the two methods became of a similar order of magnitude.

5.3 SIMULATION STUDIES AND RESULTS

Computer simulation is used to compare the performances of SRDF, SLDF, SQDF, SEDF and SRDF-B in the same variety of settings as that used in Chapter 3, Section 3.5, with the addition that the two-group case is studied, as well as the three-group situation. In all cases the group distributions are normal and the sample size from each group was 14, giving a total sample size of 28 or 42. For each set of conditions, simulations were performed for various levels of dimensionality: $p = 6, 10$ and 20. The optimisation grid for the SRDF was set equal to that used in previous chapters. Since the sample size to dimensions ratio is less than one for some simulations, the zero eigenvalues of the group covariance matrix estimates were replaced by a small quantity, sufficient to permit numerically stable covariance inversion.

There were 100 repetitions of the following experiment for each value of p and for each of the six settings (see Subsection 3.5.1). As before, random samples of size 14 from each group were drawn from specified multivariate normal distributions and were used to construct the classification rules for all five of the above methods. An additional test sample of size 100 was randomly generated from the same distributions and classified using each of the five rules given above, yielding estimates of the overall error rate for each rule. These are presented in Tables 5.2

Table 5.2: **Equal, Spherical Covariance Matrices. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.08 (.03)	.10 (.04)	.10 (.04)
SRDF-B	.08 (.03)	.09 (.03)	.10 (.04)
SQDF	.16 (.06)	.30 (.07)	.30 (.07)
SLDF	.10 (.04)	.14 (.05)	.24 (.09)
SEDF	.08 (.03)	.09 (.03)	.09 (.03)
$\bar{\lambda}^{SRDF}$	.86 (.29)	.83 (.32)	.84 (.32)
$\bar{\gamma}^{SRDF}$	.79 (.33)	.78 (.31)	.82 (.27)
$\bar{\lambda}^{SRDF-B}$	.90 (.05)	.77 (.08)	.58 (.09)
$\bar{\gamma}^{SRDF-B}$	.91 (.17)	.90 (.19)	.87 (.23)

to 5.13, along with the means and standard deviations of the selected regularisation parameters for SRDF and SRDF-B over the 100 replications. In the tables $\bar{\lambda}^{SRDF}$ and $\bar{\lambda}^{SRDF-B}$ denote the mean value of λ for SRDF and SRDF-B respectively. The mean value of γ for each method is defined similarly.

In the various conditions tested for the two- and three-group cases it is clear that SRDF and SRDF-B yield very similar average error rates over the 100 replications. In nine of the eighteen sets of simulation conditions for the two-group case represented in Tables 5.2 to 5.13, SRDF-B performs slightly better (and often with a reduced standard deviation) than SRDF in terms of their estimated error rates. In five of the sets the SRDF has a slightly lower error rate. The model selection procedures of SRDF and SRDF-B give roughly similar results regarding the selection of γ by introducing appropriate degrees of this parameter for each set of simulation conditions. Regarding the selection of λ, the two procedures can give entirely different results (e.g. Table 5.7) and yet the average error rates remain very similar. This again shows that in a number of situations the error surface is quite flat with respect to the covariance mixing parameter λ. In conclusion, neither technique is superior to the other in terms of experimental classification error rates.

In the cases where there are three groups, represented in Tables 5.8 to 5.13, the SRDF-B again performs comparably to SRDF in most settings. There are two exceptions to this, where SRDF-B performs somewhat worse than SRDF. These both occur in higher dimensional setting ($p = 20$) (see Tables 5.10 and 5.12). In

Table 5.3: **Unequal, Spherical Covariance Matrices. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.13 (.05)	.11 (.05)	.08 (.05)
SRDF-B	.12 (.04)	.10 (.05)	.10 (.10)
SQDF	.20 (.06)	.34 (.08)	.35 (.07)
SLDF	.17 (.05)	.20 (.06)	.30 (.07)
SEDF	.15 (.04)	.15 (.04)	.18 (.04)
$\bar{\lambda}^{SRDF}$	.48 (.37)	.33 (.33)	.28 (.23)
$\bar{\gamma}^{SRDF}$	.75 (.34)	.81 (.28)	.89 (.19)
$\bar{\lambda}^{SRDF-B}$	.44 (.27)	.10 (.15)	.00 (.00)
$\bar{\gamma}^{SRDF-B}$	.77 (.34)	.85 (.25)	.81 (.31)

Table 5.4: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the Low-Variance Subspace. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.03 (.03)	.05 (.04)	.12 (.06)
SRDF-B	.01 (.02)	.08 (.05)	.16 (.05)
SQDF	.02 (.02)	.14 (.08)	.28 (.07)
SLDF	.01 (.01)	.03 (.03)	.15 (.08)
SEDF	.09 (.04)	.12 (.05)	.15 (.05)
$\bar{\lambda}^{SRDF}$	.97 (.16)	.92 (.21)	.88 (.26)
$\bar{\gamma}^{SRDF}$	.23 (.32)	.22 (.29)	.43 (.29)
$\bar{\lambda}^{SRDF-B}$	.35 (.12)	.14 (.05)	.01 (.00)
$\bar{\gamma}^{SRDF-B}$	.01 (.10)	.36 (.44)	.72 (.38)

Table 5.5: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the High-Variance Subspace. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.02 (.02)	.03 (.02)	.04 (.02)
SRDF-B	.02 (.02)	.02 (.02)	.04 (.02)
SQDF	.06 (.04)	.19 (.09)	.23 (.09)
SLDF	.03 (.02)	.05 (.03)	.16 (.08)
SEDF	.02 (.02)	.02 (.02)	.04 (.02)
$\bar{\lambda}^{SRDF}$	.95 (.19)	.96 (.16)	.93 (.23)
$\bar{\gamma}^{SRDF}$	.80 (.34)	.87 (.28)	.91 (.18)
$\bar{\lambda}^{SRDF-B}$	.36 (.12)	.14 (.05)	.01 (.00)
$\bar{\gamma}^{SRDF-B}$	.67 (.13)	.75 (.11)	.83 (.11)

Table 5.6: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.17 (.07)	.13 (.06)	.05 (.03)
SRDF-B	.15 (.05)	.11 (.05)	.10 (.07)
SQDF	.16 (.05)	.19 (.08)	.20 (.05)
SLDF	.48 (.06)	.46 (.06)	.45 (.06)
SEDF	.48 (.06)	.46 (.06)	.43 (.05)
$\bar{\lambda}^{SRDF}$	.14 (.14)	.11 (.11)	.15 (.12)
$\bar{\gamma}^{SRDF}$	.13 (.21)	.47 (.32)	.65 (.30)
$\bar{\lambda}^{SRDF-B}$	.04 (.02)	.00 (.00)	.00 (.00)
$\bar{\gamma}^{SRDF-B}$	.12 (.21)	.28 (.33)	.50 (.34)

Table 5.5: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the High-Variance Subspace. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.02 (.02)	.03 (.02)	.04 (.02)
SRDF-B	.02 (.02)	.02 (.02)	.04 (.02)
SQDF	.06 (.04)	.19 (.09)	.23 (.09)
SLDF	.03 (.02)	.05 (.03)	.16 (.08)
SEDF	.02 (.02)	.02 (.02)	.04 (.02)
$\bar{\lambda}^{SRDF}$	.95 (.19)	.96 (.16)	.93 (.23)
$\bar{\gamma}^{SRDF}$	.80 (.34)	.87 (.28)	.91 (.18)
$\bar{\lambda}^{SRDF-B}$	.36 (.12)	.14 (.05)	.01 (.00)
$\bar{\gamma}^{SRDF-B}$	.67 (.13)	.75 (.11)	.83 (.11)

Table 5.6: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.17 (.07)	.13 (.06)	.05 (.03)
SRDF-B	.15 (.05)	.11 (.05)	.10 (.07)
SQDF	.16 (.05)	.19 (.08)	.20 (.05)
SLDF	.48 (.06)	.46 (.06)	.45 (.06)
SEDF	.48 (.06)	.46 (.06)	.43 (.05)
$\bar{\lambda}^{SRDF}$	.14 (.14)	.11 (.11)	.15 (.12)
$\bar{\gamma}^{SRDF}$	.13 (.21)	.47 (.32)	.65 (.30)
$\bar{\lambda}^{SRDF-B}$	.04 (.02)	.00 (.00)	.00 (.00)
$\bar{\gamma}^{SRDF-B}$	.12 (.21)	.28 (.33)	.50 (.34)

Table 5.7: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. (Two Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.04 (.03)	.06 (.04)	.04 (.04)
SRDF-B	.02 (.02)	.04 (.03)	.05 (.05)
SQDF	.02 (.02)	.08 (.07)	.11 (.04)
SLDF	.03 (.02)	.09 (.04)	.18 (.06)
SEDF	.09 (.04)	.13 (.05)	.13 (.05)
$\bar{\lambda}^{SRDF}$	.74 (.37)	.50 (.34)	.42 (.26)
$\bar{\gamma}^{SRDF}$	.34 (.31)	.52 (.35)	.74 (.31)
$\bar{\lambda}^{SRDF-B}$	.05 (.04)	.01 (.01)	.00 (.00)
$\bar{\gamma}^{SRDF-B}$	.13 (.15)	.28 (.33)	.38 (.32)

Table 5.8: **Equal, Spherical Covariance Matrices. (Three Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.12 (.04)	.13 (.04)	.15 (.06)
SRDF-B	.11 (.03)	.11 (.03)	.14 (.04)
SQDF	.22 (.06)	.39 (.07)	.41 (.07)
SLDF	.13 (.04)	.17 (.05)	.25 (.06)
SEDF	.11 (.03)	.12 (.03)	.14 (.04)
$\bar{\lambda}^{SRDF}$	.80 (.34)	.78 (.36)	.76 (.36)
$\bar{\gamma}^{SRDF}$	.77 (.30)	.78 (.30)	.82 (.24)
$\bar{\lambda}^{SRDF-B}$	.90 (.04)	.76 (.05)	.53 (.08)
$\bar{\gamma}^{SRDF-B}$	.93 (.09)	.91 (.12)	.88 (.15)

Table 5.9: **Unequal, Spherical Covariance Matrices. (Three Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.19 (.04)	.17 (.05)	.15 (.05)
SRDF-B	.19 (.05)	.17 (.05)	.17 (.10)
SQDF	.31 (.06)	.46 (.07)	.51 (.07)
SLDF	.23 (.04)	.27 (.05)	.35 (.07)
SEDF	.21 (.04)	.22 (.04)	.25 (.05)
$\bar{\lambda}^{SRDF}$	.33 (.36)	.21 (.23)	.16 (.16)
$\bar{\gamma}^{SRDF}$	.71 (.31)	.86 (.21)	.88 (.20)
$\bar{\lambda}^{SRDF-B}$	.45 (.18)	.11 (.10)	.01 (.02)
$\bar{\gamma}^{SRDF-B}$	.89 (.21)	.89 (.20)	.85 (.25)

Table 5.10: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the Low-Variance Subspace. (Three Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.05 (.03)	.10 (.05)	.23 (.06)
SRDF-B	.04 (.02)	.15 (.09)	.29 (.07)
SQDF	.10 (.04)	.30 (.08)	.46 (.06)
SLDF	.05 (.02)	.09 (.04)	.21 (.05)
SEDF	.21 (.05)	.27 (.06)	.33 (.06)
$\bar{\lambda}^{SRDF}$	.95 (.17)	.83 (.27)	.79 (.28)
$\bar{\gamma}^{SRDF}$	.02 (.08)	.04 (.13)	.20 (.23)
$\bar{\lambda}^{SRDF-B}$	.68 (.30)	.80 (.19)	.68 (.06)
$\bar{\gamma}^{SRDF-B}$	.00 (.00)	.28 (.40)	.78 (.34)

Table 5.11: **Equal, Highly Ellipsoidal Covariance Matrices with Mean Differences concentrated in the High-Variance Subspace. (Three Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.07 (.03)	.10 (.03)	.13 (.04)
SRDF-B	.07 (.03)	.10 (.03)	.13 (.04)
SQDF	.15 (.05)	.35 (.08)	.43 (.08)
SLDF	.08 (.03)	.13 (.04)	.23 (.06)
SEDF	.07 (.02)	.10 (.03)	.12 (.03)
$\bar{\lambda}^{SRDF}$	.83 (.35)	.86 (.27)	.84 (.30)
$\bar{\gamma}^{SRDF}$	.65 (.39)	.62 (.35)	.77 (.24)
$\bar{\lambda}^{SRDF-B}$	.74 (.29)	.81 (.21)	.71 (.04)
$\bar{\gamma}^{SRDF-B}$	.74 (.14)	.81 (.11)	.85 (.08)

Table 5.12: **Unequal, Highly Ellipsoidal Covariance Matrices with Zero Mean Differences. (Three Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.20 (.06)	.14 (.05)	.13 (.06)
SRDF-B	.16 (.05)	.12 (.05)	.21 (.09)
SQDF	.16 (.04)	.20 (.06)	.24 (.05)
SLDF	.60 (.05)	.59 (.05)	.59 (.06)
SEDF	.60 (.05)	.59 (.05)	.57 (.05)
$\bar{\lambda}^{SRDF}$	.03 (.05)	.04 (.06)	.08 (.07)
$\bar{\gamma}^{SRDF}$	.12 (.15)	.30 (.16)	.45 (.18)
$\bar{\lambda}^{SRDF-B}$	.02 (.06)	.00 (.01)	.00 (.00)
$\bar{\gamma}^{SRDF-B}$	.04 (.00)	.11 (.14)	.68 (.32)

Table 5.13: **Unequal, Highly Ellipsoidal Covariance Matrices with Non-zero Mean Differences. (Three Groups)** Error rate (with standard deviation) for several discriminant functions.

	Dimension: p		
	6	**10**	**20**
SRDF	.06 (.03)	.06 (.03)	.07 (.04)
SRDF-B	.05 (.04)	.05 (.04)	.06 (.04)
SQDF	.04 (.02)	.10 (.06)	.13 (.04)
SLDF	.16 (.04)	.18 (.04)	.28 (.06)
SEDF	.15 (.04)	.16 (.04)	.20 (.04)
$\bar{\lambda}^{SRDF}$	.07 (.14)	.09 (.10)	.14 (.13)
$\bar{\gamma}^{SRDF}$	.17 (.20)	.37 (.22)	.51 (.20)
$\bar{\lambda}^{SRDF-B}$	.10 (.22)	.06 (.17)	.01 (.07)
$\bar{\gamma}^{SRDF-B}$	.13 (.14)	.24 (.20)	.47 (.28)

these instances, the value of $\bar{\gamma}$ for the SRDF-B appears to be too high, which indicates inappropriate regularisation parameter estimates, and consequently a high error rate. On the whole, however, the model selection procedure of the SRDF-B performs well, and generally in agreement with the model selection procedure of the SRDF.

The standard deviations of the selected regularisation parameters tended to be smaller for SRDF-B, perhaps because of the more direct nature of the path taken to select the pair of values $(\hat{\lambda}, \hat{\gamma})$ in the parameter selection procedure in SRDF-B compared with SRDF. Furthermore, the model selection process in SRDF-B affords a unique choice of the estimated best pair of values $(\hat{\lambda}, \hat{\gamma})$, without having to break ties in an arbitrary way, as for SRDF.

In conclusion, it can be stated that the Bhattacharyya distance between groups does indeed provide information as to appropriate regularisation parameter values to use in expression (3.7). This can be used to obtain a classification rule which seeks to minimise the actual overall error rate for data from two or more specified normal distributions. Unfortunately, no tidy, direct theoretical relationship appears to exist in the literature between components of the Bhattacharyya distance and the error rate. Thus the derivation of the model selection procedure was based on empirical data and it can be seen to perform as well as the model selection procedure developed by Friedman (1989) in the SRDF method, at least under the

tested range of simulated conditions.

Computational considerations

A substantial advantage of the model selection procedure in SRDF-B over that of SRDF relates to the computation time required for each method. Table 5.14 gives approximate ratios (SRDF-B/SRDF) of CPU times for various dimensions. The actual CPU time in minutes required to estimate the regularisation parameters from samples of size 14 from each group for SRDF-B are given in brackets in the table. These are the times required to run the procedures, which are all written in MATLAB™ (1995), on a SUN Sparcstation ELC.

Table 5.14: Ratios of CPU times required for each method (SRDF-B/SRDF).

	$p = 6$	$p = 10$	$p = 20$
Two Groups	.02 (.12)	.02 (.23)	.02 (.65)
Three Groups	.02 (.38)	.02 (.67)	.03 (2.00)

These results indicate the large gain in computational efficiency in using SRDF-B over SRDF. It is expected that as the number of groups increases, the ratio of CPU times would increase, since SRDF-B deals with each pair of groups in turn. Nevertheless, the SRDF-B method would still be expected to be considerably faster than SRDF even for a large number of groups.

5.4 CASE STUDIES

The various classification rules, including SRDF-B, were tested on a number of real data sets. These case studies are performed to compliment the large simulation studies of this and previous chapters. The aim of this case study section is to focus on the performance of the SRDF-B procedure developed in this chapter. Also, the SEDF is included among the classifiers tested since it was the main subject of the work in Chapter 2 of this thesis. Comparison is restricted to the criterion of error rate, although the matter of computational efficiency has already been addressed in some previous sections. The re-sampling methods employed by SRDF render it computationally the slowest technique by far. A brief description of each data set, along with the various classifier error rates (obtained using the technique

of cross-validation), and regularisation parameter values, is given below. Cross-validation was chosen over some other re-sampling error-rate estimating methods (see, for example, Ganeshanandam and Krzanowski (1990), Koolaard and Lawoko (1993)) such as the 0.632 estimator (Efron (1983)), purely for practical reasons of computational feasibility. That is, previous work in this research project already involved cross-validated error rates, and it was decided to use the same computer programs for the case studies. The aim is to observe the effect of regularisation on the cross-validated error rate, and also to compare Friedman's original method of determining the degree of regularisation with the new method employing Bhattacharyya distance.

Insect data

Lindsey, Herzberg and Watts (1987)

Three variables were measured on each of ten insects for each of three species of a type of insect, Chaetocnema. The first variable is the width of the first joint of the first tarsus; the second is the width of the first joint of the second tarsus, and the third is the maximal width of the aedugus. The objective would be to correctly classify a given individual as belonging to one of the species. The cross-validated error rates obtained for the various methods were: SRDF .03, SRDF-B .07, SQDF .03, SLDF .07, SEDF .17. Values of $(\bar{\lambda}, \bar{\gamma})$ for the SRDF were (.97, .43) while those for SRDF-B were (.53,.50). The problem is well posed here (n/p ratio is 3.3), so the benefits of regularisation are not expected to be significant, and this is shown to be the case. All methods except the SEDF yield low error rates, since maximal eigenvalue shrinkage removes the moderately ellipsoidal nature of the group covariance matrices in the example, and decreases the rule's ability to separate the groups.

Cancer data

Hong and Yang (1991)

The cancer data set was previously analysed by Aeberhard et al. (1994) using various classification rules including the SRDF. This data set relates to three types of pathological lung cancer. Each cancer type is described by 56 variables with each variable taking on one of the integer values 1 through 4. The sample sizes from each cancer type, or group, (9, 13 and 10 respectively) are very small. Hence, this problem is extremely ill-posed. The cross-validated error rates for the various

methods were SRDF .50, SRDF-B .44, SQDF .72, SLDF .60, SEDF .40. The values of $(\bar{\lambda}, \bar{\gamma})$ were $(.52, .75)$ for the SRDF and $(.53, .95)$ for SRDF-B. It is clear that a high degree of eigenvalue shrinkage is necessary to stabilise the covariance estimates. In fact the SEDF has the lowest error rate in this case, implying that much of the information in the covariance estimates does not improve discriminability in this case, where there is such a large dimensionality and so few observations.

The magnitude of the error rate estimates for the SRDF in this example is somewhat different to that obtained by Aeberhard et al. (1994), probably due to differences in the implementation details of the method. The following error rates were obtained by Aeberhard et al.: SQDF .69, SLDF .81, SRDF .37. These differences include the particular regularisation grid specified in the model selection procedure, as well as the precise implementation of the procedure which replaces the zero eigenvalues of the covariance estimates with positive numbers sufficiently large to permit numerically stable matrix inversion (see Section 5.3). Both of these factors could affect the error rate obtained, particularly for this high-dimensional data set which has a very small n/p ratio. The SQDF and SRDF are also affected by the procedure which replaces the zero eigenvalues.

Diabetes data

Reaven and Miller (1979)

This data set comprises five variables measured on each of 145 non-obese individuals belonging to one of three groups which relate to the type of diabetes they have. The groups are: overt nonketotic diabetes (33 observations), chemical subclinical diabetes (36 observations) and the final group is termed normal, indicating no diabetes (76 observations). The problem is again well posed, and because of this, little regularisation is necessary. The model selection procedures of SRDF and SRDF-B do the right thing in this regard. It appears that eigenvalue shrinkage is not beneficial for classification, and even the small amount ($\gamma = .08$) employed by the SRDF results in a slightly greater error rate for that classification rule compared to SRDF-B and SQDF. The cross-validated error rates are: SRDF .15, SRDF-B .11 SQDF .10 SLDF .11 and SEDF .14. The values of $(\bar{\lambda}, \bar{\gamma})$ for SRDF are $(.28, .08)$, and for SRDF-B are $(.12, .00)$. The model selection procedure for SRDF-B performs very well in this case.

Kangaroo data

Andrews and Herzberg (1985)

This data set relates to three species (groups) of kangaroo. Each group is described by nine variables measuring physical characteristics of the animals. The data set was split by sex and discriminant analysis was performed on males and females separately. The sample sizes for each sex/species combination was 25 with the exception of the sample of males from group 2 which numbered 23.

The error rates for the various classifiers applied to the male kangaroo data were SRDF .32, SRDF-B .30, SQDF .34, SLDF .25 and SEDF .51. The values of $(\bar{\lambda}, \bar{\gamma})$ for SRDF are $(.86, .13)$, and for SRDF-B are $(.18, .00)$. The cross-validated error rates for the various classifiers applied to the female kangaroo data were: SRDF .25, SRDF-B .28, SQDF .40, SLDF .25 and SEDF .52. The values of $(\bar{\lambda}, \bar{\gamma})$ for SRDF are $(.82, .01)$, and for SRDF-B are $(.17, .00)$. Eigenvalue shrinkage appears not to be beneficial in this instance, but employing covariance mixing does, although there does not appear to be a clear relationship between the degree of regularisation and the error rate. The model selection procedures for SRDF and SRDF-B select similar values for the parameter γ, but not λ, although the resulting error rates for each rule are similar. The SLDF performs slightly better than the two regularised rules for this data, which is an indication of very similar group covariance matrices.

Tibetan Skull data

Morant (1923)

This data set comprises 32 observations collected from skulls in parts of Tibet. There are two types (groups) of skull represented in the sample, 17 from the Sikkim area (type A) and 15 from the province of Khams (type B). The data consist of five physical measurements made on each skull. The cross-validated error rates for the various classifiers applied to the data were: SRDF .22, SRDF-B .22, SQDF .44, SLDF .34 and SEDF .22. The values of $(\bar{\lambda}, \bar{\gamma})$ for SRDF are $(1.0, .92)$, and for SRDF-B are $(.46, .96)$. Despite being a seemingly well-posed problem, these results indicate that a high degree of eigenvalue shrinkage is beneficial, as well as a substantial degree of covariance mixing. This is perhaps an unexpected result, but it indicates that the reduction in variance achieved by regularisation can be of benefit in some situations where the sample size to dimension ratio is of a moderate magnitude.

In conclusion, these case studies give a variety of examples in which some classifiers perform better than others in some cases, and worse in others. However, the regularised rules always perform as well as any of the other three. In addition, the method proposed in this chapter, SRDF-B, performs about as well as the original rule SRDF in all instances. Thus it seems that the Bhattacharyya distance can indeed be employed to give reliable indications as to appropriate values for the regularisation parameters. While these values are not always close to those obtained by SRDF through re-sampling techniques, the assessed error rates are quite close, illustrating the fact that in many cases, as mentioned earlier, it is not the *degree* of regularisation that is important to discrimination in a given case, so much as its *presence* in some appropriate form.

Chapter 6

ANALYTIC ASSESSMENT OF REGULARISATION PARAMETERS ON THE PROBABILITY OF MISCLASSIFICATION OF THE QUADRATIC DISCRIMINANT FUNCTION

6.1 INTRODUCTION

In Chapters 3 to 5, Monte-Carlo simulation studies have been used to estimate error rates and assess the effect of the regularisation parameters on the overall error rate of the sample quadratic discriminant function with regularised covariance matrices. Ideally, one would prefer to study the effect of the regularisation parameters by using an exact analytic expression of the overall probability of misclassification of the SQDF. This is the motivation for the work done in this chapter. That is, using analytic results rather than empirical/simulation evidence, it is desired to confirm or obtain support for the results depicting the relationship between regularisation parameters and error rates, which were obtained in Chapters 3 to 5, largely from simulation experiments.

Suppose we have two multivariate normal populations Π_1 and Π_2 where the population parameters are known. The true error rates can be calculated exactly for the LDF (where the population covariances are assumed equal) for any dimension, p. In the case of unequal covariances these error rates are difficult to evaluate because percentage points for linear combinations of non-central chi-squared random variables must be calculated, Bayne and Tan (1981). Various authors have

commented on this problem, including Gilbert (1969) and McLachlan (1975). Several studies of error rate properties in the case of unequal covariance matrices have examined the special case of proportional covariances with zero off-diagonal matrices. Very few analytical results exist regarding the misclassification probabilities of the QDF, and exact expressions for the error rates do not exist, with the exception of an expression recently derived by Houshmand (1993). This expression, however, is limited to the case of two univariate populations only, and where both the means and variances of the populations must be unequal. Nevertheless, it is a manageable exact expression for the error rate of the QDF, and since the RDF is just a variant of the QDF where the population covariance estimates are replaced by regularised estimates, we can use these expressions to investigate the effect of the regularisation parameters on this error rate of the QDF.

6.2 ERROR RATES OF THE QDF IN THE LITERATURE

There has been some attempt in the literature to investigate error rates associated with the QDF. In this section, a brief summary of some relevant papers, in chronological order, is given.

(i) Han (1969) obtained the distribution of the QDF for the two population case with known (proportional) covariance matrices Σ_1 and Σ_2, such that $\Sigma_2 = \sigma^2\Sigma_1$ $(\sigma^2 > 1)$. Using asymptotic expansions this distribution was obtained for the case of unknown population means.

(ii) Gilbert (1969) investigated the performance of the LDF when the populations are normal but the covariances are unequal. It was compared against the QDF situation when all parameters are known. The error rate was one criterion upon which the comparison was made, and it was found that the QDF performed better than the LDF for larger p, and for more unequal covariances. It should be pointed out that the values of p used by Gilbert were small $(p = 1, 2)$ and moderate $(p = 6, 10)$ only. The error rate for the LDF was compared to an approximation of the error rate for the QDF where, in both cases knowledge of the parameters was assumed. The results showed the expected conclusion that pooling covariances (as in the LDF) is generally

harmful to discrimination if the covariances are in fact unequal.

(iii) McLachlan (1975) obtained the expected error rates of the SQDF in the form of asymptotic expansions for the case of two multivariate populations with unequal means and proportional covariance matrices.

(iv) Bayne and Tan (1981) found approximating methods to obtain misclassification probabilities for the general covariance case. The purpose of their paper was to study the effect of unequal covariances, and correlation between variables, on error rate. The study was limited to two bivariate populations whose mean vectors are $\boldsymbol{\mu}_1 = 0, \boldsymbol{\mu}_2 = (\mu_1, \mu_2)$, and whose covariance matrices are $\boldsymbol{\Sigma}_1 = \mathbf{I}$ and

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The matrix $\boldsymbol{\Sigma}_2$ may be diagonalised by a linear transformation of the observations. Four settings of the parameters ρ, σ_1 and σ_2 were proposed, and the distributional form of the QDF was written for each. Pearson curves were used to evaluate approximate error rates for each setting. The effects of ρ on QDF error rate was examined, for different values of $\boldsymbol{\mu}_2$, σ_1 and σ_2.

(v) Bayne, Beauchamp and Kane (1984) evaluated the error rates for the QDF via numerical integration, in the case of two bivariate normal populations with known parameters, and no conditions on the parameters.

(vi) Wakaki (1990) obtained asymptotic expansions of the distribution of the SLDF and SQDF. Comparison of the estimated error rates of each method was made in the special case of proportional covariance matrices, and in the situation where the sample sizes are equal.

(vii) Houshmand (1993) provided the expression for the exact distribution of the QDF for two univariate normal populations ($p = 1$), and hence derived the exact error rates for this case in the form of integrals which can be calculated using numerical techniques. In the case of two multivariate normal populations, Houshmand describes an existing approximation for the distribution of the QDF, and gives a new approximation. From these, the error rates may be approximated. Methods for computing the above error rates are also provided in the paper, and references therein.

6.3 COMPUTING THE ERROR RATE FOR THE QDF AND ITS DERIVATIVE IN THE UNIVARIATE CASE

In this section we outline the expressions for the error rate of the QDF in the two univariate normal population case, and its derivative with respect to the regularisation parameters. In Chapters 3 and 4 it was reported in the discussion about the overall sample error rate for various λ and γ, that the estimated error rate surface over the (λ, γ) grid was reasonably flat in the λ direction. It is of interest to examine how the true overall error rate changes with the regularisation parameters, in order to get some indication as to the effect of λ on the QDF true error rate in the two univariate population situation. This can be achieved by looking at the rate of change (i.e. derivative) of the error rates with respect to the regularisation parameters. Houshmand (1993) obtained the following expressions for P(1|2) and P(2|1), where P($i|j$) denotes the probability of classifying an observed vector $\boldsymbol{x}$ into population Π_i when it in fact belongs to population Π_j (see Chapter 1, Section 1.2). In the univariate case, population Π_i has mean μ_i and variance σ_i^2, $(i = 1, 2)$.

$$\mathrm{P}(1|2) = 1 - \int_0^{U_1} \sum_{i=0}^{\infty} \frac{z^{i-1/2}\exp\{-z/2\}}{2^{i+1/2}\Gamma(i+1/2)} \times \frac{\exp^{-.5\beta_1}(0.5\beta_1)^i}{i} dz \qquad (6.1)$$

where

$$\begin{aligned} \beta_1 &= \frac{\sigma_2^2(\mu_1 - \mu_2)^2}{(\sigma_1^2 - \sigma_2^2)^2} \\ U_1 &= \sigma_1^2(K + r)(\sigma_1^2 - \sigma_2^2)^{-1} \\ K &= \ln\{\sigma_1^2/\sigma_2^2\} \text{ for equal priors and costs of misclassification} \end{aligned}$$

and

$$r = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 - \sigma_2^2}.$$

$$\mathrm{P}(2|1) = 1 - \int_0^{U_2} \sum_{i=0}^{\infty} \frac{z^{i-1/2}\exp\{-z/2\}}{2^{i+1/2}\Gamma(i+1/2)} \times \frac{\exp^{-.5\beta_2}(0.5\beta_2)^i}{i} dz \qquad (6.2)$$

where

$$\beta_2 = \frac{\sigma_1^2(\mu_1 - \mu_2)^2}{(\sigma_1^2 - \sigma_2^2)^2}$$

and

$$U_2 = \sigma_2^2(K+r)(\sigma_1^2-\sigma_2^2)^{-1}.$$

The effects of the regularisation parameters on these error rates are studied by evaluating the derivative of the probabilities with respect to the regularisation parameters. The derivatives of P(1|2) and P(2|1) with respect to λ have been obtained, but not with respect to the eigenvalue shrinkage parameter γ, since in the univariate situation such shrinkage has no effect anyway. The primary practical use of γ lies in its application, in the multivariate situation, to the sample covariance matrix, whereas in this chapter we are dealing with population parameters which are assumed known. The expressions (6.1) and (6.2) are valid in all situations where $\mu_1 \neq \mu_2$ and $\sigma_1 \neq \sigma_2$, since they are QDF error rates. Hence the derivatives obtained are also not finite if either the population means or variances are equal, as is the case when $\lambda = 1$.

In this chapter the derivative of P(1|2) with respect to λ is evaluated for a variety of settings of μ_1, μ_2, σ_1^2 and σ_2^2 over the range $0 \leq \lambda < 1$. The four settings of the population parameters chosen represent four general classification cases in the univariate situation: means and variances similar in magnitude, means similar but variances disparate, means separate but variances similar, and both means and variances dissimilar. They serve to give an impression of the effect of the covariance mixing parameter on error rate. Four figures, (Figures 6.1 to 6.4), are presented with the results in Section 6.6, showing how the rate of change of the overall error rate $P_e = 0.5P(1|2) + 0.5P(2|1)$ is affected by λ.

The integrals in expressions (6.1) and (6.2), and similar integrals in the (derivative) expression (6.3) in Section 6.4, were computed using the algorithms of Lau (1980) and Narula and Desu (1981). To do this, a computer program was written in FORTRAN 77 and was based on one received by the author from Houshmand (1995). Additional programs were written using MATLAB™ to complete computation of expression (6.3).

The expression $\frac{dP(1|2)}{d\lambda}$ is given in the following section.

6.4 DERIVATIVE OF P(1|2) IN THE UNIVARIATE SITUATION

From Houshmand (1993), the rate of change of one component (P(1|2)) of the overall error rate with respect to the regularisation parameter λ is given by:

$$
\begin{aligned}
\frac{dP(1|2)}{d\lambda} = & \int_0^{\kappa_1} \left\{ \sum_{i=0}^{\infty} z^{i-\frac{1}{2}} \exp\left\{-\frac{1}{2}z\right\} \left(2^{i+\frac{1}{2}}\Gamma(i+\frac{1}{2})\Gamma(i+1)\right)^{-1} \right. \\
& \times \left(-\frac{1}{2} \frac{\left(-\frac{\sigma_2^2}{2} + \frac{\sigma_1^2}{2}\right)(\mu_1-\mu_2)^2}{k_1^2} + \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2(-\sigma_1^2+\sigma_2^2}{k_1^3} \right. \\
& \times \left(\exp\left\{ -\frac{1}{2} \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \right\} \right) \\
& \times \left(\frac{1}{2} \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \right)^i \\
& + 2z^{i-\frac{1}{2}} \exp\left\{-\frac{1}{2}z\right\} \\
& \times \left(\exp\left\{ -\frac{1}{2} \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \right\} \right) \\
& \times \left(\frac{1}{2} \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \right)^i \\
& \times i \left(\frac{1}{2} \frac{\left(-\frac{\sigma_2^2}{2} + \frac{\sigma_1^2}{2}\right)(\mu_1-\mu_2)^2}{k_1^2} + \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2(-\sigma_1^2+\sigma_2^2)}{k_1^3} \right. \\
& \left. \left. \left(\frac{k_1^2}{2^{i+\frac{1}{2}}\Gamma(i+\frac{1}{2})\Gamma(i+1)\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2} \right) \right\} dz \right. \\
& + \sum_{i=0}^{\infty} \left\{ \kappa_1^{i-\frac{1}{2}} \exp\left\{-\frac{1}{2}\kappa_1\right\} \right. \\
& \times \left(\exp\left\{ -\frac{1}{2} \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \right\} \right) \\
& \times \left(\frac{1}{2} \frac{\left((1-\lambda)\sigma_2^2 + \lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \right)^i \\
& \left. \left. \times \left(2^{i+\frac{1}{2}}\Gamma(i+\frac{1}{2})\Gamma(i+1) \right) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\times \left\{ \frac{\left(-\frac{\sigma_1^2}{2}+\frac{\sigma_2^2}{2}\right)k_2}{k_1} + \frac{\left((1-\lambda)\sigma_1^2+\lambda\sigma_p^2\right)}{k_1} \right. \\
&\times \left[\left(\frac{-\frac{\sigma_1^2}{2}+\frac{\sigma_2^2}{2}}{(1-\lambda)\sigma_2^2+\lambda\sigma_p^2} - \frac{\left((1-\lambda)\sigma_1^2+\lambda\sigma_p^2\right)\left(-\frac{\sigma_1^2}{2}+\frac{\sigma_2^2}{2}\right)}{\left((1-\lambda)\sigma_2^2+\lambda\sigma_p^2\right)^2} \right) \right. \\
&\times \left(\frac{(1-\lambda)\sigma_2^2+\lambda\sigma_p^2}{(1-\lambda)\sigma_1^2+\lambda\sigma_p^2} \right) \\
&\left. \left. - \left(\frac{(\mu_1-\mu_2)^2\left(-\sigma_1^2+\sigma_2^2\right)}{k_1^2} \right) \right] \right. \\
&\left. - \frac{\left((1-\lambda)\sigma_1^2+\lambda\sigma_p^2\right)k_2\left(-\sigma_1^2+\sigma_2^2\right)}{k_1^2} \right\}
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_p^2 &= \frac{(\sigma_1^2+\sigma_2^2)}{2} \\
\Gamma(t) &= \int_0^\infty \exp\{-z\}\, z^{t-1} dz \\
k_1 &= (1-\lambda)\sigma_1^2 - (1-\lambda)\sigma_2^2 \\
k_2 &= \ln\left(\frac{(1-\lambda)\sigma_1^2+\lambda\sigma_p^2}{(1-\lambda)\sigma_2^2+\lambda\sigma_p^2} \right) + \frac{(\mu_1-\mu_2)^2}{k_1} \\
\kappa_1 &= \frac{\left((1-\lambda)\sigma_1^2+\lambda\sigma_p^2\right)k_2}{k_1}.
\end{aligned}
$$

Rewriting the above expression we obtain

$$
\begin{aligned}
\frac{dP(1|2)}{d\lambda} &= \kappa_3 \int_0^{\kappa_1} \sum_{i=0}^{\infty} \frac{z^{i-\frac{1}{2}} \exp\left\{-\frac{1}{2}z\right\}}{2^{i+\frac{1}{2}}\Gamma(i+\frac{1}{2})} \times \frac{\exp\left\{-\frac{1}{2}\beta_1\right\}\left(\frac{1}{2}\beta_1\right)^i}{i} dz \\
&\quad + 2\kappa_4 \int_0^{\kappa_1} \sum_{i=0}^{\infty} \frac{z^{i-\frac{1}{2}} \exp\left\{-\frac{1}{2}z\right\}}{2^{i+\frac{1}{2}}\Gamma(i+\frac{1}{2})} \times \frac{\exp\left\{-\frac{1}{2}\beta_1\right\}\left(\frac{1}{2}\beta_1\right)^i i}{i} dz \\
&\quad + \kappa_5 \sum_{i=0}^{\infty} \frac{\kappa_1^{i-\frac{1}{2}} \exp\left\{-\frac{1}{2}\kappa_1\right\} \exp\left\{-\frac{1}{2}\beta_1\right\}\left(\frac{1}{2}\beta_1\right)^i}{2^{i+\frac{1}{2}}\Gamma(i+\frac{1}{2})i}
\end{aligned} \tag{6.3}
$$

where

$$
\begin{aligned}
\beta_1 &= \frac{\left((1-\lambda)\sigma_2^2+\lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2} \\
\beta_2 &= \frac{\left((1-\lambda)\sigma_1^2+\lambda\sigma_p^2\right)(\mu_1-\mu_2)^2}{k_1^2}
\end{aligned}
$$

$$
\begin{aligned}
\kappa_3 &= \frac{1}{2}\frac{k_3\beta_1}{(1-\lambda)\sigma_2^2+\lambda\sigma_p^2}+\frac{2\beta_1 k_3}{k_1} \\
\kappa_4 &= \frac{1}{2}\frac{-k_3}{(1-\lambda)\sigma_2^2+\lambda\sigma_p^2}-\frac{2k_3}{k_1} \\
\kappa_5 &= \frac{k_3\kappa_1}{(1-\lambda)\sigma_1^2+\lambda\sigma_p^2}+\left[\left(\frac{k_3}{(1-\lambda)\sigma_2^2+\lambda\sigma_p^2}+\frac{\left((1-\lambda)\sigma_1^2+\lambda\sigma_p^2\right)k_3}{\left((1-\lambda)\sigma_2^2+\lambda\sigma_p^2\right)^2}\right)\right. \\
&\quad \left.\times\left((1-\lambda)\sigma_2^2+\lambda\sigma_p^2\right)-2\beta_2 k_3\right]\left(\frac{1}{k_1}\right)-\frac{2\kappa_1 k_3}{k_1}\Bigg\}
\end{aligned}
$$

and k_1, k_2 and κ_1 are as defined earlier.

6.5 ERROR RATE FOR QDF: MULTIVARIATE NORMAL POPULATIONS WITH DIAGONAL COVARIANCE MATRICES

Consider the situation of two multivariate normal populations, $\Pi_1 : \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\Pi_2 : \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal matrices, with leading diagonal elements denoted as $(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{1p}^2)$ and $(\sigma_{21}^2, \sigma_{22}^2, \ldots, \sigma_{2p}^2)$ respectively. The mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are given by $(\mu_{11}, \mu_{12}, \ldots, \mu_{1p})$ and $(\mu_{21}, \mu_{22}, \ldots, \mu_{2p})$. Houshmand (1993) has given an approximation for the error rates P(1|2) and P(2|1) for the QDF in the case of $\sigma_{1i} \neq \sigma_{2i}$ and $\mu_{1i} \neq \mu_{2i}$, for all i. This is an extension of Patnaik's (1949) method of approximating the distribution of a linear combination of independent non-central Chi-square variates, which the QDF is. The technique involves the conjecture that such a linear combination as stated above may be approximated by a multiple of central Chi-square variates with ν degrees of freedom, $c\chi^2_{(\nu)}$.

Houshmand (1993) obtained the following expressions for the error rates:

$$
\mathrm{P}(1|2) = \begin{cases} 1-\Pr\left[\chi^2_{(\nu)} < (K+r-a)/c\right] & \text{if } c>0 \\ \Pr\left[\chi^2_{(\nu)} < (K+r-a)/c\right] & \text{if } c<0, \end{cases} \tag{6.4}
$$

where

$$
\begin{aligned}
K &= \ln\left\{|\boldsymbol{\Sigma}_1| \,/\, |\boldsymbol{\Sigma}_2|\right\} \\
r &= \sum_{j=1}^{p}\frac{(\mu_{1j}-\mu_{2j})^2}{\sigma_{1j}^2-\sigma_{2j}^2}
\end{aligned}
$$

$$
\begin{aligned}
a &= \frac{Z_3 Z_1 - 2Z_2^2}{Z_3} \\
c &= \frac{Z_3}{4Z_2} \\
n &= \frac{8Z_2^3}{Z_3^2} \\
Z_1 &= \sum_{j=1}^{p} \alpha_j (1 + \delta_j^2) \\
Z_2 &= 2\sum_{j=1}^{p} \alpha_j^2 (1 + 2\delta_j^2) \\
Z_3 &= \sum_{j=1}^{p} 8\alpha_j^3 (1 + 3\delta_j^2) \\
\alpha_j &= (\sigma_{1j}^2 - \sigma_{2j}^2)\sigma_{1j}^{2\,-2}
\end{aligned}
$$

and

$$
\delta_j = \sigma_{2j}^2 (\mu_{1j} - \mu_{2j})(\sigma_{1j}^2 - \sigma_{2j}^2)^{-1}.
$$

$$
\mathrm{P}(2|1) = \begin{cases} \Pr\left[\chi^2_{(\nu)} < (K + r - a)/c\right] & \text{if } c > 0 \\ 1 - \Pr\left[\chi^2_{(\nu)} < (K + r - a)/c\right] & \text{if } c < 0, \end{cases} \tag{6.5}
$$

where $K, r, a, c, n, Z_1, Z_2, Z_3$ have the same form as above except that now

$$
\alpha_j = (\sigma_{1j}^2 - \sigma_{2j}^2)\sigma_{2j}^{2\,-2}
$$

and

$$
\delta_j = \sigma_{1j}^2 (\mu_{1j} - \mu_{2j})(\sigma_{1j}^2 - \sigma_{2j}^2)^{-1}.
$$

Settings of the population parameters were used which are similar to those six conditions used in the simulation studies of Chapters 3 to 5 (see, in particular, Section 3.5). Some of the mean vectors and covariance matrices had to be altered slightly since the expressions (6.4) and (6.5) are not valid if corresponding elements of either $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, or $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are equal, which is the case for most of the six conditions. Figures 6.5 to 6.22 are displays of the overall error rate, $\mathrm{P}_e = (\mathrm{P}(1|2) + \mathrm{P}(2|1))/2$ (assuming equal prior probabilities), as it varies with respect to the covariance regularisation parameters λ and γ under the above six conditions. That is, regularisation of the same form as that in expression (3.7) is applied to

matrices Σ_1 and Σ_2. The resulting covariance matrices are used in expressions (6.4) and (6.5) above.

As mentioned before, the error rate expressions in both the univariate and multivariate cases above are for the true error rate. It would have been most appropriate to have been able to use similar expressions for the error rate conditional on the training sample since it is on the sample estimates that regularisation is employed. No exact, analytical expression exists for the error rate of the SQDF, however, as has been already stated. In any event, the effect of regularisation on the SQDF was the very focus in Chapters 3 to 5 in the various simulation studies. With this in mind, the purpose of the work in this chapter is to analytically and algebraically confirm some of the findings of previous chapters, particularly regarding appropriate magnitudes of λ and γ for given parameter settings.

6.6 RESULTS

6.6.1 Univariate populations

The results of evaluating expressions (6.1) and (6.2) for various population parameters and values of λ are now discussed. Figures 6.1 to 6.4 and Tables 6.1 to 6.4 show how the overall error rate, P_e, as well as $P(1|2)$ and its derivative vary with λ. As a general comment, it may be observed that the overall error rate is not greatly affected by lambda. Over most of the range of λ, there is usually a small rate of change in error rate with respect to λ, although in some conditions, as λ nears 1, the error rate changes more rapidly.

Figure 6.1 shows P_e changing with λ in the case where both the population means and variances are similar in magnitude: ($\mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 1$). The overall level of error rate is high here since these are difficult conditions for discrimination between the populations. Since the population means are close together, if λ is increased and thereby the variances tend to equality, then P_e increases due to the large 'overlap' between the populations. The overall error rate increases by around 15% as λ increases from 0 to 0.9, and the rate of increase is steady for the most part but decreases as λ approaches 1. From Table 6.1 it may be observed that $P(1|2)$ increases from $\lambda = 0$ to approximately $\lambda = 0.7$, then decreases rather rapidly as λ nears 1. On the other hand, $P(2|1)$ increases steadily throughout the range of λ until about $\lambda = 0.8$, when it appears to increase rapidly

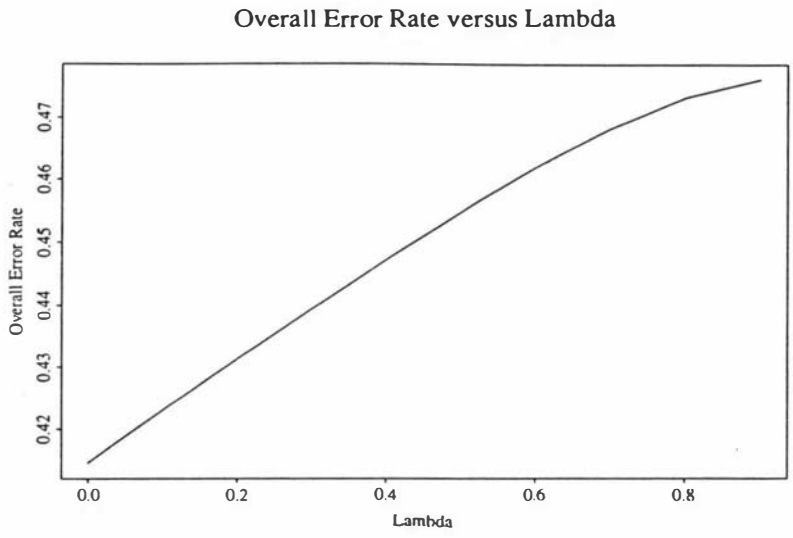


Figure 6.1: Overall error rate (P_e) versus Lambda (λ) when the two population means and variances are similar. ($p = 1, \mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 1$)

as λ approaches 1. These two error rates presumably rapidly tend to equality as λ increases from 0.9 and approaches 1. However, the computations in this range of λ are unstable because the summation to infinity in expressions (6.1) and (6.2) is difficult to obtain.

These results indicate that using covariance mixing in conditions where the population means are close together has the effect of diminishing any information (that the covariance matrices might contain) which could be used to separate the populations. Hence error rates increase. This is in agreement with conclusions from the simulation studies under similar conditions which are also difficult for discrimination.

Figure 6.2 shows the overall error rate against λ when the population means are also close together but the variances are more disparate. Once again P_e increases as λ increases from 0 to 0.9, but this time by over 40%, since making the variances more similar in magnitude increases the 'overlap' between the populations. The rate of increase of P_e is close to constant across the whole range of λ, and P(1|2) and P(2|1) behave in a similar way to the previous case, with P(1|2) peaking and dropping between 0.8 and 1, and P(2|1) increasing more rapidly in the same range of λ (Table 6.2). Once again the general conclusion can be made that if the population means are very close together, the (co)variances should not be regularised to equality.

Table 6.1: Error rates in the case of similar population means and variances. ($\mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 1$)

λ	P(1|2)	$\frac{dP(1\|2)}{d\lambda}$	P(2|1)	P_e
0.0	0.5922	0.0954	0.2370	0.4146
0.1	0.6016	0.0920	0.2446	0.4231
0.2	0.6106	0.0884	0.2521	0.4313
0.3	0.6193	0.0844	0.2595	0.4394
0.4	0.6275	0.0791	0.2669	0.4472
0.5	0.6350	0.0704	0.2744	0.4547
0.6	0.6412	0.0513	0.2823	0.4618
0.7	0.6440	-0.0063	0.2922	0.4681
0.8	0.6347	-0.2340	0.3118	0.4732
0.9	0.5814	-0.8618	0.3707	0.4761

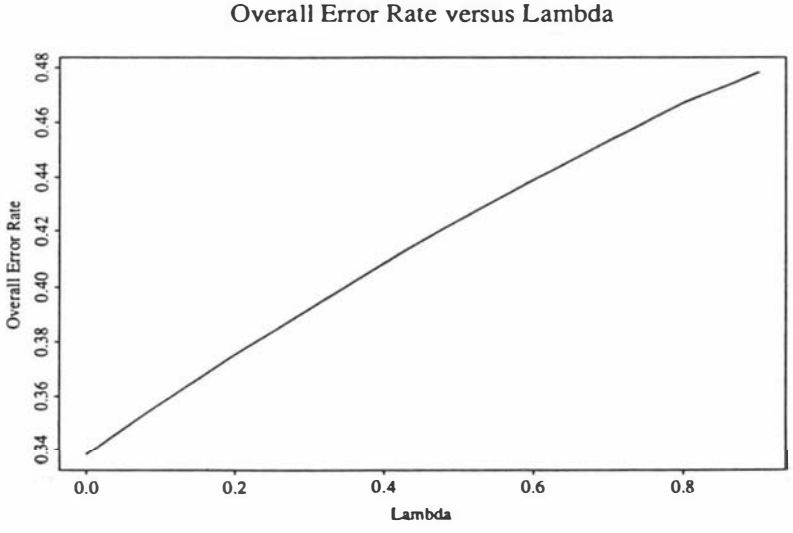


Figure 6.2: Overall error rate (P_e) versus Lambda (λ) when the two population means are similar, but their variances are disparate. ($p = 1, \mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$)

Table 6.2: Error rates in the case of similar population means, but variances disparate. ($\mu_1 = 0, \mu_2 = 0.1, \sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$)

λ	P(1|2)	$\frac{dP(1\|2)}{d\lambda}$	P(2|1)	P_e
0.0	0.5025	0.2460	0.1734	0.3379
0.1	0.5259	0.2231	0.1887	0.3573
0.2	0.5473	0.2052	0.2034	0.3753
0.3	0.5670	0.1907	0.2177	0.3923
0.4	0.5855	0.1786	0.2317	0.4086
0.5	0.6028	0.1682	0.2454	0.4241
0.6	0.6191	0.1582	0.2590	0.4391
0.7	0.6344	0.1454	0.2725	0.4534
0.8	0.6476	0.1096	0.2863	0.4669
0.9	0.6468	-0.3130	0.3094	0.4781

The third case involves population means which are reasonably far apart, and variances which are close together (Figure 6.3). The overall error rate increases only slightly (by less than 1%) as λ increases from 0 to 0.8, and appears to level off for higher values of λ. From Table 6.3 it may be seen that P(1|2) decreases at a slow but almost constant rate as λ increases to about 0.8, while P(2|1) increases at a similarly slow but very steady rate. The level of P_e is much lower than for the previous two cases, due primarily to the much greater separation between population means. Since the variances are similar to begin with, regularisation with λ does not affect the error rate much, as is evident from both Figure 6.3 and Table 6.3.

The final case looked at in this section involves two populations whose means and variances are quite dissimilar. From Figure 6.4 and Table 6.4 it is again evident that P_e (nor indeed P(1|2) or P(2|1)) change much as λ increases from 0 to 0.8, with P_e increasing by 5%. Once again, P(1|2) and P(2|1) change in opposite directions at a very slow, almost constant rate. The level of error rate is quite low due to the large separation between means. This case and the previous one illustrate the fact that (co)variance mixing using the λ parameter is less effective when there is reasonable separation between population means, but it also often affects the different components of P_e in opposite ways.

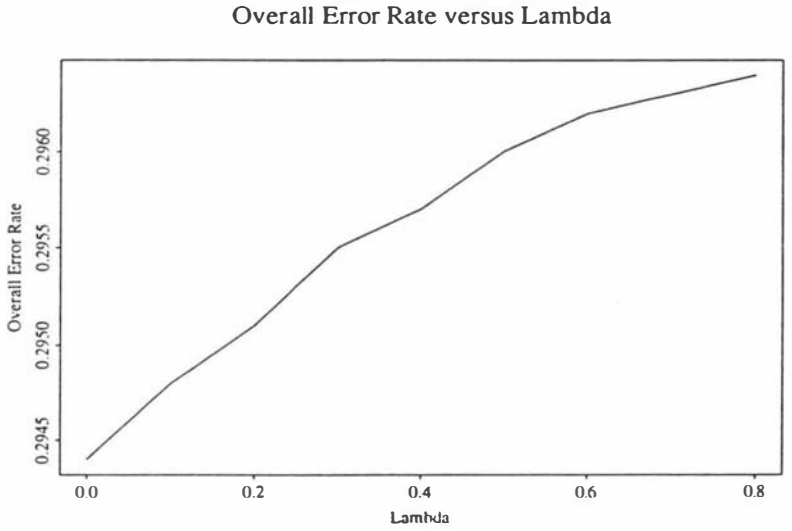


Figure 6.3: Overall error rate (P_e) versus Lambda (λ) when the two population variances are similar, but their means are disparate. ($p = 1, \mu_1 = 0, \mu_2 = 1, \sigma_1^2 = 0.75$ and $\sigma_2^2 = 1$)

Table 6.3: Error rates in the case of similar population variances, but disparate means. ($\mu_1 = 0, \mu_2 = 1, \sigma_1^2 = 0.75$ and $\sigma_2^2 = 1$)

λ	P(1|2)	$\frac{dP(1\|2)}{d\lambda}$	P(2|1)	P_e
0.0	0.3396	-0.0392	0.2491	0.2944
0.1	0.3357	-0.0401	0.2539	0.2948
0.2	0.3316	-0.0410	0.2587	0.2951
0.3	0.3275	-0.0418	0.2634	0.2955
0.4	0.3232	-0.0426	0.2682	0.2957
0.5	0.3189	-0.0434	0.2730	0.2960
0.6	0.3146	-0.0440	0.2777	0.2962
0.7	0.3101	-0.0447	0.2825	0.2963
0.8	0.3056	-0.0452	0.2872	0.2964
0.9	n/a	n/a	n/a	n/a

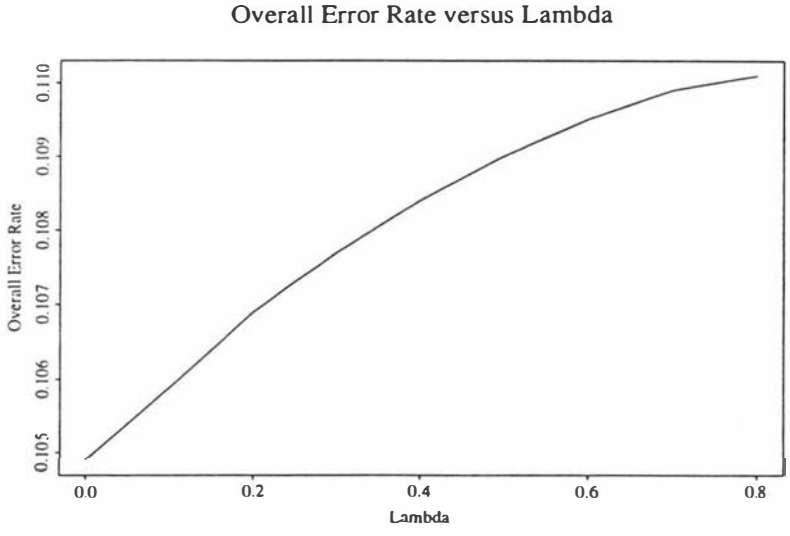


Figure 6.4: Overall error rate (P_e) versus Lambda (λ) when the two population means and variances are disparate. ($p = 1, \mu_1 = 0, \mu_2 = 3, \sigma_1^2 = 1$ and $\sigma_2^2 = 2$)

Table 6.4: Error rates for the case of disparate population means and variances. ($\mu_1 = 0, \mu_2 = 3, \sigma_1^2 = 1$ and $\sigma_2^2 = 2$)

λ	P(1|2)	$\frac{dP(1\|2)}{d\lambda}$	P(2|1)	P_e
0.0	0.1294	-0.0110	0.0803	0.1049
0.1	0.1282	-0.0129	0.0836	0.1059
0.2	0.1268	-0.0147	0.0869	0.1069
0.3	0.1253	-0.0164	0.0901	0.1077
0.4	0.1235	-0.0179	0.0932	0.1084
0.5	0.1217	-0.0194	0.0963	0.1090
0.6	0.1197	-0.0208	0.0993	0.1095
0.7	0.1175	-0.0221	0.1022	0.1099
0.8	0.1152	-0.0234	0.1050	0.1101
0.9	n/a	n/a	n/a	n/a

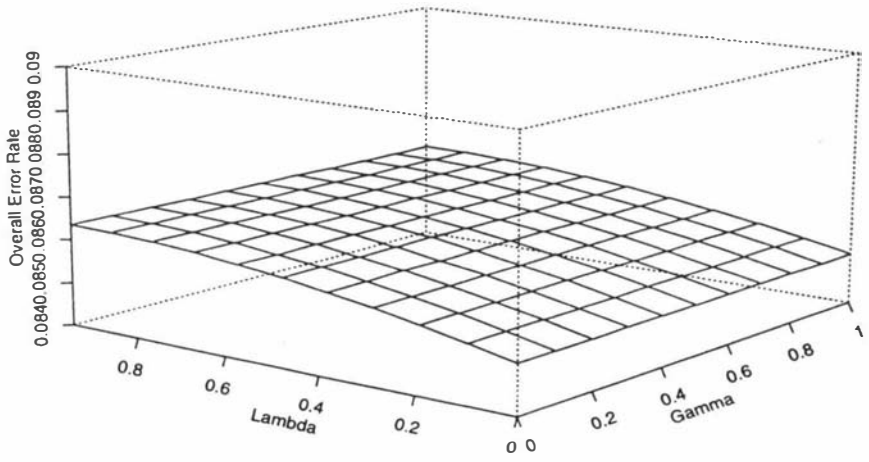


Figure 6.5: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal and spherical covariance matrices ($p = 6$). (i.e. Condition 1 in Chapter 3, Section 3.5)

6.6.2 Multivariate populations

The results of evaluating expressions (6.4) and (6.5), under the conditions stated in the previous section (Section 6.5), are now discussed. The figures show P_e plotted against λ and γ.

From Figures 6.5, 6.6 and 6.7, it is clear that in the case of similar, spherical covariance matrices, the error surface over the (λ, γ) grid is very flat. This confirms observations from the simulation studies of previous chapters. The choice of high (close to one) values for both regularisation parameters is also supported from these figures, and, on such a flat surface as this, shows how surprisingly sensitive the model selection procedures of Friedman and that in Chapter 5 (using Bhattacharyya distance) are.

It can be observed from these figures (i.e. Figures 6.5, 6.6 and 6.7) that γ is shown to have no effect on error rate. This is because the covariance matrices being used are already perfectly spherical. It has been shown from the simulation studies, however, that this condition is ideal for applying eigenvalue regularisation to the sample covariance matrix, since the bias introduced by it is towards the true value. The magnitude of the true error rate for these parameter settings (around 10%) is comparable to that of the error rates of the SRDF in the simulation study of Chapter 3 (Section 3.5).

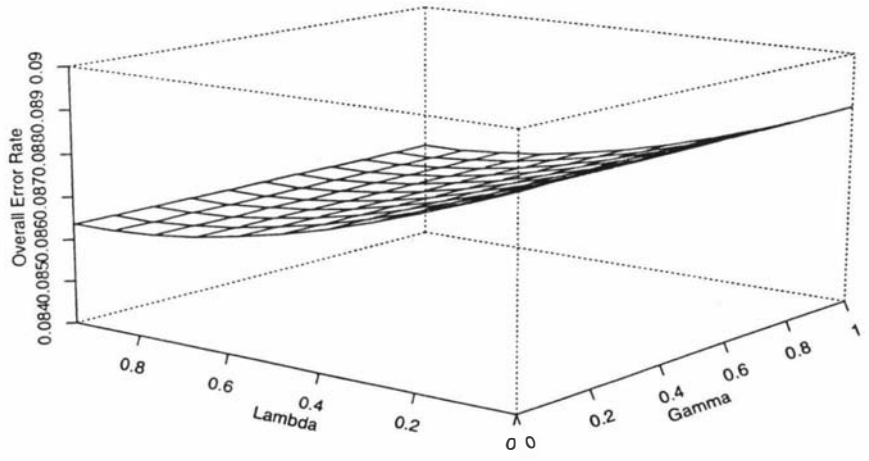


Figure 6.6: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal and spherical covariance matrices ($p = 10$). (i.e. Condition 1 in Chapter 3, Section 3.5)

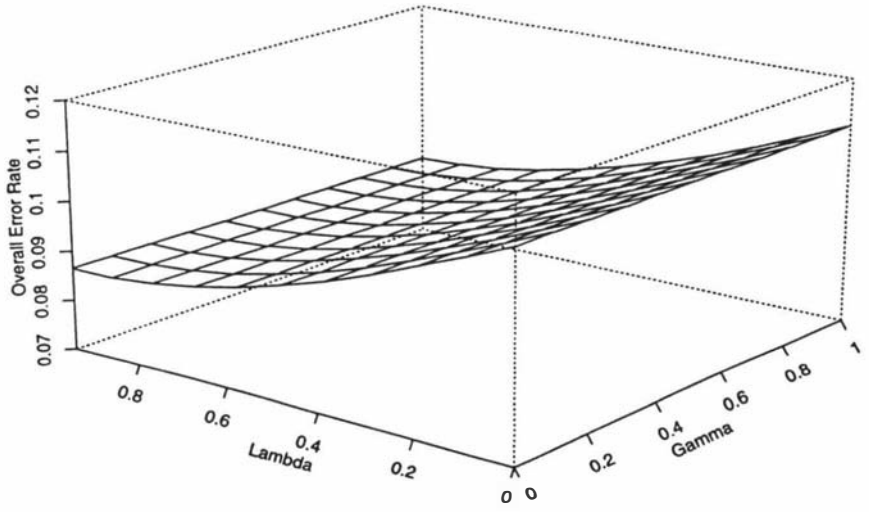


Figure 6.7: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal and spherical covariance matrices ($p = 20$). (i.e. Condition 1 in Chapter 3, Section 3.5)

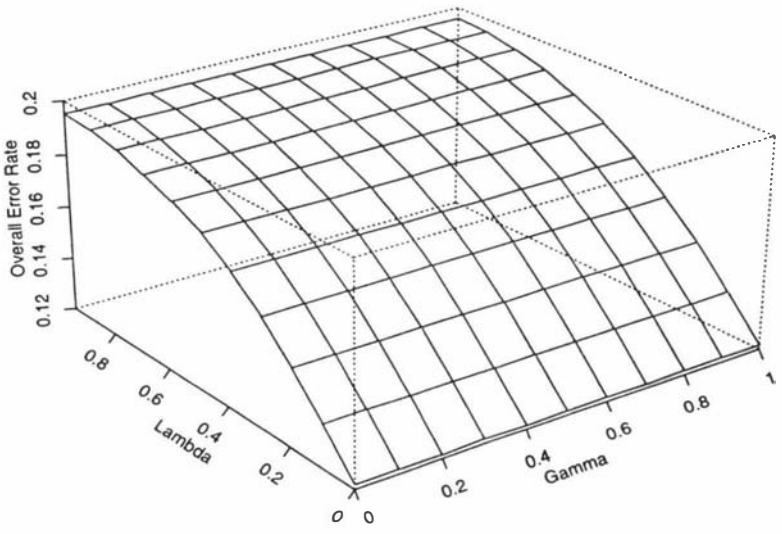


Figure 6.8: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal and spherical covariance matrices ($p = 6$). (i.e. Condition 2 in Chapter 3, Section 3.5)

For the case of unequal, spherical covariances (Figures 6.8, 6.9 and 6.10), the effect of covariance regularisation using λ is substantial. The minimum error rate occurs when λ is zero. Again, since the covariance matrices are already perfectly spherical, there is no effect of γ on the true error rate in this situation, although it is clear that in practice eigenvalue shrinkage will be beneficial since it makes the resulting matrix closer to its (true) spherical shape. The magnitude of the true error rate surface at its minimum is comparable to the minimising cross-validated error rate for the SRDF in Chapter 3, between 10% and 14%.

For the case of similar, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace (Figures 6.11, 6.12 and 6.13), the detrimental effect of γ regularisation on error rate is obvious. This result is in agreement with findings from the simulation studies of previous chapters. Since the population mean differences are in the low variance subspace, those differences are identifiable if the covariance matrices remain ellipsoidal. Eigenvalue shrinkage causes the covariance matrices to become more spherical, and the resulting increase in variance in the low variance subspace leads to the mean differences becoming less identifiable, and hence error rate increases.

Since the population covariances are very similar in this case, λ is shown to have virtually no effect on error rate, since λ shrinks the covariances to their average.

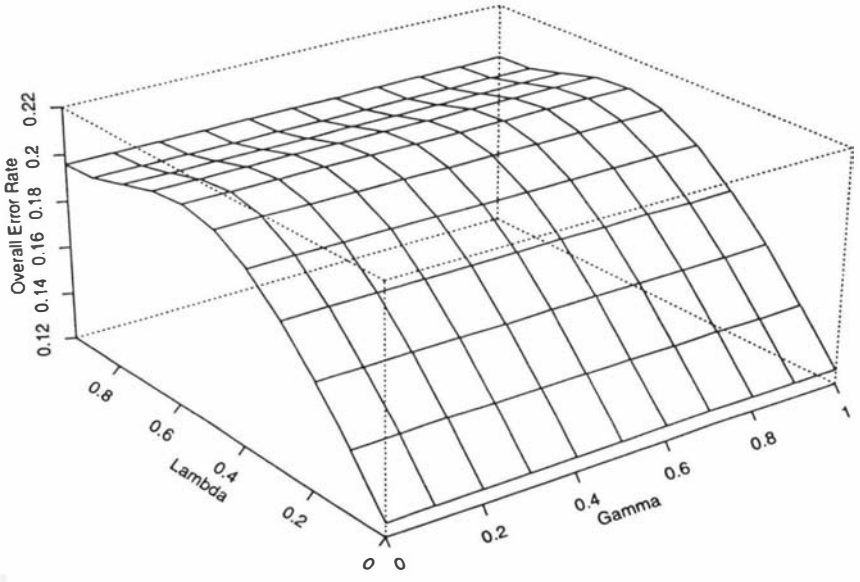


Figure 6.9: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal and spherical covariance matrices ($p = 10$). (i.e. Condition 2 in Chapter 3, Section 3.5)

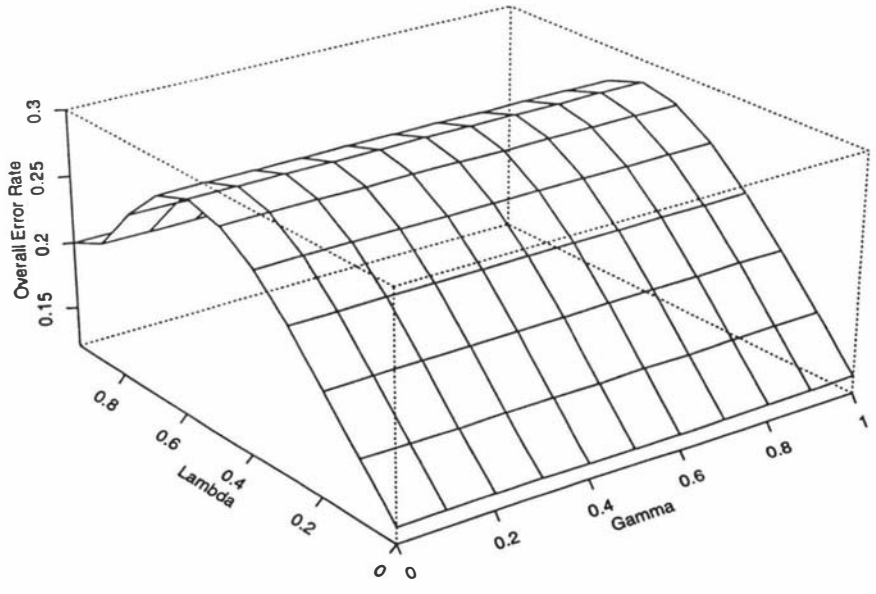


Figure 6.10: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal and spherical covariance matrices ($p = 20$). (i.e. Condition 2 in Chapter 3, Section 3.5)

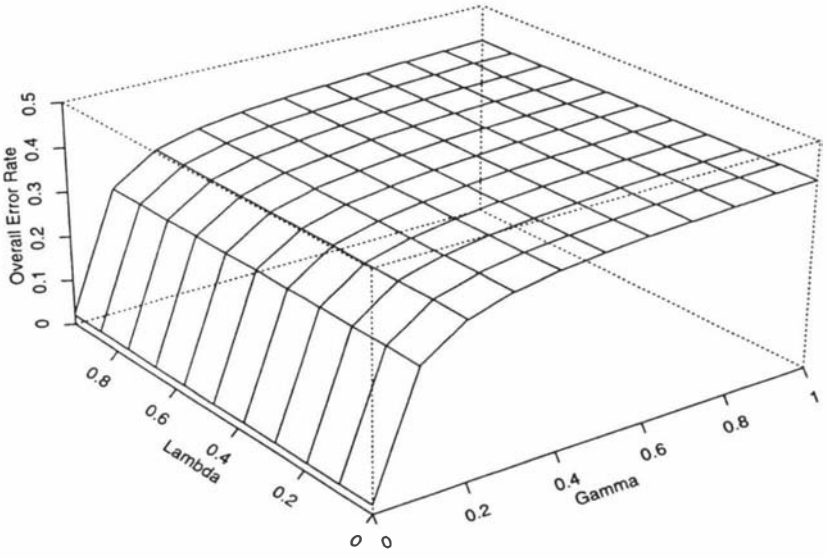


Figure 6.11: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace ($p = 6$). (i.e. Condition 3 in Chapter 3, Section 3.5)

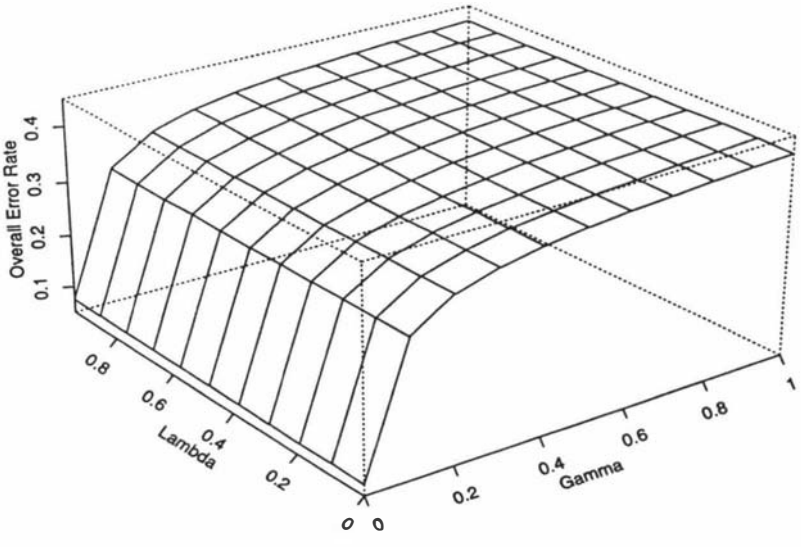


Figure 6.12: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace ($p = 10$). (i.e. Condition 3 in Chapter 3, Section 3.5)

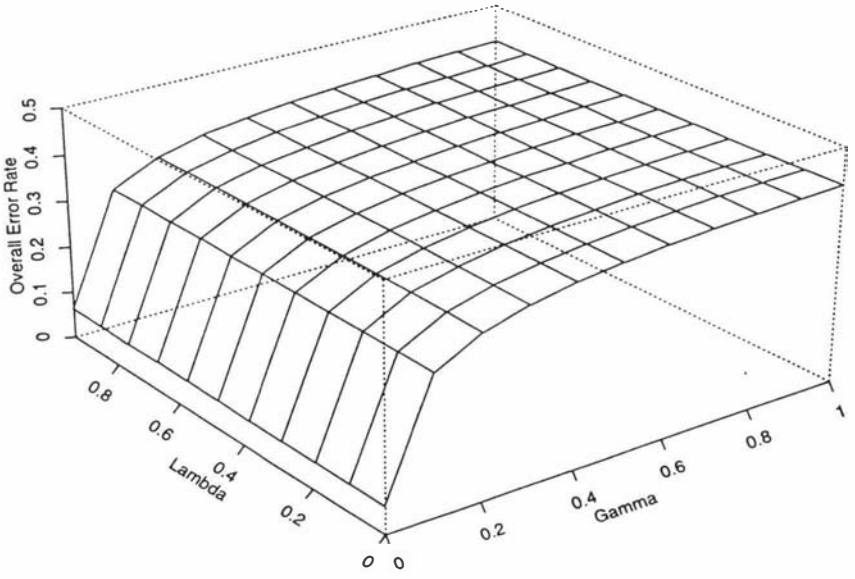


Figure 6.13: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the low variance subspace ($p = 20$). (i.e. Condition 3 in Chapter 3, Section 3.5)

However, in the sampling situation, we have seen that a high value of λ is appropriate since such shrinkage is exactly what is required. Note that the magnitude of the error rate at its minimum is again similar to that obtained in the simulation study from the minimum cross-validating error rate, around 10%.

Turning to the case of equal, highly ellipsoidal population covariance matrices, but where the mean differences are hidden in the high variance subspace (Figures 6.14, 6.15 and 6.16), the error rate surface over the (λ, γ) grid drops as γ increases. The variance-reducing effect of eigenvalue shrinkage acts primarily on the high variance subspace where the mean differences are located, to make them more identifiable for discrimination purposes.

Two other effects of the regularisation parameters exhibited by these plots (i.e. Figures 6.14, 6.15 and 6.16) differ from observations made from the simulation studies of previous chapters. Firstly, if γ remains very low, the error rate does not decrease as the covariance matrices are regularised with increasing λ, closer to the pooled covariance. In the simulation studies for these conditions, the SLDF performed much better than the SQDF, especially as p became large. The reason for this stems again from the fact that here we are dealing with population covariance matrices, which are very close together to begin with, so regularisation to the pooled covariance has little effect. In the sample situation of the simulation study,

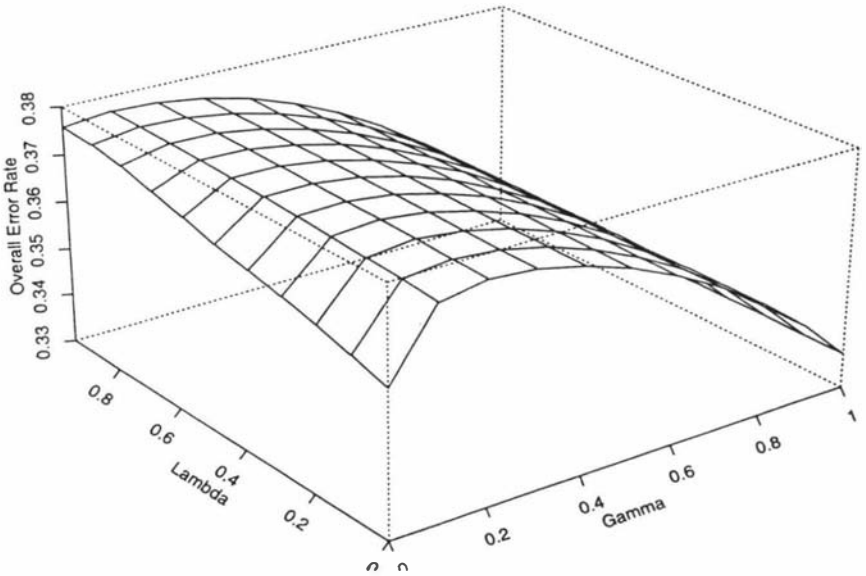


Figure 6.14: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the high variance subspace ($p = 6$). (i.e. Condition 4 in Chapter 3, Section 3.5)

the variance reducing effect of both λ and γ prove beneficial for discrimination.

The second difference is in the magnitude of the overall error rate at its minimum over the (λ, γ) grid. The minimum rate, from Figures 6.14, 6.15 and 6.16, is over 30%, whereas in the simulations studies in Chapters 3 to 5 the minimum cross-validated error rate was less than 5%. The reason for the true error rate here being so high relates to the sensitivity of the QDF to small differences in the level of variation in the high variance subspace. The discriminant function uses any disparity in the covariance matrices as an aid to discrimination. In the sample situation there can be large differences between corresponding elements of the differing covariance matrices. These differences aid in the discrimination process, and in this situation, such differences between the population covariance matrices are negligible.

In the case of unequal, highly ellipsoidal covariance matrices with zero mean differences (Figures 6.17, 6.18 and 6.19), there is a clear indication and confirmation that the appropriate values for λ and γ are close to $(0, 0)$. This corresponds to the QDF. The error rate rises markedly as λ increases, since regularising the covariances in any way towards their average results in a loss of information with which to separate the populations.

If λ remains very small while γ increases (i.e. the population covariances become

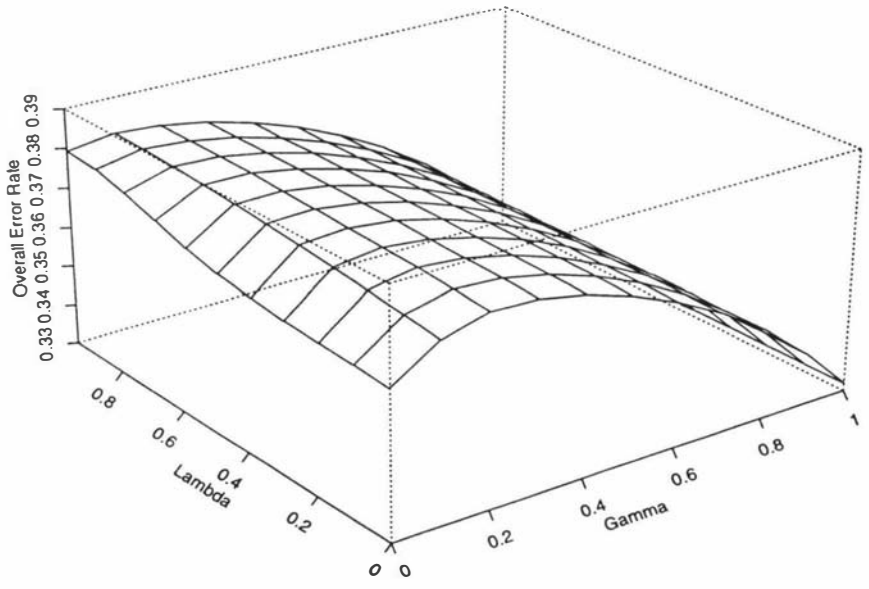


Figure 6.15: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the high variance subspace ($p = 10$). (i.e. Condition 4 in Chapter 3, Section 3.5)

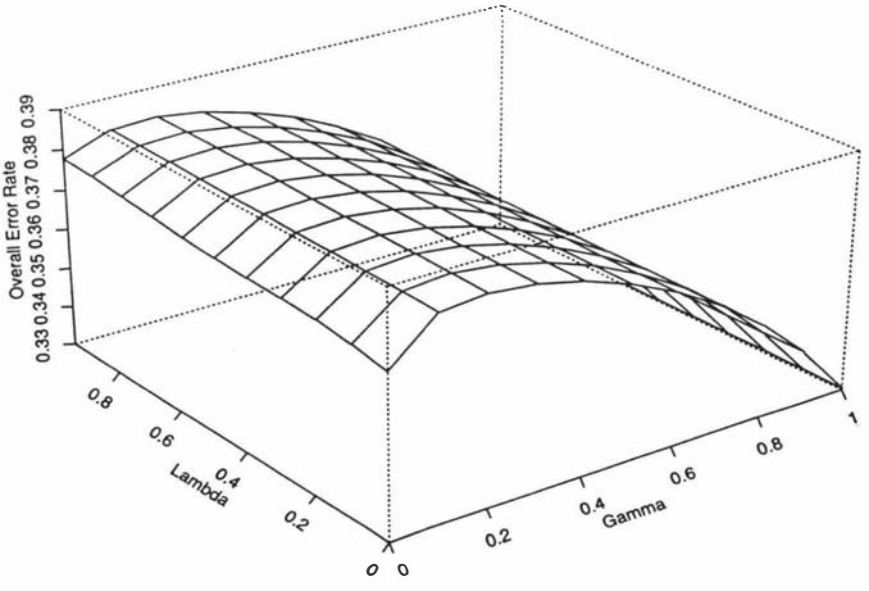


Figure 6.16: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of equal, highly ellipsoidal covariance matrices, with mean differences in the high variance subspace ($p = 20$). (i.e. Condition 4 in Chapter 3, Section 3.5)

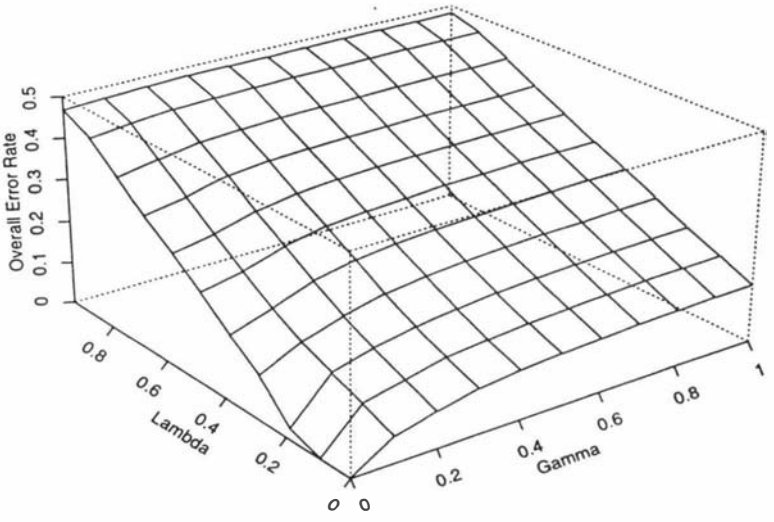


Figure 6.17: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with zero mean differences ($p = 6$). (i.e. Condition 5 in Chapter 3, Section 3.5)

spherical), the true error rate increases less dramatically, and in fact, for $p = 20$, rises only slightly. This is consistent with the findings of the simulation studies, and in fact some degree of eigenvalue shrinkage does prove to be beneficial for discrimination in practice (i.e. in the sample situation), so as to stabilise the sample covariance matrices, especially in the high dimensional setting.

The level of the true error rate at its minimum is close to zero, although it must be remembered that in this case the population means are not identical (i.e. very difficult for discrimination), but very similar, making the task of separating the groups that much easier. In the simulation study the minimum cross-validated error rate was around 14% for $p = 6, 10$, but close to zero for $p = 20$. This is consistent with the results from Friedman (1989), as well as those from Chapter 3. The covariance matrices are so different from each other that discriminating between the two populations is relatively easy when no regularisation is applied.

For the situation where the population covariances are highly ellipsoidal but the population mean differences are greater than in the previous condition, the true error rate surfaces over the (λ, γ) grid (Figures 6.20, 6.21 and 6.22) remain similar to the previous cases of zero population mean differences (Figures 6.17, 6.18 and 6.19). Hence, the optimal choice of λ and γ in this situation is again close to $(0, 0)$. In this (and in the magnitude of the error rates) there is agreement with

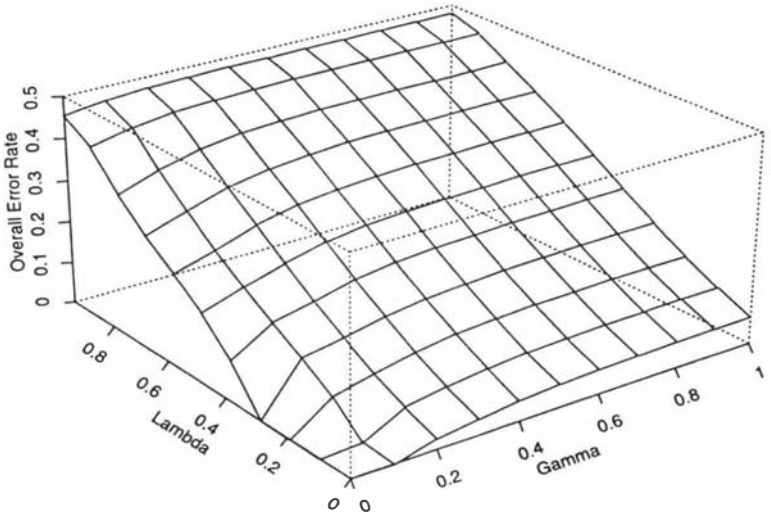


Figure 6.18: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with zero mean differences ($p = 10$). (i.e. Condition 5 in Chapter 3, Section 3.5)

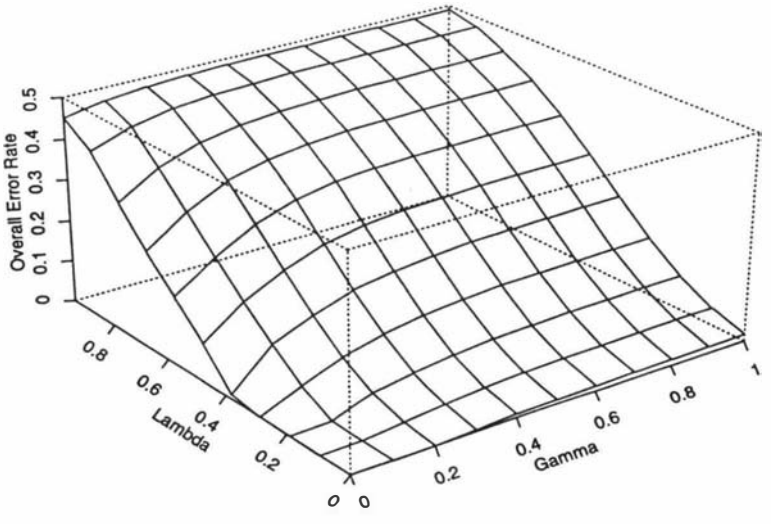


Figure 6.19: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with zero mean differences ($p = 20$). (i.e. Condition 5 in Chapter 3, Section 3.5)

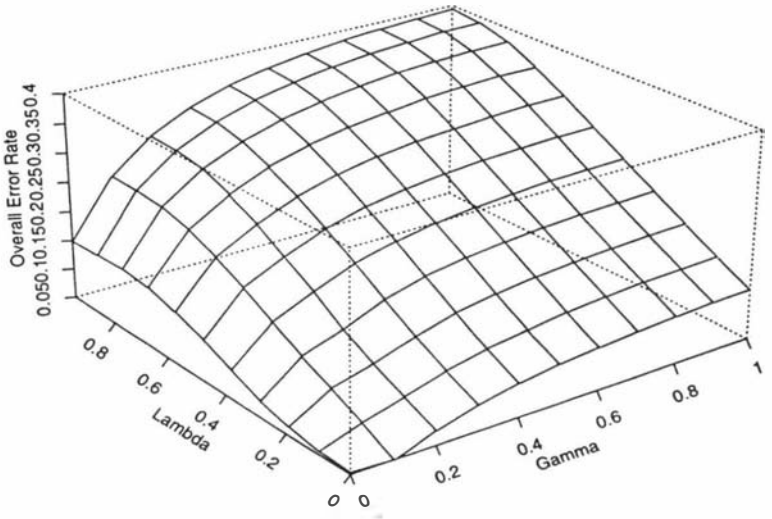


Figure 6.20: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with non-zero mean differences ($p = 6$). (i.e. Condition 6 in Chapter 3, Section 3.5)

the simulation studies of Chapters 3 to 5.

Summary

In this chapter, an attempt has been made to illustrate the effects of the regularisation parameters on the true error rate of the QDF in the univariate and multivariate situations. The primary motivation for this work was to determine if the observed (empirical) relationships between the (estimated) error rates and the regularisation parameters, as observed in previous chapters, could be confirmed by the true error rates used here. Only the covariance mixing parameter, λ, was relevant in the univariate situation, however. Expressions for the true error rate (assuming known population parameters) of the QDF in the two-population case, under conditions of unequal population means and covariances, were obtained by Houshmand (1993). For numerical purposes, in the multivariate situation, every element of each population covariance matrix had to be unequal to the corresponding element in the other covariance matrix.

Despite the fact that regularisation of the kind that is dealt with in this thesis is designed to be applied to the sample covariance matrices, many results from the simulation studies agree with the observations made in this chapter. Since there is no exact analytical expression for the conditional error rate of the SQDF, it is

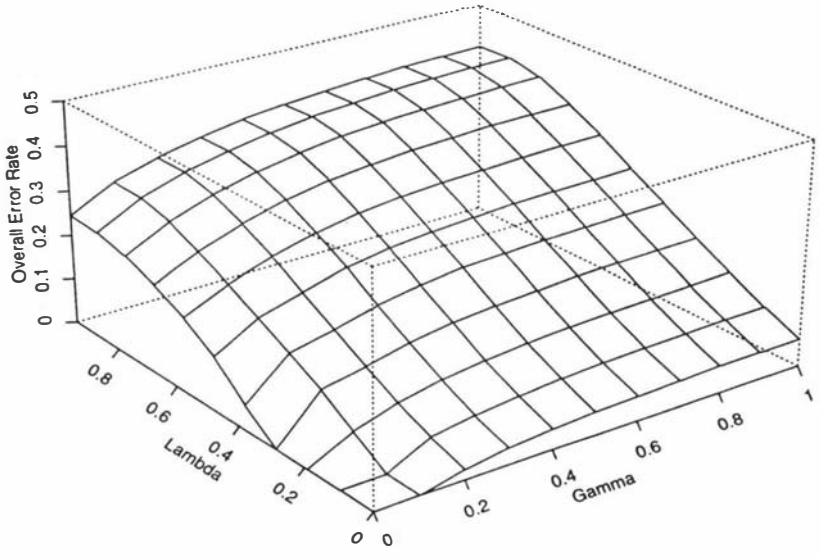


Figure 6.21: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with non-zero mean differences ($p = 10$). (i.e. Condition 6 in Chapter 3, Section 3.5)

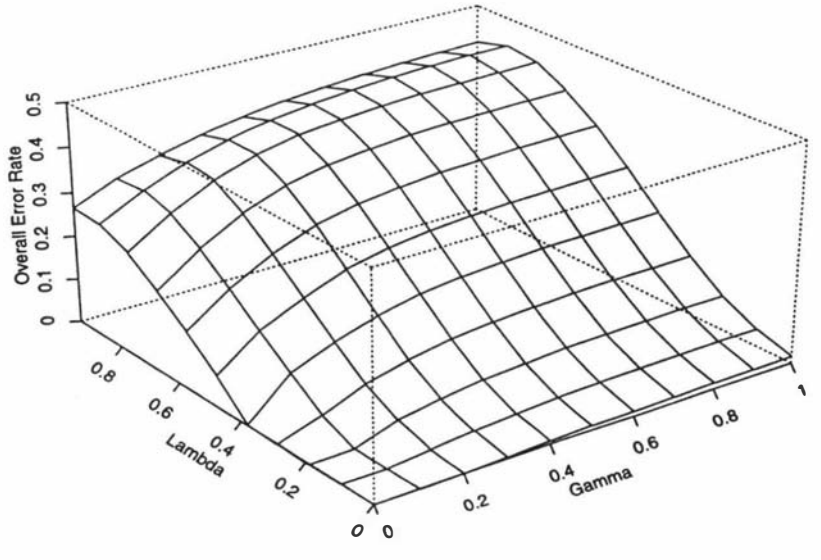


Figure 6.22: Overall error rate (P_e) versus Lambda (λ) and Gamma (γ) under conditions of unequal, highly ellipsoidal covariance matrices, with non-zero mean differences ($p = 20$). (i.e. Condition 6 in Chapter 3, Section 3.5)

necessary to employ simulation studies with which to study the effect of regularisation, as has been done in this thesis. Nevertheless, expressions such as those used in this chapter, assuming known or restricted parameter configurations, do also provide insight into the problem, as has been demonstrated here.

Chapter 7

SUMMARY

The focus in this thesis has been on addressing the problems associated with (poor) estimation of the covariance matrix in the problem of statistical discriminant analysis based on multivariate normal populations. Alternatives to the commonly used normal–based rules (i.e. sample linear discriminant function and sample quadratic discriminant function) are considered. These alternatives are more robust to the circumstances which tend to lead to poor estimation of the covariance matrix. The technique used in these alternatives is shrinkage, or regularisation, of the covariance matrix estimates towards a plausible, specified matrix.

The sample Euclidean distance function (SEDF) represents an extreme shrinkage towards the identity matrix. This function has been compared to the SLDF in several studies. Raudys and Pikelis (1980) used numerical integration assuming very restricted structures of the Σ_k (the covariance matrices of the K groups, $(k = 1 \ldots K)$). Marco et al. (1987) used Monte Carlo simulations with data generated from groups having the same specific covariance structure. Both studies showed that in terms of yielding a smaller overall misclassification error rate, the SEDF performed better than the SLDF when the dimension (p) is large in relation to the sample size. From the latter study a further conclusion was that the SEDF is preferable when the Mahalanobis distance between the groups is similar or smaller than the Euclidean distance, and when the variables in the data are mildly but positively correlated. It is shown algebraically in this thesis that the determining factor of the relative performance between the SEDF and the SLDF (i.e the relative influence of the Mahalanobis and Euclidean distances mentioned previously) is the extent and nature of the correlation among the variables. It is also shown in this thesis, via asymptotic expansions and simulation experiments, that under

both equi-correlation and auto-regressive (order 1) correlation structures for the Σ_k, negative correlations between variables lead to a large Mahalanobis distance relative to Euclidean distance. These are conditions where the SLDF would perform better than the SEDF. In particular, numerical evaluation of the asymptotic expansions for the SEDF and SLDF showed that the SEDF performed better than the SLDF in conditions of medium to high positive correlation between variables. In conditions of negative correlation between the variables, the SLDF substantially out-performed the SEDF. In the case of mild to moderate correlation there was little difference in the performances of the two classifiers.

In general, the results in this thesis from the asymptotic expansions confirm the work of Marco et al. (1987). Simulations were performed to verify the numerical evaluations of the expansions. The expansions were not evaluated for p larger than eight due to the large amount of computation time required. This is especially true for the expansion of the expected actual error rate for the SLDF, since it is rather complex. Therefore, the claim that the SEDF performs better than the SLDF when the dimension is high relative to the sample size was examined later in the thesis through further simulation studies.

The SEDF employs a crude method of regularisation of the covariance matrix estimates, yet such shrinkage is obviously beneficial in a number of situations. However, the main focus of this thesis has been the flexible regularisation facility of the sample regularised discriminant function(SRDF). Since it introduces a class of models which incorporates as special cases the SLDF, SQDF and SEDF, it ought to be a technique which yields the lowest error rate of all the rules based on the multivariate normal distribution theory. Indeed, it has been shown via simulation studies that it generally performs at least as well as the other rules, especially in the higher dimensional setting when the training sample is not large (Friedman (1989)). This is particularly true in situations when the sample size to dimension ratio is small. Further, Aeberhard et al. (1994) found the SRDF to be superior to a number of non-parametric classifiers. The results of simulations performed in this thesis under similar conditions to those in the paper by Friedman (1989) confirmed this superiority of the SRDF over the other rules, even when the group separations(i.e distances) are very small.

The success of the SRDF, however, hinges on the process which determines,

from the training data, the degree of regularisation towards the pooled sample covariance matrix, and, separately, towards (a multiple of) the identity matrix. That process involves repeated cross-validation, which is computationally intensive and which rarely leads directly to a unique "optimum" value for the regularisation parameters without a 'tie-breaking' rule. Instead, it often indicates low sensitivity to the degree of regularisation, especially with respect to regularisation towards the pooled covariance matrix. This is evident from the simulation study in Chapter 3 where regularisation rules with two different policies for breaking ties are compared. Often the degree of regularisation resulting from the different policies is quite dissimilar, yet the error rates of the constructed rules applied to a test sample of data are usually very similar.

The key to the success of the SRDF, especially when the sample size to dimension ratio is small, is the facility to regularise towards (a multiple of) the identity matrix with the γ parameter (eigenvalue shrinkage to equality). Despite the bias introduced through this facility, the reduction in variance achieved, by even a small degree of eigenvalue shrinkage, proves beneficial for discrimination in many situations, often even when the group covariance matrix eigenvalues are quite disparate. The price to be paid, however, for allowing eigenvalue shrinkage is that the SRDF lacks scale invariance. In an attempt to ascertain just how important this type of shrinkage is, a modified regularisation rule was developed and tested in a further simulation study against the SRDF and the other normal-based rules. The modified rule omits eigenvalue shrinkage but, to compensate for this in some measure, allows for a separate covariance-mixing parameter, λ_k, for each group. This would be expected to make the rule more sensitive to the data, since it sometimes occurs that the various group covariance matrices are of quite different structures, and it may be appropriate to apply covariance shrinkage to one of the group covariance matrix, but not to another. While in general it is shown that the omission of eigenvalue shrinkage clearly leads to an inferior classifier, the modified regularisation rule can result in a comparable performance to the SRDF for certain population parameter configurations.

In an article which has become known to the author at the end of this Ph.D. project, Loh (1995) studied the discrimination problem between two p – dimensional normal groups via adaptive ridge classification rules. Such a rule may be thought of as similar to the SLDF, but where the pooled covariance matrix estimate

is replaced by a regularised estimate. The regularisation used was the same as that achieved by the γ parameter in the SRDF model (i.e. towards the identity matrix). No covariance mixing was employed. A closed form solution to the adaptive parameter, which is similar to our γ, was given, in terms of the group parameters, for the case of equal group covariance matrices and equal prior probabilities. The resulting regularised value of S_p was employed to obtain the adaptive discriminant rule. A Monte Carlo simulation study compared the error rate of this rule with that of the SLDF. Two other rules were also included in the comparison. These rules involved obtaining the ridge parameter γ by re-sampling methods in a similar way to the original SRDF. The adaptive discriminant rule compared reasonably well with respect to the SLDF, but there was not much difference in performance between Loh's adaptive rule and those two which employed re-sampling methods to obtain γ. Since the justification for the adaptive ridge classification rule is asymptotic, it is not appropriate to compare it with the regularised rules we have been looking at in this thesis, which are designed to address the situation of sample sizes which are not large.

Focussing further on the model selection process of the SRDF, it is also demonstrated in this thesis that the components of the Bhattacharyya distance measure, estimated from the training sample, can give information leading to appropriate values for the regularisation parameters. These values would be more directly obtained than if computationally intensive re-sampling techniques are employed, as is the case with the original SRDF. The minimum cross-validated error-rate based on the training sample is a natural measure to use if one wishes to select the model which will yield the lowest error rate when applied to a future test sample of data from the same population. However, it is shown in this thesis (and supported by work from other researchers) that it is not necessary to determine the regularisation parameters precisely. Thus a regularisation rule is developed which bases its model selection procedure on an estimate of the Bhattacharyya distance between pairs of groups. It is shown to perform at least as well as the SRDF in most of the simulation conditions, as well as in several case studies. Computationally, the new model selection procedure is many times faster than that of the SRDF since it avoids re-sampling methods. It also leads directly to an approximate but unique regularisation model.

Finally, it is of interest to examine the effect of the regularisation parameters

λ and γ on an analytic expression of the error rate of the SQDF, which is equivalent to the SRDF when its regularisation parameters are zero. However, few of such expressions exist in the literature, and they are often complicated expressions involving approximations and limiting assumptions. Houshmand (1993) derived manageable expressions for (i) the exact overall error rate of the SQDF in the univariate case, assuming known population parameters; and (ii) the approximate overall error rate for the SQDF for multivariate normal populations, assuming known population parameters. Because of their manageable nature, these expressions were used to examine the effect on the overall error rate of the SQDF, of regularising the population covariance matrix estimates. Many of the observations from the earlier simulation studies are confirmed by these results.

As mentioned earlier, one major negative feature of the sample regularised discriminant function is its lack of scale invariance. This is certainly an area which has potential for future research. That is, to develop a scale invariant replacement for eigenvalue shrinkage, but which maintains effective covariance estimate stabilisation at high dimension and with small sample size. There is also scope for research into replacements for the identity matrix as a matrix to regularise towards. Possible options are the matrices $\boldsymbol{\Sigma}_A$ (equi-correlation) and $\boldsymbol{\Sigma}_B$ (AR(1) covariance structure) from Chapter 2, which appear to be robust enough. Alternatively, one could let the data choose, among many options, which matrix the covariance matrices should be regularised to. Certain types of discriminant analysis problems, where substantial prior information is available on the structure of the data, would be candidates for this type of approach.

The final problem which requires further research is the matter of choosing regularisation parameters using the Bhattacharyya distance, or some other distance measures. Although the heuristic algorithms developed in this thesis have been shown to work surprisingly well, it is necessary to obtain analytic results to support these empirical results. This is no trivial task, however, and is clearly an area of considerable potential for future research.

Bibliography

Aeberhard, S., Coomans, D. and de Vel, O. (1994). Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition*, **27**(8), 1065-1077.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* Second Edition. New York: Wiley.

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* New York: Springer-Verlag.

Basu, A. P., and Odell, P. L. (1974). Effects of intraclass correlation among training samples on the misclassification probabilities of Bayes' procedure. *Pattern Recognition*, **6**, 13-16.

Bayne, C. K., and Tan, W. Y. (1981). QDF misclassification probabilities for known population parameters. *Commun. Statist.-Theory Meth.*, **A10**, 2315-2326.

Bayne, C. K., Beauchamp, J. J., Kane, V. E., and McCabe, G. P. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Comput. Statist. Data Anal.*, **1**, 257-273.

Bayne, C. K., Beauchamp, J. J., and Kane, V. E. (1984). Misclassification probabilities for second-order discriminant functions used to classify bivariate normal populations. *Commun. Statist.-Simula.*, **13**, 669-673.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhya*, **A7**, 401-406.

Cochran, W. G. (1962). On the performance of the linear discriminant function (report on a discussion of a paper by W. G. Cochran). *Bull. Internatl. Stat. Inst.*, **35**, 157-158.

Dey, D. K., and Srinivasan, C. (1985). Estimation of a covarian matrix under Stein's loss. *Ann. Statist.*, **13**, 1581-1591.

Efron, B., and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.*, **4**, 22-32.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, **78**, 316-331.

Estes, S. E. (1965). Measurement selection for linear discriminant used in pattern classification. Unpublished Ph.D. thesis, Stanford University.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179-188.

Friedman, J.H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165-175.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition.* First Edition. New York: Academic Press.

Fukunaga, K. and Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI-11**, 873-885.

Ganeshanandam, S. and Krzanowski, W. J. (1990). Error-rate estimation in two-group discriminant analysis using the linear discriminant function. *J. Statist. Comput. Simul.*, **36**, 157-175.

Gilbert, E.S. (1969). The effect of unequal variance covariance matrices on Fisher's linear discriminant function. *Biometrics*, **25**, 505-515.

Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition*, **10**, 211-222.

Greene, T. and Rayens, W. (1989). Partially pooled covariance matrix estimation in discriminant analysis. *Comm. Statist. Theory Meth.*, **18** (10), 3679-3702.

Haff, L.R. (1979). Estimation of the inverse covariance matrix: random mixtures of the inverse Wishart matrix and the identity matrix. *Ann. Statist.*, **7**, 1264-1276.

Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.*, **8**, 586-597.

Han, C.P. (1969). Distribution of discriminant function when covariance matrices are proportional. *Ann. Math. Statist.*, **40**, 979-985.

Hand, D.J. (1986). Recent advances in error rate estimation. *Pattern Recognition Letters*, **4**, 335-346.

Hills, M. (1966). Allocation rules and their error rates (with discussion). *J.R. Statist. Soc.*, **B 28**, 1-31.

Hong, Z. Q. and Yang, J. Y. (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, **24**, 317-324.

Houshmand, A. A. (1993). Misclassification probabilities for quadratic discriminant function. *Commun. Statist.-Simula.*, **22**, 81-98.

Houshmand, A. A. (1995). Personal communication.

Jain, A. K. (1976). On an estimate of the Bhattacharyya distance. *IEEE Trans. Syst. Man Cybern.*, **SMC-6**, 763-766.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symp.*, (Vol. 1). Berkeley: University of California Press, pp. 361-379.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Tech.*, **COM-15**, 52-60.

Koolaard, J. P. and Lawoko, C. R. O. (1993). Estimating error rates in discriminant analysis with correlated training observations: a simulation study. *J. Statist. Comput. Simul.*, **48**, 81-99.

Koolaard, J. P. and Lawoko, C. R. O. (1994). Some results on the error rates of the Euclidean and linear discriminant functions. *Proceedings of the ORSNZ/NZSA Conference, Massey University, Palmerston North, New Zealand* (August 1994). pp 327-332.

Koolaard, J. P., Lawoko, C. R. O. and Ganesalingam, S. (1996). Regularized discriminant (classification) analysis involving Bhattacharya distance measure. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 2, Poster, pp 35-43.

Koolaard, J. P., Ganesalingam, S. and Lawoko, C. R. O. (1996). Comparison of regularised discriminant analysis with the standard discrimination methods. Paper presented to the International Biometrics Conference (IBC '96), Amsterdam, the Netherlands (July 1996). Also submitted to the Journal of Classification.

Koolaard, J. P. and Lawoko, C. R. O. (1996). The linear and Euclidean discriminant functions: a comparison via asymptotic expansions and simulation study. *Commun. Statist.- Theory Meth.* (To appear).

Lachenbruch, P.A. (1975). *Discriminant Analysis.* New York: Hafner Press.

Lachenbruch, P.A. and Mickey, M.R. (1968). Estimation of error rates in discriminant Analysis. *Technometrics*, **10**, 1-11.

Lau, Chi-Leung (1980). A simple series for the incomplete gamma integral. *Applied Statistics*, Algorithm AS 147 **29(1)**.

Lawoko, C.R.O. and McLachlan, G.J. (1983). Some asymptotic results on the effect of autocorrelation on the error rates of the sample linear discriminant function. *Pattern Recognition*, **16**, 119-121.

Lawoko, C. R. O., and Koolaard, J. P. (1996). Applications of regularised discriminant (classification) functions in the classification of objects: a discussion of potential applications to remote sensing. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 1, pp 177-184.

Lim, T. K. (1992). Comparison of the Euclidean and linear discriminant functions in statistical discriminant analysis. Unpublished M.Sc. thesis, Massey University, New Zealand.

Lin, H.E. (1979). Classification rules based on U-statistics. *Sankhya*, **B 41**, 41-52.

Lin, S. P. (1978). An improved procedure for the estimation of a correlation matrix. Dep. Math., Memphis State Univ., Memphis, TN. Tech. Rep. 78-7.

Lindsey, J.C., Herzberg, A.M. and Watts, D.G. (1987). A method for cluster analysis based on projections and quantile-quantile plots. *Biometrics*, 43, 327-341.

Loh, W. L. (1995). On linear discriminant analysis with adaptive ridge classification rules. *J. Multivar. Stat.*, 53, 264-278.

Marco, V.R., Young, D.M., and Turner, D.W. (1987). The Euclidean distance classifier: an alternative to the linear discriminant function. *Commun. Statist.-Simula.*, **16**, 485-505.

Marks, S. and Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *J. Amer. Statist. Assoc.*, **69**, 555-559.

MATLAB(1992). MATLAB reference guide. The MathWorks, Inc., Natick, Massachusetts.

McLachlan, G.J. (1972). An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function. *Austral. J. Statist.*, **14**, 68-72.

McLachlan, G.J. (1973). An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. *Austral. J. Statist.*, **15**, 210-214.

McLachlan, G.J. (1974a). The asymptotic distributions of the conditional error rate and risk in discriminant analysis. *Biometrika*, **61**, 131-135.

McLachlan, G.J. (1974b). An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics*, **30**, 239-249.

McLachlan, G.J. (1975). Some expected values for the error rates of the sample quadratic discriminant function. *Austral. J. Statist.*, **17**, 161-165.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Morant, G.M. (1923). A first study of the Tibetan skull. *Biometrika*, **14**, 193-260.

NAG(1983). *NAG Fortran Manual, Mark 10.* Numerical Algorithms Group Ltd. Oxford.

Narula, S.C. and Desu, M.M. (1981). Computation of probability of a non-central Chi-square distribution. *Applied Statistics*, Algorithm AS 170 **30(3)**.

Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.*, **34**, 1286-1301. Correction (1968). *Ann. Math. Statist.*, **39**, 1358-1359.

Patnaik, P.B. (1949). The non-central χ^2 and F-distributions and their applications. *Biometrika*, **36**, 202-232.

Peck, R. and van Ness, J. (1982). The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI-4**, 530-537.

Raudys, S. J. and Pikelis, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI-2**, 242-252.

Rayens, W and Greene, T. (1991). Covariance pooling and stabilization for classification. *Comput. Statist. Data Anal.*, **11**, 17-42.

Reaven, G.M. and Miller, R.G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, **16**, 17-24.

Seber, G.A.F. (1984). *Multivariate Observations.* New York: Wiley.

Sitgreaves, R. (1961). Some results on the W-classification statistic. In Studies in Item analysis and Prediction, (Editor: H. Solomon). Stanford: Stanford University Press, pp. 241-251.

Stein, C., Efron, B. and Morris, C. (1972). Improving the usual estimator of a normal covariance matrix. *Technical Report No. 37.* Stanford: Department of Statistics, Stanford University.

Stein, C. (1975). Estimation of a covariance matrix. Rietz Lecture notes, Annu. Meeting Amer. Statist. Assoc., Atlanta, GA.

Tubbs, J. D. (1980). Effect of autocorrelated training samples on Bayes' probabilities of misclassification. *Pattern Recognition*, **12**, 351-354.

Wahl, P.W. and Kronmal, R.A. (1977). Discriminant functios when covariance matrices are unequal and sample sizes are moderate. *Biometrics*, **33**, 479-484.

Wakaki, H. (1990). Comparison of linear and quadratic discriminant functions. *Biometrika*, **77**, 227-229.

Appendix A

ASYMPTOTIC EXPANSIONS FOR THE CONDITIONAL ERROR RATE OF LINEAR DISCRIMINANT FUNCTION UNDER CONDITIONS OF "EQUIVALENCE" .

A.1 Covariance matrix of the form $\Sigma = \Sigma_A$.

The conditional probability of misclassifying an observation from population 1 into population 2 for the LDF is $P_{21(A)}^{LDF}$ (Equation 2.10, Chapter 2). The Taylor Series expansion of this to first order approximation is

$$\begin{aligned}
\Phi(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \mathbf{S}_p) &= \Phi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\
&\quad + \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial \bar{x}_{1j}} (\bar{x}_{1j} - \mu_{1j}) + \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial \bar{x}_{2j}} (\bar{x}_{2j} - \mu_{2j}) \\
&\quad + \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial s_{ij}} (s_{ij} - \sigma_{ij}) \\
&\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}} (\bar{x}_{1i} - \mu_{1i}) (\bar{x}_{1j} - \mu_{1j}) \\
&\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i} \partial \bar{x}_{2j}} (\bar{x}_{2i} - \mu_{2i}) (\bar{x}_{2j} - \mu_{2j}) \\
&\quad + \tfrac{1}{2} \sum_{k=1}^{p} \sum_{l=1}^{p} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial s_{kl} \partial s_{ij}} (s_{kl} - \sigma_{kl}) (s_{ij} - \sigma_{ij}) \\
&\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{2j}} (\bar{x}_{1i} - \mu_{1i}) (\bar{x}_{2j} - \mu_{2j}) \\
&\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial s_{ij}} (\bar{x}_{1i} - \mu_{1i}) (s_{ij} - \sigma_{ij})
\end{aligned}$$

$$+\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial s_{ij}}\left(\bar{x}_{2i}-\mu_{2i}\right)\left(s_{ij}-\sigma_{ij}\right)$$

where

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{2j}}=\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial s_{ij}}=\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial s_{ij}}=0.$$

Taking expected values ($\mathrm{E}(.)$ denoting expectation) of the expansion yields

$$\begin{aligned}\mathrm{E}\Phi\left(\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_2,\mathbf{S}_p\right) &= \Phi\left(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}\right)\\ &+\sum_{j=1}^{p}\frac{\partial\Phi(.)}{\partial\bar{x}_{1j}}\mathrm{E}\left(\bar{x}_{1j}-\mu_{1j}\right)+\sum_{j=1}^{p}\frac{\partial\Phi(.)}{\partial\bar{x}_{2j}}\mathrm{E}\left(\bar{x}_{2j}-\mu_{2j}\right)\\ &+\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial\Phi(.)}{\partial s_{ij}}\mathrm{E}\left(s_{ij}-\sigma_{ij}\right)\\ &+\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}\mathrm{cov}\left(\bar{x}_{1i},\bar{x}_{1j}\right)\\ &+\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}\mathrm{cov}\left(\bar{x}_{2i},\bar{x}_{2j}\right)\\ &+\tfrac{1}{2}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}\mathrm{cov}\left(s_{kl},s_{ij}\right)\end{aligned}\qquad(A.1)$$

where

$$\mathrm{E}\left(\bar{x}_{1j}-\mu_{1j}\right)=\mathrm{E}\left(\bar{x}_{2j}-\mu_{2j}\right)=\mathrm{E}\left(s_{ij}-\sigma_{ij}\right)=0.$$

Therefore the following quantities are required to be obtained

$$\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}},\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}},\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}.$$

Under "equivalence", $\boldsymbol{\mu}_1=(m,m,\ldots,m)$ and $\boldsymbol{\mu}_2=\mathbf{0}$. Since $\boldsymbol{\mu}_2=\mathbf{0}$, equation (2.10) may be written as

$$P_{21(A)}^{LDF}=\Phi\left(-A\right)$$

where

$$A=\left[\bar{\boldsymbol{x}}_1{}'\mathbf{S}_p{}^{-1}\boldsymbol{\Sigma}\mathbf{S}_p{}^{-1}\bar{\boldsymbol{x}}_1\right]^{-1/2}\left[\boldsymbol{\mu}_1-\tfrac{1}{2}\bar{\boldsymbol{x}}_1{}'\mathbf{S}_p{}^{-1}\bar{\boldsymbol{x}}_1\right].$$

This expression will be used to obtain the desired quantities.

A.1.1 Obtaining $\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}$.

$$\begin{aligned}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}} &= \frac{\partial\Phi(.)}{\partial\bar{x}_{1i}}\left[-\phi(-A)\frac{\partial A}{\partial\bar{x}_{1j}}\right] \\ &= -\left[\phi(-A)\frac{\partial^2 A}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}} - A\phi(-A)\frac{\partial A}{\partial\bar{x}_{1i}}\frac{\partial A}{\partial\bar{x}_{1j}}\right] \\ &= -\phi(-A)\left[\frac{\partial^2 A}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}} - A\frac{\partial A}{\partial\bar{x}_{1i}}\frac{\partial A}{\partial\bar{x}_{1j}}\right]. \qquad (A.2)\end{aligned}$$

Now

$$\begin{aligned}\frac{\partial A}{\partial\bar{\boldsymbol{x}}_1} &= -\left[\bar{\boldsymbol{x}}_1'\mathrm{S}_p^{-1}\boldsymbol{\Sigma}\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1\right]^{-3/2}\mathrm{S}_p^{-1}\boldsymbol{\Sigma}\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1\left(\boldsymbol{\mu}_1 - \tfrac{1}{2}\bar{\boldsymbol{x}}_1\right)'\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1 \\ &\quad + \left[\bar{\boldsymbol{x}}_1'\mathrm{S}_p^{-1}\boldsymbol{\Sigma}\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1\right]^{-1/2}\mathrm{S}_p^{-1}\left(\boldsymbol{\mu}_1 - \bar{\boldsymbol{x}}_1\right)\end{aligned}$$

where

$$\bar{\boldsymbol{x}}_1'\mathrm{S}_p^{-1}\boldsymbol{\Sigma}\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1 = \sum_{w=1}^{p}\bar{x}_{1w}\left[\sum_{v=1}^{p}\left(\sum_{u=1}^{p}\bar{x}_{1u}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right] = \mathcal{B}_1$$

and

$$\mathrm{S}_p^{-1}\boldsymbol{\Sigma}\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1\left(\boldsymbol{\mu}_1 - \tfrac{1}{2}\bar{\boldsymbol{x}}_1\right)^{prime}\mathrm{S}_p^{-1}\bar{\boldsymbol{x}}_1 =$$

$$\begin{bmatrix} \left(\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{1l}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}\bar{x}_{1k}\right)\right)\left(\sum_{v=1}^{p}\left(\sum_{w=1}^{p}\left(\mu_{1w} - \frac{1}{2}\bar{x}_{1w}\right)s^{wv}\right)\bar{x}_{1v}\right) \\ \vdots \\ \left(\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{pl}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}\bar{x}_{1k}\right)\right)\left(\sum_{v=1}^{p}\left(\sum_{w=1}^{p}\left(\mu_{1w} - \frac{1}{2}\bar{x}_{1w}\right)s^{wv}\right)\bar{x}_{1v}\right) \end{bmatrix} = \mathcal{B}_2 \qquad (A.3)$$

and

$$\mathrm{S}_p^{-1}\left(\boldsymbol{\mu}_1 - \bar{\boldsymbol{x}}_1\right) = \begin{bmatrix} \sum_{l=1}^{p}s^{1l}\left(\mu_{1l} - \bar{x}_{1l}\right) \\ \vdots \\ \sum_{l=1}^{p}s^{pl}\left(\mu_{1l} - \bar{x}_{1l}\right) \end{bmatrix} = \mathcal{B}_3. \qquad (A.4)$$

Therefore

$$\frac{\partial A}{\partial\bar{\boldsymbol{x}}_1} = -\left(\mathcal{B}_1\right)^{-3/2}\left(\mathcal{B}_2\right) + \left(\mathcal{B}_1\right)^{-1/2}\left(\mathcal{B}_3\right) \qquad (A.5)$$

and hence

$$\begin{aligned}\frac{\partial A}{\partial\bar{x}_{1j}} &= \left\langle -\left(\mathcal{B}_1\right)^{-3/2} \times \left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{pl}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}\bar{x}_{1k}\right)\right\}\right. \\ &\quad \left. \times \left\{\sum_{v=1}^{p}\left(\sum_{w=1}^{p}\left(\mu_{1w} - \tfrac{1}{2}\bar{x}_{1w}\right)s^{wv}\right)\bar{x}_{1v}\right\}\right\rangle \\ &\quad + \left\langle\left(\mathcal{B}_1\right)^{-1/2}\left[\sum_{l=1}^{p}s^{jl}\left(\mu_{1l} - \bar{x}_{1l}\right)\right]\right\rangle. \qquad (A.6)\end{aligned}$$

Note that $\frac{\partial A}{\partial \bar{x}_{1i}}$ is defined similarly to $\frac{\partial A}{\partial \bar{x}_{1j}}$ but with j replaced by i. Label the quantity $\frac{\partial A}{\partial \bar{x}_{1i}}$ as equation (A.6b).

Represent the right-hand side of equation (A.6) as $\langle a_1 \times a_2 \times a_3 \rangle + \langle b_1 \rangle$, where

$$\begin{aligned}
a_1 &= -(\mathcal{B}_1)^{-3/2}, \\
a_2 &= \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{pl} \sigma_{lu} \right) \left(\sum_{k=1}^{p} s^{uk} \bar{x}_{1k} \right), \\
a_3 &= \sum_{v=1}^{p} \left(\sum_{w=1}^{p} \left(\mu_{1w} - \tfrac{1}{2} \bar{x}_{1w} \right) s^{wv} \right) \bar{x}_{1v} \\
&= \sum_{v=1}^{p} \left(\sum_{w=1}^{p} \mu_{1w} s^{wv} \right) - \tfrac{1}{2} \sum_{v=1}^{p} \bar{x}_{1v} \left(\sum_{w=1}^{p} \bar{x}_{1w} s^{wv} \right), \\
b_1 &= (\mathcal{B}_1)^{-1/2} \left[\sum_{l=1}^{p} s^{jl} (\mu_{1l} - \bar{x}_{1l}) \right].
\end{aligned}$$

If $a = a_1 \times a_2 \times a_3$, write

$$\frac{\partial^2 A}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}} = \frac{\partial}{\partial \bar{x}_{1i}} \left(\frac{\partial A}{\partial \bar{x}_{1j}} \right) = \frac{\partial a}{\partial \bar{x}_{1i}} + \frac{\partial \mathbf{b}_1}{\partial \bar{x}_{1i}}. \tag{A.7}$$

Now

$$\frac{\partial a}{\partial \bar{x}_{1i}} = a_1 a_2 \frac{\partial a_3}{\partial \bar{x}_{1i}} + a_1 a_3 \frac{\partial a_2}{\partial \bar{x}_{1i}} + a_2 a_3 \frac{\partial a_1}{\partial \bar{x}_{1i}}$$

where

$$\begin{aligned}
\frac{\partial a_1}{\partial \bar{x}_{1i}} &= \tfrac{3}{2} (\mathcal{B}_1)^{-5/2} \times \frac{\partial \mathcal{B}_1}{\partial \bar{x}_{1i}}, \\
\frac{\partial a_2}{\partial \bar{x}_{1i}} &= \sum_{u=1}^{p} s^{ui} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right), \\
\frac{\partial a_3}{\partial \bar{x}_{1i}} &= \sum_{w=1}^{p} \mu_{1w} s^{wi} - \tfrac{1}{2} \sum_{w=1}^{p} \bar{x}_{1w} s^{wi} - \tfrac{1}{2} \bar{x}_{1i} s^{ii}.
\end{aligned}$$

The asymptotic expansion being derived is to be evaluated at the point where $\bar{\boldsymbol{x}}_1 = \boldsymbol{\mu}_1$. Replacing $\bar{\boldsymbol{x}}_1$ with $\boldsymbol{\mu}_1$, some of the above expressions may be simplified:

$$A = \frac{m}{2} \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right\}^{-1/2} \left\{ \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right\} \tag{A.8}$$

$$\begin{aligned}
a_1 &= -\left\{ m^2 \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-3/2} \\
&= -(\mathcal{B}_1)^{-3/2}, \\
a_2 &= m \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(\sum_{k=1}^{p} s^{uk} \right),
\end{aligned}$$

$$
\begin{aligned}
a_3 &= m^2 \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} - \tfrac{1}{2} m^2 \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \\
&= \tfrac{1}{2} m^2 \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv}, \\
\frac{\partial \mathcal{B}_1}{\partial \bar{x}_{1i}} &= m \left\{ \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right] + \left[\sum_{v=1}^{p} s^{iv} \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right] \right\}, \\
\frac{\partial a_2}{\partial \bar{x}_{1i}} &= \sum_{u=1}^{p} s^{ui} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right), \\
\frac{\partial a_3}{\partial \bar{x}_{1i}} &= m \sum_{w=1}^{p} s^{wi} - \tfrac{1}{2} m \sum_{w=1}^{p} s^{wi} - \tfrac{1}{2} m s^{ii} \\
&= \tfrac{1}{2} m \left(\sum_{w=1}^{p} s^{wi} - s^{ii} \right).
\end{aligned}
$$

Thus, at the point where $\bar{\boldsymbol{x}}_1 = \boldsymbol{\mu}_1$,

$$
\begin{aligned}
\frac{\partial a}{\partial \bar{x}_{1i}} &= -\frac{1}{2m} \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-3/2} \\
&\quad \times \left\{ \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(\sum_{k=1}^{p} s^{uk} \right) \right\} \left\{ \sum_{w=1}^{p} s^{wi} - s^{ii} \right\} \\
&\quad + \frac{1}{2m} \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-3/2} \\
&\quad \times \left\{ \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right\} \left\{ \sum_{u=1}^{p} s^{ui} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \right\} \\
&\quad + \frac{3}{4m} \left\{ \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(\sum_{k=1}^{p} s^{uk} \right) \right\} \left\{ \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right\} \\
&\quad \times \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-5/2} \\
&\quad \times \left\{ \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right] + \left[\sum_{v=1}^{p} s^{iv} \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right] \right\}. \qquad \text{(A.9)}
\end{aligned}
$$

- Now to find $\frac{\partial b_1}{\partial \bar{x}_{1i}}$, and write it as at the point $\bar{\boldsymbol{x}}_1 = \boldsymbol{\mu}_1$.

$$
\begin{aligned}
\frac{\partial b_1}{\partial \bar{x}_{1i}} &= (\mathcal{B}_1)^{-1/2} \frac{\partial}{\partial \bar{x}_{1i}} \left[\sum_{l=1}^{p} s^{jl} (\mu_{1l} - \bar{x}_{1l}) \right] + \frac{\partial}{\partial \bar{x}_{1i}} (\mathcal{B}_1)^{-1/2} \left[\sum_{l=1}^{p} s^{jl} (\mu_{1l} - \bar{x}_{1l}) \right] \\
&= (\mathcal{B}_1)^{-1/2} [-s^{ji}] + \left[-\tfrac{1}{2} (\mathcal{B}_1)^{-3/2} \frac{\partial \mathcal{B}_1}{\partial \bar{x}_{1i}} \right] \left[\sum_{l=1}^{p} s^{jl} (\mu_{1l} - \bar{x}_{1l}) \right] \\
&= -s^{ji} \left\{ m^2 \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-1/2} \\
&= -\frac{s^{ji}}{m} \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-1/2}. \qquad \text{(A.10)}
\end{aligned}
$$

Substitute equations (A.9) (after factorising) and (A.10) into (A.7) to obtain $\frac{\partial^2 A}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}}$ at the point where $\bar{x}_1 = \mu_1$.

$$\begin{aligned}
\frac{\partial^2 A}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}} = & -\frac{s^{ji}}{m} \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-1/2} \\
& + \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-3/2} \\
& \times \left\langle \frac{3}{4m} \left\{ \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(\sum_{k=1}^{p} s^{uk} \right) \right\} \left\{ \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right\} \right. \\
& \times \left\{ \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right] \right\}^{-1} \\
& \times \left\{ \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right] + \left[\sum_{v=1}^{p} s^{iv} \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right] \right\} \\
& - \frac{1}{2m} \left\{ \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(\sum_{k=1}^{p} s^{uk} \right) \right\} \left\{ \sum_{w=1}^{p} s^{wi} - s^{ii} \right\} \\
& \left. - \frac{1}{2m} \left\{ \sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right\} \left\{ \sum_{u=1}^{p} s^{ui} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \right\} \right\rangle \qquad \text{(A.11)}
\end{aligned}$$

Substituting equations (A.8), (A.6), (A.6b) and (A.11) into equation (A.2) gives the first desired quantity of equation (A.1), namely

$$\frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}} = -\phi\left(-\mathcal{P}_1\right)\left(\mathcal{P}_2 - \mathcal{P}_1 \times \mathcal{P}_3 \times \mathcal{P}_4\right) \qquad \text{(A.12)}$$

where

$$\begin{aligned}
\mathcal{P}_1 &= A \text{ (equation(A.8))} \\
\mathcal{P}_2 &= \text{equation (A.11)} \\
\mathcal{P}_3 &= \text{equation (A.6)} \\
\mathcal{P}_4 &= \text{equation (A.6b).}
\end{aligned}$$

A.1.2 Obtaining $\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}$.

$$\begin{aligned}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}} &= \frac{\partial\Phi(.)}{\partial\bar{x}_{2i}}\left[-\phi(-A)\frac{\partial A}{\partial\bar{x}_{2j}}\right] \\ &= -\phi(-A)\left[\frac{\partial^2 A}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}} - A\frac{\partial A}{\partial\bar{x}_{2i}}\frac{\partial A}{\partial\bar{x}_{2j}}\right]. \qquad \text{(A.13)}\end{aligned}$$

Again write equation (2.10) as

$$P_{21(A)}^{LDF} = \Phi(-A)$$

where

$$A = \left[(\bar{x}_1 - \bar{x}_2)' \mathrm{S}_p^{-1}\Sigma\mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2)\right]^{-1/2}\left[\mu_1 - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)\right]' \mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2). \qquad \text{(A.14)}$$

Now

$$\begin{aligned}\frac{\partial A}{\partial\bar{x}_2} &= \left[(\bar{x}_1 - \bar{x}_2)' \mathrm{S}_p^{-1}\Sigma\mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2)\right]^{-3/2}\mathrm{S}_p^{-1}\Sigma\mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2) \\ &\quad \times \left(\mu_1 - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)\right)' \mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2) \\ &\quad + \left[(\bar{x}_1 - \bar{x}_2)' \mathrm{S}_p^{-1}\Sigma\mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2)\right]^{-1/2}\mathrm{S}_p^{-1}(-\mu_1 + \bar{x}_2)\end{aligned}$$

where

$$(\bar{x}_1 - \bar{x}_2)' \mathrm{S}_p^{-1}\Sigma\mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2) =$$

$$\sum_{w=1}^{p}(\bar{x}_{1w} - \bar{x}_{2w})\left[\sum_{v=1}^{p}\left(\sum_{u=1}^{p}(\bar{x}_{1u} - \bar{x}_{2u})s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right] = \mathcal{B}_4 \qquad \text{(A.15)}$$

and

$$\mathrm{S}_p^{-1}\Sigma\mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2)\left[\mu_1 - \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)\right]' \mathrm{S}_p^{-1}(\bar{x}_1 - \bar{x}_2) =$$

$$\begin{bmatrix} \left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{1l}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}(\bar{x}_{1k} - \bar{x}_{2k})\right)\right\} \\ \times\left\{\sum_{v=1}^{p}\left(\sum_{w=1}^{p}\left(\mu_{1w} - \frac{1}{2}(\bar{x}_{1w} - \bar{x}_{2w})\right)s^{wv}\right)(\bar{x}_{1v} - \bar{x}_{2v})\right\} \\ \vdots \\ \left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{pl}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}(\bar{x}_{1k} - \bar{x}_{2k})\right)\right\} \\ \times\left\{\sum_{v=1}^{p}\left(\sum_{w=1}^{p}\left(\mu_{1w} - \frac{1}{2}(\bar{x}_{1w} - \bar{x}_{2w})\right)s^{wv}\right)(\bar{x}_{1v} - \bar{x}_{2v})\right\}\end{bmatrix} = \mathcal{B}_5 \qquad \text{(A.16)}$$

and

$$\mathrm{S}_p^{-1}(-\mu_1 + \bar{x}_2) = \begin{bmatrix}\sum_{u=1}^{p}s^{1u}(-\mu_{1u} + \bar{x}_{2u}) \\ \vdots \\ \sum_{u=1}^{p}s^{pu}(-\mu_{1u} + \bar{x}_{2u})\end{bmatrix} = \mathcal{B}_6. \qquad \text{(A.17)}$$

Write the expression for $\frac{\partial A}{\partial \bar{x}_{2j}}$ at the point where $\bar{\boldsymbol{x}}_1 = \boldsymbol{\mu}_1$:

$$\begin{aligned}\frac{\partial A}{\partial \bar{x}_{2j}} &= \left\langle \left\{ m \sum_{w=1}^{p} (\mathcal{B}_7) - \sum_{w=1}^{p} \bar{x}_{2w} (\mathcal{B}_7) \right\}^{-3/2} \right. \\ &\quad \times \left[\sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(m \sum_{k=1}^{p} s^{uk} - \sum_{k=1}^{p} s^{uk} \bar{x}_{2k} \right) \right] \\ &\quad \left. \times \left[\tfrac{1}{2} m \sum_{v=1}^{p} \sum_{w=1}^{p} (m + \bar{x}_{2w}) s^{wv} - \tfrac{1}{2} \sum_{v=1}^{p} \bar{x}_{2v} \left(\sum_{w=1}^{p} (m + \bar{x}_{2w}) s^{wv} \right) \right] \right\rangle \\ &\quad + \left\langle \left\{ m \sum_{w=1}^{p} (\mathcal{B}_7) - \sum_{w=1}^{p} \bar{x}_{2w} (\mathcal{B}_7) \right\}^{-1/2} \right. \\ &\quad \left. times \left\{ \sum_{u=1}^{p} s^{ju} \bar{x}_{2u} - m \sum_{u=1}^{p} s^{ju} \right\} \right\rangle \end{aligned} \tag{A.18}$$

where

$$\mathcal{B}_7 = \sum_{v=1}^{p} \left\{ m \left(\sum_{u=1}^{p} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) - \left(\sum_{u=1}^{p} \bar{x}_{2u} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \right\}.$$

Now the right-hand side of equation (A.18) can be expressed as

$$\langle a_4 \times a_5 \times a_6 \rangle + \langle b_2 \times b_3 \rangle .$$

and so

$$\begin{aligned}\frac{\partial^2 A}{\partial \bar{x}_{2i} \partial \bar{x}_{2j}} &= \frac{\partial}{\partial \bar{x}_{2i}} \left(\frac{\partial A}{\partial \bar{x}_{2j}} \right) \\ &= \left(a_4 a_5 \frac{\partial a_6}{\partial \bar{x}_{2i}} + a_4 a_6 \frac{\partial a_5}{\partial \bar{x}_{2i}} + a_5 a_6 \frac{\partial a_4}{\partial \bar{x}_{2i}} \right) \\ &\quad + \left(b_2 \frac{\partial b_3}{\partial \bar{x}_{1i}} + b_3 \frac{\partial b_2}{\partial \bar{x}_{1i}} \right). \end{aligned} \tag{A.19}$$

where

$$\frac{\partial a_4}{\partial \bar{x}_{2i}} = \frac{\partial \mathcal{B}_4^{-3/2}}{\partial \bar{x}_{2i}} = -\tfrac{3}{2} (\mathcal{B}_4)\}^{-5/2} \times \frac{\partial \mathcal{B}_4}{\partial \bar{x}_{2i}},$$

where

$$\begin{aligned}\frac{\partial \mathcal{B}_4}{\partial \bar{x}_{2i}} &= \bar{x}_{2i} \sum_{v=1}^{p} s^{iv} \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) - m \sum_{w=1}^{p} \sum_{v=1}^{p} s^{iv} \left(\sum_{u=1}^{p} \sigma_{vu} s^{uw} \right) \\ &\quad - \sum_{v=1}^{p} \left\{ m \sum_{u=1}^{p} s^{uv} \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) - \left(\sum_{u=1}^{p} \bar{x}_{2u} s^{uv} \right) \left(\sum_{u=1}^{p} \sigma_{vu} s^{ui} \right) \right\}, \\ \frac{\partial a_5}{\partial \bar{x}_{2i}} &= \frac{\partial}{\partial \bar{x}_{2i}} \left[\sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) \left(m \sum_{k=1}^{p} s^{uk} - \sum_{k=1}^{p} s^{uk} \bar{x}_{2k} \right) \right] - \sum_{u=1}^{p} \left(\sum_{l=1}^{p} s^{jl} \sigma_{lu} \right) s^{ui}, \end{aligned}$$

$$\begin{aligned}
\frac{\partial a_6}{\partial \bar{x}_{2i}} &= \frac{\partial}{\partial \bar{x}_{2i}}\left[\tfrac{1}{2}m\sum_{v=1}^{p}\left(\sum_{w=1}^{p}\left(m+\bar{x}_{2w}\right)s^{wv}\right)-\tfrac{1}{2}\sum_{v=1}^{p}\bar{x}_{2v}\left(\sum_{w=1}^{p}\left(m+\bar{x}_{2w}\right)s^{wv}\right)\right] \\
&= \tfrac{1}{2}m\sum_{v=1}^{p}s^{iv}-\tfrac{1}{2}m\sum_{w=1}^{p}s^{wi}-\tfrac{1}{2}\sum_{w=1}^{p}\bar{x}_{2w}s^{wi}-\tfrac{1}{2}\bar{x}_{2i}\sum_{w=1}^{p}s^{wi}, \\
\frac{\partial b_2}{\partial \bar{x}_{2i}} &= \frac{\partial\left(\mathcal{B}_4\right)^{-1/2}}{\partial \bar{x}_{2i}}=-\tfrac{1}{2}\left(\mathcal{B}_4\right)^{-3/2}\times\frac{\partial \mathcal{B}_4}{\partial \bar{x}_{2i}}, \\
&\text{and} \\
\frac{\partial b_3}{\partial \bar{x}_{2i}} &= \frac{\partial}{\partial \bar{x}_{2i}}\left[\sum_{u=1}^{p}s^{ju}\bar{x}_{2u}-m\sum_{u=1}^{p}s^{ju}\right]=s^{ji}.
\end{aligned}$$

Since $\bar{x}_2=\mu_2=0$, some of the above components of $\frac{\partial^2 A}{\partial \bar{x}_{2i}\partial \bar{x}_{2j}}$ may be simplified:

$$\begin{aligned}
a_4 &= \left\{m^2\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-3/2}, \\
a_5 &= m\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{jl}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}\right), \\
a_6 &= \tfrac{1}{2}m^2\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}, \\
b_2 &= \left\{m^2\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-1/2}, \\
b_3 &= -m\sum_{u=1}^{p}s^{ju}, \\
\mathcal{B}_4 &= m^2\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right), \\
\frac{\partial \mathcal{B}_4}{\partial \bar{x}_{2i}} &= -\sum_{v=1}^{p}\left\{m\sum_{u=1}^{p}s^{uv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{ui}\right)\right\}-m\sum_{w=1}^{p}\sum_{v=1}^{p}s^{iv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right) \\
&= -m\left\{\sum_{v=1}^{p}\sum_{u=1}^{p}s^{uv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{ui}\right)+\sum_{w=1}^{p}\sum_{v=1}^{p}s^{iv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}, \\
\frac{\partial a_6}{\partial \bar{x}_{2i}} &= \tfrac{1}{2}m\left(\sum_{v=1}^{p}s^{iv}-\sum_{w=1}^{p}s^{wi}\right).
\end{aligned}$$

Inserting all the above into equation (A.19) yields

$$\begin{aligned}
\frac{\partial^2 A}{\partial \bar{x}_{2i}\partial \bar{x}_{2j}} &= \frac{1}{2m}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-3/2} \\
&\quad\times\left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p}s^{jl}\sigma_{lu}\right)\left(\sum_{k=1}^{p}s^{uk}\right)\right\}\left\{\sum_{u=1}^{p}s^{iv}-\sum_{w=1}^{p}s^{wi}\right\} \\
&\quad-\frac{1}{2m}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-3/2}
\end{aligned}$$

$$
\begin{aligned}
&\times \left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\}\left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p} s^{jl}\sigma_{lu}\right)s^{ui}\right\} \\
&+\frac{3}{4m}\left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p} s^{jl}\sigma_{lu}\right)\left(\sum_{k=1}^{p} s^{uk}\right)\right\}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\} \\
&\times \left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-5/2} \\
&\times \left\{\sum_{v=1}^{p}\sum_{u=1}^{p} s^{uv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{ui}\right)+\sum_{w=1}^{p}\sum_{v=1}^{p} s^{iv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\} \\
&+\frac{s^{ji}}{m}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-1/2} \\
&-\frac{1}{2m}\left\{\sum_{u=1}^{p} s^{ju}\right\}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-3/2} \\
&\times \left\{\sum_{v=1}^{p}\sum_{u=1}^{p} s^{uv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{ui}\right)+\sum_{w=1}^{p}\sum_{v=1}^{p} s^{iv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\} \\
=\ &\frac{1}{2m}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-3/2} \\
&\times \left\langle\left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p} s^{jl}\sigma_{lu}\right)\left(\sum_{k=1}^{p} s^{uk}\right)\right\}\left\{\sum_{v=1}^{p} s^{iv}-\sum_{w=1}^{p} s^{wi}\right\}\right. \\
&-\left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\}\left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p} s^{jl}\sigma_{lu}\right)s^{ui}\right\} \\
&+\tfrac{3}{2}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-1} \\
&\times \left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p} s^{jl}\sigma_{lu}\right)\left(\sum_{k=1}^{p} s^{uk}\right)\right\}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\} \\
&\times \left\{\sum_{v=1}^{p}\sum_{u=1}^{p} s^{uv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{ui}\right)+\sum_{w=1}^{p}\sum_{v=1}^{p} s^{iv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}. \\
&+2s^{ji}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}-\left\{\sum_{u=1}^{p} s^{ju}\right\} \\
&\times \left.\left\{\sum_{v=1}^{p}\sum_{u=1}^{p} s^{uv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{ui}\right)+\sum_{w=1}^{p}\sum_{v=1}^{p} s^{iv}\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}\right\rangle \qquad \text{(A.20)}
\end{aligned}
$$

The other components of equation (A.13), evaluated at the point where $\bar{\boldsymbol{x}}_1 = \boldsymbol{\mu}_1$ and $\bar{\boldsymbol{x}}_2 = \boldsymbol{\mu}_2$ are

$$
\begin{aligned}
\frac{\partial A}{\partial \bar{x}_{2j}} \ =\ &\tfrac{1}{2}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-3/2} \\
&\times \left\{\sum_{u=1}^{p}\left(\sum_{l=1}^{p} s^{jl}\sigma_{lu}\right)\left(\sum_{k=1}^{p} s^{uk}\right)\right\}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\}
\end{aligned}
$$

$$-\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(\sum_{u=1}^{p}\sigma_{vu}s^{uw}\right)\right\}^{-1/2}\left\{\sum_{u=1}^{p}s^{ju}\right\} \qquad \text{(A.21)}$$

Note that $\frac{\partial A}{\partial \bar{x}_{2i}}$ is defined similarly to $\frac{\partial A}{\partial \bar{x}_{2j}}$ but with j replaced by i. Label the quantity $\frac{\partial A}{\partial \bar{x}_{2i}}$ as equation (A.21b). Substituting equations (A.20), (A.21), (A.21b) and (A.8) into (A.13) yields the second desired quantity of equation (A.1), namely

$$\frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i}\partial \bar{x}_{2j}} = -\phi\left(-\mathcal{P}_1\right)\left(\mathcal{P}_5 - \mathcal{P}_1 \times \mathcal{P}_6 \times \mathcal{P}_7\right) \qquad \text{(A.22)}$$

where

$$\begin{aligned}
\mathcal{P}_1 &= A \text{ (equation(A.8))} \\
\mathcal{P}_5 &= \text{equation (A.20)} \\
\mathcal{P}_6 &= \text{equation (A.21)} \\
\mathcal{P}_7 &= \text{equation (A.21b).}
\end{aligned}$$

A.1.3 Obtaining $\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}$.

Under equivalence and with $\Sigma = \Sigma_A$, again write equation (2.10) as

$$P_{21(A)}^{LDF} = \Phi(-A),$$

and at the point where $\bar{x}_1 = \mu_1$ and $\bar{x}_2 = \mu_2$, write A as

$$A = \tfrac{1}{2}m\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw} + \rho\sum_{u\neq v=1}^{p} s^{uw}\right)\right]^{-1/2} \tag{A.23}$$

Now

$$\begin{aligned}\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}} &= \frac{\partial}{\partial s_{kl}}\left\{\frac{\partial\Phi(-A)}{\partial s_{ij}}\right\} \\ &= \frac{\partial}{\partial s_{kl}}\left\{-\tfrac{1}{2}m\phi(-A)\times\mathcal{B}_9\right\} \\ &= -\tfrac{1}{2}m\left[\phi(-A)\mathcal{B}_8 + \mathcal{B}_{10}\times\mathcal{B}_9\right] \end{aligned} \tag{A.24}$$

where

$$\begin{aligned}\mathcal{B}_8 &= \frac{\partial\mathcal{B}_9}{\partial s_{kl}}, \\ \mathcal{B}_9 &= \frac{\partial}{\partial s_{ij}}\left\{\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)(\mathcal{B}_{11})^{-1/2}\right\}, \\ \mathcal{B}_{10} &= \frac{\partial\phi(-A)}{\partial s_{kl}}, \end{aligned}$$

and

$$\mathcal{B}_{11} = \sum_{w=1}^{p}\sum_{u=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw} + \rho\sum_{u\neq v=1}^{p} s^{uw}\right).$$

Now

$$\mathcal{B}_9 = \left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\frac{\partial(\mathcal{B}_{11})^{-1/2}}{\partial s_{ij}} + (\mathcal{B}_{11})^{-1/2}\frac{\partial}{\partial s_{ij}}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\}$$

where

$$\begin{aligned}\frac{\partial(\mathcal{B}_{11})^{-1/2}}{\partial s_{ij}} &= -\tfrac{1}{2}(\mathcal{B}_{11})^{-3/2}\frac{\partial}{\partial s_{ij}}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw} + \rho\sum_{u\neq v=1}^{p} s^{uw}\right)\right\} \\ &= -\tfrac{1}{2}(\mathcal{B}_{11})^{-3/2}\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\left(\sum_{u=1}^{p} s^{uv}\right)\left(\frac{\partial s^{vw}}{\partial s_{ij}} + \rho\sum_{u\neq v=1}^{p}\frac{\partial s^{uw}}{\partial s_{ij}}\right)\right. \\ &\quad \left. + \left(s^{vw} + \rho\sum_{u\neq v=1}^{p} s^{uw}\right)\left(\sum_{u=1}^{p}\frac{\partial s^{uv}}{\partial s_{ij}}\right)\right] \end{aligned}$$

$$= -\tfrac{1}{2}(\mathcal{B}_{11})^{-3/2} w_0 \sum_{w=1}^{p}\sum_{v=1}^{p}\left[\left(\sum_{u=1}^{p} s^{uv}\right)\right.$$
$$\left\{(s^{vi}s^{jw}+s^{vj}s^{iw})+\rho\sum_{u\neq v=1}^{p}(s^{ui}s^{jw}+s^{uj}s^{iw})\right\}$$
$$\left.+\left(s^{vw}+\rho\sum_{u\neq v=1}^{p}s^{uw}\right)\left(\sum_{u=1}^{p}(s^{ui}s^{jv}+s^{uj}s^{iv})\right)\right],$$

$$\frac{\partial}{\partial s_{ij}}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right\} = w_0\sum_{v=1}^{p}\sum_{w=1}^{p}(s^{wi}s^{jv}+s^{wj}s^{iv})$$

and

$$w_0 = \begin{cases} -0.5 & \text{if } i=j \\ -1 & \text{if } i\neq j. \end{cases}$$

Write

$$\mathcal{B}_9 = w_0(a_8)^{-3/2}\left[-\tfrac{1}{2}a_7a_9+a_8a_{10}\right] \tag{A.25}$$

where

$$a_7 = \sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}$$
$$a_8 = \sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\rho\sum_{u\neq v=1}^{p}s^{uw}\right)$$
$$a_{10} = \sum_{v=1}^{p}\sum_{w=1}^{p}(s^{wi}s^{jv}+s^{wj}s^{iv})$$
$$a_9 = \sum_{w=1}^{p}\sum_{v=1}^{p}\left[\left(\sum_{u=1}^{p}s^{uv}\right)(a_{9a}+a_{9b})+a_{9c}a_{9d}\right]$$

where

$$a_{9a} = s^{vi}s^{jw}+s^{vj}s^{iw}$$
$$a_{9b} = \rho\sum_{u\neq v=1}^{p}(s^{ui}s^{jw}+s^{uj}s^{iw})$$
$$a_{9c} = s^{vw}+\rho\sum_{u\neq v=1}^{p}s^{uw}$$
$$a_{9d} = \sum_{u=1}^{p}(s^{ui}s^{jv}+s^{uj}s^{iv}).$$

To find $\mathcal{B}_8$, write

$$\mathcal{B}_8 = \frac{\partial^2}{\partial s_{kl}\partial s_{ij}}\left\{\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\rho\sum_{u\neq v=1}^{p}s^{uw}\right)\right]^{-1/2}\right\}$$

$$
\begin{aligned}
&= \frac{\partial \mathcal{B}_9}{\partial s_{kl}} \\
&= \frac{\partial}{\partial s_{kl}} \left\{ w_0 \left(a_8\right)^{-3/2} \left[-\tfrac{1}{2} a_7 a_9 + a_8 a_{10} \right] \right\} \\
&= w_0 \left[\left(a_8\right)^{-3/2} \frac{\partial}{\partial s_{kl}} \left\{ -\tfrac{1}{2} a_7 a_9 + a_8 a_{10} \right\} \right. \\
&\quad \left. + \left(-\tfrac{1}{2} a_7 a_9 + a_8 a_{10} \right) \frac{\partial \left(a_8\right)^{-3/2}}{\partial s_{kl}} \right]
\end{aligned}
$$

where

$$
\frac{\partial}{\partial s_{kl}} \left\{ -\tfrac{1}{2} a_7 a_9 + a_8 a_{10} \right\} = -\tfrac{1}{2} \left[a_7 \frac{\partial a_9}{\partial s_{kl}} + a_9 \frac{\partial a_7}{\partial s_{kl}} \right] + a_8 \frac{\partial a_{10}}{\partial s_{kl}} + a_{10} \frac{\partial a_8}{\partial s_{kl}}
$$

and

$$
\frac{\partial \left(a_8\right)^{-3/2}}{\partial s_{kl}} = -\tfrac{3}{2} \left(a_8\right)^{-5/2} \frac{\partial a_8}{\partial s_{kl}}.
$$

Now

$$
\begin{aligned}
\frac{\partial a_7}{\partial s_{kl}} &= w_1 \sum_{v=1}^{p} \sum_{w=1}^{p} \left(s^{wk} s^{lv} + s^{wl} s^{kv} \right), \\
\frac{\partial a_8}{\partial s_{kl}} &= w_1 \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vk} s^{lw} + s^{vl} s^{kw} + \rho \sum_{u \neq v=1}^{p} \left(s^{uk} s^{lw} + s^{ul} s^{kw} \right) \right) \right. \\
&\quad \left. + \left(s^{vw} + \rho \sum_{u \neq v=1}^{p} s^{uw} \right) \sum_{u=1}^{p} \left(s^{uk} s^{lv} + s^{ul} s^{kv} \right) \right], \\
\frac{\partial a_{10}}{\partial s_{kl}} &= w_1 \sum_{v=1}^{p} \sum_{w=1}^{p} \left[s^{wi} \left(s^{jk} s^{lv} + s^{jl} s^{kv} \right) + s^{jv} \left(s^{wk} s^{li} + s^{wl} s^{ki} \right) \right. \\
&\quad \left. + s^{wj} \left(s^{ik} s^{lv} + s^{il} s^{kv} \right) + s^{iv} \left(s^{wk} s^{lj} + s^{wl} s^{kj} \right) \right] \\
\frac{\partial a_9}{\partial s_{kl}} &= \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\frac{\partial a_{9a}}{\partial s_{kl}} + \frac{\partial a_{9b}}{\partial s_{kl}} \right) \right. \\
&\quad + \left(a_{9a} + a_{9b} \right) \left(\sum_{u=1}^{p} \frac{\partial s^{uv}}{\partial s_{kl}} \right) \\
&\quad \left. + a_{9c} \frac{\partial a_{9d}}{\partial s_{kl}} + a_{9d} \frac{\partial a_{9c}}{\partial s_{kl}} \right]
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial a_{9a}}{\partial s_{kl}} &= w_1 \left[s^{vi} \left(s^{jk} s^{lw} + s^{jl} s^{kw} \right) + s^{jw} \left(s^{vk} s^{li} + s^{vl} s^{ki} \right) \right. \\
&\quad \left. + s^{vj} \left(s^{ik} s^{lw} + s^{il} s^{kw} \right) + s^{iw} \left(s^{vk} s^{lj} + s^{vl} s^{kj} \right) \right],
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial a_{9b}}{\partial s_{kl}} &= w_1 \rho \sum_{u \neq v=1}^{p} \left[s^{ui} \left(s^{jk} s^{lw} + s^{jl} s^{kw} \right) + s^{jw} \left(s^{uk} s^{li} + s^{ul} s^{ki} \right) \right. \\
&\quad \left. + s^{uj} \left(s^{ik} s^{lw} + s^{il} s^{kw} \right) + s^{iw} \left(s^{uk} s^{lj} + s^{ul} s^{kj} \right) \right], \\
\frac{\partial a_{9c}}{\partial s_{kl}} &= w_1 \left[\left(s^{vk} s^{lw} + s^{vl} s^{kw} \right) + \rho \sum_{u \neq v=1}^{p} \left(s^{uk} s^{lw} + s^{ul} s^{kw} \right) \right], \\
\frac{\partial a_{9d}}{\partial s_{kl}} &= w_1 \sum_{u=1}^{p} \left[s^{ui} \left(s^{jk} s^{lv} + s^{jl} s^{kv} \right) + s^{jv} \left(s^{uk} s^{li} + s^{ul} s^{ki} \right) \right. \\
&\quad \left. + s^{uj} \left(s^{ik} s^{lv} + s^{il} s^{kv} \right) + s^{iv} \left(s^{uk} s^{lj} + s^{ul} s^{kj} \right) \right], \\
\sum_{u=1}^{p} \frac{\partial s^{uv}}{\partial s_{kl}} &= w_1 \sum_{u=1}^{p} \left(s^{uk} s^{lv} + s^{ul} s^{kv} \right)
\end{aligned}
$$

and

$$
w_1 = \begin{cases} -0.5 & \text{if } k = l \\ -1 & \text{if } k \neq l. \end{cases}
$$

Write $\mathcal{B}_8$ as

$$
\begin{aligned}
\mathcal{B}_8 &= w_0 \left[(a_8)^{-3/2} \left(-\tfrac{1}{2} \left[a_7 \frac{\partial a_9}{\partial s_{kl}} + a_9 \frac{\partial a_7}{\partial s_{kl}} \right] + a_8 \frac{\partial a_{10}}{\partial s_{kl}} + a_{10} \frac{\partial a_8}{\partial s_{kl}} \right) \right. \\
&\quad \left. + \left(-\tfrac{1}{2} a_7 a_9 + a_8 a_{10} \right) \left(-\tfrac{3}{2} \right) (a_8)^{-5/2} \frac{\partial a_8}{\partial s_{kl}} \right] \qquad \text{(A.26)}
\end{aligned}
$$

Express $\mathcal{B}_{10}$ as

$$
\begin{aligned}
\mathcal{B}_{10} &= \frac{\partial \phi(-A)}{\partial s_{kl}} \\
&= -A \phi(-A) \frac{\partial A}{\partial s_{kl}} \\
&= -\tfrac{1}{2} m \left(\sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right) \left[\sum_{w=1}^{p} \sum_{v=1}^{p} \left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vw} + \rho \sum_{u \neq v=1}^{p} s^{uw} \right) \right]^{-1/2} \times \phi(-A) \left(\tfrac{1}{2} m \right) \\
&\quad \times \frac{\partial}{\partial s_{kl}} \left\{ \left(\sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right) \left[\sum_{w=1}^{p} \sum_{v=1}^{p} \left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vw} + \rho \sum_{u \neq v=1}^{p} s^{uw} \right) \right]^{-1/2} \right\} \\
&= -\tfrac{1}{4} m^2 \left(\sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right) \left[\sum_{w=1}^{p} \sum_{v=1}^{p} \left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vw} + \rho \sum_{u \neq v=1}^{p} s^{uw} \right) \right]^{-1/2} \\
&\quad \times \phi(-A) w_1 \left[\sum_{w=1}^{p} \sum_{v=1}^{p} \left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vw} + \rho \sum_{u \neq v=1}^{p} s^{uw} \right) \right]^{-3/2}
\end{aligned}
$$

$$
\begin{aligned}
&\times \left\langle -\tfrac{1}{2}\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\right. \\
&\times \left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\sum_{u=1}^{p} s^{uv}\left(s^{vk}s^{lw}+s^{vl}s^{kw}+\rho\sum_{u\neq v=1}^{p}\left(s^{uk}s^{lw}+s^{ul}s^{kw}\right)\right)\right.\right. \\
&\left.\left.+\left(s^{vw}+\rho\sum_{u\neq v=1}^{p} s^{uw}\right)\sum_{u=1}^{p}\left(s^{uk}s^{lv}+s^{ul}s^{kv}\right)\right]\right\} \\
&\left.+\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw}+\rho\sum_{u\neq v=1}^{p} s^{uw}\right)\right]\left[\sum_{v=1}^{p}\sum_{w=1}^{p}\left(s^{wk}s^{lv}+s^{wl}s^{kv}\right)\right]\right\rangle \\
= \; & -\tfrac{1}{4}m^2\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\phi(-A)w_1\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw}+\rho\sum_{u\neq v=1}^{p} s^{uw}\right)\right]^{-2} \\
&\times \left\langle -\tfrac{1}{2}\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\sum_{u=1}^{p} s^{uv}\left(s^{vk}s^{lw}+s^{vl}s^{kw}\right)\right.\right.\right. \\
&\left.+\rho\sum_{u\neq v=1}^{p}\left(s^{uk}s^{lw}+s^{ul}s^{kw}\right)\right. \\
&\left.\left.+\left(s^{vw}+\rho\sum_{u\neq v=1}^{p} s^{uw}\right)\sum_{u=1}^{p}\left(s^{uk}s^{lv}+s^{ul}s^{kv}\right)\right]\right\} \\
&+\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw}+\rho\sum_{u\neq v=1}^{p} s^{uw}\right)\right] \\
&\left.\times\left[\sum_{v=1}^{p}\sum_{w=1}^{p}\left(s^{wk}s^{lv}+s^{wl}s^{kv}\right)\right]\right\rangle \qquad \text{(A.27)}
\end{aligned}
$$

Thus

$$\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}} = -\tfrac{1}{2}m\left[\phi(-A)\times\mathcal{B}_8+\mathcal{B}_{10}\times\mathcal{B}_9\right] \qquad \text{(A.28)}$$

where A is as in equation (A.23), $\mathcal{B}_8$ is as in equation (A.26), $\mathcal{B}_9$ is as in equation (A.25) and $\mathcal{B}_{10}$ is as in equation (A.27).

A.2 Covariance matrix of the form $\Sigma = \Sigma_B$.

Under equivalence and with $\Sigma = \Sigma_B$, again write equation (2.10) as

$$P_{21(A)}^{LDF} = \Phi(-A),$$

and at the point where $\bar{x}_1 = \mu_1$ and $\bar{x}_2 = \mu_2$, write A as

$$A = \tfrac{1}{2}m\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw} + \sum_{u\neq v=1}^{p} s^{uw}\rho^{|v-u|}\right)\right]^{-1/2} \tag{A.29}$$

The only partial derivative term which differs from the expansion in the previous section is that involving differentiation with respect to elements of Σ, that is, $\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}$. In an analogous expression to equation (A.24),

$$\begin{aligned}\frac{\partial^2\Phi(-A)}{\partial s_{kl}\partial s_{ij}} &= \frac{\partial}{\partial s_{kl}}\left\{\frac{\partial\Phi(-A)}{\partial s_{ij}}\right\} \\ &= \frac{\partial}{\partial s_{kl}}\left\{-\tfrac{1}{2}m\phi(-A)\times C_2\right\} \\ &= -\tfrac{1}{2}m\left[\phi(-A)\,C_1 + C_3\times C_2\right] \end{aligned} \tag{A.30}$$

where

$$\begin{aligned} C_1 &= \frac{\partial C_2}{\partial s_{kl}}, \\ C_2 &= \frac{\partial}{\partial s_{ij}}\left\{\left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)(C_4)^{-1/2}\right\} \\ C_3 &= \frac{\partial\phi(-A)}{\partial s_{kl}}, \end{aligned}$$

and

$$C_4 = \sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p} s^{uv}\right)\left(s^{vw} + \sum_{u\neq v=1}^{p} s^{uw}\rho^{|v-u|}\right).$$

Now

$$C_2 = \left(\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right)\frac{\partial(C_4)^{-1/2}}{\partial s_{ij}} + (C_4)^{-1/2}\frac{\partial}{\partial s_{ij}}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p} s^{wv}\right\}$$

where

$$\begin{aligned}
\frac{\partial (C_4)^{-1/2}}{\partial s_{ij}} &= -\tfrac{1}{2}(C_4)^{-3/2}\frac{\partial}{\partial s_{ij}}\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right\} \\
&= -\tfrac{1}{2}(C_4)^{-3/2}\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\left(\sum_{u=1}^{p}s^{uv}\right)\left(\frac{\partial s^{vw}}{\partial s_{ij}}+\sum_{u\neq v=1}^{p}\frac{\partial s^{uw}}{\partial s_{ij}}\rho^{|v-u|}\right)\right. \\
&\quad \left.+\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\left(\sum_{u=1}^{p}\frac{\partial s^{uv}}{\partial s_{ij}}\right)\right] \\
&= -\tfrac{1}{2}(C_4)^{-3/2}w_0\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\left(\sum_{u=1}^{p}s^{uv}\right)\right. \\
&\quad \left\{\left(s^{vi}s^{jw}+s^{vj}s^{iw}\right)+\sum_{u\neq v=1}^{p}\left(\left(s^{ui}s^{jw}+s^{uj}s^{iw}\right)\rho^{|v-u|}\right)\right\} \\
&\quad \left.+\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\left(\sum_{u=1}^{p}\left(s^{ui}s^{jv}+s^{uj}s^{iv}\right)\right)\right], \\
\frac{\partial}{\partial s_{ij}}\left\{\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right\} &= w_0\sum_{v=1}^{p}\sum_{w=1}^{p}\left(s^{wi}s^{jv}+s^{wj}s^{iv}\right)
\end{aligned}$$

and

$$w_0 = \begin{cases} -0.5 & \text{if } i=j \\ -1 & \text{if } i\neq j. \end{cases}$$

Write

$$C_2 = w_0\,(d_2)^{-3/2}\left[-\tfrac{1}{2}d_1d_3+d_2d_4\right] \tag{A.31}$$

where

$$\begin{aligned}
d_1 &= \sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv} \\
d_2 &= \sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right) \\
d_4 &= \sum_{v=1}^{p}\sum_{w=1}^{p}\left(s^{wi}s^{jv}+s^{wj}s^{iv}\right) \\
d_3 &= \sum_{w=1}^{p}\sum_{v=1}^{p}\left[\left(\sum_{u=1}^{p}s^{uv}\right)\left(d_{3a}+d_{3b}\right)+d_{3c}d_{3d}\right]
\end{aligned}$$

where

$$\begin{aligned}
d_{3a} &= s^{vi}s^{jw}+s^{vj}s^{iw} \\
d_{3b} &= \sum_{u\neq v=1}^{p}\left(s^{ui}s^{jw}+s^{uj}s^{iw}\right)\rho^{|v-u|}
\end{aligned}$$

$$
\begin{aligned}
d_{3c} &= s^{vw} + \sum_{u \neq v=1}^{p} s^{uw} \rho^{|v-u|} \\
d_{3d} &= \sum_{u=1}^{p} \left(s^{ui} s^{jv} + s^{uj} s^{iv} \right).
\end{aligned}
$$

To find C_1, write

$$
\begin{aligned}
C_1 &= \frac{\partial^2}{\partial s_{kl} \partial s_{ij}} \left\{ \left(\sum_{v=1}^{p} \sum_{w=1}^{p} s^{wv} \right) \left[\sum_{w=1}^{p} \sum_{v=1}^{p} \left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vw} + \sum_{u \neq v=1}^{p} s^{uw} \rho^{|v-u|} \right) \right]^{-1/2} \right\} \\
&= \frac{\partial C_2}{\partial s_{kl}} \\
&= \frac{\partial}{\partial s_{kl}} \left\{ w_0 \left(d_2 \right)^{-3/2} \left[-\tfrac{1}{2} d_1 d_3 + d_2 d_4 \right] \right\} \\
&= w_0 \left[\left(d_2 \right)^{-3/2} \frac{\partial}{\partial s_{kl}} \left\{ -\tfrac{1}{2} d_1 d_3 + d_2 d_4 \right\} \right. \\
&\quad \left. + \left(-\tfrac{1}{2} d_1 d_3 + d_2 d_4 \right) \frac{\partial \left(d_2 \right)^{-3/2}}{\partial s_{kl}} \right]
\end{aligned}
$$

where

$$
\frac{\partial}{\partial s_{kl}} \left\{ -\tfrac{1}{2} d_1 d_3 + d_2 d_4 \right\} = -\tfrac{1}{2} \left[d_1 \frac{\partial d_3}{\partial s_{kl}} + d_3 \frac{\partial d_1}{\partial s_{kl}} \right] + d_2 \frac{\partial d_4}{\partial s_{kl}} + d_4 \frac{\partial d_2}{\partial s_{kl}}
$$

and

$$
\frac{\partial d_2^{-3/2}}{\partial s_{kl}} = -\tfrac{3}{2} \left(d_2 \right)^{-5/2} \frac{\partial d_2}{\partial s_{kl}},
$$

where

$$
\begin{aligned}
\frac{\partial d_1}{\partial s_{kl}} &= w_1 \sum_{v=1}^{p} \sum_{w=1}^{p} \left(s^{wk} s^{lv} + s^{wl} s^{kv} \right), \\
\frac{\partial d_2}{\partial s_{kl}} &= \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(\frac{\partial s^{vw}}{\partial s_{kl}} + \sum_{u \neq v=1}^{p} \frac{\partial s^{uw}}{\partial s_{kl}} \rho^{|v-u|} \right) \right. \\
&\quad \left. + \left(s^{vw} + \sum_{u \neq v=1}^{p} s^{uw} \rho^{|v-u|} \right) \left(\sum_{u=1}^{p} \frac{\partial s^{uv}}{\partial s_{kl}} \right) \right] \\
&= w_1 \sum_{w=1}^{p} \sum_{v=1}^{p} \left[\left(\sum_{u=1}^{p} s^{uv} \right) \left(s^{vk} s^{lw} + s^{vl} s^{kw} + \sum_{u \neq v=1}^{p} \left(s^{uk} s^{lw} + s^{ul} s^{kw} \right) \rho^{|v-u|} \right) \right. \\
&\quad \left. + \left(s^{vw} + \sum_{u \neq v=1}^{p} s^{uw} \rho^{|v-u|} \right) \sum_{u=1}^{p} \left(s^{uk} s^{lv} + s^{ul} s^{kv} \right) \right], \\
\frac{\partial d_4}{\partial s_{kl}} &= w_1 \sum_{v=1}^{p} \sum_{w=1}^{p} \left[s^{wi} \left(s^{jk} s^{lv} + s^{jl} s^{kv} \right) + s^{jv} \left(s^{wk} s^{li} + s^{wl} s^{ki} \right) \right.
\end{aligned}
$$

$$+ s^{wj}\left(s^{ik}s^{lv} + s^{il}s^{kv}\right) + s^{iv}\left(s^{wk}s^{lj} + s^{wl}s^{kj}\right)\Big]$$

and

$$\begin{aligned}\frac{\partial d_3}{\partial s_{kl}} &= \sum_{w=1}^{p}\sum_{v=1}^{p}\Bigg[\left(\sum_{u=1}^{p} s^{uv}\right)\left(\frac{\partial d_{3a}}{\partial s_{kl}} + \frac{\partial d_{3b}}{\partial s_{kl}}\right)\\ &\quad + (d_{3a} + d_{3b})\left(\sum_{u=1}^{p}\frac{\partial s^{uv}}{\partial s_{kl}}\right) + d_{3c}\frac{\partial d_{3d}}{\partial s_{kl}} + d_{3d}\frac{\partial d_{3c}}{\partial s_{kl}}\Bigg]\end{aligned}$$

where

$$\begin{aligned}\frac{\partial d_{3a}}{\partial s_{kl}} &= w_1\left[s^{vi}\left(s^{jk}s^{lw} + s^{jl}s^{kw}\right) + s^{jw}\left(s^{vk}s^{li} + s^{vl}s^{ki}\right)\right.\\ &\quad \left. + s^{vj}\left(s^{ik}s^{lw} + s^{il}s^{kw}\right) + s^{iw}\left(s^{vk}s^{lj} + s^{vl}s^{kj}\right)\right],\\ \frac{\partial d_{3b}}{\partial s_{kl}} &= w_1\sum_{u\neq v=1}^{p}\left[\rho^{|v-u|}\left\langle s^{ui}\left(s^{jk}s^{lw} + s^{jl}s^{kw}\right) + s^{jw}\left(s^{uk}s^{li} + s^{ul}s^{ki}\right)\right.\right.\\ &\quad \left.\left. + s^{uj}\left(s^{ik}s^{lw} + s^{il}s^{kw}\right) + s^{iw}\left(s^{uk}s^{lj} + s^{ul}s^{kj}\right)\right\rangle\right],\\ \frac{\partial d_{3c}}{\partial s_{kl}} &= w_1\left[s^{vk}s^{lw} + s^{vl}s^{kw} + \sum_{u\neq v=1}^{p}\left\{\rho^{|v-u|}\left(s^{uk}s^{lw} + s^{ul}s^{kw}\right)\right\}\right],\\ \frac{\partial d_{3d}}{\partial s_{kl}} &= w_1\sum_{u=1}^{p}\left[s^{ui}\left(s^{jk}s^{lv} + s^{jl}s^{kv}\right) + s^{jv}\left(s^{uk}s^{li} + s^{ul}s^{ki}\right)\right.\\ &\quad \left. + s^{uj}\left(s^{ik}s^{lv} + s^{il}s^{kv}\right) + s^{iv}\left(s^{uk}s^{lj} + s^{ul}s^{kj}\right)\right],\end{aligned}$$

and

$$\sum_{u=1}^{p}\frac{\partial s^{uv}}{\partial s_{kl}} = w_1\sum_{u=1}^{p}\left(s^{uk}s^{lv} + s^{ul}s^{kv}\right),$$

where

$$w_1 = \begin{cases} -0.5 & \text{if } k = l \\ -1 & \text{if } k \neq l. \end{cases}$$

Write C_1 and C_3 as

$$\begin{aligned}C_1 &= w_0\left[d_2^{-3/2}\left(-\tfrac{1}{2}\left[d_1\frac{\partial d_3}{\partial s_{kl}} + d_3\frac{\partial d_1}{\partial s_{kl}}\right] + d_2\frac{\partial d_4}{\partial s_{kl}} + d_4\frac{\partial d_2}{\partial s_{kl}}\right)\right.\\ &\quad \left. + \left(-\tfrac{1}{2}d_1d_3 + d_2d_4\right)\left(-\tfrac{3}{2}\right)d_2^{-5/2}\frac{\partial d_2}{\partial s_{kl}}\right] \qquad \text{(A.32)}\end{aligned}$$

$$C_3 = \frac{\partial\phi(-A)}{\partial s_{kl}}$$

$$
\begin{aligned}
&= -A\phi(-A)\frac{\partial A}{\partial s_{kl}} \\
&= -\tfrac{1}{2}m\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right]^{-1/2} \\
&\quad\times\phi(-A)\left(\tfrac{1}{2}m\right) \\
&\quad\times\frac{\partial}{\partial s_{kl}}\left\{\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right]^{-1/2}\right\} \\
&= -\tfrac{1}{4}m^2\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right]^{-1/2} \\
&\quad\times\phi(-A)w_1\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right]^{-3/2} \\
&\quad\times\left\langle-\tfrac{1}{2}\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\right. \\
&\quad\times\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\sum_{u=1}^{p}s^{uv}\left(s^{vk}s^{lw}+s^{vl}s^{kw}+\sum_{u\neq v=1}^{p}\rho^{|v-u|}\left(s^{uk}s^{lw}+s^{ul}s^{kw}\right)\right)\right.\right. \\
&\quad\left.\left.+\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\sum_{u=1}^{p}\left(s^{uk}s^{lv}+s^{ul}s^{kv}\right)\right]\right\} \\
&\quad\left.+\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right]\left[\sum_{v=1}^{p}\sum_{w=1}^{p}\left(s^{wk}s^{lv}+s^{wl}s^{kv}\right)\right]\right\rangle \\
&= -\tfrac{1}{4}m^2\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\phi(-A)w_1\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right]^{-2} \\
&\quad\times\left\langle-\tfrac{1}{2}\left(\sum_{v=1}^{p}\sum_{w=1}^{p}s^{wv}\right)\left\{\sum_{w=1}^{p}\sum_{v=1}^{p}\left[\sum_{u=1}^{p}s^{uv}\left(s^{vk}s^{lw}+s^{vl}s^{kw}\right)\right.\right.\right. \\
&\quad\left.+\sum_{u\neq v=1}^{p}\rho^{|v-u|}\left(s^{uk}s^{lw}+s^{ul}s^{kw}\right)\right. \\
&\quad\left.\left.+\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\sum_{u=1}^{p}\left(s^{uk}s^{lv}+s^{ul}s^{kv}\right)\right]\right\} \\
&\quad+\left[\sum_{w=1}^{p}\sum_{v=1}^{p}\left(\sum_{u=1}^{p}s^{uv}\right)\left(s^{vw}+\sum_{u\neq v=1}^{p}s^{uw}\rho^{|v-u|}\right)\right] \\
&\quad\left.\times\left[\sum_{v=1}^{p}\sum_{w=1}^{p}\left(s^{wk}s^{lv}+s^{wl}s^{kv}\right)\right]\right\rangle
\end{aligned}
\tag{A.33}
$$

Thus

$$
\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}} = -\tfrac{1}{2}m\left[\phi(-A)\times C_1 + C_3\times C_2\right] \tag{A.34}
$$

where A is as in equation (A.29), C_1 is as in equation (A.32), C_2 is as in equation (A.31) and C_3 is as in equation (A.33).

All the required quantities in equation (A.1) have been obtained and thus the asymptotic expansion is derived for the conditional error rate of the Linear Discriminant Function under conditions of "equivalence" (See Marco, Young and Turner (1987)) for two forms of covariance matrix: *(i)* $\Sigma = \Sigma_A$ and *(ii)* $\Sigma = \Sigma_B$. These expansions were evaluated for various values of p, Σ (determined by ρ) and Mahalanobis distance Δ between the population means (determined by m).

Appendix B

HEURISTIC ALGORITHM FOR MODEL SELECTION PROCEDURE USING BHATTACHARYYA DISTANCE

Outlined below is the algorithm used to select $\hat{\gamma}$ and $\hat{\lambda}$ for the model selection procedure using Bhattacharyya distance which was presented in Chapter 5.

(The first section sets a minimum value for γ in extreme cases of high dimension and small sample size, so that only values greater than θ are considered)

If training sample size is less than 2×dimension
    set minimum γ value, (θ), equal to 0.04,
    or 0.08 if the dimension is large (> 10).
    If training sample size is less than dimension
        increase the minimum γ value, (θ), still further,
        depending on the magnitude of the dimension.
    end if
end if

(Calculating $\widehat{B1}$ and $\widehat{B2}$ for various values of γ from θ to 1 in increments of 0.04)

Loop for $\gamma = \theta$ to 1 in steps of 0.04
- Regularise sample covariance matrices using γ
- Check the eigenvalues of the sample covariance matrices, and replace any eigenvalues less than some threshhold (10^{-4}) with that threshhold, to permit stable inversion.

- Calculate and store the values of $\widehat{B1}$, $\widehat{B2}$ and $\widehat{B1}/\widehat{B2}$

end Loop

(Calculating various measures from the stored values of $\widehat{B1}$, $\widehat{B2}$ and $\widehat{B1}/\widehat{B2}$)

Calculate the following:

- range of $\widehat{B1}$ values ($range_{\widehat{B1}}$).
- mean of $\widehat{B1}$ values ($mean_{\widehat{B1}}$).
- ratio of largest $\widehat{B1}$ value to smallest $\widehat{B1}$ value ($r_{\widehat{B1}}$).
- range of $\widehat{B2}$ values ($range_{\widehat{B2}}$).
- maximum of $\widehat{B2}$ values ($max_{\widehat{B2}}$).
- ratio of largest $\widehat{B2}$ value to smallest $\widehat{B2}$ value ($r_{\widehat{B2}}$).

(Obtaining the appropriate value of γ, $\hat{\gamma}$, using five different decision paths)

If $r_{\widehat{B1}} < 2.8$ AND $mean_{\widehat{B1}} \leq 2$ AND $max_{\widehat{B2}} \leq (.3p - .8)$
(i.e. if $\widehat{B1}$ is small and not greatly affected by γ, and if the effect of γ on $\widehat{B2}$ is small.)
Select the value of γ which gives the largest ratio of $\widehat{B1}$ to $\widehat{B2}$
else if ($r_{\widehat{B1}} < 2.8$ or $range_{\widehat{B1}} < 1.5$) AND ($mean_{\widehat{B1}} \leq 2$) AND ($max_{\widehat{B2}} \leq (.3p - .8)$)
(i.e. if $\widehat{B1}$ is small and not greatly affected by γ, and if the effect of γ on $\widehat{B2}$ is large.)
Select the value of γ which gives the smallest ratio of $\widehat{B1}$ to $\widehat{B2}$
else if ($r_{\widehat{B1}} < 2.8$ AND ($mean_{\widehat{B1}} > 2$)
(i.e. if $\widehat{B1}$ is large but not greatly affected by γ)
Select the value of γ corresponding to an average value of $\widehat{B1}/\widehat{B2}$
else if ($r_{\widehat{B1}} > 2.8$ AND ($max_{\widehat{B2}} < (.3p - .8)$)
(i.e. if $\widehat{B1}$ is greatly affected by γ, and if the effect of γ on $\widehat{B2}$ is small)
Select the value of γ which maximises $\widehat{B2}$
else if ($r_{\widehat{B1}} > 2.8$ AND ($max_{\widehat{B2}} > (.3p - .8)$)
(i.e. if $\widehat{B1}$ is greatly affected by γ, and if the effect of γ on $\widehat{B2}$ is large)
Choose a value of γ whereby $\widehat{B1}$ is maximised subject to $\widehat{B2}$ remaining small
end If

(Obtaining various quantities to be used to obtain the estimate for λ, $\hat{\lambda}$)

- Calculate initial value of λ with $\hat{\lambda} = \exp\left\{-\widehat{B2}_{\gamma=\theta}\right\}$
- Calculate $y = \log\left(\widehat{B1}/\widehat{B2}\right)$ when $\gamma = 1$.

(Let $\bar{y}$ denote the average value of y over all pairs of groups.)

- Calculate $z = \sum_{i=1}^{p} |e_{1i} - e_{2i}|$ where e_{1i} is the ith eigenvalue of Σ_1.

(Let $\bar{z}$ denote the average value of z over all pairs of groups.)

(Refining the choice of $\hat{\lambda}$ to obtain $\hat{\lambda}$)

If $\bar{y} > 1$ *(i.e. $\bar{y}$ large, indicating covariance matrices are similar to each other.)*

If $\bar{z} > 17$ *(i.e. $\bar{z}$ large, indicating covariance matrices are not in fact) dissimilar.)*

$\hat{\lambda}' = \hat{\lambda}^2$ *(adjust $\hat{\lambda}$ towards zero.)*

if $\bar{y} < 1.6$ AND $1 < \bar{z} < 3$ *(Indicating similar covariance matrices)*

$\hat{\lambda}' = \hat{\lambda}^{1/(\bar{y}-.6)}$ *(Adjust $\hat{\lambda}$ upwards)*

if $\bar{y} > 1.6$ AND $1 < \bar{z} < 3$

$\hat{\lambda}' = \hat{\lambda}^{1/(\bar{y}-1)}$

if $\bar{y} < 1.6$ AND $3 < \bar{z} < 17$

$\hat{\lambda}' = \hat{\lambda}^{1/(\bar{y}+2)}$

else

$\hat{\lambda}' = \hat{\lambda}^{1/\bar{y}}$

end If

else if $0.5 < \bar{y} < 1$

$\hat{\lambda}' = \hat{\lambda}^{1/\bar{y}^{1.6}}$

else if $\bar{y} < 0.5$

$\hat{\lambda}' = \hat{\lambda}^2$

end If

Appendix C

List of Additional Publications by Author Including Papers Presented at Conferences

Copies of these papers appear in the following pages.

1. **Koolaard**, J. P. and Lawoko, C. R. O. (1993). Estimating error rates in discriminant analysis with correlated training observations: a simulation study. *J. Statist. Comput. Simul.* 48, 81-99.

2. **Koolaard**, J. P. and Lawoko, C. R. O. (1994). Some results on the error rates of the Euclidean and linear discriminant functions. *Proceedings of the ORSNZ/NZSA Conference, Massey University, Palmerston North, New Zealand* (August 1994). pp 327-332.

3. **Koolaard**, J. P. (1995). Covariance Shrinkage in Discriminant Analysis. Paper presented to the A. C. Aitken Centenary Conference, Dunedin, New Zealand (August 1995). [Winner of SPSS Statistics Prize for best statistics paper presented by a student.]

4. **Koolaard**, J. P., Lawoko, C. R. O. and Ganesalingam, S. (1996). Regularized discriminant (classification) analysis involving Bhattacharya distance measure. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 2, Poster, pp 35-43.

5. **Lawoko**, C. R. O., and Koolaard, J. P. (1996). Applications of regularised discriminant(classification) functions in the classification of objects: a discussion of potential applications to remote sensing. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 1, pp 177-184.

6. **Koolaard**, J. P., Ganesalingam, S. and Lawoko, C. R. O. (1996). Comparison of regularised discriminant analysis with the standard discrimination methods. Paper presented to the International Biometrics Conference (IBC '96), Amsterdam, the Netherlands (July 1996). Also submitted to the Journal of Classification.

7. **Koolaard**, J. P. and Lawoko, C. R. O. (1996). The linear and Euclidean discriminant functions: a comparison via asymptotic expansions and simulation study. *Commun. Statist.- Theory Meth.*, (To appear).

A Comparison of the Euclidean and Linear Discriminant Functions

J P Koolaard and C R O Lawoko
Department of Statistics
Massey University
Palmerston North
New Zealand

Abstract

The linear discriminant function is a very popular technique for statistical discrimination because of its robustness. However, it has been demonstrated recently that the much simpler Euclidean distance classifier can out-perform the linear discriminant function in certain situations. In this article we present further results on the relative performances of the two discriminant functions. Whilst in previous work most of the comparisons have been based on simulation studies and numerical integration of error rates, in this article we base our comparisons on asymptotic expansions of error rates as well as some simulation experiments.

1. Background and Motivation

In statistical discriminant analysis the Linear Discriminant Function (LDF) which is based on assumptions of multivariate normality and equal covariance matrices is quite popular because of its robustness and simplicity. Clearly, there are situations when the LDF is inappropriate, and related competitors are the quadratic discriminant function (QDF) and the Euclidean Distance Classifier (EDC). For the two-population situation which we consider here, these rules are as follows:

Suppose the two populations have means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, so that the sample estimators of these parameters (from training data) are $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, $\mathbf{S}_1$ and $\mathbf{S}_2$. The sample versions of these discriminant functions (i.e. SQDF, SLDF, SEDC) are:

(i) **SQDF:** Classify a new object with observation vector $\mathbf{x}$ as belonging to population 1 if $Q(\mathbf{x}) > \log_e k$,

where
$$Q(\mathbf{x}) = \frac{1}{2}\log_e\{|\mathbf{S}_2| + |\mathbf{S}_1|\} - \frac{1}{2}\{\mathbf{x}'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}\}$$
$$- 2\mathbf{x}'(\mathbf{S}_1^{-1}\bar{\mathbf{x}}_1 - \mathbf{S}_2^{-1}\bar{\mathbf{x}}_2) + \bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}\bar{\mathbf{x}}_2' - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1}\bar{\mathbf{x}}_2,$$

and k is some appropriately chosen constant. Clearly, if $Q(\mathbf{x}) < \log_e k$ we allocate $\mathbf{x}$ to population 2.

(ii) **SLDF:** If it can be established that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ say (or it is assumed so), then one should use the SLDF, whereby an object with observation $\mathbf{x}$ is allocated to population 1 if $L(\mathbf{x}) > \log_e k$,

where $L(x) = (\bar{x}_1 - \bar{x}_2)' S^{-1} \{x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\}$

(otherwise it is allocated to population 2).

(iii) **SEDC**: If $\Sigma = I$ in the LDF situation or the information in the covariance matrix is deliberately ignored for the purpose of discrimination, then the SEDC should be used. That is, allocate an object with observation x to population 1 if $E(x) > \log_e k$,

where $E(x) = (\bar{x}_1 - \bar{x}_2)' \{(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\}$,

(otherwise allocate it to population 2).

There has been considerable interest in the literature in the relative performances of these discriminant functions. These comparisons have usually been based on various measures of estimates of error rates (probabilities of misclassifications) since direct evaluations of these probabilities have proved algebraically intractable. Articles which provide relevant background for this study are:

(i) Raudys and Pikelis (1980) who performed a simulation study to compare the SLDF, SQDF, SEDC and a variant of the SLDF for independent measurements (i.e. off diagonal elements of Σ being set to zero). They evaluated the relative performances of these discriminant functions when the populations are spherically normal. Since computations of reliable estimators of error rates have been traditionally difficult, they used numerical integration techniques in evaluating the integrals in the definitions of the probabilities of misclassification. They concluded that the simpler SEDC performed better than the SLDC when p is large relative to n. In fact the SEDC was found to perform at least as well as the SLDF even for non-spherical covariance structures.

(ii) Marco, Young and Turner (1987) compared the SLDF and SEDC under conditions derived to make the two classifiers "equivalent" or "non equivalent". They defined the LDF and EDC as "equivalent" if they have the same (true) error rates (i.e. assuming known population parameters). Their conclusion, based on simulation studies only, was that the EDC generally performed better than the LDF except when the Mahalanobis distance between the two populations (i.e. Δ) was substantially larger than the corresponding Euclidean distance. Also, the SEDC performed at least as well as the SLDF when the population parameters were set so as to achieve either equivalence or non equivalence of the classifiers.

(iii) Other related work include Peck and Van Ness (1982) and Van Ness (1979), among others; see Kim (1992).

The motivation for this work arose from the fact that no "easily-computable" asymptotic results appear to be available in the literature on the relative performance of these discriminant functions. Most of the available results are based on simulation studies or 'brute force' extensive numerical integration of very complicated probability functions (following basic definitions of the error rates).

2. Asymptotic Expansions and Evaluations

The asymptotic expected error rates were obtained using Taylor series expansions of the conditional error rates (i.e. conditional on $\bar{x}_1$, $\bar{x}_2$ and S) and taking expectations over the distributions of $\bar{x}_1$, $\bar{x}_2$ and S. In particular, if H(.) is a differentiable function of parameters $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_s)$, where $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_s)$ are consistent estimators of

$(\beta_1, \beta_2, ..., \beta_s)$ then the Taylor series expansion of E(H) about the point $(\beta_1, \beta_2, ..., \beta_s)$ can be expressed as

$$E(H) \approx H(\beta_1, \beta_2, ..., \beta_s) + \sum_{j=1}^{s} \frac{\partial H}{\partial \hat{\beta}_j} E(\hat{\beta}_j - \beta_j)$$

$$+ \frac{1}{2} \sum_{i,j} \frac{\partial^2 H}{\partial \hat{\beta}_i \partial \hat{\beta}_j} E\{(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)\}$$

For our expansions $H = \Phi(.)$, the standard normal distribution function, and $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_s$ are the elements of $\bar{x}_1, \bar{x}_2, S$. The expansions are evaluated at the point (μ_1, μ_2, Σ).

In this article, we evaluate expected actual (unconditional) and expected plug-in error rates, so that H(.) takes the following forms (for misclassifying an object from population 1 into population 2):

(i) Actual error rate for LDF: $H(.) = \Phi(-D_1/D_2)$

where $D_1 = [\mu_1 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]' S^{-1}(\bar{x}_1 - \bar{x}_2)$, $D_2 = \{(\bar{x}_1 - \bar{x}_2)' S^{-1} \Sigma S^{-1}(\bar{x}_1 - \bar{x}_2)\}^{1/2}$.

(ii) Actual error rate for EDC: $H(.) = \Phi(-D_3/D_4)$,

where $D_3 = [\mu_1 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2]' (\bar{x}_1 - \bar{x}_2)$, $D_4 = \{(\bar{x}_1 - \bar{x}_2)' \Sigma (\bar{x}_1 - \bar{x}_2)\}^{1/2}$.

(iii) Plug-in error rate for LDF: $H(.) = \Phi(-D_5)$,

where $D_5 = -\frac{1}{2}\{(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)\}^{1/2}$.

(iv) Plug-in error rate for EDC: $H(.) = \Phi(-D_6/D_7)$,

where $D_6 = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2)$

$D_7 = \{(\bar{x}_1 - \bar{x}_2)' S(\bar{x}_1 - \bar{x}_2)\}^{1/2}$.

Results in Okamoto (1963) were used in obtaining some of the preliminary results in the asymptotic expansions. One of the results is that if Σ is symmetric and invertible and $\Sigma = \{\sigma_{rs}\}$, $\Sigma^{-1} = \{\sigma^{ij}\}$ then

$$\frac{\partial(\sigma^{ij})}{\partial \sigma_{rs}} = -\frac{1}{(1 + \delta_{rs})}(\sigma^{ir}\sigma^{sj} + \sigma^{is}\sigma^{rj}) \quad (r \le s),$$

where δ_{rs} is the Kronecker delta.

Each asymptotic expansion takes a slightly different form, depending on (i) the structure of Σ assumed, (ii) whether the expansions are obtained under "equivalence" or "non-equivalence" conditions and (iii) whether the expansion is for the LDF or the EDC. For example, the expansion of the conditional error rate associated with the LDF under "equivalence" conditions is of the form

$$p_{LDF} = \Phi\left[-\frac{m}{2}\{\sum_w \sum_v (\sum_u s^{uv})(\sum_u \sigma_{vu} s^{uw})\}^{-1/2} \{\sum_k \sum_l s^{lk}\}\right]$$

$$+ \frac{1}{2n_i} \sum_i \sum_j \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}} \sigma_{ij} + \frac{1}{2n_2} \sum_i \sum_j \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i} \partial \bar{x}_{2j}} \sigma_{ij}$$

$$+ \frac{1}{2} \frac{(n_1 + n_2)}{(n_1 + n_2 - 2)^2} \sum_k \sum_l \sum_i \sum_j \frac{\partial^2 \Phi(.)}{\partial s_{kl} \partial s_{ij}} (\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}),$$

where the quantities $\frac{\partial^2 \Phi(.)}{\partial a \partial b}$ are obtained separately for each assumed structure of μ_1, μ_2 and Σ for any variables 'a' and 'b'..

In comparing the asymptotic expected actual error rates for the LDF and EDC, various 'settings' of certain parameters were used. For example, the values of one parameter (denoted here by 'm' ... see below) were chosen so that Δ^2 was the same in

both cases of equivalence and non-equivalence. Two structures of Σ were considered, which are denoted by $\Sigma = \Sigma_A$ and $\Sigma = \Sigma_B$, where

$$\Sigma_A = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ & & & \ddots & \rho \\ \rho & \cdots & \cdots & \rho & 1 \end{bmatrix} \quad \text{and} \quad \Sigma_B = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ & & & \ddots & \\ & & & & \rho \\ \rho^{p-1} & & & \rho & 1 \end{bmatrix}$$

Numerical Evaluations

In evaluating the results, appropriate values of ρ (both positive and negative) were used. Meanwhile, for the situation of "equivalence" between LDF and EDC (where $\mu_1 = (m, m, ..., m)'$ and $\mu_2 = 0'$) the value of m is calculated as

$$m = \sqrt{\{\Delta^2 / \sum_i \sum_j \sigma^{ij}\}}$$

and for "non-equivalence" (where $\mu_1 = (m^*, 0, 0, ..., 0)'$ and $\mu_2 = 0'$), the value of m^* is calculated as $m^* = \sqrt{\Delta^2/\sigma^{11}}$. The sample sizes were taken to be equal at $n_1 = n_2 = 50$, and the dimensions of the observations were taken to be $p = 4$ and $p = 8$. Although lots of results under various conditions have been obtained (and are available from the authors), in this article we concentrate on comparing the performances of the LDF and EDC under "equivalence" and also on determining (through simulations) whether the asymptotic expansions are accurate. Afterall, the situation of "equivalence" provides the fairest scenario for comparing EDC and LDF. The discussions presented here are basic summaries of general trends and results, since limitation of space does not allow detailed discussion of peculiarities etc. These will be available elsewhere.

3. Discussion of Results

We shall refer to the various error rates as follows:

- e_L^*, e_E^* = true error rates (i.e. for known population parameter values) for the LDF (e_L^*) and EDC (e_E^*)
- e_L, e_E = asymptotic expected (unconditional) error rates
- $\hat{e}_L, \hat{e}_E$ = asymptotic expected (unconditional) plug-in error rates
- e_{SL}, e_{SE} = error rates from simulation experiments

Note that although we obtained several estimates of the error rates from the simulation experiments (e.g. cross-validation, bootstrap, resubstitution) previous work (e.g. Ganeshanandam and Krzanowski (1990)) suggest that cross-validation is a good estimator to use. Thus the comments on the results for simulated error rates are based on cross-validation. We have omitted stating well known results such as error rates generally increase with p or decrease with Δ. A sample page of the tables of results is presented below. The results are now discussed under separate categories:

(a) e_L vs e_E and $\hat{e}_L$ vs $\hat{e}_E$ under equivalence with $\Sigma = \Sigma_A$:

Note that $e_L^* = e_E^* = \Phi(-\Delta/2)$. In this case e_E provides a reasonably good estimate of e_E^* (especially for $\rho > 0$). It is quite clear that the asymptotic expansion given by e_L tends to substantially underestimate the true error rate for small Δ, and as Δ increases it tends to overestimate it. On the other hand, e_E remains reasonably good for all values

of Δ. Meanwhile, simulation experiments suggest that the error rates associated with SEDC are slightly less than those for SLDF in general, in this situation.

When we consider the performance of the plug-in error rates ($\hat{e}_L$ and $\hat{e}_E$) as estimators of e_L and e_E, we observe firstly the well known result that $\hat{e}_L$ underestimates e_L. Our expansions show that a similar statement can be made about $\hat{e}_E$ an e_E. Note, however, that $\hat{e}_E$ provides a much better estimator of e_E, in that the underestimation is of a smaller magnitude. For both SLDF and SEDC the asymptotic expected error rates decrease with $\rho(> 0)$ while the asymptotic expected plug-in error rates increase. For small Δ, there is little difference in the performances of $\hat{e}_L$ and $\hat{e}_E$ (as estimates of e_L and e_E respectively), but as Δ increases, the relative performance of $\hat{e}_E$ improves considerably.

(b) e_L vs e_E and $\hat{e}_L$ vs $\hat{e}_E$ under equivalence with $\mathbf{\Sigma} = \mathbf{\Sigma}_B$:

In this case, the relative performance of e_E to e_L depends on the values of the parameters. For example, when ρ is small and positive and Δ is small to moderate, the e_L appears to be a better estimate but as Δ and ρ increase the relative performance of e_E improves slightly. In general, e_L has a tendency to overestimate e_L^* when Δ is large, and to underestimate it when Δ is small. Meanwhile e_E tends to overestimate e_E^* at all times.

When ρ becomes negative, the performance of e_E deteriorates considerably (especially for large p) and the asymptotic approximation is clearly not appropriate in this situation. On the other hand, e_L is much more stable to changes in ρ and p, suggesting that the information contained in ρ is clearly very relevant for discrimination in this case.

Turning our attention to the plug-in error rates, it is very clear that for positive ρ, $\hat{e}_E$ is a much better estimator of e_E when compared to $\hat{e}_L$ as an estimator of e_L. In fact $\hat{e}_L$ deteriorates considerably, especially for large p and (ironically) for large Δ. When ρ is negative, both plug-in error rates do not perform well at all, and totally break down for high negative correlation.

The simulation results indicate that e_L and especially e_E perform reasonably well when ρ is positive, but when ρ is negative, e_E, $\hat{e}_L$ and $\hat{e}_E$ yield poor approximations.

References

[1] Ganeshanandam, S and Krzanowski, W J (1990). Error-rate estimation in two group Discriminant Analysis using the Linear Discriminant Function. *J. Statist. Comput. Simul.* Vol 36, pp 157-175.

[2] Marco, V R, Young, D M and Turner, D W (1987). The Euclidean Classifier: an alternative to linear discriminant function. *Commun. Statistics - Simul.* 16, 485-505.

[3] Okamoto, M (1963). An asymptotic expansion of the distribution of the linear discriminant function. *Ann. Math. Statist.* 34, 1286-1301. Correction: *Ann Math. Statist.* 39, 1358-1359.

[4] Peck, R and Van Ness, J (1982). The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans. on Pattern Anal. Machine Intell.* PAMI-4, 530-537.

[5] Raudy, S and Pikelis, V (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. on Pattern Anal. Machine Intell.* PAMI-2, 242-252.

[6] Kim, T K. Comparison of the Euclidean and Linear Discriminant Functions in Statistical Discriminant Analysis. Unpublished MSc dissertation, Massey University, Palmerston North, New Zealand.

Table: The 'true', expected actual, expected plug-in and simulated error rates of the EDC and LDF under the case of 'equivalence' with $\Sigma = \Sigma_A$.

Δ^2	ρ	p = 4				p = 8			
		true	actual	plug-in	simul	true	actual	plug-in	simul
		e_E^*	e_E	$\hat{e}_E$	e_{SE}	e_E^*	e_E	$\hat{e}_E$	e_{SE}
		e_L^*	e_L	$\hat{e}_L$	e_{SL}	e_L^*	e_L	$\hat{e}_L$	e_{SL}
0.5	0.0	**0.3618**	**0.3788**	**0.3470**	.38	**0.3618**	**0.4001**	**0.3261**	.40
		0.3618	0.3597	0.3373	.38	0.3618	0.3695	0.3037	.40
	0.2	**0.3618**	**0.3669**	**0.3510**	.37	**0.3618**	**0.3671**	**0.3426**	.37
		0.3618	0.3572	0.3378	.38	0.3618	0.3624	0.3046	.39
	0.4	**0.3618**	**0.3641**	**0.3554**	.36	**0.3618**	**0.3639**	**0.3524**	.37
		0.3618	0.3521	0.3381	.38	0.3618	0.3498	0.3052	.39
	0.65	**0.3618**	**0.3631**	**0.3593**	.37	**0.3618**	**0.3631**	**0.3586**	.36
		0.3618	0.3331	0.3384	.39	0.3618	0.3051	0.3059	.40
1.0	0.0	**0.3085**	**0.3205**	**0.2994**	.32	**0.3085**	**0.3347**	**0.2857**	.33
		0.3085	0.3110	0.2867	.32	0.3085	0.3245	0.2562	.33
	0.2	**0.3085**	**0.3125**	**0.3020**	.31	**0.3085**	**0.3128**	**0.29670.**	.33
		0.3085	0.3092	0.2873	.32	0.3085	0.3194	2574	.33
	0.4	**0.3085**	**0.3107**	**0.3050**	.31	**0.3085**	**0.3107**	**0.3032**	.31
		0.3085	0.3057	0.2877	.33	0.3085	0.3109	0.2580	.34
	0.65	**0.3085**	**0.3101**	**0.3076**	.31	**0.3085**	**0.3102**	**0.3074**	.31
		0.3085	0.2930	0.2882	.32	0.3085	0.3812	0.2590	.34
2.0	0.0	**0.2398**	**0.2481**	**0.2351**	.25	**0.2398**	**0.2571**	**0.2268**	.25
		0.2398	0.2461	0.2196	.26	0.2398	0.2633	0.1902	.26
	0.2	**0.2398**	**0.2431**	**0.2367**	.24	**0.2398**	**0.2434**	**0.2336**	.24
		0.2398	0.2448	0.2204	.25	0.2398	0.2596	0.1917	.26
	0.4	**0.2398**	**0.2420**	**0.2386**	.25	**0.2398**	**0.2421**	**0.2377**	.25
		0.2398	0.2425	0.2209	.25	0.2398	0.2541	0.1926	.27
	0.65	**0.2398**	**0.2416**	**0.2402**	.24	**0.2398**	**0.2418**	**0.2403**	.24
		0.2398	0.2345	0.2214	.25	0.2398	0.2355	0.1939	.26
2.5	0.0	**0.2146**	**0.2219**	**0.2112**	.22	**0.2146**	**0.2296**	**0.2043**	.22
		0.2146	0.2220	0.1951	.23	0.2146	0.2400	0.1661	.23
	0.2	**0.2146**	**0.2178**	**0.2126**	.21	**0.2146**	**0.2181**	**0.2101**	.21
		0.2146	0.2208	0.1959	.22	0.2146	0.2367	0.1677	.23
	0.4	**0.2146**	**0.2168**	**0.2141**	.22	**0.2146**	**0.2170**	**0.2135**	.22
		0.2146	0.2188	0.1964	.23	0.2146	0.2321	0.1686	.23
	0.65	**0.2146**	**0.2165**	**0.2155**	.21	**0.2146**	**0.2167**	**0.2156**	.21
		0.2146	0.2121	0.1970	.22	0.2146	0.2164	0.1699	.23

Covariance Shrinkage in Discriminant Analysis

J.P.Koolaard
Department of Statistics
Massey University
Palmerston North

Abstract

Friedman (1989) proposed Regularised Discriminant Analysis (RDA) as a compromise between normal-based Linear and Quadratic Discriminant Analyses by considering alternatives to the usual maximum likelihood estimates for the covariance matrices. These alternatives are characterised by two (regularisation) parameters, the values of which are customized to individual situations by jointly minimising a sample-based estimate of future misclassification risk. This technique offers sizeable gains in classification accuracy in many circumstances, although it is computationally intensive.

To further investigate some aspects of the operation and performance of RDA, a series of simulation studies were implemented which establish key factors in RDA's success, and demonstrate that the advantage which RDA enjoys over Linear and Quadratic Discriminant Analyses can be noticeable even if the sample size to dimension (number of feature variables) ratio is quite large.

Because of the computational burden inherent in RDA, and with regard to criticisms of the technique by Rayens and Greene (1991), it was investigated whether information about appropriate values for the two regularisation parameters could be gleaned by examining the behaviour of the Bhattacharyya Distance between the various populations. A classification rule for the two (normal) population case which uses regularisation parameters obtained from the Bhattacharyya distance (and which is computationally much faster than Friedman's RDA) is presented and compared with the original RDA.

1.1 Introduction

A purpose of classification or discriminant analysis is to assign objects to one of several (K) groups based on a set of measurements $\mathbf{X}=(X_1,X_2,\ldots,X_p)$ (where p denotes the dimensionality of the data) obtained from each object or observation. An object is assumed to be a member of exactly one of the groups, and an error is incurred if it is assigned to a different one.

The most common discriminant rules are based on the multivariate normal distribution. Assuming we have K (normal) groups each with population mean $\boldsymbol{\mu}_k$ and covariance Σ_k

(k=1, ..., K), and π_k is the prior probability of observing a member of that group, the classification rule is to assign an object to group k^*, where

$$d_{k^*}(\mathbf{X}) = \min d_k(\mathbf{X}) \ (1 \le k \le K). \tag{1}$$

Here $d_k(\mathbf{X})$ is defined as

$$d_k(\mathbf{X}) = (\mathbf{X}-\boldsymbol{\mu}_k)' \, \Sigma_k^{-1} (\mathbf{X}-\boldsymbol{\mu}_k) + \ln |\Sigma_k| - 2 \ln \pi_k, \tag{2}$$

which is often called the discriminant score for the kth group.

Equations (1) and (2) define the quadratic discriminant function (QDF) since the regions of the measurement space corresponding to each group assignment are separated by quadratic boundaries.The special case occurring when all of the class covariance matrices are presumed to be equal, i.e.

$$\Sigma_k = \Sigma \quad (1 \le k \le K), \tag{3}$$

is called the linear discriminant function (LDF).

This paper is concerned with the problems associated with estimating the group population covariance matrices, Σ_k $(1 \le k \le K)$. Quadratic discriminant analysis (QDA) requires approximately normal group conditional densities and reasonably large training sample sizes, n_k, before it can be expected to work well. This is due to it's sensitivity to the quality of the parameter estimates, particularly the sample covariance matrix S_k,

$$S_k = \frac{1}{n_k} \sum_{v=1}^{n_k} (\mathbf{X}_v - \bar{\mathbf{X}}) \, (\mathbf{X}_v - \bar{\mathbf{X}}_k)' \tag{4}$$

which is the unbiased and consistent sample estimates of Σ_k. Linear discriminant analysis (LDA) is more robust to non-normality, and requires less parameter estimation than QDA. However, poor estimates of the pooled covariance matrix (i.e. Σ in equation (3)) are possible, particularly if the size of the training sample, N ($N = \sum_{k=1}^{K} n_k$), is small in relation to the dimension of the measurement space, p. The covariance matrix estimates can be highly variable in this situation, and Friedman (1989) showed the effect of this phenomenon on discriminant analysis by representing the group covariance matrices by their spectral decompositions

$$\Sigma_k = \sum_{i=1}^{p} e_{ik} \, \boldsymbol{\eta}_{ik} \, \boldsymbol{\eta}'_{ik} \tag{5}$$

where e_{ik} is the ith eigenvalue of Σ_k and $\boldsymbol{\eta}_{ik}$ is its corresponding eigenvector. The discriminant score (2) can thus be written as

$$d_k(\mathbf{X}) = \sum_{i=1}^{p} \frac{[\boldsymbol{\eta}'_{ik} \, (\mathbf{X} - \boldsymbol{\mu}_k)]^2}{e_{ik}} + \sum_{i=1}^{p} \ell n \, e_{ik} - 2 \, \ell n \, \pi_k \tag{6}$$

It is clear from (6) that small eigenvalues and their eigenvectors will have a large effect on this quantity. It is well-known that sample based estimates of the Σ_k produce biased estimates of the eigenvalues with the bias being more pronounced when the eigenvalues of the population parameters (Σ_k) are similar, especially for small training sample size. When $n_k \le p$, the smallest eigenvalues of the S_k are zero, with obvious consequence for the sample discriminant score, which is equation (5) but where e_{ik} is replaced by the ith eigenvalue of S_k , and $\boldsymbol{\eta}_{ik}$ becomes its

corresponding eigenvector. Thus the importance of the low variance subspace spanned by the eigenvectors corresponding to the smallest sample eigenvalues is greatly exaggerated. In fact most of the variation in the sample discriminant score is associated with directions of low sample variance in the measurement space.

2.1 Regularised estimates of Σ_k in discrimination

One can reduce the variance associated with sample-based estimates of Σ_k by biasing the estimates away from the usual sample values and towards values which are more realistic in practice. Regularisation parameters may be introduced which control the amount of biasing, and the sample data can give information to estimate these parameters.

If one introduces a regularisation parameter, λ, which controls the degree of shrinkage of the individual group conditional covariance matrix estimates, the S_k, to the pooled estimate S_p, the following set of alternatives may be obtained:

$$\hat{\Sigma}_k(\lambda) = \frac{(1-\lambda)(n_k-1)S_k + \lambda S_p}{(1-\lambda)(n_k-1) + \lambda(N-K)}. \qquad (7)$$

Now λ takes on values $0 \leq \lambda \leq 1$ and it is evident that if $\hat{\Sigma}_k(\lambda)$ is used to estimate Σ_k in (2), the scenario $\lambda=0$ simply yields QDA, while one can obtain LDA by setting $\lambda=1$.

Equation (7) may not provide for sufficient regularisation, especially if the total sample size, N, is less than or comparable in size to p. In these cases even for LDA, the number of parameters to be estimated is close to, or less than, the number of observations available. One usually wants to avoid this scenario in practice, however. Also, biasing the S_k to the pooled estimate may not be appropriate in some situations.

Friedman (1989), therefore, introduced further regularisation of the S_k to obtain

$$\hat{\Sigma}_k(\lambda,\gamma) = (1 - \gamma)\,\hat{\Sigma}_k(\lambda) + \gamma\frac{tr[\hat{\Sigma}_k(\lambda)]}{p} \qquad (8)$$

where $tr[\hat{\Sigma}_k(\lambda)]$ is the trace of the matrix $\hat{\Sigma}_k(\lambda)$ in (7), I is a pxp identity matrix and γ is the additional parameter which regulates shrinkage towards a multiple of the identity matrix (the multiplier simply being the average eigenvalue of $\hat{\Sigma}_k(\lambda)$). Shrinking in this way acts counter to the bias, described earlier, produced by sample estimation of the eigenvalues by decreasing the larger eigenvalues of $\hat{\Sigma}_k(\lambda)$ and increasing the smaller ones.

Friedman proposed that the regularised sample group covariance matrices, $\hat{\Sigma}_k(\lambda,\gamma)$, be used as the estimate for Σ_k in (1) and (2) for discriminant analysis. As $0 \leq \lambda,\gamma \leq 1$, a technique is required to select an appropriate λ,γ combination for use in the model, and Friedman employed one which selects the combination that minimises an estimate of the future error rate (See section 2.2 below). He termed this procedure regularised discriminant analysis (RDA).

RDA provides a rich class of regularisation alternatives. The possible λ,γ combinations may be thought of as a plane with four corners. The bottom left vertex ($\lambda=0, \gamma=0$) corresponds to QDA, ($\lambda=1, \gamma=0$) gives LDA, ($\lambda=1, \gamma=1$) yields a discriminant rule based on minimum euclidean distance between groups, while ($\lambda=0, \gamma=1$) yields a weighted minimum euclidean distance rule where the group weights are inversely proportional to the average variance of the measurement variables in the group, i.e. $tr[S_k]/p$. If γ is fixed at zero and λ varied, intermediate rules between QDA and LDA are obtained. If λ is fixed at 1 and γ increased from 0, one obtains an analogy to ridge regression for LDA.

2.2 Selecting λ and γ values and tie-breaking

In practice, optimal values for the regularisation parameters λ and γ are not known before hand, and Friedman suggests they be estimated from the training data. The selected λ, γ combination is that which gives rise to the minimum cross-validated estimate of the error rate associated with the regularised discriminant rule.

A grid of points is chosen on the λ, γ plane ($0 \leq \lambda,\gamma \leq 1$), containing typically between 25 and 50 points. Using the λ, γ values to create the classification rule at each point, cross-validation is used to estimate the misclassification risk for each combination of (λ, γ), and the point $(\hat{\lambda},\hat{\gamma})$ with the lowest estimated error rate is used as an estimate of the optimal values of λ and γ. This two-parameter optimisation problem would require excessive computation were it to be implemented in a straight-forward way. However, Friedman developed updating formulas for the computation of the regularised sample covariance matrix and its inverse when a different observation is successively omitted from the sample, as during cross-validation.

Rayens and Greene (1991) noted two criticisms of the model selection procedure of Friedman. Firstly, it was stated that the minimum cross-validated estimate of the misclassification risk is often constant for a range of (λ, γ) combinations. Hence the optimal choice of λ and γ for the model will often not be uniquely determined. Friedman employed a strategy of maximum regularisation where, for all point yielding the minimum error rate on the (λ, γ) grid, that point $(\hat{\lambda},\hat{\gamma})$ is selected which gives rise to the largest value of γ for the largest value of λ. Secondly, Rayens and Greene (1991) demonstrated a situation that can and does occur where only a very small proportion of the sample data influences in any way the optimal choices of λ and γ, and the remainder of the sample observations are correctly classified for almost all points on the λ, γ plane. This occurs especially when the groups are well separated.

Friedman (1989) performed a simulation study to compare RDA with QDA and LDA in terms of their simulated overall error rates. The simulation conditions represented a wide range of situations in terms of the general structure of the group means and covariance matrices. Some of these conditions were chosen because they were expected to be unfavourable to RDA in that any regularisation away from QDA or LDA would be detrimental to the discrimination process. Other conditions were chosen because they were expected to be favourable to regularisation. The six

conditions, defined in terms of the population covariance matrices and means, which are also those employed in the following simulation studies in this paper, are:

1) Equal spherical population covariance matrices. A spherical matrix is one where all the eigenvalues are similar in magnitude.

2) Unequal, spherical population covariance matrices.

3) Equal, highly ellipsoidal population covariance matrices with group mean differences in the low variance subspace. By ellipsoidal we mean that there is a large difference in magnitude between the smallest and largest eigenvalues.

4) Equal, highly ellipsoidal population covariance matrices with group mean differences in the high variance subspace.

5) Unequal, highly ellipsoidal population covariance matrices with zero mean differences.

6) Equal, highly ellipsoidal population covariance matrices with non-zero mean differences.

2.3 Selection of values for the regularisation parameters when the choice is not uniquely determined by the minimum cross-validated error rate

In the previous section we noted that the optimal choice of $(\hat{\lambda},\hat{\gamma})$ is very often not uniquely determined. It is of interest to study the effect of a different procedure than that employed by Friedman (1989) for selecting the values to use for the regularisation parameters. A simulation study has been performed under the same conditions as in the previous section but employing a policy of minimum regularisation in the advent of the minimum cross-validated error rate not being uniquely determined. If there is more than one point on the (λ, γ) grid associated with the minimum cross-validated error rate, that point is chosen having the smallest γ value for the smallest λ value. This method will be denoted RDA1 and is compared with RDA which follows the opposite policy of maximum regularisation to break ties. The other discriminant rules are also included for comparison. In all cases there are 3 populations or groups, and sample sizes are set to be just larger than the dimension p in each case, so as to avoid singularity in the group covariance matrix estimates. The (λ, γ) grid of points consists of 25 points and is defined to be the same as that used in Friedman's study. Results for each set of simulation conditions are in Tables (1) to (6).

The first and major finding from the present study comparing RDA1 with RDA is that the cross-validated error rate surface over the λ, γ plane is often very flat at its minimum. In such situations the error rate estimate will be very similar under both methods for dealing with ties, even though the assessed $\hat{\lambda}$ and $\hat{\gamma}$ values are quite different. This would indicate that employing a policy of minimum regularisation does not have much effect on the performance of RDA in most of the parameter settings considered, and indicates the degree of homogeneity in the cross-validated error rate response surface over the λ, γ plane. In particular, the choice of λ can be considerably less precise than the choice of γ in determining the performance of the rule in terms of its error rate.

Table 1
Equal, Spherical Covariance Matrices

	p=6	p=10	p=20
misclassification risk			
RDA	.11 (.04)	.12 (.04)	.12 (.04)
RDA1	.12 (.03)	.14 (.04)	.12(.03)
LDA	.13 (.04)	.14 (.04)	.15 (.04)
QDA	.24 (.06)	.32 (.07)	.41 (.07)
EDC	.11 (.04)	.11 (.03)	.11 (.03)
Average regularisation parameter values			
RDA λ	.87 (.29)	.85 (.30)	.80 (.34)
RDA γ	.78 (.34)	.81 (.26)	.81 (.24)
RDA1 λ	.15 (.26)	.20 (.33)	.24 (.33)
RDA1 γ	.67 (.32)	.69 (.30)	.80 (.25)

Table 2
Unequal, Spherical Covariance Matrices

	p=6	p=10	p=20
misclassification risk			
RDA	.11 (.04)	.12 (.04)	.12 (.04)
RDA1	.12 (.03)	.14 (.04)	.12 (.03)
LDA	.13 (.04)	.14 (.04)	.15 (.04)
QDA	.24 (.06)	.32 (.07)	.41 (.07)
EDC	.11 (.04)	.11 (.03)	.11 (.03)
Average regularisation parameter values			
RDA λ	.87 (.29)	.85 (.30)	.80 (.34)
RDA γ	.78 (.34)	.81 (.26)	.81 (.24)
RDA1 λ	.15 (.26)	.20 (.33)	.24 (.33)
RDA1 γ	.67 (.32)	.69 (.30)	.80(.25)

Table 3
Equal, Highly Ellipsoidal Covariance Matrices
(Mean Differences in Low Variance Subspace)

	p=6	p=10	p=20
misclassification risk			
RDA	.07 (.05)	.12 (.04)	.15 (.04)
RDA1	.08 (.04)	.13 (.05)	.16 (.04)
LDA	.06 (.03)	.11 (.04)	.14 (.04)
QDA	.14 (.05)	.29 (.06)	.39 (.06)
EDC	.24 (.06)	.29 (.06)	.32 (.05)
Average regularisation parameter values			
RDA λ	.87 (.24)	.89 (.23)	.87 (.19)
RDA γ	.05 (.14)	.04 (.11)	.04 (.09)
RDA1 λ	.41 (.28)	.56 (.30)	.73 (.27)
RDA1 γ	.02 (.07)	.03 (.11)	.02 (.07)

Table 4
Equal, Highly Ellipsoidal Covariance Matrices
(Mean Differences in High VarianceSubspace)

	p=6	p=10	p=20
misclassification risk			
RDA	.06 (.03)	.10 (.03)	.11 (.03)
RDA1	.07 (.03)	.10 (.03)	.11 (.03)
LDA	.07 (.03)	.12 (.04)	.14 (.04)
QDA	.16 (.06)	.30 (.08)	.42 (.06)
EDC	.06 (.03)	.10 (.03)	.11 (.03)
Average regularisation parameter values			
RDA λ	.85 (.31)	.86 (.29)	.79 (.33)
RDA γ	.58 (.37)	.62 (.33)	.67 (.27)
RDA1 λ	.15 (.25)	.26 (.32)	.32 (.34)
RDA1 γ	.50 (.35)	.55 (.26)	.67 (.27)

Table 5
Unequal, Highly Ellipsoidal Covariance Matrices
(with Zero Mean Differences)

	p=6	p=10	p=20
misclassification risk			
RDA	.20 (.06)	.12 (.05)	.03 (.02)
RDA1	.18 (.06)	.11 (.04)	.03 (.02)
LDA	.60 (.06)	.59 (.06)	.58 (.05)
QDA	.17 (.05)	.14 (.06)	.14 (.04)
EDC	.60 (.06)	.59 (.06)	.58 (.05)
Average regularisation parameter values			
RDA λ	.04 (.07)	.04 (.06)	.04 (.06)
RDA γ	.12 (.15)	.25 (.16)	.35 (.18)
RDA1 λ	.01 (.04)	.01 (.04)	.02 (.05)
RDA1 γ	.10 (.14)	.26 (.15)	.26 (.15)

Table 6
Unequal, Highly Ellipsoidal Covariance Matrices
(with Non-zero Mean Differences)

	p=6	p=10	p=20
misclassification risk			
RDA	.06 (.04)	.06 (.04)	.02 (.02)
RDA1	.05 (.02)	.05 (.04)	.01 (.01)
LDA	.17 (.05)	.18 (.04)	.21 (.04)
QDA	.04 (.03)	.05 (.04)	.06 (.04)
EDC	.16 (.04)	.17 (.04)	.17 (.04)
Average regularisation parameter values			
RDA λ	.10 (.20)	.10 (.14)	.07 (.06)
RDA γ	.19 (.27)	.29 (.22)	.35 (.19)
RDA1 λ	.01 (.03)	.02 (.04)	.00 (.00)
RDA1 γ	.10 (.13)	.22 (.15)	.27 (.09)

In conclusion, altering the way ties are broken in the search for the optimum values of λ and γ does not have a great influence on the performance of RDA. Some of the parameter configurations looked at would favour a greater degree of regularisation and some a lesser degree, but the difference in error rates was slight.

2.4 Usefulness of RDA for various ratios of sample size to dimension

From the study by Friedman (1989) as well as in the previous section it is clear that RDA has proved itself at least equal to but usually superior to the other classification rules under a fairly wide range of situations. The superiority is greatest in the larger dimensional settings (p>10). The comparisons with QDA and LDA indicate that the advantage RDA has over the other classification rules is a result of allowing for eigenvalue shrinkage. A question which becomes of interest is: to what extent do the benefits of regularisation, in particular eigenvalue shrinkage, diminish as the sample size to dimensionality ratio increases?

A (further) simulation study was implemented in the manner of Friedman (1989) (and the previous section), using the same six simulation conditions. In those studies, the ratio of training sample size to dimensionality ($\frac{n}{p}$) is around 2 or less. We investigate the performance of RDA relative to the other classification rules over a wider range of $\frac{n}{p}$ ratios. It would be anticipated that eigenvalue shrinkage would no longer be useful for discriminating once the training sample size increases past some point sufficiently larger than p. The various $\frac{n}{p}$ ratios employed were 1.2, 1.5, 2, 3, 5, 10 for dimensions 6, 10 and 20. The (λ, γ) grid of values for use in the model selection procedure of RDA is defined by the outer product of λ= (0, .25, .5, .75, 1) and γ = (0, .25, .5, .75, 1). The entire training sample is 3n in each case, the test sample is 200, and 50 replications of each experiment were performed. Average error rate (with standard deviation in brackets) are given for each classification rule. The results are given in Tables 7 to 12.

Eigenvalue shrinkage appears to enhance the classification process under conditions of equal, spherical covariance matrices only for small $\frac{n}{p}$ ratio ($\frac{n}{p} < 3$). For larger ratios the advantage RDA enjoys over the other methods disappears. QDA shows the most dramatic improvement in error rate as the $\frac{n}{p}$ ratio increases, owing to improved parameter estimates through larger sample size.

In the situation of unequal, spherical population covariance matrices RDA proved superior for all n/p ratios studied, especially for smäller n/p ratios, indicating the benefit of eigenvalue shrinkage which biases the covariance estimates towards the appropriate value (a multiple of the identity matrix) in these circumstances.

Table 7
Misclassification Risk for Various n/p Ratios
Equal Spherical Covariance Matrices

	p=6			p=10			p=20		
	Ratio 1.2:1	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*
RDA	.22	.20	.16	.20	.15	.10	.13	.10	.09
	(.04)	(.03)	(.03)	(.05)	(.03)	(.03)	(.03)	(.02)	(.02)
LDA	.30	.25	.18	.28	.26	.18	.28	.24	.19
	(.06)	(.02)	(.02)	(.05)	(.04)	(.03)	(.03)	(.03)	(.02)
QDA	.53	.34	.17	.52	.35	.14	.55	.37	.12
	(.07)	(.06)	(.02)	(.07)	(.05)	(.03)	(.04)	(.04)	(.02)

Table 8
Misclassification Risk for Various n/p Ratios
Unequal, Spherical Covariance Matrices

	p=6			p=10			p=20		
	Ratio 1.2:1	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*
RDA	.22	.20	.16	.20	.15	.10	.13	.10	.09
	(.04)	(.03)	(.03)	(.05)	(.03)	(.03)	(.03)	(.02)	(.02)
LDA	.30	.25	.18	.28	.26	.18	.28	.24	.19
	(.06)	(.02)	(.02)	(.05)	(.04)	(.03)	(.03)	(.03)	(.02)
QDA	.53	.34	.17	.52	.35	.14	.55	.37	.12
	(.07)	(.06)	(.02)	(.07)	(.05)	(.03)	(.04)	(.04)	(.02)

Table 9
Misclassification Risk for Various n/p Ratios
Equal, Highly Ellipsoidal Covariance Matrices
(Mean Differences in Low Variance Subspace)

	p=6			p=10			p=20		
	Ratio 1.2:1	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*
RDA	.12	.08	.05	.16	.12	.10	.18	.14	.11
	(.06)	(.04)	(.02)	(.04)	(.04)	(.03)	(.04)	(.03)	(.02)
LDA	.10	.07	.04	.14	.11	.08	.17	.14	.11
	(.03)	(.02)	(.01)	(.03)	(.02)	(.02)	(.03)	(.02)	(.02)
QDA	.41	.15	.05	.44	.24	.09	.49	.32	.14
	(.09)	(.05)	(.02)	(.09)	(.05)	(.02)	(.06)	(.04)	(.02)

Table 10
Misclassification Risk for Various n/p Ratios
Equal, Highly Ellipsoidal Covariance Matrices
(Mean Differences in High Variance Subspace)

	p=6			p=10			p=20		
	Ratio 1.2:1	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*
RDA	.08	.07	.06	.10	.10	.11	.12	.10	.09
	(.03)	(.02)	(.02)	(.03)	(.02)	(.04)	(.03)	(.02)	(.02)
LDA	.12	.09	.05	.14	.11	.08	.16	.13	.13
	(.03)	(.03)	(.01)	(.03)	(.03)	(.02)	(.03)	(.02)	(.03)
QDA	.43	.18	.06	.45	.23	.09	.48	.30	.10
	(.10)	(.05)	(.01)	(.07)	(.05)	(.02)	(.05)	(.04)	(.03)

Table 11
Misclassification Risk for Various n/p Ratios
Unequal, Highly Ellipsoidal Covariance Matrices
(with Zero Mean Differences)

	p=6			p=10			p=20		
	Ratio 1.2:1	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*
RDA	.34	.19	.10	.15	.09	.06	.03	.02	.00
	(.11)	(.05)	(.04)	(.06)	(.03)	(.03)	(.02)	(.02)	(.00)
LDA	.61	.59	.62	.59	.59	.61	.58	.59	.62
	(.05)	(.05)	(.04)	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)
QDA	.39	.18	.08	.29	.10	.02	.20	.04	.00
	(.09)	(.04)	(.02)	(.09)	(.03)	(.01)	(.07)	(.02)	(.00)

Table 12
Misclassification Risk for Various n/p Ratios
Unequal, Highly Ellipsoidal Covariance Matrices
(with Non-zero Mean Differences)

	p=6			p=10			p=20		
	Ratio 1.2:1	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*	*Ratio 1.2:1*	*Ratio 2:1*	*Ratio 10:1*
RDA	.14	.07	.02	.09	.04	.01	.03(.0	.01	.00
	(.04)	(.03)	(.01)	(.05)	(.03)	(.01)	2)	(.01)	(.00)
LDA	.21	.18	.13	.24	.20	.16	.22	.18	.15
	(.05)	(.04)	(.03)	(.04)	(.04)	(.03)	(.04)	(.03)	(.03)
QDA	.25	.05	.02	.19	.04	.01	.14	.01	.00
	(.12)	(.02)	(.01)	(.10)	(.02)	(.01)	(.06)	(.01)	(.00)

Eigenvalue shrinkage proves to be of no benefit when the population covariance matrices are equal but highly ellipsoidal with mean differences in the low variance measurement subspace, at least for the n/p ratios studied ($\frac{n}{p} > 1.2$). This is because if the covariance matrix eigenvalues are biased towards equality, the variance in all subspaces is equalised and hence in this case the mean differences will become obscured. Conversely, when the mean differences are exhibited in the high variance subspace, eigenvalue shrinkage proves useful in reducing the variance in those subspaces where mean differences are exhibited. RDA has a lower error rate than those rules with no eigenvalue shrinkage for $\frac{n}{p}$ ratio less than 3. At $\frac{n}{p}$ =3 and larger, LDA performs as well as RDA.

In the case of unequal, highly ellipsoidal population covariance matrices with either zero or non-zero differences between the means, a small amount of eigenvalue shrinkage enables RDA to out-perform QDA, but only when the sample size is less than twice the dimension. In this case, eigenvalue shrinkage is generally not desirable since the covariance matrices provide substantial information needed for discrimination. A small degree of eigenvalue shrinkage is beneficial in counteracting eigenvalue bias (see section 1.1) in those situations of small $\frac{n}{p}$ ratio. For larger $\frac{n}{p}$ ratios QDA's performance is comparable to that of RDA, indicating eigenvalue shrinkage loses its

effectiveness. While the average γ value used in RDA is usually small, there is substantial variation, indicating that under these fairly difficult discrimination conditions (especially zero mean differences), selection of γ is sensitive to peculiarities in the data.

In conclusion, this simulation study underlines the usefulness of the eigenvalue shrinkage technique as employed by RDA. The advantage that it affords over the other rules is strongest when the training sample size from each group is small in relation to the dimensionality, p. Furthermore, often that advantage remains, even when the sample size increases to several times that of the dimension.

Model Selection Using Bhattacharyya Distance

3.1 Introduction

In section 2.2, several weaknesses in the model selection procedure of Regularised Discriminant Analysis as developed by Friedman (1989) were noted by Rayens and Greene(1991). These included the fact that the regularisation parameters were often determined by a small fraction of the data points available, and that in many instances (especially with smaller sample sizes) there will not be a unique choice of the parameters (λ,γ) for the model. Furthermore, despite the development of computationally efficient algorithms to enhance the attractiveness of what is inherently a computationally intensive model, the computation time is still rather high from the author's experience using MATLAB^{TM} on a multiprocessing SUN Sparcstation ELC. Therefore it is of interest to explore other ways of arriving at appropriate regularisation parameter values in place of minimising the cross-validated error rate at a range of points over the λ,γ grid.

Distance measures have often been considered as alternatives to error rates. For example, Jain (1976) investigated the behaviour of an estimate of the Bhattacharyya distance when used as a criterion in variable selection. It was shown that the bias and variance of the estimate is related to the number of training samples and parameter values of the distribution. Kailath (1967) addressed the problem that minimising the error rate to determine optimum classification can be difficult to accomplish in practice. He investigated the idea of using simpler, albeit sub-optimal performance measures instead of the error rate, and compared the Bhattacharyya distance with an often-used measure, the divergence, which is closely related to Shannon's logarithmic measure of information. Not only is the Bhattacharyya Distance easier to evaluate than the divergence, but in some examples in the study it was found to perform at least as well as the divergence in minimising the probability of misclassification. Kailath obtained an upper bound on the the probability of misclassification in terms of the Bhattacharyya distance in the case of equal prior probabilities of the distributions. Note that Kailath only treated the case of two populations. Also, all his work assumed knowledge of the parameters, whereas, as we shall see later, if one has to use sample estimates of the parameters, the link between Bhattacharyya distance and error rate is a lot less clear. Also, Fukunaga and Hayes (1989) obtained an upper bound, in terms of the Bhattacharyya distance, on the Bayes error for classifying between two Gaussian distributions .

The Bhattacharyya distance between two multivariate normal density functions with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is

$$B = B1 + B2 \tag{9}$$

where

$$B1 = \frac{1}{8}(\mu_1 - \mu_2)' \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_1 - \mu_2)$$

and

$$B2 = \frac{1}{2} \ell n \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|}\sqrt{|\Sigma_2|}}.$$

The first term of the expression, B1, is similar to the well-known Mahalanobis distance between the densities. It measures the distance between the two distributions caused by the mean shift. The second term B2 utilises the determinants of each distribution's covariance as well as that of the average group covariance matrix. It gives a measure of the the difference between the two distributions due to the covariance shift.

Fukunaga and Hayes (1989) derived expressions for the expected bias and variance of the terms B1 and B2 and showed that the bias of term B1 is proportional to $\frac{p}{n}$ (n = sample size). i.e. increases as the ratio $\frac{p}{n}$ increases. They also showed that the bias of term B2 is proportional to $\frac{(p+1)p}{n}$. In other words, estimates of this distance measure become increasingly biased as the ratio $\frac{p}{n}$ increases, with term B2 more seriously affected than term B1. Thus in high dimensional space the bias present in the Bhattacharyya distance estimate is dominated by the bias inherent in estimation of term B2. They also showed that as the dimensionality increases, an increasingly large ratio of $\frac{n}{p}$ is needed to maintain a constant expected value of B.

With the above knowledge of the Bhattacharyya distance function between two Gaussian distributions, it is plausible to expect that some degree of regularisation of the covariance, such as is provided for by the two-parameter model in equation (8), would improve the estimation of the Bhattacharyya distance. The reason for this stems from the accepted knowledge that covariance estimates based on equation (4) yield eigenvalue estimates which are biased. The largest ones are biased towards high values and the smallest ones are biased towards values which are too low. This bias will be worse in the situation where the true population eigenvalues are approximately equal, but in all cases this bias becomes more pronounced as the ratio of sample size to dimension decreases.

The term B2 of the Bhattacharyya distance is most vulnerable to such bias occurring, being a ratio of determinants of sample covariance estimates, and eigenvalue shrinkage ought to prove useful in counteracting bias-induced anomalies in estimates of B2, particularly as p becomes large.

3.2 Behaviour of Bhattacharyya Distance with Regularised Covariances

Kailath (1967) admitted that it was too much to hope for to obtain a strong relationship between distance measures and error rate, but he nevertheless was able to obtain several useful theoretical results, assuming known population parameters. In the present covariance regularisation context with two parameters controlling shrinkage, as in equations (7) and (8), it is also too much to hope for to expect that the $(\hat{\lambda},\hat{\gamma})$ combination which maximises the Bhattacharyya distance for a given set of data will also yield a classification rule which minimises the future misclassification risk. Instead, from the example below, we can detect no such relationship between sample Bhattacharyya distance and minimum error rate. The figure shows the components B1, B2 of the Bhattacharyya distance at a range of points over the λ,γ grid. The cross-validated error rate (e_{cv}) at each point is also stated to give an indication of where the range in which the minimum actual error rate lies. The data set consisted of samples of size 13 from each of two normal populations with means and covariances as in Table 4 in section 2.3.

Table 13. Example of λ,γ Grid of Bhattacharyya Distances (e_{cv}, B1, B2)

$\gamma = 1$	*0.08*, 3.84, 0.05	*0.08*, 3.84, 0.00	*0.08*, 3.84, 0.00
$\gamma = 0.5$	*0.04*, 2.93, 0.10	*0.04*, 2.93, 0.01	*0.04*, 2.93, 0.00
$\gamma = 0$	*0.15*, 2.73, 0.59	*0.08*, 2.73, 0.05	*0.08*, 2.73, 0.00
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$

It is evident from Table 13 that the largest value of B=B1+B2 will always occur on the axis λ=0 on the λ,γ grid; i.e. no regularisation of the individual covariance matrices towards the average covariance. This is the case for samples from any two normal distributions.

There are several reasons for this:

1) The value of B1 is not affected by the value of λ since the central component of it, $[\hat{\Sigma}_1(\lambda,\gamma)+\hat{\Sigma}_2(\lambda,\gamma)]$ /2, is nothing but the value towards which the individual covariances are biased anyway by the use of λ. Note that $(S_1(\lambda)+S_2(\lambda))/2$ always reduces to $(S_1+S_2)/2$:

$$(S_1(\lambda)+S_2(\lambda))/2 = \tfrac{1}{2}(1-\lambda)S_1 + \tfrac{1}{2}\lambda S_p + \tfrac{1}{2}(1-\lambda)S_2 + \tfrac{1}{2}\lambda S_p = (S_1+S_2)/2$$

where $S_p = \tfrac{1}{2}S_1+\tfrac{1}{2}S_2$.

2) The value of B2 decreases monotonically as λ increases, for fixed γ, and when λ=1, $\hat{\Sigma}_1(\lambda,\gamma)$ and $\hat{\Sigma}_2(\lambda,\gamma)$ are both equal to $(1-\gamma)S_p + \gamma\frac{tr[S_p]}{p}I$, where S_p is the pooled between-groups sample covariance matrix. Hence the numerator and denominator of B2 are equal and the term becomes zero.

3) Term B2 is always non-negative since for two p-dimensional positive definite matrices, A and B,

$$\sqrt{|A|}\sqrt{|B|} < \left|\frac{A + B}{2}\right|.$$

4) The value of B2 decreases monotonically as γ increases from 0 to 1, for fixed λ. An intuitive reason for this is as follows. The ratio

$$\frac{\left|\frac{\hat{\Sigma}_1 + \hat{\Sigma}_2}{2}\right|}{\sqrt{|\hat{\Sigma}_1|}\sqrt{|\hat{\Sigma}_2|}}$$

is a measure of the covariance shift between the two distributions and as the eigenvalues of the separate covariances are increasingly biased towards equality, the distributions become more similar in shape.

3.3 Model Selection

Since the regularisation parameter λ does not affect term B1, and monotonically diminishes term B2 as it increases, it is evident that an appropriate value for it in a given situation cannot be determined from information about B. Re-sampling methods can be employed to give a unique choice for λ. However these methods are computationally intensive, and since both the terms B1 and B2 exhibit the same behaviour in relation to λ for all values of γ, it is sensible to first turn our attention to choosing a value for γ so as to narrow down the search area for λ on the (λ,γ) plane.

Selection of the parameter γ

Increasing the value of the eigenvalue shrinkage parameter γ typically decreases the term B1, but not always, and the trend is not always monotonic. However from point 4 above we see that B2 exhibits only monotonic behaviour in relation to γ. So it seems sensible to first look at the behaviour of B1 for a range of γ.

From the empirical data we can identify three scenarios relating to B1:
1) Magnitude of B1 small, and not greatly affected by the value of γ changing between 0 and 1.
2) Magnitude of B1 large and not greatly affected by the value of γ changing between 0 and 1.
3) The effect on B1 of γ changing between 0 and 1 is large.

Now from the behaviour of primarily B1, and secondarily B2, calculated for various γ over $0 \leq \gamma \leq 1$, the following decision paths are proposed for the selection of an appropriate γ.

Under scenario 1 above, B1 is not providing much information as to an appropriate value of γ, so look at the effect of various γ on B2. If it is large, choose that γ which gives a minimal value of B1/B2, since in this case a dominant covariance shift over mean shift would seem to be important

for enhancing classification. If γ also has little effect on B2, choose that γ which leads to a maximal value of B1/B2.

Scenario 2 above indicates these are good conditions for classification due to the large Mahalanobis distance (B1) for all values of γ. Some average value of γ will suffice.

Under scenario 3, if γ has little effect on B2, it is clearly desirable to select that γ yielding a large value of B1. However if B2 is greatly affected by γ also, some greater degree of reduction in the variance of the system (by increasing γ a little) is desirable for classification purposes, whilst still maintaining a sizeable Mahalanobis distance (B1) between the groups.

The above guidelines lead to a simple flow chart for the selection of γ to use in equations (7) and (8) based on the three scenarios above and followed by the selection of λ using a re-sampling technique. The critical values at each decision stage have been arrived at empirically through observing the values of B1 and B2 for various random samples from various normal populations. The six simulation conditions proposed by Friedman (1989), and used in the present paper (section 2.2), offer a comprehensive set of group population distributions and n/p ratios from which to estimate these critical values.

Selection of the parameter λ

For the selection of the regularisation parameter λ, only the term B2 can be employed since B1 is constant over all values of λ for a given value of γ. However, the decrease in B2 from its maximum value at λ=0 to λ=1, when it is zero, is monotonic. Bootstrapping is used to estimate that upper bound on B2 for the selected value of γ ($\hat{\gamma}$) and λ=0, and this estimate is compared to the full-sample estimate of B2 for that same degree of regularisation and a unique value of the parameter λ obtained.

The magnitude of B2 when λ=0 and γ=1 gives further indication as to the similarity or dissimilarity of the group covariance estimates, and hence also indication as to an appropriate value of λ. Under this situation of maximal eigenvalue shrinkage the determinants of the group covariances are reduced to their average eigenvalue raised to the power of the dimension, p. If the group covariances are similar, the average of their eigenvalues will be similar in magnitude and the fraction in term B2 will be close to one, resulting in the value of B2 itself being close to zero. This being the case, the selected $\hat{\lambda}$, is raised to a power 1/k where k is proportional to 1/B2(0,1) (B2(a,b) denotes the value of B2 when λ=a and γ=b).

Thus model selection using the Bhattacharyya distance consists of the following steps:

1) Evaluate B1 and B2 from the available data for varying degrees of covariance eigenvalue shrinkage (a range of γ), but using no covariance mixing (λ=0).

2) Select $\hat{\gamma}$ using decision flow chart that implements the guidelines of this section .

3) Using the amount of eigenvalue shrinkage determined by the selected parameter value $\hat{\gamma}$, estimate the upper bound of the range of B2 using the re-sampling technique of bootstrapping.

The re-sampling technique is therefore used only at one point on the (λ,γ) plane. This contrasts with Friedman's RDA where a sample-reuse method (cross-validation) is performed at each of a whole grid of typically between 25 and 50 points. No matrix updating formulas are therefore required in this case which results in a greatly reduced computational burden (see section 3.6).

3.4 Discussion

The simulations performed with Friedman's RDA in earlier section have enabled us to observe that for a number of different simulation conditions, there is no unique selection of $\hat{\lambda}$ and $\hat{\gamma}$, using the criteria of minimum cross-validated error rate, and indeed altering the rule for the breaking of such ties had little effect on the overall performance of the procedure. In other words, the *degree* of regularisation (either covariance mixing or eigenvalue shrinkage, or both) is often not as important as its *presence* in any form.Thus it seems that complex methods to obtain a precise selection of $\hat{\lambda}$ and $\hat{\gamma}$ are not warranted.

Another conclusion from the simulation studies is that as the sample size to dimension ratio decreases, a degree of eigenvalue shrinkage using γ (i.e. $\gamma>0$) becomes more necessary to counteract the bias in the eigenvalues of the estimated covariances. Also, an increasing amount of regularisation away from $\gamma=0$ is required as p increases, even for those conditions where any shrinkage of the eigenvalues to equality would appear to be strongly counter-productive. (See, for example, Table (18) where the average $\hat{\gamma}$ value increases with dimension to substantial levels, even though no regularisation, or QDA, would seem to be the best option in these conditions.) The benefits of a decrease in variance from such shrinkage is proven to outweigh any introduced bias. The proposed method of selecting γ from the Bhattacharyya distance therefore only considers values of γ in the range $\theta \leq \gamma \leq 1$, where $\theta \geq 0$ but usually fairly close to zero and where θ depends on both the magnitude of p and the sample size to dimensionality ratio.

A goal of this model selection procedure using the Bhattacharyya distance is to provide a much faster algorithm to that proposed by Friedman using cross-validation. Also, the model selection procedure should choose appropriate levels of covariance mixing and eigenvalue shrinkage so that the classification rule obtained is comparable in performance to Friedman's RDA.

3.5 Simulation Studies

Computer simulation is used to compare the performance RDA, LDA, QDA, EDC and RDA-B (which denotes Regularised Discriminant Analysis using the Bhattacharyya distance measure to select the model) in the same variety of settings as that used by Friedman (1989), except that only two groups are present instead of three. In all cases the group distributions are normal and the total sample size from those distributions was 28, 14 from each group. For each set of conditions, simulations were performed for various levels of dimensionality: p=6, 10 and 20. The optimisation

grid for RDA was set as in Friedman's study. Since the sample size to dimensions ratio is less than one for some simulations, the zero eigenvalues of the group covariance matrix estimates were replaced by a small quantity, sufficient to permit numerically stable covariance inversion.

There were 100 repetitions of the experiment for each of the six settings. As before, random samples were drawn from specified multivariate normal distributions and were used to construct the classification rules for all five of the above methods. An additional test sample of size 100 was randomly generated from the same distributions and classified using each of the five rules obtained, yielding estimates of the error rate for each rule. These are presented in Tables (13) to (18), along with the mean and standard deviation of the selected regularisation parameters for RDA and RDA-B over the 100 replications. $\bar{\lambda}_{(RDA)}$ and ($\bar{\lambda}_{(RDA\text{-}B)}$ denote the mean value of λ for RDA and RDA-B respectively. The mean value of γ for each method is defined similarly.

Table 14

Equal Spherical Covariance Matrices (k=2 groups)

	p = 6	p = 10	p = 20
RDA	.08 (.03)	.09 (.03)	.11 (.04)
RDA-B	.08 (.03)	.09 (.03)	.10 (.04)
QDA	.16 (.06)	.29 (.07)	.32 (.06)
LDA	.10 (.04)	.14 (.05)	.24 (.07)
EDC	.08 (.03)	.09 (.03)	.10 (.03)
$\bar{\lambda}_{(RDA)}$	.86 (.30)	.94 (.18)	.94 (.20)
$\bar{\gamma}_{(RDA)}$	.73 (.34)	.86 (.22)	.76 (.28)
$\bar{\lambda}_{(RDA\text{-}B)}$	.85 (.21)	.84 (.21)	.84 (.16)
$\bar{\gamma}_{(RDA\text{-}B)}$	.94 (.11)	.93 (.07)	.84 (.26)

Table 15

Unequal Spherical Covariance Matrices (k=2 groups)

	p = 6	p = 10	p = 20
RDA	.11 (.04)	.11 (.04)	.08 (.05)
RDA-B	.11 (.04)	.09 (.05)	.10 (.09)
QDA	.20 (.05)	.32 (.08)	.35 (.07)
LDA	.15 (.04)	.20 (.06)	.32 (.07)
EDC	.13 (.04)	.15 (.05)	.18 (.05)
$\bar{\lambda}_{(RDA)}$	.46 (.37)	.35 (.35)	.28 (.27)
$\bar{\gamma}_{(RDA)}$	.80 (.31)	.77 (.30)	.88 (.20)
$\bar{\lambda}_{(RDA\text{-}B)}$	.15 (.21)	.09 (.11)	.04 (.03)
$\bar{\gamma}_{(RDA\text{-}B)}$	.72 (.37)	.86 (.24)	.77 (.34)

Table 16

Equal, Highly Ellipsoidal Covariance Matrices (k=2 groups)

(Mean differences in low variance subspace)

	p = 6	p = 10	p = 20
RDA	.02 (.04)	.05 (.03)	.13 (.05)
RDA-B	.01 (.01)	.06 (.05)	.13 (.05)
QDA	.02 (.02)	.16 (.08)	.28 (.07)
LDA	.01 (.01)	.03 (.02)	.15 (.07)
EDC	.09 (.04)	.10 (.04)	.16 (.05)
$\bar{\lambda}_{(RDA)}$	.96 (.17)	.96 (.14)	.87 (.28)
$\bar{\gamma}_{(RDA)}$	.20 (.36)	.29 (.31)	.49 (.31)
$\bar{\lambda}_{(RDA-B)}$	.94 (.05)	.84 (.15)	.81 (.19)
$\bar{\gamma}_{(RDA-B)}$	.01 (.10)	.36 (.44)	.68 (.39)

Table 17

Equal, Highly Ellipsoidal Covariance Matrices (k=2 groups)

(Mean differences in high variance subspace)

	p = 6	p = 10	p = 20
RDA	.03 (.02)	.03 (.02)	.05 (.03)
RDA-B	.02 (.02)	.02 (.02)	.03 (.02)
QDA	.07 (.04)	.19 (.09)	.23 (.08)
LDA	.03 (.02)	.06 (.04)	.15 (.07)
EDC	.03 (.02)	.03 (.02)	.04 (.02)
$\bar{\lambda}_{(RDA)}$	1.0 (.00)	.94 (.23)	.94 (.21)
$\bar{\gamma}_{(RDA)}$	.89 (.26)	.95 (.14)	.82 (.26)
$\bar{\lambda}_{(RDA-B)}$	.91 (.08)	.89 (.08)	.87 (.10)
$\bar{\gamma}_{(RDA-B)}$	.69 (.14)	.75 (.11)	.82 (.11)

Table 18

Unequal Highly Ellipsoidal Covariance Matrices (k=2 groups)

(Zero mean differences)

	p = 6	p = 10	p = 20
RDA	.18 (.08)	.13 (.06)	.05 (.03)
RDA-B	.18 (.06)	.10 (.05)	.05 (.04)
QDA	.17 (.06)	.22 (.09)	.20 (.05)
LDA	.47 (.06)	.47 (.07)	.44 (.06)
EDC	.47 (.05)	.46 (.05)	.43 (.05)
$\bar{\lambda}_{(RDA)}$	.13 (.12)	.12 (.12)	.15 (.11)
$\bar{\gamma}_{(RDA)}$	.12 (.26)	.39 (.29)	.67 (.29)
$\bar{\lambda}_{(RDA-B)}$	.16 (.09)	.10 (.06)	.04 (.03)
$\bar{\gamma}_{(RDA-B)}$	.19 (.31)	.33 (.35)	.63 (.34)

Table 19

Unequal Highly Ellipsoidal Covariance Matrices (k=2 groups)

(non zero mean differences)

	p = 6	p = 10	p = 20
RDA	.03 (.03)	.06 (.04)	.04 (.05)
RDA-B	.02 (.03)	.03 (.02)	.02 (.02)
QDA	.02 (.02)	.07 (.05)	.11 (.04)
LDA	.03 (.02)	.10 (.05)	.19 (.08)
EDC	.10 (.04)	.12 (.04)	.13 (.04)
$\bar{\lambda}_{(RDA)}$	.76 (.34)	.49 (.35)	.46 (.26)
$\bar{\gamma}_{(RDA)}$	.25 (.34)	.46 (.35)	.76 (.30)
$\bar{\lambda}_{(RDA\text{-}B)}$	.18 (.14)	.11 (.07)	.05 (.03)
$\bar{\gamma}_{(RDA\text{-}B)}$	.16 (.23)	.37 (.38)	.35 (.30)

3.6 Results

In all the various conditions tested it is clear that RDA and RDA-B yield very similar error rates over the 100 replications. There are 18 sets of simulations represented in tables (13) to (18). In ten of these cases, RDA-B performs slightly better (and with a reduced standard deviation) than RDA in terms of estimated error rate, and in two of the cases RDA has a slightly lower error rate. Thus overall, neither technique is superior to the other in terms of experimental classification error rates. The average regularisation parameter values for RDA and RDA-B show that for both methods, the model selection procedures tend to do the right thing by introducing appropriate degrees of each type of regularisation for the various simulation conditions.

The standard deviations of the selected regularisation parameters tended to be smaller for RDA-B, perhaps because of the more direct nature of the path taken to select the pair of values $(\hat{\lambda}, \hat{\gamma})$ in the parameter selection procedure in RDA-B compared with RDA. Furthermore, the model selection process in RDA-B affords a unique choice of the estimated best pair of values $(\hat{\lambda}, \hat{\gamma})$, without having to break ties in an arbitrary way as for RDA.

In conclusion, it can be established that the Bhattacharyya distance between groups does indeed provide information as to appropriate regularisation parameter values to use in equation (8). This can be used to obtain a classification rule which seeks to minimise the actual error rate for data from two specified normal distributions. Unfortunately, no tidy, direct theoretical relationship exists between components of the Bhattacharyya distance and the error rate. Instead we have derived the model selection procedure based on empirical data and it can be seen to perform as well, at least under the tested range of simulated conditions, as the model selection procedure developed by Friedman (1989) in the RDA method.

Finally, a substantial advantage of the model selection procedure in RDA-B over that of RDA relates to the computation time required for each. The table below gives approximate ratios (RDA-B/RDA) of CPU times for various dimensions.

p=6	p=10	p=20
.15	.12	.08

These results indicate the the gain in computational efficiency in using RDA-B over RDA.

References

Friedman, J.H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84**, 165-175.

Fukunaga, K. and Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-11**, 873-885.

Jain, A. K. (1976). On an estimate of the Bhattacharyya distance. *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 763-766.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Tech.* **COM-15**, 52-60.

Rayens, W and Greene, T. (1991). Covariance pooling and stabilization for classification. *Comput. Statist. Data Anal.* **11**, 17-42.

REGULARIZED DISCRIMINANT(CLASSIFICATION) ANALYSIS INVOLVING BHATTACHARYYA DISTANCE MEASURE

J.P. Koolaard, C.R.O. Lawoko[1], S. Ganesalingam
Department of Statistics, Massey University
Private Bag 11222, Palmerston North, New Zealand
Telephone 06 350 4261 Fax 06 350 5611 Email: C.Lawoko@massey.ac.nz

ABSTRACT

Friedman (1989) proposed a Regularised Discriminant Function (RDF) as a compromise between normal-based Linear and Quadratic Discriminant Functions, by considering alternatives to the usual maximum likelihood estimates for the covariance matrices. These alternatives are characterised by two (regularisation) parameters, the values of which are customized to individual situations by jointly minimising a sample-based(cross-validated) estimate of future misclassification risk. This technique appears to provide considerable gains in classification accuracy in many circumstances, although it is computationally intensive.

Because of the computational burden inherent in RDF, and with regard to criticisms of the technique by Rayens et. al. (1991), we investigated whether information about appropriate values of the two regularisation parameters could be gleaned by examining the behaviour of the Bhattacharyya Distance between the various populations. A classification rule for the two (normal) population case which uses regularization parameters obtained from the Bhattacharyya distance (and which is computationally much faster than Friedman's RDF) is presented and compared with the original RDF.

1. INTRODUCTION

Regularized discriminant analysis was introduced by Friedman (1989) as an alternative to the common normal-theory-based discriminant functions, such as the nearest-mean (euclidean distance) classifier(EDF), the linear discriminant function(LDF) and the quadratic discriminant function(QDF). Simulation results by various authors suggest that regularized discriminant analysis can perform much better than these other normal-theory based discriminant functions (see, for example, Friedman (1989), Rayens et. al. (1991)). Experiences of the authors of this article also confirm these results. Meanwhile, a recent article by Aeberhard et. al. (1994) reported results which found that the regularized discriminant function (RDF) performed much better than seven other discriminant functions, including several non-parametric ones.

To introduce the notation, suppose we have multivariate (p-dimensional) measurements ($\mathbf{x}$) on each object (pixel), where each object belongs to one of K classes. In order to apply normal-theory based classification, it is usually assumed (correctly or incorrectly) that the multivariate normal distribution can adequately describe the distribution of measurements from each class. Let us denote the population mean and covariance function for group i by

[1] Author to whom correspondence should be addressed. Address after February 1, 1996: Faculty of Business, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia.

μ_i and Σ_i respectively. These are usually estimated by the sample mean $\bar{x}_i$ and sample covariance matrix S_i, from the training sample. If it is assumed that the covariance matrices Σ_i are equal for all K classes, then the common value Σ is estimated by

$$S_p = \sum_{i=1}^{K} (n_i/n)\ S_i \qquad (1)$$

On the basis of these estimates, the sample classification functions based on the normal distribution would allocate a pixel to class k according to the following rules:

(i) Sample euclidean distance function (SEDF, or the nearest mean classifier)

$$\text{SEDF(k)} = \min \text{ (over i) } \{(x - \bar{x}_i)'(x - \bar{x}_i)\}$$

(ii) Sample linear discriminant function (SLDF)

$$\text{SLDF(k)} = \min \text{ (over i) } \{\bar{x}_i' S_p^{-1} x - \tfrac{1}{2} \bar{x}_i S_p^{-1} \bar{x}_i + \ln \pi_i$$

(iii) Sample quadratic discriminant function (SQDF)

$$\text{SQDF(k)} = \min \text{ (over i) } \{(x - \bar{x}_i)' S_i^{-1}(x - \bar{x}_i) + \ln|S_i| - 2 \ln \pi_i\} \qquad (2)$$

The motivation for developing the RDF was partly to do with the well-known problems associated with estimating Σ_i by S_i, which leads to relatively poor performance of the SQDF, especially when the sample size to dimension ratio (i.e. n_i: p ratio) is small. [This problem is discussed in Friedman (1989), and also in an article at this conference (Lawoko et. al. (1996)]. This led Friedman (1989) to propose a regularization parameter, λ, which controls the regularization of S_i to S_p, thereby controlling the degree to which Σ_i is estimated by pooled information from the several S_i matrices, or by each S_i separately. Thus the initial proposal for the sample RDF (SRDF) is to use the SQDF in (2), but with S_i estimated by $\hat{\Sigma}_i(\lambda)$, defined as

$$\hat{\Sigma}_i(\lambda) = \frac{(1-\lambda)(n_i - 1) S_i + \lambda S_p}{(1-\lambda)(n_i - 1) + \lambda(N-K)}, \qquad (3)$$

where $N = n_1 + n_2 + \ldots + n_K$, and $0 \le \lambda \le 1$.

Because of well-known problems of bias associated with estimating the eigenvalues of the covariance matrix (and since there are situations where one wants to perform a discriminant analysis when $n_i \geq p$), the above regularization may still not be adequate. Thus Friedman proposed further regularization beyond that in (3) by providing an option for regularizing the eigenvalues of $\hat{\Sigma}_i(\lambda)$ towards equality using a second regularization parameter, γ. Consequently, the estimate of Σ_i used is given by

$$\hat{\Sigma}_i(\lambda, \gamma) = (1 - \gamma)\, \hat{\Sigma}_i(\lambda) + \frac{\gamma}{p} \text{tr}[\hat{\Sigma}_i(\lambda)]\, I, \qquad (4)$$

where $\hat{\Sigma}_i(\lambda)$ is given in (3) and I is the identity matrix. Note that this shrinkage has the effect of decreasing the larger eigenvalues and increasing the smaller ones, to counter the bias in sample estimates of the eigenvalues of covariance matrices.

2. PROBLEMS ASSOCIATED WITH IMPLEMENTING RDF

For the RDF to be implemented, λ and γ (plus all the other parameters) need to be estimated from the data. Friedman proposed that the (λ, γ) value chosen should be that which minimizes the cross-validated error rate of a training sample through a grid-search procedure. In spite of Friedman's tremendous work in deriving matrix algebraic relationships which reduce the computational burden significantly, this is still a very computationally intensive procedure. It is therefore of interest to consider alternative methods of estimating the appropriate (λ, γ) combination for a given set of data.

Another problem associated with the RDF, which was discussed in detail by Rayens et. al. (1991) is that since the estimated value of (λ, γ) is obtained on the basis of error rates (i.e. empirical misclassification rates), those objects (pixels) which are correctly classified for most (λ, γ) values in the grid do not contribute to the estimation of λ and γ. It follows that in many practical situations, only a very small fraction of the training data may determine the values of λ and γ. It is therefore of interest to investigate if other methods which use all the training data to estimate λ and γ may perform better. Incidentally, the regularization method developed by Rayens et. al. (1989, 1991) uses all the data to estimate their regularisation parameters (which are not λ and γ), although they use Friedman's γ in addition to their parameters in the 1991 paper.

One further reason for considering alternative ways of estimating λ and γ relates to the empirical evidence that the error rate surface seems to be fairly flat in a very wide neighbourhood of the minimum. Our own research into alternative ways of "breaking ties" (in the case of several local minima) suggest that any of the values of $(\hat{\lambda}, \hat{\gamma})$ which determine the local minima could be chosen without any serious changes to the performance of the RDF. It follows (from the "inexactness" of the values of $\hat{\lambda}$ and $\hat{\gamma}$ required for successful implementation of the RDF), that it may not be necessary to go through the (required) intensive computation in the cross-validation method (and still get the RDF to perform reasonably well).

In view of these issues discussed in this section, we investigated the use of the Bhattacharyya distance measure, as an alternative to the cross-validated error rate, in determining the optimal values of $\hat{\lambda}$ and $\hat{\gamma}$.

3. SOME PROPERTIES OF THE BHATTACHARYYA DISTANCE MEASURE

Distance measures have often been considered as alternatives to error rates as a criterion for choosing among various options. For example, Jain (1976) investigated the behaviour of an estimate of the Bhattacharyya distance when used as a criterion in variable selection. It was shown that the bias and variance of the estimate is related to the number of training samples and parameter values of the distribution. Kailath (1967) addressed the problem that minimising the error rate to determine optimum classification can be difficult to accomplish in practice. He investigated the idea of using simpler, albeit sub-optimal performance measures instead of the error rate, and compared the Bhattacharyya distance with an often-used measure, the divergence, which is closely related to Shannon's logarithmic measure of information. Not only is the Bhattacharyya distance easier to evaluate than the divergence, but

in some examples in the study it was found to perform at least as well as the divergence in minimising the probability of misclassification. Kailath obtained an upper bound on the probability of misclassification in terms of the Bhattacharyya distance, in the case of equal prior probabilities of the distributions. Note that Kailath (1967) only treated the case of two populations. That work also assumed knowledge of the parameters, whereas, as we shall see later, if one has to use sample estimates of the parameters, the link between Bhattacharyya distance and error rate is a lot less clear. Also, Fukunaga et. al. (1989) obtained an upper bound, in terms of the Bhattacharyya distance, on the Bayes error for classifying between two Gaussian distributions.

The Bhattacharyya distance between two multivariate normal density functions with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is

$$B = B1 + B2$$

where

$$B1 = \frac{1}{8}(\mu_1 - \mu_2)' \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_1 - \mu_2)$$

and

$$B2 = \frac{1}{2} \ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|}\sqrt{|\Sigma_2|}} \quad . \tag{5}$$

The first term of the expression, B1, is similar to the well-known Mahalanobis distance between the densities. It measures the distance between the two distributions caused by the mean shift. The second term B2 utilises the determinants of each covariance matrix as well as that of the average (class) covariance matrix. It gives a measure of the difference between the two distributions due to the covariance shift. Fukunaga et. al. (1989) derived expressions for the expected bias and variance of the terms B1 and B2 and showed that the bias of term B1 is inversely proportional to (n/p) (n_i = n = sample size). i.e. decreases as the ratio (n/p) increases. They also showed that the bias of term B2 is inversely proportional to $\frac{n}{(p+1)p}$. In other words, estimates of this distance measure become increasingly biased as the ratio (n/p) decreases, with term B2 more seriously affected than B1. Thus, when the (n/p) ratio increases the bias present in the Bhattacharyya distance estimate is dominated by the bias inherent in estimation of the term B2. They also showed that as the dimensionality increases, an increasingly large ratio of (n/p) is needed to maintain a constant expected value of B.

With the above knowledge of the Bhattacharyya distance function between two Gaussian distributions, it is plausible to expect that some degree of regularization of the covariances, such as is provided for by the two-parameter model in equation (4), would improve the estimation of the Bhattacharyya distance. The reason for this stems from the accepted knowledge that sample covariance estimates S_i yield eigenvalue estimates which are biased. The largest eigenvalues are biased towards high values and the smallest ones are biased towards values which are too low. This bias will be worse in the situation where the true population eigenvalues are approximately equal, but in all cases this bias becomes more pronounced as the ratio of sample size to dimension decreases. The term B2 of the Bhattacharyya distance is most vulnerable to such bias occurring, being a ratio of determinants of sample covariance estimates, and eigenvalue shrinkage (regularization) ought to prove useful in counteracting bias-induced anomalies in estimates of B2, particularly as p

becomes large (relative to n_i).

4. BEHAVIOUR OF BHATTACHARYYA DISTANCE WITH REGULARIZED COVARIANCES, AND CHOOSING λ AND γ

Kailath (1967) admitted that one would not expect to obtain a strong relationship between distance measures and error rate. Nevertheless, the author was able to obtain several useful theoretical results, assuming known population parameters. In the present covariance regularization context with two parameters controlling shrinkage, as in equations (3) and (4), it would be too optimistic to expect that the $(\hat{\lambda}, \hat{\gamma})$ combination which maximises the Bhattacharyya distance for a given set of data will also yield a classification rule which minimises the future misclassification risk. Instead, from the example below, we can detect no strong relationship between sample Bhattacharyya distance and minimum error rate. That is, Table 1 shows the components B1 and B2 of the Bhattacharyya distance at a range of points over the λ,γ grid. The cross-validated error rate (e_{cv}) at each point is also stated to give an indication of where the range in which the minimum error rate lies. The data set consisted of samples of size 13 from each of two normal populations with means and covariances as in Table 2 (Condition II).

Table 1. Example of λ,γ Grid of Bhattacharyya Distances (e_{cv}, B1, B2)

γ = 1	*0.08*, 3.84, 0.05	*0.08*, 3.84, 0.00	*0.08*, 3.84, 0.00
γ = 0.5	*0.04*, 2.93, 0.10	*0.04*, 2.93, 0.01	*0.04*, 2.93, 0.00
γ = 0	*0.15*, 2.73, 0.59	*0.08*, 2.73, 0.05	*0.08*, 2.73, 0.00
	λ = 0	λ = 0.5	λ = 1

It is evident from Table 1 that the largest value of B=B1+B2 will always occur on the axis λ=0 on the λ,γ grid; i.e. no regularisation of the individual covariance matrices towards the average covariance. This is the case for samples from any two normal distributions. There are several reasons for this:

1) The value of B1 is not affected by the value of λ since the central component of it, $[\hat{\Sigma}_1(\lambda,\gamma)+\hat{\Sigma}_2(\lambda,\gamma)]/2$, is not changed by λ for a fixed value of γ.

2) The value of B2 decreases monotonically as λ increases, for fixed γ. And when λ=1, $\hat{\Sigma}_1(\lambda,\gamma)$ and $\hat{\Sigma}_2(\lambda,\gamma)$ are both equal to $(1-\gamma)\,S_p + \gamma\frac{tr[S_p]}{p}\,\mathbf{I}$. Hence the numerator and denominator of B2 are equal and the term becomes zero.

3) Term B2 is always non-negative since for two p-dimensional positive definite matrices, **A** and **B**,

$$\sqrt{|\mathbf{A}|}\sqrt{|\mathbf{B}|} < \left|\frac{\mathbf{A}+\mathbf{B}}{2}\right| .$$

4) The value of B2 decreases monotonically as γ increases from 0 to 1, for fixed λ. Since B2 is fundamentally a measure of the covariance shift between the two distributions, if the eigenvalues of the separate covariances are increasingly biased

towards equality, the distributions become more similar in shape.

Since the regularisation parameter λ does not affect term B1, and monotonically decreases as the term B2 increases, it would seem that an appropriate value for it in a given situation cannot be determined from information about B. Re-sampling methods can be employed to give a unique choice for λ for given sets of data. However these methods are computationally intensive, and since both the terms B1 and B2 exhibit similar behaviour in relation to λ for all values of γ, it is sensible to first choose a value for γ so as to narrow down the search area for λ on the (λ,γ) grid. Also, the bias inherent in the estimate of B1 would be expected to be less than that in estimates of B2, so that the principle upon which selection of the regularization parameter γ is made involves giving B1 greater importance than B2. Thus, in general the aim is to choose that γ which gives a large or maximal value of B1 or B1/B2. In situations where the distribution means are close together and B1 is small, a minimal value of B1/B2 is needed since in this case a dominant covariance shift over mean shift would seem to be important in enhancing classification.

The technique which employs the Bhattacharyya distance to select λ and γ is denoted as RDF-B. A detailed description of the (heuristic) algorithm for choosing λ and γ is not possible in this article (space limitations) but the steps involved will be discussed at the conference presentation, and will be reported elsewhere.

5. DISCUSSION AND SIMULATION STUDIES

As mentioned earlier, the simulations performed to investigate the behaviour of Friedman's RDF have enabled us to observe that for a number of different simulation conditions, there is no unique selection of $\hat{\lambda}$ and $\hat{\gamma}$, using the criteria of minimum cross-validated error rate. Indeed, altering the rule for the breaking of such ties had little effect on the overall performance of the procedure. In other words, the *degree* of regularisation (either covariance mixing or eigenvalue shrinkage, or both) is often not as important as its *presence* in any form. Thus it would seem that complex methods to obtain a precise selection of $\hat{\lambda}$ and $\hat{\gamma}$ are not warranted.

Another conclusion from the simulation studies is that as the sample size to dimension (n:p) ratio decreases, a degree of eigenvalue shrinkage using γ (i.e. $\gamma>0$) becomes more necessary in order to counteract the bias in the eigenvalues of the estimated covariances. Also, an increasing amount of regularisation away from $\gamma=0$ appears to be required as p increases, even for those conditions where any shrinkage of the eigenvalues to equality would appear to be strongly counter-productive. The benefits of a decrease in variance from such shrinkage appears to outweigh any introduced bias. The proposed method of selecting γ from the Bhattacharyya distance therefore only considers values of γ in the range $\theta \leq \gamma \leq 1$, where $\theta > 0$ but usually fairly close to zero and where θ depends on both the magnitude of p and the sample size to dimensionality ratio.

A goal of this model selection procedure using the Bhattacharyya distance is to provide a much faster algorithm to that proposed by Friedman using cross-validation. Also, the model selection procedure should choose appropriate levels of covariance mixing and eigenvalue shrinkage so that the classification rule obtained is comparable in performance to Friedman's

RDF. In the case of discriminating between more than two classes, one can either use the average regularisation parameter values of all possible pairs of classes, or introduce a separate λ and γ value for each class. This matter is currently under investigation by the authors.

Computer simulation was used to compare the performances of RDF (Friedman's RDF method) and RDF-B (which denotes RDF using the Bhattacharyya distance measure to select the model) in the same variety of settings as that used by Friedman (1989), except that only two classes are present instead of three. Further details of these six settings or conditions are given in the article by Lawoko et. al., in the "Proceedings" of this conference. In all cases the class distributions are normal and the total sample size from those distributions was 28 i.e. 14 from each class. For each set of conditions, simulations were performed for various levels of dimensionality: p=6, 10 and 20. The optimisation grid for RDF was set as in Friedman's study. Since the sample size to dimensions ratio is less than one for some simulations, the zero eigenvalues of the class covariance matrix estimates were replaced by a small quantity, sufficient to permit numerically stable covariance inversion(as done in Friedman(1989)).

Table 2.

	p = 6			p = 10			p = 20		
Condition*	I	II	III	I	II	III	I	II	III
RDF	.08 (.03)	.03 (.02)	.18 (.08)	.09 (.03)	.03 (.02)	.13 (.06)	.11 (.04)	.05 (.03)	.05 (.03)
RDF-B	.08 (.03)	.02 (.02)	.18 (.06)	.09 (.03)	.02 (.02)	.10 (.05)	.10 (.04)	.03 (.02)	.05 (.04)
$\bar{\lambda}$ (RDF)	.86 (.30)	1.0 (.00)	.13 (.12)	.94 (.18)	.94 (.23)	.12 (.12)	.94 (.20)	.94 (.21)	.15 (.11)
γ (RDF)	.73 (.34)	.89 (.26)	.12 (.26)	.86 (.22)	.95 (.14)	.39 (.29)	.76 (.28)	.82 (.26)	.67 (.29)
$\bar{\lambda}$ (RDF-B)	.85 (.21)	.91 (.08)	.16 (.09)	.84 (.21)	.89 (.08)	.10 (.06)	.84 (.16)	.87 (.10)	.04 (.03)
γ (RDF-B)	.94 (.11)	.69 (.14)	.19 (.31)	.93 (.07)	.75 (.11)	.33 (.35)	.84 (.26)	.82 (.11)	.63 (.34)

* Condition I: Equally spherical covariance matrices.
Condition II: Equal, highly ellipsoidal covariance matrices, with mean difference in the high variance subspace
Condition III: Unequal highly ellipsoidal covariance matrices with zero mean differences

There were 100 repetitions of the experiment for each of the six settings. As before, randomsamples were drawn from specified multivariate normal distributions and were used to construct the classification rules. An additional test sample of size 100 was randomly generated from the same distributions and classified using the two rules, yielding estimates of the error rate for each rule. These are presented in Table 2 (with sample standard deviations in brackets). The mean and standard deviation of the selected regularization parameters for RDF and RDF-B over the 100 replications are also given, with $\bar{\lambda}_{(RDF)}$ and $\bar{\lambda}_{(RDF\text{-}B)}$ denoting the mean values of λ for RDF and RDF-B respectively. The corresponding values for γ for each method are defined similarly.

6. RESULTS

In all the various conditions tested it is clear that RDF and RDF-B yield very similar error rates over the 100 replications, which are generally better or comparable to the best of the SEDF, SLDF, and SQDF. Thus the average regularization parameter values for RDF and

RDF-B show that for both methods, the model selection procedures tend to do the right thing by introducing appropriate degrees of each type of regularization for the various simulation conditions. The standard deviations of the selected regularization parameters tended to be smaller for RDF-B, perhaps because of the more direct nature of the path taken to select the pair of values $(\hat{\lambda}, \hat{\gamma})$ in the parameter selection procedure in RDF-B ,compared with RDF. Furthermore, the model selection process in RDF-B provides a unique choice of the estimated best pair of values $(\hat{\lambda}, \hat{\gamma})$, without having to break ties in an arbitrary way as in the RDF.

In conclusion, it can be established that the Bhattacharyya distance between two classes does indeed provide adequate information about the appropriate regularization parameter values to use in equations (3) and (4). This can be used to obtain a classification rule which approximately minimises the error rate for data from two specified normal distributions. Unfortunately, no tidy, direct theoretical relationship exists between components of the Bhattacharyya distance and the error rate. Instead we have derived the model selection procedure based on empirical evidence and observations. It does perform approximately as well, however, as the model selection procedure developed by Friedman (1989) in the original RDF method.

Finally, a substantial advantage of the model selection procedure in RDF-B over that of RDF relates to the computation time required for each. The table below gives approximate ratios (RDF-B/RDF) of CPU times for various dimensions.

p=6	p=10	p=20
.15	.12	.08

These results indicate the gain in computational efficiency in using RDF-B over RDF.

REFERENCES

Friedman, J.H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84**, 165-175.

Fukunaga, K. and Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-11**, 873-885.

Jain, A. K. (1976). On an estimate of the Bhattacharyya distance. *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 763-766.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Tech.* **COM-15**, 52-60.

Lawoko, C and Koolaard, J.P. (1996). Applications of regularized discriminant (classification) functions in the classification of objects: a discussion of potential application to remote sensing. Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, March (1996).

Rayens, W and Greene, T. (1991). Covariance pooling and stabilization for classification. *Comput. Statist. Data Anal.* **11**, 17-42.

APPLICATIONS OF REGULARIZED DISCRIMINANT (CLASSIFICATION) FUNCTIONS IN THE CLASSIFICATION OF OBJECTS: A DISCUSSION OF POTENTIAL APPLICATION TO REMOTE SENSING

Charles R O Lawoko[1] & John P Koolaard
Department of Statistics, Massey University
Private Bag 11222, Palmerston North, New Zealand
Telephone 06 350 4261 Fax 06 350 5611 Email: C.Lawoko@massey.ac.nz

ABSTRACT

In this article we discuss the performance and properties of regularized discriminant functions (RDFs). These are classification functions which can be intermediary among the common discriminant functions, namely the nearest-mean (euclidean distance) method, linear discriminant function, and the quadratic discriminant function. It has been demonstrated by several researchers that the RDF can out-perform most of the common discriminant functions, including non-parametric ones; see, for example, Friedman (1989), Rayens et. al. (1989, 1992), and Aeberhard et. al. (1994).

In spite of its impressive performance, the RDF has some drawbacks, like computational intensity and lack of scale invariance. We report and discuss results from simulation experiments which investigated some of these properties. Alternative means of estimating the regularization parameters are also introduced and discussed. It must be pointed out that because of space limitations, very few results have been included in this article, although more results(from remotely sensed and GIS data) will be discussed in the presentation.

1. INTRODUCTION

Consider the problem of classification or discriminant analysis, where we want to classify an object (pixel) to one of several (K) groups (classes) based on multivariate (p-dimensional) data ($\mathbf{x}$) on each object. Common supervised classification methods based on the normal distribution are the Nearest-mean classifier or the Euclidean discriminant function (EDF), the Linear discriminant function (LDF), and the Quadratic discriminant function (QDF). Specifically, suppose the population mean and covariance matrices for class i (i = 1, 2, ..., K) are μ_i and Σ_i respectively, and π_i is the prior probability of class i. Assuming normality, the three discriminant functions allocate a pixel with observation vector $\mathbf{x}$ to class k according to the following well-known rules:

(i) EDF: $\text{EDF(k)} = \min \text{ (over i)}\{(\mathbf{x} - \mu_i)' (\mathbf{x} - \mu_i)\}$

(ii) LDF: $\text{LDF(k)} = \min \text{ (over i)}\{\mu_i' \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mu_i' \Sigma^{-1} \mu_i + \ln\pi_i\}$

(iii) QDF: $\text{QDF(k)} = \min \text{ (over i)}\{\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \ln|\Sigma_i| - 2 \ln\pi_i\}$ (1)

Clearly, the unknown population parameters in expression (1) have to be estimated from training samples. Usually μ_i and Σ_i are estimated by unbiased and consistent estimators,

---

[1]Author to whom correspondence should be addressed. Address after February 1, 1996: Faculty of Business, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland, 4001, Australia.

which are the sample mean ($\bar{x}_i$) and the sample covariance matrix (S_i). In the case of LDF, the common value of Σ is estimated by the pooled sample covariance matrix, S_p.In practice therefore,the sample discriminant functions used are:

(i) Sample EDF: SEDF(k) = min (over i)$\{(x - \bar{x}_i)' (x - x_i)\}$

(ii) Sample LDF: SLDF(k) = min (over i)$\{\bar{x}_i' S_p^{-1} x - \frac{1}{2} \bar{x}_i' S_p^{-1} \bar{x}_i + \ln\pi_i\}$

(iii) Sample QDF: SQDF(k) = min (over i)$\{(x - \bar{x}_i)' S_i^{-1}(x - \bar{x}_i) + \ln|S_i| - 2 \ln \pi_i\}$ (2)

(see, for example, McLachlan (1992) or Fukunaga (1990)).

In practical applications, the SEDF is used only under very special circumstances because of the restrictive assumption that $\Sigma = I$ (the identity matrix). Meanwhile, the SLDF has been shown to be quite robust to violations of the (required) $\Sigma_i = \Sigma$ assumption, and to non-normality of data. Consequently the SLDF is a quite popular classification procedure. Meanwhile the SQDF, which should be the most widely used, suffers from the fact that it has a large number of unknown parameters which must be estimated from the training data. It is also quite sensitive to violations of the normality assumption. This means that the SQDF requires very high n_i: p ratios for successful implementation. For example, it has been shown that SLDF can out-perform the SQDF for small to moderate (n_i: p) ratios, even if the covariance matrices are quite different. This matter is of relevance to the classification of remotely sensed data because, even if there are usually a considerable abundance of data, there are usually problems with finding good training data for some groups.

The relative underperformance of the SQDF is partly due to the estimation of Σ_i. This can be demonstrated by representing Σ_i by its spectral decomposition. That is, rewrite Σ_i as (see, for example, Friedman (1989)):

$$\Sigma_i = \sum_{j=1}^{p} e_{ji} v_{ji} v_{ji}', \quad \text{so that} \quad \Sigma_i^{-1} = \sum_{j=1}^{p} \frac{v_{ji} v_{ji}'}{e_{ji}} \qquad (3)$$

where e_{ji} is the j^{th} eigenvalue of Σ_i and v_{ji} is its corresponding eigenvector.

Thus the QDF discriminant score in (1) becomes

$$QDF(k) = \sum_{j=1}^{p} \frac{[v_{ji}'(x - \mu_i)]^2}{e_{ji}} + \sum_{j=1}^{p} \ln e_{ji} - 2 \ln\pi_i \qquad (4)$$

Expression (4) demonstrates the fact that small eigenvalues may have a disproportionately large effect on the discriminant score. This problem is worsened when Σ_i is estimated by S_i because the eigenvalues of S_i are well-known to be biased estimates of the eigenvalues of Σ_i, and the bias is usually more pronounced if the eigenvalues of Σ_i are similar. The problem is further compounded by the facts that the largest eigenvalues are biased upwards (i.e. high) and the smallest ones biased towards values which are even smaller. Also, the problem worsens as the (n_i: p) ratio decreases. The consequences of all this is that the importance of the eigenvalues and vectors associated with the low-variance subspace (i.e. small eigenvalues) in a classification problem is greatly exaggerated. Thus, as noted by Friedman (1989) "... most of the variance incurred in estimating discriminant scores is associated with direction of low sample variance in the measurement space". One approach

to address this problem is to employ a regularization method which works by biasing sample estimates away from their usual sample-based values towards what one believes to be more plausible values, which serves to reduce the variance associated with S_i . These methods were used by Friedman (1989) and Greene et. al. (1989), leading to two different versions of regularized discriminant functions (RDFs), which are effectively "middle-of-the-road" classification functions, between EDF, LDF and QDF.

2. REGULARISED DISCRIMINANT FUNCTIONS (RDFs)

A researcher who is applying normal-theory-based classification might suspect that $\Sigma_i = \Sigma$ for all i. One approach which enables the researcher to decide between SLDF and SQDF would be to initially perform a test of H_0: $\Sigma_1 = \Sigma_2 = ... = \Sigma_k = \Sigma$ (say), and use either the SLDF or the SQDF depending on the outcome of the test. An alternative approach is to introduce a regularization parameter ω, which controls the degree of regularization (shrinkage) of the S_i to S_p. Thus S_i in the SQDF in expressions (2) is replaced by

$$\hat{\Sigma}_i(\omega) = \omega_i S_i + (1 - \omega) S_p \quad (0 \leq \omega \leq 1), \tag{5}$$

where ω is determined from the data. Note that at one extreme ($\omega = 1$) $\hat{\Sigma}_i(\omega) = S_i$, and at the other extreme ($\omega = 0$), $\hat{\Sigma}_i(\omega)$ is S_p. Variations of this middle-of-the-road type of discriminant function were developed independently by Friedman (1989) and Greene et. al. (1989). The two regularized discriminant functions will now be introduced separately.

(i) Green and Raynes (1989)

In their paper, these authors obtained empirical Bayes formulation for estimating Σ_i. That is, assuming that the training data from group i are i.i.d. $N_p(\mu_i, \Sigma_i)$, it follows that (conditionally on Σ_i)

$$(n_i - 1) S_i \sim W_p (\Sigma_i, (n_i - 1)), \tag{6}$$

where $W_p(.)$ denotes the central Wishart distribution with parameter matrix Σ_i and degrees of freedom $(n_i - 1)$. They then assume a conjugate prior distribution for Σ_i, which is the inverted Wishart distribution. That is, that Σ_i are mutually independent with

$$\Sigma_i \sim W_p^{-1} ((\alpha - p - 1)\Psi, \alpha) \tag{7}$$

where $\alpha > p + 1$, Ψ is the matrix of hyperparameters and α represents the degree of "concentration" of Σ_i around Ψ. In particular, it can be established that

$$E(\Sigma_i) = \Psi, \text{ and for } \alpha > p + 3 \quad \text{and} \quad 1 \leq h, j, k, l \leq p,$$

$$\text{cov}\,[(\Sigma_i)_{hj}, (\Sigma_i)_{kl}] = \frac{(\alpha-p-1)}{(\alpha-p)(\alpha-p-3)} [\Psi_{hk} \Psi_{jl} + \Psi_{hl} \Psi_{kj}] + \frac{2}{(\alpha-p)(\alpha-p-3)} \Psi_{hj}\Psi_{kl} \tag{8}$$

After some algebra and further results, it can be shown that the empirical Bayes estimate of Σ_i for known α is:

$$\hat{\Sigma}_i(\alpha) = \frac{d_i}{d_i + \alpha - p - 1} S_i + \frac{(\alpha - p - 1)}{(d_i + \alpha - p - 1)} S_p(\alpha), \tag{9}$$

where $d_i = (n_i - 1)$. The unknown parameter α is estimated by either conditionally

maximizing the marginal likelihood of $S_1, S_2, ..., S_p$ over α or using a method-of-moments type estimator. Details of this nontrivial computational task are given in their paper.

(ii) Friedman (1989)

Friedman proposed a regularization parameter, λ, which controls the shrinkage (regularization) of S_i to S_p, whereby Σ_i in (1) is estimated by

$$\hat{\Sigma}_i(\lambda) = \frac{(1-\lambda)(n_i - 1) S_i + \lambda S_p}{(1-\lambda)(n_i - 1) + \lambda(N-K)}, \tag{10}$$

where $N = n_1 + n_2 + ... + n_K$, and $0 \leq \lambda \leq 1$.

Further regularisation can be achieved by shrinking the eigenvalues of each $\hat{\Sigma}_i(\lambda)$ towards equality, so that the resulting estimates (of eigenvalues) become multiples of the identity matrix. The consequence of implementing these two dimensions of regularization is to replace Σ_i in (1) by

$$\hat{\Sigma}_i(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}_i(\lambda) + \frac{\gamma}{p} \text{tr}[\hat{\Sigma}_i(\lambda)]\mathbf{I} \tag{11}$$

Thus λ controls regularization of S_i to S_p while γ (simultaneously) controls regularization to $\text{tr}[\hat{\Sigma}_i(\lambda)]/p$, the average of the p eigenvalues of $\hat{\Sigma}_i(\lambda)$ in (10). The appropriate values of γ and λ need to be determined from the data, and the approach proposed by the author is to choose a (λ, γ) combination which minimizes the cross-validated estimate of future (expected) misclassification error, on the basis of available training data. Implementation of this cross-validation strategy is a computationally intensive problem, which Friedman simplifies to a limited extent by deriving some algebraic results. In spite of this simplification, it is still a rather slow process.

(iii) Rayens and Green (1992)

As a consequence of the ideas in Friedman's article, Rayens et. al. (1992) modified their regularization method to take into consideration another regularization parameter (like γ in Friedman's paper). They also proposed an alternative cross-validation approach for estimating their first regularization parameter α, following a result which arises out of using the Kullback-Leibler distance measure for discrimination. Once again, major computational complications have to be addressed.

3. CONSEQUENCES OF APPLYING (FRIEDMAN'S) RDF

Simulation experiments done by us and the various authors mentioned in the previous section indicate that the RDF can perform impressively better than the other discriminant functions (i.e. SEDF, SLDF and SQDF). This is not (intuitively) entirely surprising since the RDF (Friedman's RDF in particular) can be any one of the three discriminant functions or something (better, in terms of error rates) in-between. Note that RDF = QDF when $\lambda = 0$ and $\gamma = 0$, RDF = LDF when $\lambda = 1$ and $\gamma = 0$, and RDF = EDF when $\lambda = 1$, $\gamma = 1$. Note, however, that Aeberhard et. al. (1994) compared RDF against seven other classification functions (including non-parametric functions) and found that the RDF was clearly the most powerful classifier overall. Some simulation results are presented (Table 1) without discussion (space

limitations). Note that all the simulation experiments reported here (unless otherwise stated) were done under similar combinations of parameters (conditions) to those of Friedman (1989). These six conditions were:

(i) COND-1: Equal and spherical covariance matrices (favourable to RDF) [A spherical matrix is one in which all the eigenvalues are similar].
(ii) COND-2: Unequal and spherical covariance matrices (even more favourable to the RDA).
(iii) COND-3: Equal but highly ellipsoidal covariance matrices (difficult for RDA). Here the mean differences between the classes is located in the low-variance subspace.
(iv) COND-4: Same as COND-3, but the mean difference between the classes is located in high-variance subspace.
(v) COND-5: Very unequal and highly ellipsoidal covariance matrices. Here the class means are identical.
(vi) COND-6: Same as COND-5, but with unequal class means.

The Tables report the average (out of 100 simulations) simulated error rates, and the corresponding standard error (in brackets). The average values of the regularization parameters are also given in some cases, with standard errors. Note that in all the Tables RDF-1 denotes the original RDF as proposed by Friedman, while RDF-2 and RDF-M are modified versions of RDF-1, which will be introduced later.

4. PROBLEMS ASSOCIATED WITH IMPLEMENTATION OF (FRIEDMAN'S) RDF

In spite of the impressive performance of the RDF, there are some fundamental problems associated with implementing it, and some of its properties require further investigation. This article reports and discusses the findings of investigations of some of these properties and problems.

(i) Lack of scale invariance

As discussed by Friedman, an important drawback of RDF is that it is not generally scale invariant. Thus changing the relative scales of the measurements or their linear combinations will (in general) change the classification rule and results. This is primarily due to the regularization involving the γ parameter, which shrinks the eigenvalues. In particular, if $\gamma = 0$, RDF is scale invariant.

Scale invariance is considered to be a fairly important property of discriminant functions, and it is unfortuante that this property is lost by the γ-regularization. Hence an obvious question is whether a similar level of success with some kind of limited regularization can be achieved, without losing the invariance property. In this study, we investigated the performance of a regularization method which employed **different** degrees of regularization to shrink each Σ_i to Σ (i.e. different λ_i values in $\hat{\Sigma}_i(\lambda_i)$), but did not use the second regularization parameter, γ. The motivation for this approach was to remove the regularization parameter associated with eigen values (γ), and yet compensate the process by allowing different levels of shrinkage of Σ_i to Σ. A second aspect of this option is that using only one value of λ for **all**

Σ_i may be too restrictive. Thus Σ_i in (1) would be estimated by

$$\hat{\Sigma}_i(\lambda_i) = \frac{(1-\lambda_i)(n_i-1)\hat{\Sigma}_i + \lambda_i \hat{\Sigma}_p}{(1-\lambda_i)(n_i-1) + \lambda_i (N-K)} \tag{12}$$

Thus this method involves K regularization parameters, with $\lambda_i = 0$ for all i, resulting in $\hat{\Sigma}_i(\lambda_i) = S_i$ (SQDF), and $\lambda_i = 1$ for all i resulting in $\hat{\Sigma}_i(\lambda_i) = S_p$(SLDF). Note that each λ_i is obtained independently of the others since each λ_i is chosen to minimize the error rate in class i. This method will be denoted in this article as RDF-Modified (RDF-M).

Simulation results presented in Table 2 suggest that RDF-M is not as successful as the original RDF (denoted in this article as RDF-1). This indicates that eigenvalue shrinkage, in spite of the problems it creates, is quite necessary. There are also some peculiar results associated with RDF-M which are still under investigation. For example, sometimes there is considerable imbalance in the distribution of the error rates among the populations.

The conclusion from this simulation study is that it is not possible to do without the eigenvalue shrinkage, and yet maintain the impressive performance of the RDF. In view of this, it is relevant to consider the importance (or contribution) of the matrix of shrinkage (in this case I) to the entire problem. For example, is there a more appropriate shrinkage matrix (instead of I), which could be determined by the data? Friedman has already alluded to this matter, and one of the authors is currently investigating this problem.

(ii) Necessity of regularization for large data sets

One of the motivations for considering RDF is to do with the fact that for small $(n_i:p)$ ratio the SQDF does not perform very well. Also, the discussion in the previous section indicated that γ-regularisation is crucial. A relevant question then is to what extent the benefits of regularization (especially γ-regularization) diminish as the ratio n_i: p increases. That is, is there a point at which (because of massive amounts of data) we can do without regularization (especially in view of its restrictive non-invariance property). Theory and the justification for regularization (as suggested earlier) suggest that there must be a point (of the n_i: p ratio) where there are no serious benefits for regularization, especially when one considers also its computational requirements.

The simulation experiment was designed to determine the performance of the RDF against the other discriminant functions, as the n_i: p ratio increases. Results are presented in Table 3.

(iii) Only vague values of λ (especially) and γ are necessary

From our experiences (as well as others') it is clear that the surface representing the cross-validated error rate is fairly constant for a wide range of (λ, γ) combinations. Thus the optimal choice of (λ, γ) is not unique, and as it turns out, a wide range of (λ, γ) combinations could do the job equally well, in any given situation. Questions which arise from these empirical observations are the following.

(a) In using RDF is it still necessary to use the "maximum regularization" strategy in the event of several local minima? If there are such ties one option (as was apparently used by Friedman) is to choose the largest value of γ from among tied grid points with the largest value of λ ("maximum regularization") (RDF-1 in Table 1). An alternative

TABLE 1

COND-1

	p=6	p=10	p=20
misclassification risk			
RDF-1	.11 (.04)	.12 (.04)	.12 (.04)
RDF-2	.12 (.03)	.14 (.04)	.12(.03)
LDF	.13 (.04)	.14 (.04)	.15 (.04)
QDF	.24 (.06)	.32 (.07)	.41 (.07)
EDF	.11 (.04)	.11 (.03)	.11 (.03)
Average regularisation parameter values			
RDF-1 λ	.87 (.29)	.85 (.30)	.80 (.34)
RDF-1 γ	.78 (.34)	.81 (.26)	.81 (.24)
RDF-2 λ	.15 (.26)	.20 (.33)	.24 (.33)
RDF-2 γ	.67 (.32)	.69 (.30)	.80 (.25)

COND-3

	p=6	p=10	p=20
misclassification risk			
RDF-1	.07 (.05)	.12 (.04)	.15 (.04)
RDF-2	.08 (.04)	.13 (.05)	.16 (.04)
LDF	.06 (.03)	.11 (.04)	.14 (.04)
QDF	.14 (.05)	.29 (.06)	.39 (.06)
EDF	.24 (.06)	.29 (.06)	.32 (.05)
Average regularisation parameter values			
RDF-1 λ	.87 (.24)	.89 (.23)	.87 (.19)
RDF-1 γ	.05 (.14)	.04 (.11)	.04 (.09)
RDF-2 λ	.41 (.28)	.56 (.30)	.73 (.27)
RDF-2 γ	.02 (.07)	.03 (.11)	.02 (.07)

COND-5

	p=6	p=10	p=20
misclassification risk			
RDF-1	.20 (.06)	.12 (.05)	.03 (.02)
RDA-2	.18 (.06)	.11 (.04)	.03 (.02)
LDF	.60 (.06)	.59 (.06)	.58 (.05)
QDF	.17 (.05)	.14 (.06)	.14 (.04)
EDC	.60 (.06)	.59 (.06)	.58 (.05)
Average regularisation parameter values			
RDF-1 λ	.04 (.07)	.04 (.06)	.04 (.06)
RDF-1 γ	.12 (.15)	.25 (.16)	.35 (.18)
RDF-2 λ	.01 (.04)	.01 (.04)	.02 (.05)
RDF-2 γ	.10 (.14)	.26 (.15)	.26 (.15)

COND-6

	p=6	p=10	p=20
misclassification risk			
RDF-1	.06 (.04)	.06 (.04)	.02 (.02)
RDF-2	.05 (.02)	.05 (.04)	.01 (.01)
LDF	.17 (.05)	.18 (.04)	.21 (.04)
QDF	.04 (.03)	.05 (.04)	.06 (.04)
EDF	.16 (.04)	.17 (.04)	.17 (.04)
Average regularisation parameter values			
RDF-1 λ	.10 (.20)	.10 (.14)	.07 (.06)
RDF-1 γ	.19 (.27)	.29 (.22)	.35 (.19)
RDF-2 λ	.01 (.03)	.02 (.04)	.00 (.00)
RDF-2 γ	.10 (.13)	.22 (.15)	.27 (.09)

TABLE 2

	ERROR RATES		AVERAGE REGULATION PARAMETERS				
	RDF-1	RDF-M	RDF-1		RDF-M		
			λ	γ	λ_1	λ_2	λ_3
COND-1(p=6)	0.11 (0.04)	0.14 (0.04)	0.87 (0.29)	0.78 (0.34)	0.79 (0.35)	0.91 (0.25)	0.92 (0.21)
COND-2(p=6)	0.14 (0.04)	0.24 (0.07)	0.37 (0.38)	0.78 (0.31)	0.70 (0.35)	0.77 (0.34)	0.43 (0.39)
COND-3(p=10)	0.12 (0.04)	0.14 (0.05)	0.89 (0.23)	0.04 (0.11)	0.79 (0.33)	0.95 (0.17)	0.87 (0.27)
COND-4(p=20)	0.11 (0.03)	0.15 (0.05)	0.79 (0.33)	0.67 (0.27)	0.80 (0.28)	0.87 (0.24)	0.88 (0.23)
COND-5 (p=10)	0.12 (0.05)	0.39 (0.11)	0.04 (0.06)	0.25 (0.16)	0.03 (0.08)	0.07 (0.09)	0.30 (0.15)
COND-6 (p=20)	0.02 (0.02)	0.22 (0.13)	0.07 (0.06)	0.35 (0.19)	0.07 (0.09)	0.13 (0.14)	0.89 (0.23)

TABLE 3

COND-1

	p=6			p=10			p=20		
n_i: p	*1.2:1*	*2:1*	*10:1*	*1.2:1*	*2:1*	*10:1*	*1.2:1*	*2:1*	*10:1*
RDF-1	.22 (.04)	.20 (.03)	.16 (.03)	.20 (.05)	.15 (.03)	.10 (.03)	.13 (.03)	.10 (.02)	.09 (.02)
LDF	.30 (.06)	.25 (.02)	.18 (.02)	.28 (.05)	.26 (.04)	.18 (.03)	.28 (.03)	.24 (.03)	.19 (.02)
QDF	.53 (.07)	.34 (.06)	.17 (.02)	.52 (.07)	.35 (.05)	.14 (.03)	.55 (.04)	.37 (.04)	.12 (.02)

COND-3

	p=6			p=10			p=20		
n_i: p	*1.2:1*	*2:1*	*10:1*	*1.2:1*	*2:1*	*10:1*	*1.2:1*	*2:1*	*10:1*
RDF-1	.12 (.06)	.08 (.04)	.05 (.02)	.16 (.04)	.12 (.04)	.10 (.03)	.18 (.04)	.14 (.03)	.11 (.02)
LDF	.10 (.03)	.07 (.02)	.04 (.01)	.14 (.03)	.11 (.02)	.08 (.02)	.17 (.03)	.14 (.02)	.11 (.02)
QDF	.41 (.09)	.15 (.05)	.05 (.02)	.44 (.09)	.24 (.05)	.09 (.02)	.49 (.06)	.32 (.04)	.14 (.02)

COND-5

	p=6			p=10			p=20		
n_i: p	*1.2:1*	*2:1*	*10:1*	*1.2:1*	*2:1*	*10:1*	*1.2:1*	*2:1*	*10:1*
RDF-1	.34 (.11)	.19 (.05)	.10 (.04)	.15 (.06)	.09 (.03)	.06 (.03)	.03 (.02)	.02 (.02)	.00 (.00)
LDF	.61 (.05)	.59 (.05)	.62 (.04)	.59 (.04)	.59 (.04)	.61 (.04)	.58 (.04)	.59 (.05)	.62 (.04)
QDF	.39 (.09)	.18 (.04)	.08 (.02)	.29 (.09)	.10 (.03)	.02 (.01)	.20 (.07)	.04 (.02)	.00 (.00)

would be to use a "minimum regularization" strategy, which chooses the smallest value of γ for the smallest value of λ (RDF-2 in Table 1).

The simulation experiments to compare the two stategies (reported in Table 1) show that the error rates were quite similar for the two strategies in spite of the quite different (λ, γ) values (compare RDF-1 and RDF-2).

(b) An issue which arises immediately from part (a) is: why bother with all the computer-intensive cross-validation technique if we only need rough guesses of (λ, γ) values. To address this issue, we have investigated the use of empirical (and relatively crude) rules for finding a (roughly) optimal (λ, γ) combination using the Bhattacharyya distance measure. We have obtained empirical rules which compete quite favourably with the cross-validation technique for two populations. These results are reported elsehwere, and extensions and refinements are currently being done.

(iv) Estimation of regularisation parameters may involve very few training data

The criterion for estimating the regularization parameters of the RDF, as suggested by Friedman, is to choose the values of λ and γ which minimize the cross-validated error rate. As argued by Rayens et. al. (1989), this means that only a small fraction of the training data may contribute towards determining the values of λ and γ, since most of the training data should be correctly classified. It follows that a criterion which uses all the data in choosing λ and γ may be preferable, and Rayens et. al. (1989) indeed demonstrate that such a criterion can outperform the misclassification rate criterion. We have investigated the use of the Bhattacharyya distance as an alternative criterion for choosing λ and γ and find that in situations where it is appropriate, it can perform as well as the RDF, with the added bonus that it requires only 10-20% of the computation time. A detailed description of the Bhattacharyya distance methodology is presented in another article.

ACKNOWLEDGEMENTS

We would like to acknowledge discussions with Dr S Ganesalingam (Statistics, Massey University) , and data from John Dymond(Landcare, Palmerston North).

REFERENCES

Friedman, J.H. (1989). Regularized discriminant analysis. Journal of the American Statistical Association *Theory and Methods*, Vol. 84, No. 405, pp 165-175.

Greene, T. and Rayens, W.S. (1989). Partially pooled covariance matrix estimation in discriminant analysis. Communications in Statistics *Theory and Methods*, Vo. 18, No. 10, pp3679-3702.

Rayens, W., and Green, T. (1991). Covariance pooling and stablization for classification. *Computational Statistics and Data Analysis,* Vol. 11, pp 17-42.

McLachlan, G.J. (1992). Discriminant analysis and statistical pattern recognition. John Wiley & Sons Inc.

Fukunaga, K. (1990). Statistical Pattern Recognition. Academic Press Inc.

COMPARISON OF REGULARISED DISCRIMINANT ANALYSIS WITH THE STANDARD DISCRIMINATION METHODS

BY

J P KOOLAARD
S GANESALINGAM*

Department of Statistics
Massey University
Private Bag 11222
Palmerston North
NEW ZEALAND

Tel: (06) 350 4257 Fax: (06) 350 2261
Email: S.Ganesalingam@massey.ac.nz

C R O LAWOKO

Faculty of Business,
Queensland University of Technology
GPO Box 2434
Brisbane, QLD 4001
AUSTRALIA

Paper presented at IBC '96 – Amsterdam, July 1996

*Author to whom correspondence should be made

COMPARISON OF REGULARISED DISCRIMINANT ANALYSIS WITH THE STANDARD DISCRIMINATION METHODS

ABSTRACT

The objective of the classical discriminant analysis problem is to classify a p-variate observation vector **x** as having come from one of K populations. It is well known that the linear discriminant function (LDF) and quadratic discriminant function (QDF) and euclidean distance based discriminant function (EDF) are the standard type of discriminant functions employed in practice. Under the assumption of normality these discriminant functions behave reasonably well in a variety of situations.

When the size of the training set is small when compared to the dimension, the performance is degraded because these methods use unstable sample mean vectors and in particular covariance matrices. Friedman (1989) and Greene and Rayens (1989) proposed different methods for addressing the problem of unstable covariance matrices. This article details a critical comparison of the standard approaches with a Friedman's newly proposed regularized discriminant function (RDF) and its implementation difficulties. The article also discusses the implementation of an extension/adaptation of Friedman's RDF.

It was noted that if the RDF is used in higher dimensional situations this is likely to reduce the overall error rate when compared to the application of the other standard discriminant functions.

1. INTRODUCTION

The objective of the classical discriminant analysis problem is to classify a p-variate observation $\mathbf{x} = (x_1, x_2, ..., x_p)$ as having come from one of the several (K) groups or classes. For example, in plant taxonomy a botanist may wish to classify a new specimen as one of several recognized species of a flower. In educational psychology a candidate for admission to a school or study program must be assigned to categories of the sort "admit", "admit conditionally" or "admission denied" on the basis of a vector of test scores, grades and ratings. In routine banking or commercial finance an officer or analyst may wish to classify loan applications as low or high credit risks on the basis of the elements of certain accounting statements. In each case the decision-maker wishes to classify from simple functions of the observation vector rather than complicated regions in the higher dimensional space of the original vector.

Now let us consider the standard approaches for classification of an unknown observation to one of K populations. The responses of the independent observations are described by multi-normal random variables with mean vectors $\mu_1, \mu_2, ..., \mu_k$ and variance-covariance matrices $\Sigma_1, \Sigma_2, ..., \Sigma_k$. If the parameters are known, and assuming given prior probabilities of population memberships and a specified matrix of misclassification costs, the Bayes rule is based upon the likelihood ratio $f_1(\mathbf{x})/f_2(\mathbf{x})$ for all pairs of populations (see, for example, Anderson (1984)). This leads to the linear discriminant function (LDF). Another competitor to the LDF, but also linear in nature is the well known euclidean distance based discriminant function (EDF), see for example Macro, Young and Turner (1987). The EDF ignores the information given by the covariance matrix Σ_i, while forming the discriminant function. It has been shown that the EDF performs better in many circumstances than the LDF (Koolaard and Lawoko (1996)).

The quadratic discriminant function (QDF) requires approximately normal group conditional densities and reasonably large training sample sizes before it can be expected to perform well in discrimination. The LDF is more robust to non-normality and requires less parameter estimation than the QDF. However, problems with obtaining good estimates of the within-groups covariance matrices can affect both these discriminant functions, in particular when the size, n_k of the training sample from group k is small in relation to the dimension of the measurement space, p.

Friedman (1989) proposed Regularised Discriminant Analysis as a compromise between normality based LDF and QDF by considering alternatives to the usual maximum likelihood estimate for the covariance matrices. These alternatives are characterized by two regularisation parameters, the values of which are customized to individual situations by jointly minimising a sample-based estimate of future misclassification error. This technique seems to offer a significant gain in classification accuracy in many circumstances, although it is computationally intensive.

In this paper we shall term the discriminant function proposed by Friedman the regularised discriminant function (RDF), and we examine by extensive simulations, the performance of the sample regularised discriminant function (SRDF) and some modifications of it with the more common sample based rules: SLDF, SEDF and SQDF, for various combinations of population parameters. Note that it is the sample based rules that we are dealing with throughout this paper.

2. THE SRDF AND PROBLEMS ASSOCIATED WITH ITS IMPLEMENTATION

The regularized discriminant function (SRDF) was introduced by Friedman (1989) as an alternative to the common normal theory based discriminant functions. With the SRDF, a two parameter family of estimates of the variance covariance matrix Σ_i of the i^{th} population, is considered, where one parameter λ controls shrinkage of the heteroscedastic estimates towards a common (usually pooled) estimate. The other parameter γ controls shrinkage towards a multiple of a specified covariance matrix such as the identity matrix. Through these two parameters, a fairly rich class of regularised discriminant rules can be provided. Further, with these two parameters assessed from the training set by minimizing the cross validated estimate of the overall error rate, a compromise between sample normal based linear and quadratic analysis is determined automatically from the available data.

Simulation results by various authors suggest that the SRDF can perform much better than the other standard approaches based on normal theory (see for example, Friedman, 1989 and Rayens et al. 1991). Our results also confirm these findings. A recent article by Aeberhard et al. (1994) reported that the SRDF performed better than seven other discriminant functions including several non-parametric ones.

To introduce notation, suppose we have (p-dimensional) multivariate measurements $\mathbf{x}$ on each object, (for example, patient, plant, pixel), where each object belongs to one of K distinct sub-populations or groups. In order to apply standard classification approaches it is usually assumed that measurements from each group follow a multi-normal distribution.

Let us denote the population mean and covariance matrix of group i by μ_i, Σ_i respectively, i = 1 ... K. These parameters are usually estimated by $\bar{\mathbf{x}}_i$ and $\mathbf{S}_i$ using the training sample of size n_i. If it is assumed that the covariance matrices Σ_i are equal for all K classes, then the common value Σ is estimated by

$$\mathbf{S}_p = \frac{1}{n} \sum_{i=1}^{K} n_i \mathbf{S}_i$$

where

$$n = \sum_{i=1}^{K} n_i .$$

On the basis of these estimates the normal theory based discriminant rules allocate an object to class k as follows:

i) SEDF(k) : Min (over i) $\{(x-\bar{x}_i)'(x-\bar{x}_i) - 2\ln \pi_i\}$

ii) SLDF(k) : Min (over i) $\{(x-\bar{x}_i)'S_p^{-1}(x-\bar{x}_i) - 2\ln \pi_i\}$

iii) SQDF(k) : Min (over i) $\{(x-\bar{x}_i)'S_i^{-1}(x-\bar{x}_i) + \ln |S_i| - 2 \ln \pi_i\}$ (2.1)

The covariance matrix estimates can be highly variable and Friedman (1989) showed the effect of this phenomenon on discriminant analysis by replacing the group covariance matrices by their spectral decompositions. The covariance matrix for group k can be written

$$\Sigma_k = \sum_{i=1}^{p} \varepsilon_{ik}\, \eta_{ik}\, \eta_{ik}'$$

where ε_{ik} is the ith eigenvalue of Σ_k and η_{ik} is its corresponding eigenvectors. The discriminant rule, for (iii) above will give the discriminant score as:

$$\sum_{i=1}^{p} \frac{[\eta_{ik}'(x-\bar{x}_k)]^2}{\varepsilon_{ik}} + \sum_{i=1}^{p} \ln \varepsilon_{ik} - 2 \ln \pi_k \qquad (2.2)$$

for an observation vector x belonging to group k.

It is clear from the above expression (2.2) that the small eigenvalues and their eigen vectors will have a large effect on the discriminant score. It is well known that sample based estimates S_k of the Σ_k produce biased estimates of the eigenvalues with the bias being more pronounced when the eigenvalues of the population parameters Σ_k are similar especially for small training sample size.

The motivation for developing SRDF was partly to do with the above mentioned bias problems. Friedman (1989) proposed a regularisation of S_i to S_p, thereby controlling the degree to which Σ_i is estimated by pooled information from the several S_i matrices or by each S_i separately. Thus the initial proposal for the SRDF is to use the SQDF in (2.1), but with S_i replaced by

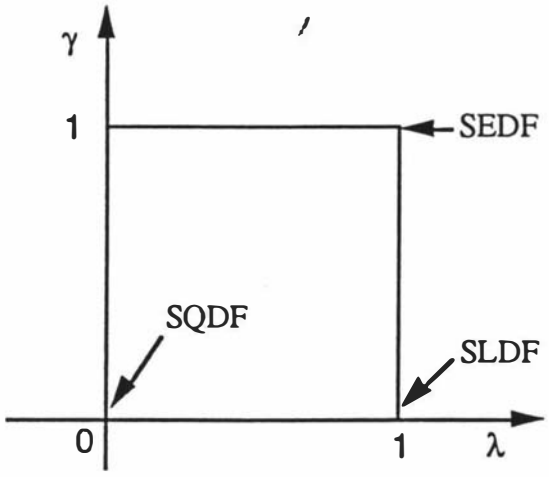
$$\hat{\Sigma}_i(\lambda) = \frac{(1-\lambda)(n_i - 1)\, S_i + \lambda S_p}{(1-\lambda)(n_i-1) + \lambda (n-k)}; \qquad 0 \le \lambda \le 1 \qquad (2.3)$$

Even this may not provide sufficient regularisation for a stable covariance estimate especially when the total sample size n is less than or comparable in size to the dimension p. Also, biasing the group covariance estimates to the pooled covariance matrix may not be appropriate in some situations. Thus Friedman introduces further regularisation by providing an option for regularising the eigenvalues of $\hat{\Sigma}_i(\lambda)$, using a second regularisation parameter γ, $(0 < \gamma < 1)$. Consequently the estimate of Σ_i used is given by

$$\hat{\Sigma}_i(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_i(\lambda) + \frac{\gamma}{p}\,\mathrm{trace}[\hat{\Sigma}_i(\lambda)]\,I \qquad (2.4)$$

where $\hat{\Sigma}_i(\lambda)$ is given in (2.3) and **I** is the identity matrix. Note that shrinkage using γ have the effect of decreasing the larger eigenvalues and increasing the smaller ones to counter the bias in the sample estimates of the eigenvalues of the covariance matrices.

The possible (λ,γ) combinations may be thought of as a plane with four corners.



The bottom left vertex ($\lambda = 0, \gamma = 0$) corresponds to the SQDF, (λ=1, γ=0) gives the SLDF, (λ=1, γ=1) yields a discriminant function based on minimum euclidean distance between groups, while (λ=0, γ=1) yields a weighted minimum euclidean distance function where the group weights are inversely proportional to the average variance of the measurement variables in the group, that is, $\mathrm{trace}[S_k]/p$. If γ is fixed at zero and λ is varied, intermediate rules between the SQDF and the SLDF are obtained. If λ is fixed at 1 and γ increased from 0, one obtains an analogy to ridge regression for the SLDF.

3. SELECTING λ AND γ VALUES AND TIE-BREAKING

In practice, optimal values for the regularisation parameters λ and γ are not known beforehand, and Friedman suggests they be estimated from the training data. The selected λ, γ combination is that which gives rise to the minimum cross-validated estimate of the overall error rate associated with the regularised discriminant rule.

A grid of points is chosen on the λ, γ plane ($0 \le \lambda, \gamma \le 1$), containing typically between 25 and 50 points. Using the λ, γ values to create the classification rule at each point, cross-validation is used to estimate the misclassification risk for each combination of (λ, γ), and the point (λ, γ) with the lowest estimated error rate is used as an estimate of the optimal values of λ and γ. This two-

parameter optimisation problem would require excessive computation were it to be implemented in a straightforward way. However, Friedman developed updating formulas for the computation of the regularised sample covariance matrix and its inverse when a different observation is successively omitted from the sample, as during cross-validation.

Rayens and Greene (1991) noted two criticisms of the model selection procedure of Friedman. Firstly, it was stated that the minimum cross-validated estimate of the misclassification risk is often constant for a range of (λ, γ) combinations. Hence the optimal choice of λ and γ for the model will often not be uniquely determined. Friedman employed a strategy of maximum regularisation where, for all points yielding the minimum error rate on the (λ, γ) grid, that point (λ, γ) is selected which gives rise to the largest value of γ for the largest value of λ. Secondly, Rayens and Greene (1991) demonstrated a situation that can and does occur where only a very small proportion of the sample data influences in any way the optimal choices of λ and γ, and the remainder of the sample observations are correctly classified for almost all points on the λ, γ plane. This occurs especially when the groups are well separated.

Friedman (1989) performed a simulation study to compare the SRDF with SQDF and SLDF in terms of their estimated overall error rates. The simulation conditions represented a wide range of situations in terms of the general structure of the group means and covariance matrices. Some of these conditions were chosen because they were expected to be unfavourable to the SRDF in that any regularisation away from the SQDF or SLDF would be detrimental to the discrimination process. Other conditions were chosen because they were expected to be favourable to regularisation. The six conditions, defined in terms of the population covariance matrices and means, which are also those employed in the following simulation studies in this paper, are:

1) Equal spherical population covariance matrices. A 'spherical' matrix is one where all the eigenvalues are similar in magnitude.
2) Unequal, spherical population covariance matrices.
3) Equal, highly ellipsoidal population covariance matrices with group mean differences in the low variance subspace. 'Ellipsoidal' in this case implies that there is a large difference in magnitude between the smallest and largest eigenvalues.
4) Equal, highly ellipsoidal population covariance matrices with group mean differences in the high variance subspace.
5) Unequal, highly ellipsoidal population covariance matrices with zero mean differences.
6) Equal, highly ellipsoidal population covariance matrices with non-zero mean differences.

3.1 Selection of values for the regularisation parameters when the choice is not uniquely determined by the minimum cross-validated error rate

In the previous section we noted that the optimal choice $(\hat{\lambda},\hat{\gamma})$ of (λ,γ) is very often not uniquely determined. It is of interest to study the effect of a different procedure than that employed by Friedman (1989) for selecting the values to use for the regularisation parameters. A simulation study has been performed under the same conditions as in the previous section but employing a policy of minimum regularisation in the advent of the minimum cross-validated error rate not being uniquely determined. If there is more than one point on the (λ, γ) grid associated with the minimum cross-validated error rate, that point is chosen having the smallest γ value for the smallest λ value. This method will be denoted SRDF-1 and is compared with SRDF which follows the opposite policy of maximum regularisation to break ties. In all cases there are 3 populations or groups, and sample sizes are set to be just larger than the dimension p in each case, so as to avoid singularity in the group covariance matrix estimates. The (λ, γ) grid of points consists of 25 points and is defined to be the same as that used in Friedman's study. Results comparing the estimated error rates ($\bar{e}$, averaged over 100 simulations) of the SRDF and SRDF-1 rules for four of the simulation conditions (described in Section 3) are in Table 1. Also shown are the average values of the two SRDF regularisation parameters λ and γ. The standard errors of $\bar{e}$ for both SRDF and SRDF-1 ranged from 0.002 to 0.006, according to the magnitude of $\bar{e}$, while the standard errors of $\bar{\lambda}$ and $\bar{\gamma}$ ranged from 0.01 to 0.03.

Table 1 – mean values of e, λ, γ (that is, $\bar{e}[\bar{\lambda}, \bar{\gamma}]$) for various values of p.

Condition 1	**p=6**	**p=10**	**p=20**
SRDF	.11 [.87,.78]	.12 [.85,.81]	.12 [.80,.81]
SRDF-1	.12 [.15,.67]	.14 [.20,.69]	.12 [.24,.80]
Condition 3			
SRDF	.07 [.87,.05]	.12 [.89,.04]	.15 [.87,.04]
SRDF-1	.08 [.41,.02]	.13 [.56,.03]	.16 [.73,.02]
Condition 4			
SRDF	.06 [.85,.58]	.10 [.86,.62]	.11 [.79,.67]
SRDF-1	.07 [.15,.50]	.10 [.26,.55]	.11 [.32,.67]
Condition 5			
SRDF	.20 [.04,.12]	.12 [.04,.25]	.03 [.04,.35]
SRDF-1	.18 [.01,.10]	.11 [.01,.26]	.03 [.02,.26]

The first and major finding from the present study comparing SRDF-1 with the SRDF is that the cross-validated error rate surface over the λ, γ plane is often very flat at its minimum. In such situations the error rate estimate will be very similar under both methods for dealing with ties, even though the assessed λ and γ values are quite different. This would indicate that employing a policy of minimum regularisation does not have much effect on the performance of the SRDF in most of the parameter settings considered, and indicates the degree of homogeneity in the cross-validated error rate response surface over the λ, γ plane. In particular, the choice of λ can be considerably less precise than the choice of γ in determining the performance of the rule in terms of its error rate. In conclusion, altering the way ties are broken in the search for the optimum values of λ and γ does not have a great influence on the performance of the SRDF. Some of the parameter configurations here favour a greater degree of regularisation and some a lesser degree, but the difference in error rates was slight.

4. USEFULNESS OF SRDF FOR VARIOUS RATIOS OF SAMPLE SIZE TO DIMENSION

From the study by Friedman (1989) as well as in the previous section it is clear that the SRDF has proved itself at least equal to but usually superior to the other classification rules under a fairly wide range of situations. The superiority is greatest in the larger dimensional settings ($p>10$). The comparisons with the SQDF and SLDF indicate that the advantage the SRDF has over the other classification rules is a result of allowing for eigenvalue shrinkage. A question which becomes of interest is: to what extent do the benefits of regularisation, in particular eigenvalue shrinkage, diminish as the sample size to dimensionality ratio increases?

A (further) simulation study was implemented in the manner of Friedman (1989) (and the previous section), using the same six simulation conditions. In those studies, the ratio of training sample size to dimensionality (n_i/p, denoted r throughout this section) is approximately between 0.5 and 2. We investigate the performance of the RDF relative to the other classification rules over a wider range of values of r. It would be anticipated that eigenvalue shrinkage would no longer be useful for discriminating once the training sample size increases past some point sufficiently larger than p. The various r values ratios employed were 1.2, 1.5, 2, 3, 5, 10 for dimensions 6, 10 and 20. The (λ, γ) grid of values for use in the model selection procedure of the RDF is defined by the outer product of λ= (0, .25, .5, .75, 1) and γ= (0, .25, .5, .75, 1). The entire training sample is 3n in each case, the test sample is 200, and 50 replications of each experiment were performed. Average error rate, $\bar{e}$ is given for each classification rule. Three sets of results are shown in Table 2 but comment is made on each of the six simulation conditions. The standard error of $\bar{e}$ is in the range 10^{-3} to 10^{-2}.

Table 2. Values of $\bar{e}$ for various combinations of r and p

Condition 2.	p=6					p=10					p=20				
	r value					r value					r value				
	1.2	1.5	2	5	10	1.2	1.5	2	5	10	1.2	1.5	2	5	10
SRDF	0.22	0.20	0.20	0.17	0.16	0.20	0.17	0.15	0.13	0.10	0.13	0.12	0.10	0.12	0.09
SLDF	0.30	0.26	0.25	0.20	0.18	0.28	0.26	0.26	0.21	0.18	0.28	0.26	0.24	0.20	0.19
SQDF	0.53	0.43	0.34	0.19	0.17	0.52	0.43	0.35	0.19	0.14	0.55	0.47	0.37	0.18	0.12
SEDF	0.23	0.22	0.22	0.19	0.18	0.24	0.22	0.21	0.20	0.18	0.23	0.22	0.21	0.19	0.18
Condition 4															
SRDF	0.08	0.07	0.07	0.06	0.06	0.10	0.10	0.10	0.08	0.11	0.12	0.12	0.10	0.09	0.09
SLDF	0.12	0.10	0.09	0.06	0.05	0.14	0.13	0.11	0.08	0.08	0.16	0.15	0.13	0.10	0.13
SQDF	0.43	0.29	0.18	0.07	0.06	0.45	0.32	0.23	0.12	0.09	0.48	0.39	0.30	0.16	0.10
SEDF	0.07	0.07	0.07	0.06	0.06	0.10	0.10	0.09	0.10	0.09	0.12	0.12	0.11	0.10	0.11
Condition 5															
SRDF	0.34	0.29	0.19	0.14	0.10	0.15	0.12	0.09	0.03	0.06	0.03	0.02	0.02	0.00	0.00
SLDF	0.61	0.60	0.59	0.60	0.62	0.59	0.58	0.59	0.60	0.61	0.58	0.57	0.59	0.61	0.62
SQDF	0.39	0.25	0.18	0.10	0.08	0.29	0.17	0.10	0.03	0.02	0.20	0.10	0.04	0.00	0.00
SEDF	0.59	0.59	0.59	0.60	0.62	0.59	0.58	0.59	0.59	0.61	0.57	0.59	0.59	0.61	0.61

Eigenvalue shrinkage appears to enhance the classification process under conditions of equal, spherical covariance matrices only for $r < 3$. For larger ratios the advantage the SRDF enjoys over the other methods disappears. The SQDF shows the most dramatic improvement in error rate as the r ratio increases, owing to improved parameter estimates through larger sample size.

In the situation of unequal, spherical population covariance matrices SRDF proved superior for all r values studied, especially the smaller values, indicating the benefit of eigenvalue shrinkage which biases the covariance estimates towards the appropriate value (a multiple of the identity matrix) in these circumstances.

Eigenvalue shrinkage proves to be of no benefit when the population covariance matrices are equal but highly ellipsoidal with mean differences in the low variance measurement subspace. This is because if the covariance matrix eigenvalues are biased towards equality, the variance in all subspaces is equalised and hence in this case the mean differences will become obscured. Conversely, when the mean differences are exhibited in the high variance subspace, eigenvalue shrinkage proves useful in reducing the variance in those subspaces where mean differences are exhibited. The SRDF has a lower error rate than those rules with no eigenvalue shrinkage for r less than 3. At $r = 3$ and larger, SLDF performs as well as the SRDF.

In the case of unequal, highly ellipsoidal population covariance matrices with either zero or non-zero differences between the means, a small amount of eigenvalue shrinkage enables the SRDF to out-perform the SQDF, but only when the sample size is less than twice the dimension. In this case, eigenvalue shrinkage is generally not desirable since the covariance matrices provide substantial information needed for discrimination. A small degree of eigenvalue shrinkage is beneficial in counteracting eigenvalue bias in those situations where r is small. For larger values of r the SQDF's performance is comparable to that of the SRDF, indicating eigenvalue shrinkage loses its effectiveness. While the average γ value used in the SRDF is usually small, there is substantial variation, indicating that under these fairly difficult discrimination conditions (especially zero mean differences), selection of γ is sensitive to peculiarities in the data.

In conclusion, this simulation study underlines the usefulness of the eigenvalue shrinkage technique as employed by the SRDF. The advantage that it affords over the other rules is strongest when the training sample size from each group is small in relation to the dimensionality, p. Furthermore, often that advantage remains, even when the sample size increases to several times that of the dimension.

5. INVARIANCE

In the simulation results reported in the earlier sections it was noted that the regularised discriminant function is not generally scale invariant. The cause of this is the presence of the eigenvalue shrinkage parameter γ. Thus it is of interest to examine the effect of removing this parameter from the model and comparing the performance of the resulting discriminant functions with the SRDF. This would result in a reduced set of regularised models between the SQDF and the SLDF only. In this situation $\hat{\Sigma}_i(\lambda) = \dfrac{(1-\lambda)(n_i-1)\hat{\Sigma}_i + \lambda\hat{\Sigma}_p}{(1-\lambda)(n_i-1) + \lambda(n-i)}$.

As mentioned in Section 2, this set of alternatives is rather restrictive. Further the resulting model may not provide appropriate regularisation if the group covariance matrices are of quite a different nature. In such a situation, it may be useful if each covariance matrix is shrunk to the pooled estimate by an appropriate degree, again estimable from the training data. Using such shrinkage could go some way to overcome the problem of inappropriate regularisation, as the model would be more sensitive to variations in the 'shape' between the various populations. In the single parameter regularisation model, it may occur that in the selection of λ, a large proportion of the training observations misclassified by cross validation come from one group. This may be in part due to the shrinkage being inappropriate for that group but appropriate for the other groups. The following model is proposed to obtain regularised and group covariance estimates:

$$\hat{\Sigma}_i(\lambda_i) = \frac{(1-\lambda_i)(n_i-1)\hat{\Sigma}_k + \lambda_i\hat{\Sigma}_p}{(1-\lambda_i)(n_i-1) + \lambda_i(n-i)}$$

where i=1 ..... K groups and $\hat{\Sigma}_p$ is the estimate of the pooled covariance matrix. Observe that the k regularisation parameters λ_i (i=1 ... K) control the degree of shrinkage of the individual group covariance matrix estimates towards the pooled estimate. The value $\lambda_i=0$ gives $\hat{\Sigma}_i(\lambda_i) = \hat{\Sigma}_i$ and $\lambda_i=1$ yields $\hat{\Sigma}_i(\lambda_i) = \hat{\Sigma}_p$. Each λ_i is obtained by minimizing the group conditional cross-validated error rate over the range $0 \leq \lambda_i \leq 1$, i = 1 ..... K. Each $\hat{\Sigma}_i = S_i$ in the SQDF in equation (iii) of (2.1) is replaced by $\hat{\Sigma}_i(\lambda_i)$ for discriminant analysis. We shall call the resulting rule SRDF-M, (modified SRDF). Section 6 reports on a further simulation study to investigate the relative performance of the scale invariant SRDF-M compared with the SRDF and other standard classification rules.

6. EFFECT OF OMITTING EIGENVALUE SHRINKAGE PARAMETER

Monte Carlo simulation studies were performed again under the same conditions as in the previous two studies reported in Sections 4 and 5. Results for three conditions only are given in Table 3, comparing the performance of SRDF-M with the other approaches. Once again the standard errors of the $\bar{e}$ values fall in the range 10^{-3} to 10^{-2}. The average regularisation parameter values in Table 3 have standard errors in the range 0.01 to 0.03. Discrimination in each situation is between three groups, hence for the SRDF-M method the average values of λ for each group are given $(\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3)$. Also given are the average minimising cross-validated error rates for each group $(\bar{e}_{cv(1)}, \bar{e}_{cv(2)}, \bar{e}_{cv(3)})$.

One immediately observes that having the option to use the regularisation parameter γ and shrink the covariance matrix eigenvalues to equality undoubtedly enhances discrimination in many situations; and not only when the populations are spherical. This type of shrinkage reduces the variance which, despite the introduced bias, is beneficial for discrimination especially in the high dimensional setting. This extra variance reduction factor, apparently explains why the minimum cross-validated error rate for SRDF-M underestimates the actual error rate (assessed from the test sample) by a greater degree than for the SRDF.

The magnitude of the minimum cross-validated error rate over the whole training sample for SRDF-M is at a comparable level to those for the SRDF, meaning it is the actual error rate which is usually higher for SRDF-M. There is also often a large variation in the corresponding (cross-validated) error rates for each group. In some case the average minimum error rate for one group was twice as large as for another, and was extremely variable.

Table 3 – Comparison of error rates and parameter values for SRDF and SRDF-M and other rules

Condition 1	**p=6**	**p=10**	**p=20**
SRDF: $\bar{e}\,[\bar{\lambda}, \bar{\gamma}]$	.11 [.87,.78]	.12 [.85,.81]	.12 [.80,.81]
SRDF-M: $\bar{e}\,[\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3]$	.14 [.79,.91,.92]	.17 [.81,.93,.87]	.16 [.84,.90,.83]
SLDF	.13	.16	.15
SQDF	.23	.39	.42
SEDF	.11	.12	.12
SRDF: $\bar{e}_{cv}$	.09	.10	.10
SRDF-M: $\bar{e}_{cv(1)}$,$\bar{e}_{cv(2)}$,$\bar{e}_{cv(3)}$	.17,.09,.09	.17,.12,.10	.21,.13,.12
Condition 4			
SRDF: $\bar{e}\,[\bar{\lambda}, \bar{\gamma}]$	.06 [.85,.58]	.10 [.86,.62]	.11 [.79,.67]
SRDF-M: $\bar{e}\,[\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3]$	.08 [.86,.88,.86]	.14 [.80,.88,.87]	.15 [.80,.87,.88]
SLDF	.07	.13	.14
SQDF	.16	.36	.38
SEDF	.07	.11	.11
SRDF: $\bar{e}_{cv}$	.04	.07	.10
SRDF-M: $\bar{e}_{cv(1)}$,$\bar{e}_{cv(2)}$,$\bar{e}_{cv(3)}$	.07,.06,.06	.15,.09,.09	.16,.09,.11
Condition 6			
SRDF: $\bar{e}\,[\bar{\lambda}, \bar{\gamma}]$	.06 [.10,.19]	.06 [.10,.29]	.02 [.07,.35]
SRDF-M: $\bar{e}\,[\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3]$	.13 [.11,.14,.88]	.21 [.11,.18,.85]	.22 [.07,.13,.89]
SLDF	.20	.21	.20
SQDF	.06	.10	.06
SEDF	.20	.20	.17
SRDF: $\bar{e}_{cv}$	.04	.03	.01
SRDF-M: $\bar{e}_{cv(1)}$,$\bar{e}_{cv(2)}$,$\bar{e}_{cv(3)}$	.05,.04,.01	.09,.06,.01	.02,.01,.00

When the group covariances are spherical and set to be equal, SRDF-M yielded error rate estimates of between 30% and 40% higher than the SRDF. Under these conditions, eigenvalue shrinkage (to equality) clearly enhances discrimination. Hence the SEDF performs well. The mean minimum cross-validated error rate over all groups underestimated the actual error rate by around 25% for $p \leq 10$, but by only about 5% for $p = 20$. The group conditional mean minimum cross-validated error rates differed significantly, and their standard deviations were also large.

If the group covariances are spherical but unequal, SRDF-M gives error rate estimates around 70% higher than for SRDF, and worse for larger dimensions. It is clear that under such conditions, eigenvalue shrinkage is very desirable in order to reduce variation in the higher dimensions. The mean minimum cross-validated error rate over all groups underestimated the actual

misclassification risk by 30%-40%, although observations from the higher variance groups were more frequently misclassified.

The SRDF-M performs comparably to the SRDF under conditions of equal but highly ellipsoidal group covariances. This is not surprising since eigenvalue shrinkage is counterproductive in this situation. In the case where the group mean differences are concentrated in the low variance subspace, and therefore more pronounced the $\bar{\lambda}_i$, (i = 1, ..., k) values are very close to one, and the performance of SRDF-M approaches that of the SLDF, which is the optimum rule in these conditions.

When the group means are concentrated in the high variance subspace, SRDF-M is less successful compared to the SRDF. The high degree of covariance shrinkage towards the identity matrix enhances discrimination, because of the reduction in variance achieved. This is why the SEDF performs as well as the SRDF under these conditions, each outperforming SRDF-M by about 40%. The minimum cross-validated error rate for SRDF-M underestimates the actual error rate by between 10% and 20% when the group mean differences are more exposed in the low variance subspace, and between 30% and 40% when the means are obscured by high variance.
The final situation looked at is when the group covariances are unequal and highly ellipsoidal. The SRDF-M does not perform well here. In fact the minimising cross-validated error rate severely underestimates the actual error rate for SRDF-M, especially for the high dimensional settings. This is a curious phenomenon which exhibits itself strongly only in these simulation conditions where the groups have high and unequal variance. The reduction in variance obtained by eigenvalue shrinkage is not the complete explanation for otherwise SRDF-M should perform comparably to the SQDF, but it does not. It should be noted that the error rate estimates for SRDF-M also have unusually high variance under these conditions. An explanation is that under these conditions the best rules are those where λ is close to zero with low variability. Since each λ_i is obtained from such a small number of data points, its variability is high.

One other feature of the performance of SRDF-M under these conditions is that λ_3 is much higher than λ_1 or λ_2. Now it happens that group 3 does not have quite the same extreme ellipsoidal nature of the other two groups. Significant shrinkage of the group 3 covariance matrix to the pooled covariance appears to lead to observations from that group becoming indistinguishable (to the classification rule) from those of the other high variance groups. It is noted that if a policy of minimum regularisation is used to break ties (similar to that employed by SRDF-1 in Section 3.1), SRDF-M is enhanced because smaller values of λ_i are selected.

In conclusion, the proposed regularisation model SRDF-M was not as successful as the SRDF. This clearly shows the importance of eigenvalue shrinkage, especially when p is large. The attempt to make SRDF-M more sensitive by employing a separate λ for each group caused other problems in certain circumstances as described above. If a solution to the problem of lack of scale

invariance is to be found, other techniques need to be devised to replace eigenvalue shrinkage while ensuring the accuracy of classification attained by the SRDF is not compromised.

7. CASE STUDIES

Case Study 1: The data considered here consists of 3 variables measured on each of 10 insects Chaetocnema, (Lindsey et al. 1987). The variables are as follows, all measured in microns.

x_1: the width of the first joint of the first tarsus
x_2: the width of the first joint of the second tarsus
x_3: the maximal width of the aedegus

Each insect was classified to one of the species according to its measurements of x_1, x_2 and x_3 an the error rate for each rule was assessed using the technique of cross-validation.

The error rates for SRDF, SQDF, SLDF and SEDF were 0.03, 0.03, 0.07, 0.17 respectively. Th values of $\bar{\lambda}$ and $\bar{\gamma}$ for the RDF were 0.97 and 0.43 respectively. The model selection procedures fc the SRDF chooses a value of λ close to 1 on average, and still yields a similar error rate to that c the SQDF. This is to be expected since r is relatively large (r = 3.3). Hence shrinking to the poole covariance makes little difference in this example. On the other hand, the degree of eigenvalu shrinkage should not be large, as evidenced by the poor performance of SEDF. This was expecte due to the high ellipsoidal nature of the covariance estimates.

Case Study 2: The data considered here relates to three types of pathological lung cancer, (Hon and Yang, 1991). Each type is described by 56 variables, the variables taking on integer values 1- The number of training samples are very small: 9 from the first, 13 from the second and 10 from th third type of cancer (group), rendering the problem very ill-posed. Each patient was classified t one of the types of cancer according to his/her measurements of the 56 variables and the error ra for each rule was assessed using the technique of cross validation. Note that in this data set r $\simeq$. which is small. The error rates for SRDF, SQDF and SEDF were .375, .688, and .813 respectivel That is, SRDF correctly classify 62.5% of the observations, while other two rules only about 31 and 19% respectively.

The results of this case study show that if the SRDF is used in higher dimensional situations, this likely to reduce the error rate when compared to the application of other rules. If the grou covariance matrices are identical, it is clear that the, SLDF will be the only method capable of ou performing SRDF (Aeberhard et al. (1994)).

REFERENCES

Aeberhard, S., Coomans, D. and de Vel, O. (1994) Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition* **27**, 1065-1077.

Anderson, T.W. (1984) An Introduction to Multivariate Statistical Analysis. Second Edition, New York, Wiley.

Friedman, J.H. (1989) Regularized Discriminant Analysis. *Journal of American Statistical Association* **84**, 165-175.

Greene, T. and Rayens, W. (1989). Partially pooled covariance estimation in Discriminant Analysis. *Commun. Statist. - Theory Meth.* **18**, 3679-3702.

Hong, Z. Q. and Yang, J. Y. (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition* **24**, 317-324.

Koolaard, J. and Lawoko, C. R. O. (1996). The linear and Euclidean discriminant functions: a comparison via asymptotic expansions and simulation study. To appear in Communications in Statistics.

Lindsey, J. C., Herzberg, A. M. and Watts, D. G. (1987). A methods for cluster analysis based on projections and quantile–quantile plots. *Biometrics* **43**, 327-341.

Macro, V. R., Young, D. M. and Turner, D. W. (1987). The Euclidean Classifier: an alternative to linear discriminant function. *Commun. Statist-Simul.* **16**, 485-505.

Rayens, W. and Greene, T. (1991) Covariance pooling and stabilization for classification, *Computational Statistics and Data Analysis* **11**, 17-42.

THE LINEAR AND EUCLIDEAN DISCRIMINANT FUNCTIONS: A COMPARISON VIA ASYMPTOTIC EXPANSIONS AND SIMULATION STUDY

J. P. Koolaard
Biometrics Unit, NZ Institute for Crop and Food Research,
Private Bag 4005, Levin, New Zealand.

C. R. O. Lawoko[1]
Faculty of Business, Queensland University of Technology,
Gardens Point Campus, GPO Box 2434, Brisbane,
QLD 4001, Australia.

(This work was completed while the authors were at the Department of Statistics, Massey University, Palmerston North, New Zealand.)

Key Words and Phrases: linear discriminant function; Euclidean distance classifier; error rate; asymptotic expansion.

ABSTRACT

This article considers the problem of statistical classification involving multivariate normal populations and compares the performance of the linear discriminant function (LDF) and the Euclidean distance function (EDF). Although the LDF is quite popular and robust, it has been established (Marco, Young and Turner, 1989) that under certain non-trivial conditions, the EDF is "equivalent" to the LDF, in terms of equal probabilities of misclassification (error rates). Thus it follows that under those conditions the sample EDF could perform better than the sample LDF, since the sample EDF involves estimation of fewer parameters. Simulation results, also from the above paper, seemed to support this hypothesis. This article compares the two sample discriminant functions through asymptotic expansions of error rates, and identifies situations when the sample EDF should perform better than the sample LDF. Results from simulation experiments are also reported and discussed.

[1]Author to whom all correspondence and enquiries should be addressed

1. INTRODUCTION

In parametric statistical discriminant analysis, the linear discriminant function (LDF), which is based on assumptions of multivariate normality and equal covariance matrices, is quite popular because of its robustness and simplicity. Clearly, there are situations when the LDF is inappropriate and related competitors like the quadratic discriminant function (QDF), Euclidean distance function (EDF) or regularized discriminant function (RDF) may be used instead; see, for example, McLachlan (1992, Chapters 3 and 5) and Friedman (1989). In this article we are concerned with the performances of the LDF and the EDF, following results in Marco, Young and Turner (1987). In their article, Marco, Young and Turner describe and discuss in detail the two discriminant functions and the error rates associated with them. Before discussing this article (plus related ones) in detail, we briefly introduce the two discriminant functions and the relevant notation.

Suppose that two multivariate normal populations have p-dimensional mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and $(p \times p)$ covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. The usual (consistent) sample estimators of these parameters (from training data) are denoted by $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively. For the two-population situation the sample versions of these discriminant functions (i.e. sample linear discriminant function, SLDF, and sample Euclidean discriminant function, SEDF) can be expressed as follows:

1. SLDF: If it can be established that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, say (or it is assumed so) then one would use the SLDF which allocates an object with observation $\boldsymbol{x}$ to population 1 if

$$\mathcal{L}(\boldsymbol{x}) \; > \; \log_e k, \tag{1}$$

where

$$\mathcal{L}(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\mathbf{S}^{-1}(\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)),$$

and k is some appropriately chosen constant (see McLachlan (1992, Chapter 3) for example). If $\mathcal{L}(\boldsymbol{x}) > \log_e k$, $\boldsymbol{x}$ is allocated to population 2.

2. SEDF: If $\Sigma = \mathbf{I}$ in the LDF or the information in the covariance matrix is deliberately ignored for the purpose of discrimination, one gets the SEDF, which allocates an object with observation $\boldsymbol{x}$ to population 1, if

$$\mathcal{E}(\boldsymbol{x}) \; > \; \log_e k, \tag{2}$$

where

$$\mathcal{E}(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2))$$

(otherwise it is allocated to population 2).

There has been considerable interest in the literature in the relative performances of these discriminant functions. These comparisons have usually been based on various measures of estimates of error rates (probabilities of misclassifications) since direct algebraic evaluations of these probabilities have proved intractable. The following sources provide relevant background for this study.

1. Raudys and Pikelis (1980), who performed a simulation study to compare the SLDF, SEDF, the sample quadratic discriminant function (SQDF), and a variant of the SLDF for independent measurements (i.e. with off-diagonal elements of Σ being set to zero); see McLachlan (1992, Section 4.6) for the details about the SQDF. The relative performances of these discriminant functions when the populations are spherically normal were evaluated. Since computations of reliable estimators of error rates have been traditionally difficult, numerical integration techniques were used in evaluating the integrals in the definitions of the probabilities of misclassification. It was concluded that the simpler SEDF performed better than the SLDF when p is large relative to n. In fact the SEDF was found to perform at least as well as the SLDF even for non-spherical covariance structures.

2. Marco, Young and Turner (1987) compared the SLDF and SEDF under conditions derived to make the two classifiers "equivalent" or "non

equivalent". The LDF and EDF were defined as "equivalent" if they have the same true error rates (i.e. assuming known population parameters). The conclusion, based on simulation studies only, was that the SEDF generally performed better than the SLDF except when the Mahalanobis distance (Δ) between the two populations was substantially larger than the corresponding Euclidean distance. Also, the SEDF performed at least as well as the SLDF when the population parameters were set so as to achieve either equivalence or non equivalence of the classifiers.

This article follows directly from the Marco, Young and Turner paper, and some results in their paper will be discussed in more detail later.

3. Other related work include Peck and van Ness (1982), van Ness (1979), Lim (1992), Friedman (1989), Greene and Rayens (1989), and Rayens and Greene (1991). The last three articles are concerned with regularised discriminant analysis whereby, depending on the values of other parameters an (effectively) quadratic discriminant function may become a linear (or even Euclidean) discriminant function. Although preliminary empirical and simulation results in those articles suggest that these regularised discriminant functions (RDFs) can perform surprisingly better than the other discriminant functions, using RDFs is a highly computer-intensive procedure and their properties are still being evaluated by researchers (mainly through simulation experiments). In particular, regularised discriminant functions are not yet abundantly available in commercial statistical software.

In view of the limited knowledge about, and lack of availability of software for RDFs, it is still relevant to investigate the relative performances of the SEDF and SLDF. As mentioned earlier, previous articles have reported comparisons based on simulation experiments and 'brute force' numerical integrations of very complicated probability functions (following basic definitions of error rates or probabilities of misclassification). In this article, following arguments from Marco, Young and Turner (1987), we highlight

situations where the SEDF performs better than the SLDF. We also report results from asymptotic expansions of the error rates associated with these discriminant functions, and results from some simulation studies.

2. IMPLICATIONS OF RESULTS FROM MARCO, YOUNG AND TURNER (1987)

This paper derives conditions under which the LDF and the EDF are "equivalent" (i.e. have the same error rates for known population parameters). The authors also report results of simulation studies to compare the SLDF and SEDF not only under conditions of equivalence but also under certain situations of "non-equivalence".

To be specific, if $k = 0$ in the expressions (1) and (2) then it is well known that the "true" error rate (i.e. when all population parameters are known) for allocating an object from population i to population j $(j \neq i)$ by the LDF is

$$P_{ij}^{LDF} = \Phi(-\Delta/2) \qquad (i \neq j = 1,2) \tag{3}$$

where

$$\Delta = [(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)]^{1/2}$$

is the Mahalanobis distance between the two populations and $\Phi(.)$ is the standard normal distribution function. The corresponding error rate for the EDF (details in Marco, Young and Turner) is

$$P_{ij}^{EDF} = \Phi\left(\frac{-\frac{1}{2}(\mu_1 - \mu_2)'(\mu_1 - \mu_2)}{[(\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2)]^{1/2}}\right) \tag{4}$$

The overall error rates are obtained by summing the errors for $j = 1,2$ in expressions (3) and (4). Since $k = 0$ here, these overall error rates (to be denoted here by P_L and P_E) would be equal to the corresponding error rates given in those expressions. That is,

$$P_L = P_{ij}^{LDF} \qquad (j \neq i = 1,2)$$

and

$$P_E = P_{ij}^{LDF} \qquad (j \neq i = 1,2). \tag{5}$$

Marco, Young and Turner proved the following related results.

(i) Let $\mathbf{V}$ be a $p \times p$ full rank matrix and $\mathbf{F}$ any $p \times 1$ matrix with pseudo inverse $\mathbf{F}^+$. If $\mathbf{F}\mathbf{F}^+$ and $\mathbf{V}^{-1}$ commute then

$$\mathbf{F}'[\mathbf{V}^{-1}\mathbf{F}]^{1/2} = (\mathbf{F}^+\mathbf{V}\ \mathbf{F}^{+\prime})^{-1} \tag{6}$$

(ii) If we add the requirement that $\mathbf{V}$ be symmetric, then

$$[\mathbf{F}'\mathbf{V}^{-1}\mathbf{F}]^{1/2} = \frac{\mathbf{F}'\mathbf{F}}{[\mathbf{F}'\mathbf{V}\mathbf{F}]^{1/2}} \tag{7}$$

(iii) If we set $\mathbf{F} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\mathbf{V} = \boldsymbol{\Sigma}$ in result (ii) where $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\boldsymbol{\Sigma}$ satisfy the requirements for results (i) and (ii), then $P_L = P_E$.

The authors argue that in view of result (iii) "... in many practical situations the SEDF might perform better than the SLDF since considerably fewer parameters must be estimated for the SEDF". Thus, since "... the performance of the SLDF deteriorates significantly as the dimension becomes large relative to the training sample sizes, the computationally simpler SEDF may be the preferred discrimination algorithm in this situation". In view of this last argument, the authors conjectured that "the SEDF may perform as well as the SLDF even for (some) 'non-equivalent' situations". The authors then performed a simulation experiment for a very special structure of $\boldsymbol{\Sigma}$ and concluded that there were indeed situations when the SEDF performed better than the SLDF. They found that the "improvement of the SEDF over the SLDF is highly dependent on the ratio of Mahalanbis distance to Euclidean distance". In particular, "whenever this ratio is small, the SEDF tends to outperform the SLDF, (and) when the ratio is large the reverse is true".

One possible explanation for the observed relative behaviours of the two discriminant functions follows from Peck and Van Ness (1982) who conjectured that all this is due to the relative effects of the errors in estimating $\boldsymbol{\Sigma}$ to that in estimating $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and the relative seriousness of these effects depends on the sizes of the Mahalanobis and Euclidean distances; see

the original article or Marco, Young and Turner, for further details and illustrations. Of course, in "non-equivalent" situations when the LDF has a lower (true) error rate than the EDF, it would be expected that the SLDF would perform better than the SEDF.

On the matter of when the EDF performs better than the LDF, consider the proof of result (iii) in Marco, Young and Turner (1987), where, if the conditions for the result are satisfied, then

$$\frac{(\mu_1 - \mu_2)'(\mu_1 - \mu_2)}{[(\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2)]^{1/2}} = [(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)]^{1/2}. \quad (8)$$

By swapping Σ and Σ^{-1} in the above result, we get the equivalent result that

$$\frac{\Delta_E^4}{\Delta^2} = \frac{[(\mu_1 - \mu_2)'(\mu_1 - \mu_2)]^2}{[(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)]} = (\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2), \quad (9)$$

where Δ_E is the Euclidean distance between the two populations. Thus the size of the ratio between the distance functions can be reduced to an explicit function of the elements of μ_1, μ_2 and Σ.

Without loss of generality, one can set $\mu_2 = (0, 0, \ldots, 0)'$. Marco, Young and Turner (1987) set the values of $\mu_1 = (m, m, \ldots, m)'$ under "equivalence" and $\mu_1 = (m^*, 0, 0, \ldots, 0)'$ under "non-equivalence", where m and m^* are appropriately chosen scalars so that the Mahalanobis distances can be set equal (under equivalence and non-equivalence) for purposes of comparison. In this article, we concentrate on the "equivalence" situation since it provides fair comparison between the two discriminant functions (both being optimal Bayes procedures for known population parameters under "equivalence"). Since $\mu_2 = (0, 0, \ldots, 0)'$ and $\mu_1 = (m, m, \ldots, m)'$ in this situation, it follows that $\Delta_E^2 = pm^2$ and

$$\frac{\Delta_E^2}{\Delta^2} = \left(\frac{\Delta_E^4}{\Delta^2}\right) \times \frac{1}{\Delta_E^2} = (\mu_1'\Sigma\mu_1)/pm^2 = \sum_{i,j}\sigma_{ij}/p, \quad (10)$$

where $\Sigma = \{\sigma_{ij}\}$.

Thus the only factors which determine the size of the ratio of the two distance functions are the elements of the covariance matrix, Σ. If, as in

Marco, Young and Turner (1987), standardisation is done and the covariance matrix is effectively a correlation matrix, then it follows that in general high positive correlations yield large values of $\sum_{i,j} \sigma_{ij}$. If there is no standardisation of the observation vectors then large (small) variances and/or large positive (negative) correlations would result in large (small) values of $\sum_{i,j} \sigma_{ij}$.

Note that in their discussions Marco, Young and Turner (1987) refer to the size of the ratio of Δ^2 to Δ_E^2, which is the reciprocal of the ratio in (10). In terms of the ratio in (10) these authors' simulation experiments suggest that: SEDF performs better (worse) than SLDF when $\sum_{i,j} \sigma_{ij}$ is large (small). It can be concluded that it is the type and extent of correlations (or covariances) among the observations which determine this observed behaviour.

3. ASYMPTOTIC EXPANSIONS AND EVALUATIONS

In this article the asymptotic expected error rates were obtained using Taylor series expansions of the conditional error rates (i.e. conditional on $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$ and $\mathbf{S}$) and taking expectations over the distributions of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{S}$. In particular, if $\mathcal{H}(.)$ is a differentiable function of parameters $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s)$, where $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s)$, are consistent estimators of $(\beta_1, \beta_2 \ldots, \beta_s)$, then the Taylor series expansion of $E(\mathcal{H})$ about the point $(\beta_1, \beta_2 \ldots, \beta_s)$ can be expressed as

$$E(\mathcal{H}) \approx \mathcal{H}(\beta_1, \beta_2, \ldots, \beta_s) + \sum_{j=1}^{s} \frac{\partial \mathcal{H}(.)}{\partial \hat{\beta}_j} E(\hat{\beta}_j - \beta_j) + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \mathcal{H}(.)}{\partial \hat{\beta}_i \partial \hat{\beta}_j} E(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j). \quad (11)$$

For our expansions $\mathcal{H}(.) = \Phi(.)$, the standard normal distribution function, and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s$ are the elements of $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\mathbf{S}$. The expansions are evaluated at the point $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$.

The two asymptotic error rates considered here are the expected "actual" (i.e. unconditional) and the expected "plug-in" (i.e. conditional) error rates. For these error rates, the function $\mathcal{H}(.)$ takes the following forms (for misclassification of an object from population 1 to population 2), where the

two subscripts 'A' and 'P' refer to the "actual" and "plug-in" error rates respectively:

$$P_{21(A)}^{LDF} = \Phi\left(-\frac{[\boldsymbol{\mu}_1 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)]'\mathbf{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right),$$

$$P_{21(A)}^{EDF} = \Phi\left(-\frac{[\boldsymbol{\mu}_1 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)]'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{\Sigma}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right),$$

$$P_{21(P)}^{LDF} = \Phi(-\tfrac{1}{2}[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\mathbf{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2})$$

and

$$P_{21(P)}^{EDF} = \Phi\left(-\frac{1}{2}\frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\mathbf{S}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}}\right). \tag{12}$$

Corresponding expressions for misclassifying an object from population 2 to population 1 are similar. Several results in Okamoto (1963) were used in obtaining the asymptotic expansions.

In a series of papers, McLachlan (1972, 1973, 1974a, 1974b) obtained asymptotic expansions of error rates for the SLDF. No such results appear to have been obtained for the SEDF. We believe this is partly due to the fact that for the SLDF the function $\mathcal{H}(.)$ can be reduced to a relatively simple function (usually referred to as "canonical form") through a linear transformation of the observation vector. This simplifies the algebra considerably, and makes the final result dependent on only a few parameters (see, for example, McLachlan (1972,1973)). Unfortunately, no similar trick can be used for the $\mathcal{H}(.)$ function for the SEDF. The canonical form that has been traditionally adopted (after the transformation) has been $\boldsymbol{\mu}_1 = (\Delta, 0, 0, \ldots, 0)'$, $\boldsymbol{\mu}_2 = (\Delta, 0, 0, \ldots, 0)'$ and $\boldsymbol{\Sigma} = \mathbf{I}$, which would not allow us to investigate the distinction between SLDF and SEDF. Hence, such an investigation would require that a particular structure of $\boldsymbol{\Sigma}$ be assumed. Consequently each asymptotic expansion takes a different form, depending on (i) the assumed structure of $\boldsymbol{\Sigma}$, (ii) whether the expansion is obtained under "equivalence" or "non-equivalence", (iii) whether the expansion is for the SLDF or the SEDF,

and also (iv) whether the expansion is for the "actual" or "plug-in" error rate.

For example, the expansion (up to first order) of the conditional error rate associated with the LDF under "equivalence" conditions is of the form

$$\begin{aligned} P_{LDF} &= \Phi\left(-\frac{m}{2}\sum_w\sum_v(\sum_u s^{uv})(\sum_u \sigma_{vu}s^{uw})^{-1/2}\{\sum_k\sum_l s^{lk}\}\right) \\ &+ \frac{1}{2n_i}\sum_i\sum_j \frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}\sigma_{ij} + \frac{1}{2n_2}\sum_i\sum_j \frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}\sigma_{ij} \\ &+ \frac{1}{2}\frac{(n_1+n_2)}{(n_1+n_2-2)^2}\sum_k\sum_l\sum_i\sum_j \frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}(\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}) \quad (13) \end{aligned}$$

where the quantities $\frac{\partial^2\Phi(.)}{\partial\theta_1\partial\theta_2}$ are obtained separately for each assumed structure of μ_1, μ_2 and Σ, for any variables θ_1 and θ_2. The full algebraic expressions of the asymptotic expansions are too complicated to be put in this paper. Interested readers can get them from the authors. However, for the purpose of completeness and to give some idea about computational requirements, we give partial details of one of the expansions in the appendix.

Two different structures of Σ were considered, and they will be denoted as Σ_A and Σ_B, where

$$\Sigma_A = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \vdots \\ \vdots & \vdots & & \ddots & \\ \rho & \rho & \dots & & 1 \end{bmatrix} \text{ and } \Sigma_B = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \dots & \vdots \\ \vdots & \vdots & & \ddots & \\ \rho^{p-1} & \rho^{p-2} & \dots & & 1 \end{bmatrix}. \quad (14)$$

4. NUMERICAL EVALUATIONS AND DISCUSSION

Appropriate values of ρ (both positive and negative) and other parameters were chosen for the numerical evaluation of the asymptotic expansions. We follow the work of Marco, Young and Turner (1987) where, in the situation of "equivalence" the value of the parameter m is given by

$$m = \sqrt{\{\Delta^2/\sum_i\sum_j \sigma^{ij}\}}. \quad (15)$$

For "non-equivalence", its counterpart (i.e. m^*) is given by

$$m^* = \sqrt{\{\Delta^2/\sigma^{11}\}}. \tag{16}$$

The sample sizes were taken to be equal at $n_1 = n_2 = 50$, and dimensions of the observation vectors used were $p = 4$ and $p = 8$. We present and discuss here a very limited set of results; several other results obtained under much more extensive conditions are available from the authors.

One table (TABLE I) is presented to illustrate the order of magnitudes of the various error rates. In the notation and discussions about TABLE I the various error rates are referred to as follows:

e_L^*, e_E^*:	true error rates (i.e. for known population parameter values) for the SLDF (subscript L) and SEDF (subscript E).
e_L, e_E:	asymptotic expected actual (i.e. unconditional) error rates.
$\hat{e}_L, \hat{e}_E$:	asymptotic expected plug-in (i.e. conditional) error rates.
e_{SL}, e_{SE}:	mean cross-validation error rates from 100 simulation experiments using computer-generated data.

Note that several other estimates of error rates (e.g. bootstrap and resubstitution) were obtained from the simulation experiments. However, previous work (e.g. Ganeshanandam and Krzanowski, 1990) suggest that the cross-validation error rate is one of the better and reliable ones to use. Also, although results for the asymptotic expected plug-in error rates (i.e. $\hat{e}_L$ and $\hat{e}_E$) are given, it is well-known that this particular error rate is biased (usually too optimistic). Henceforth, results about this error rate will not be discussed or referred to.

It is easier to visualise the relative performances of the SLDF and SEDF, through a graphic presentation of their error rates. Define the differences between the estimated and true error rates as:

λ_L = $e_L - e_L^*$ = difference between the expected error rate and true error rate for the SLDF.

λ_E = $e_E - e_E^*$ = difference between the expected error rate and the true error rate for the SEDF.

λ_{SL} = $e_{SL} - e_L^*$ = difference between the simulated error rate and the true error rate for the SLDF.

λ_{SE} = $e_{SE} - e_E^*$ = difference between the simulated error rate and the true error rate for the SEDF.

Graphical displays of values of $|\lambda_L|, |\lambda_E|, |\lambda_{SL}|$ and $|\lambda_{SE}|$ for various values of the Mahalanobis distance (Δ) and levels of correlation among the observations are presented in FIGS. 1 to 3. Results for positive autocorrelation structures are presented in FIG. 1 ($\Sigma = \Sigma_A$) and FIG. 2 ($\Sigma = \Sigma_B$), while results for negative autocorrelation between neighbouring observations with $\Sigma = \Sigma_B$ are presented in FIG. 3. Since FIGS. 1 to 3 show absolute differences between the error rates, they hide any bias that an estimator might tend to have. Consequently, we have provided FIG. 4 which displays values of $\lambda_L, \lambda_E, \lambda_{SL}$ and λ_{SE}, to illustrate this bias issue.

The main features of these plots and results are the following:

- For positive ρ (FIGS. 1 and 2) it is interesting to note that $|\lambda_E|$ tends to decrease as ρ increases while $|\lambda_L|$ tends to increase.

- It was hypothesized in Marco, Young and Turner (1987) that the SEDF performs better than the SLDF if the ratio Δ^2/Δ_E^2 is large. It was also established in Section 2 that this condition can be reduced to the size of $\sum_{ij} \sigma_{ij}$. Since large ρ means large $\sum_{ij} \sigma_{ij}$, a comparison of the plots of $|\lambda_L|$ and $|\lambda_E|$ for a given value of ρ indicates that the asymptotic expected error rates provide support for those arguments and conjectures. Plots of λ_L and λ_E in FIG. 4 also support these results.

- The plots in FIGS. 1 and 2 might appear to suggest that the expected error rate associated with SLDF tends to initially decrease with Δ and then increase as Δ increases further. The plots in FIG. 4 clarify this

matter, since λ_L initially underestimates the true error rate (when Δ is small) and this estimation improves as Δ increases until it overestimates the true error rate for very large Δ.

- The plots for the simulated error rates in FIGS. 1 and 2 suggest that for positive ρ, $|\lambda_{SE}|$ tends to be smaller than $|\lambda_{SL}|$, and FIG. 3 suggests that for negative ρ, the reverse happens. Note that although it is the absolute values of the simulated error rates which are plotted in FIGS. 1 to 3, plots of λ_{SE} and λ_{SL} would be very similar, indicating that the simulated error rates tend to be generally larger than the true error rates (not surprisingly).

- From FIG. 4, an interesting difference between λ_L and λ_E is that as ρ increases λ_E decreases from positive values towards zero. Meanwhile, λ_L decrease from positive values through zero, to negative values.

- When we compare the results for $\Sigma = \Sigma_A$ with those for $\Sigma = \Sigma_B$ we find the corresponding values of $|\lambda_L|$, $|\lambda_E|$, $|\lambda_{SE}|$ and $|\lambda_{SL}|$ are quite similar. In fact, it can be seen from the orders of magnitude of these differences in error rates that the estimation of the error rates provided by the asymptotic expansions are quite reasonable in both cases. This is confirmed by the simulated error rates being of similar order of magnitude. It appears however, that when ρ is negative and Δ is large the approximation provided by e_E is quite inaccurate. The problem is worsened as p increases. This asymptotic expansion is therefore not recommended for approximating the error rate under this situation. Note, however, that the simulated error rates are also unusually large under this situation (FIG. 3).

5. CONCLUSION

The sample linear discriminant function (SLDF) is still the most popular classifier among users of discrimination procedures, in spite of its drawbacks. Meanwhile, the sample Euclidean discriminant function (SEDF), which is a

simpler version of the SLDF, has been shown to (surprisingly) perform better than the SLDF under some circumstances. It is established algebraically in this article that the relative performances of the SLDF and SEDF are determined by the type and extent of correlations (or covariances) among the observations. This result explains and supports previously published conjectures and simulation results on this matter (Peck and Van Ness (1982); Marco et al (1987)).

Asymptotic expansions of the error rates associated with the two discriminant functions are given for two specific structures of the covariance matrix, Σ. Although several expansions are available in the literature for the SLDF, similar expansions are not available for the SEDF, because of the fact that its error rate function cannot be reduced to a function of only a few parameters. Consequently, any asymptotic expansion of the error rate for the SEDF (and hence comparison with the SLDF) is likely to be "messy", and feasible for particular structures of Σ only. Two such structures of Σ are adopted in this article, and the asymptotic expansions (and subsequent numerical evaluations) provide support for the earlier conjectures and simulation results about the relative performances of SLDF and SEDF.

Comparisons of the asymptotic expansions with the simulated cross-validated error rates indicate that the asymptotic expansions are quite reasonable, except under certain parameter configurations, which are identified in this article. This article also identifies situations when the two estimated error rates provide biased estimates of the true error rates.

APPENDIX

ASYMPTOTIC EXPANSION FOR THE CONDITIONAL ERROR RATE OF LINEAR DISCRIMINANT FUNCTION UNDER CONDITIONS OF "EQUIVALENCE" WITH COVARIANCE MATRIX OF THE FORM $\Sigma = \Sigma_A$.

The probability, conditional on the samples, that the Linear Discriminant Function misclassifies an observation from group 1 into group 2 is:

$$P_{21(A)}^{LDF} = \Phi\left(-\frac{[\mu_1 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]'S^{-1}(\bar{x}_1 - \bar{x}_2)}{[(\bar{x}_1 - \bar{x}_2)'S^{-1}\Sigma S^{-1}(\bar{x}_1 - \bar{x}_2)]^{1/2}}\right) \qquad (A.1)$$

For the error rates of the LDF and EDF to be equivalent, the population means must be set so that

$$\begin{aligned} \boldsymbol{\mu}_1 &= (m, m, \ldots, m)' \\ \boldsymbol{\mu}_2 &= (0, 0, \ldots, 0)'. \end{aligned} \tag{A.2}$$

The Taylor Series expansion (up to first order approximation) is

$$\begin{aligned} \Phi(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \mathbf{S}) &= \Phi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\ &\quad + \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial \bar{x}_{1j}} (\bar{x}_{1j} - \mu_{1j}) + \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial \bar{x}_{2j}} (\bar{x}_{2j} - \mu_{2j}) \\ &\quad + \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial s_{ij}} (s_{ij} - \sigma_{ij}) \\ &\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{1j}} (\bar{x}_{1i} - \mu_{1i}) (\bar{x}_{1j} - \mu_{1j}) \\ &\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i} \partial \bar{x}_{2j}} (\bar{x}_{2i} - \mu_{2i}) (\bar{x}_{2j} - \mu_{2j}) \\ &\quad + \tfrac{1}{2} \sum_{k=1}^{p} \sum_{l=1}^{p} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial s_{kl} \partial s_{ij}} (s_{kl} - \sigma_{kl}) (s_{ij} - \sigma_{ij}) \\ &\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{2j}} (\bar{x}_{1i} - \mu_{1i}) (\bar{x}_{2j} - \mu_{2j}) \\ &\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial s_{ij}} (\bar{x}_{1i} - \mu_{1i}) (s_{ij} - \sigma_{ij}) \\ &\quad + \tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i} \partial s_{ij}} (\bar{x}_{2i} - \mu_{2i}) (s_{ij} - \sigma_{ij}) \end{aligned}$$

where

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial \bar{x}_{2j}} = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{1i} \partial s_{ij}} = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 \Phi(.)}{\partial \bar{x}_{2i} \partial s_{ij}} = 0. \tag{A.3}$$

Taking expected values of the expansion yields

$$\begin{aligned} \mathrm{E}\{\Phi(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \mathbf{S})\} &= \Phi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\ &\quad + \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial \bar{x}_{1j}} \mathrm{E}(\bar{x}_{1j} - \mu_{1j}) + \sum_{j=1}^{p} \frac{\partial \Phi(.)}{\partial \bar{x}_{2j}} \mathrm{E}(\bar{x}_{2j} - \mu_{2j}) \end{aligned}$$

$$+\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial\Phi(.)}{\partial s_{ij}}\mathrm{E}\,(s_{ij}-\sigma_{ij})$$
$$+\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}\mathrm{cov}\,(\bar{x}_{1i},\bar{x}_{1j})$$
$$+\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}\mathrm{cov}\,(\bar{x}_{2i},\bar{x}_{2j})$$
$$+\tfrac{1}{2}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}\mathrm{cov}\,(s_{kl},s_{ij}) \qquad \text{(A.4)}$$

where

$$\mathrm{E}\,(\bar{x}_{1j}-\mu_{1j})=\mathrm{E}\,(\bar{x}_{2j}-\mu_{2j})=\mathrm{E}\,(s_{ij}-\sigma_{ij})=0.$$

Thus the following quantities need to be obtained

$$\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}},\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}},\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}}.$$

Under "equivalence" $\mu_2 = 0$, and we may write equation (A.1) as

$$P_{21(A)}^{LDF}=\Phi\,(-B)$$

where

$$B=[\bar{x}_1'\mathrm{S}^{-1}\Sigma\mathrm{S}^{-1}\bar{x}_1]^{-1/2}\,[\mu_1-\tfrac{1}{2}\bar{x}_1'\mathrm{S}^{-1}\bar{x}_1]\,.$$

This expression is used to obtain the desired quantities:

$$\frac{\partial^2\Phi(.)}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}} = -\phi(-B)\left[\frac{\partial^2 B}{\partial\bar{x}_{1i}\partial\bar{x}_{1j}}-B\frac{\partial B}{\partial\bar{x}_{1i}}\frac{\partial B}{\partial\bar{x}_{1j}}\right] \qquad \text{(A.5)}$$

$$\frac{\partial^2\Phi(.)}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}} = -\phi(-B)\left[\frac{\partial^2 B}{\partial\bar{x}_{2i}\partial\bar{x}_{2j}}-B\frac{\partial B}{\partial\bar{x}_{2i}}\frac{\partial B}{\partial\bar{x}_{2j}}\right] \qquad \text{(A.6)}$$

$$\frac{\partial^2\Phi(.)}{\partial s_{kl}\partial s_{ij}} = -\phi(-B)\left[\frac{\partial^2 B}{\partial s_{kl}\partial s_{ij}}-B\frac{\partial B}{\partial s_{kl}}\frac{\partial B}{\partial s_{ij}}\right]. \qquad \text{(A.7)}$$

The asymptotic expansion for the probability in equation (A.1) would be obtained after collection and evaluation of all these expressions.

ACKNOWLEDGEMENT

The authors acknowledge and thank the Associate Editor and referee for their suggestions which resulted in improvements of the presentation of this article.

BIBLIOGRAPHY

Friedman, J.H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165-175.

Ganeshanandam, S. and Krzanowski, W.J. (1990). Error-rate estimation in two-group discriminant analysis using the linear discriminant function. *J. Statist. Comput. Simul.*, **36**, 157-175.

Greene, T. and Rayens, W.S. (1989). Partially pooled covariance matrix estimation in discriminant analysis. *Commun. Statist. - Theory Method.*, **18**, 3679-3702.

Lim, T. (1992). Comparison of the euclidean and linear discriminant functions in statistical discriminant analysis. Unpublished MSc thesis, Massey University, New Zealand.

Marco, V.R., Young, D.M. and Turner, D. W. (1987). The euclidean distance classifier: an alternative to the linear discriminant function. *Commun. Statist. - Simula.*, **16**, 485-505.

McLachlan, G.J. (1972). An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function. *Australian J. Statist.*, **14**, 68-72.

McLachlan, G.J. (1973). An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. *Australian J. Statist.*, **15**, 210-214.

McLachlan, G.J. (1974a). An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics*, **30**, 239-249.

McLachlan, G.J. (1974b). The asymptotic distributions of the conditional error rate and risk in discriminant analysis. *Biometrika*, **61**, 131-135.

McLachlan, G.J. (1992). Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons Inc. New York.

Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.*, **34**, 1286-1301. Correction (1968). *Ann. Math. Statist.*, **39**, 1358-1359.

Peck, R. and Van Ness, J.W. (1982). The use of shrinkage estimators in

linear discriminant analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI-4**, 530-537.

Raudys, S.J., and Pikelis, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, **PAMI-2**, 242-252.

Rayens, W.S. and Greene, T. (1991). Covariance pooling and stabilization for classification. *Comput. Statist. Data Anal.*, **11**, 17-42.

Van Ness, J.W. (1979). On the effects of dimension and discriminant analysis for unequal covariance populations. *Technometrics*, **21**, 119-127.

TABLE I

$p = 4$				$p = 8$			
true	actual	plug-in	simulated	true	actual	plug-in	simulated
$\mathrm{EDF}(e_E^*)$ $\mathrm{LDF}(\mathbf{e}_L^*)$	$\mathrm{EDF}(e_E)$ $\mathrm{LDF}(\mathbf{e}_L)$	$\mathrm{EDF}(\hat{e}_E)$ $\mathrm{LDF}(\hat{\mathbf{e}}_L)$	$\mathrm{EDF}(e_{SE})$ $\mathrm{LDF}(\mathbf{e}_{SL})$	$\mathrm{EDF}(e_E^*)$ $\mathrm{LDF}(\mathbf{e}_L^*)$	$\mathrm{EDF}(e_E)$ $\mathrm{LDF}(\mathbf{e}_L)$	$\mathrm{EDF}(\hat{e}_E)$ $\mathrm{LDF}(\hat{\mathbf{e}}_L)$	$\mathrm{EDF}(e_{SE}$ $\mathrm{LDF}(\mathbf{e}_{SL}$
.3721	.7889	.6551	.4370(.062)	.3706	**	**	.4594(.059
.3618	**.3643**	**.3250**	**.3766(.068)**	**.3618**	**.3815**	**.2875**	**.3932(.068**
.3654	.4549	.3841	.4008(.052)	.3644	.6034	.4346	.4286(.057
.3618	**.3614**	**.3350**	**.3858(.064)**	**.3618**	**.3729**	**.3009**	**.3986(.067**
.3618	.3788	.3470	.3828(.055)	.3618	.4001	.3261	.3858(.060
.3618	**.3597**	**.3373**	**.3852(.065)**	**.3618**	**.3695**	**.3037**	**.3936(.063**
.3634	.3673	.3541	.3696(.064)	.3636	.3725	.3417	.3616(.054
.3618	**.3548**	**.3378**	**.3812(.065)**	**.3618**	**.3615**	**.3044**	**.3832(.059**
.3642	.3659	.3597	.3646(.069)	.3660	.3692	.3549	.3750(.057
.3618	**.3400**	**.3378**	**.3786(.066)**	**.3618**	**.3361**	**.3046**	**.3984(.072**
.3222	.6022	.5125	.3948(.061)	.3203	**	.9497	.4292(.053
.3085	**.3167**	**.2790**	**.3288(.052)**	**.3085**	**.3395**	**.2350**	**.3266(.058**
.3133	.3636	.3264	.3476(.059)	.3119	.4717	.3595	.3632(.060
.3085	**.3128**	**.2837**	**.3192(.057)**	**.3085**	**.3282**	**.2524**	**.3360(.060**
.3085	.3205	.2994	.3156(.059)	.3085	.3347	.2857	.3364(.052
.3085	**.3110**	**.2867**	**.3136(.057)**	**.3085**	**.3245**	**.2562**	**.3368(.061**
.3106	.3139	.3052	.3174(.058)	.3109	.3176	.2973	.3184(.051
.3085	**.3076**	**.2873**	**.3280(.061)**	**.3085**	**.3189**	**.2570**	**.3410(.059**
.3117	.3135	.3095	.3162(.054)	.3140	.3170	.3079	.3070(.061
.3085	**.2980**	**.2874**	**.3250(.053)**	**.3085**	**.3022**	**.2573**	**.3394(.057**
.2328	.3848	.3365	.2952(.058)	.2302	.6843	.5695	.3564(.060
.2146	**.2290**	**.1762**	**.2226(.052)**	**.2146**	**.2590**	**.1394**	**.2334(.052**
.2209	.2539	.2291	.2496(.053)	.2190	.3046	.2458	.2618(.057
.2146	**.2240**	**.1915**	**.2252(.055)**	**.2146**	**.2442**	**.1612**	**.2350(.059**
.2146	.2219	.2112	.2200(.048)	.2146	.2296	.2043	.2384(.054
.2146	**.2220**	**.1951**	**.2262(.052)**	**.2146**	**.2400**	**.1661**	**.2478(.055**
.2173	.2201	.2158	.2234(.050)	.2178	.2225	.2122	.2368(.050
.2146	**.2200**	**.1959**	**.2258(.045)**	**.2146**	**.2367**	**.1671**	**.2430(.055**
.2188	.2208	.2189	.2286(.050)	.2219	.2246	.2201	.2274(.049
.2146	**.2153**	**.1959**	**.2266(.052)**	**.2146**	**.2281**	**.1676**	**.2384(.059**

:ates that for those conditions the asymptotic expansions yield estimates out of bounds for probabilities.

TABLE I. The true, expected actual, expected plug-in and mean simulated (with standard deviation) error rates of the SEDF and SLDF in the case of 'equivalence' with $\Sigma = \Sigma_B$ and various ρ.

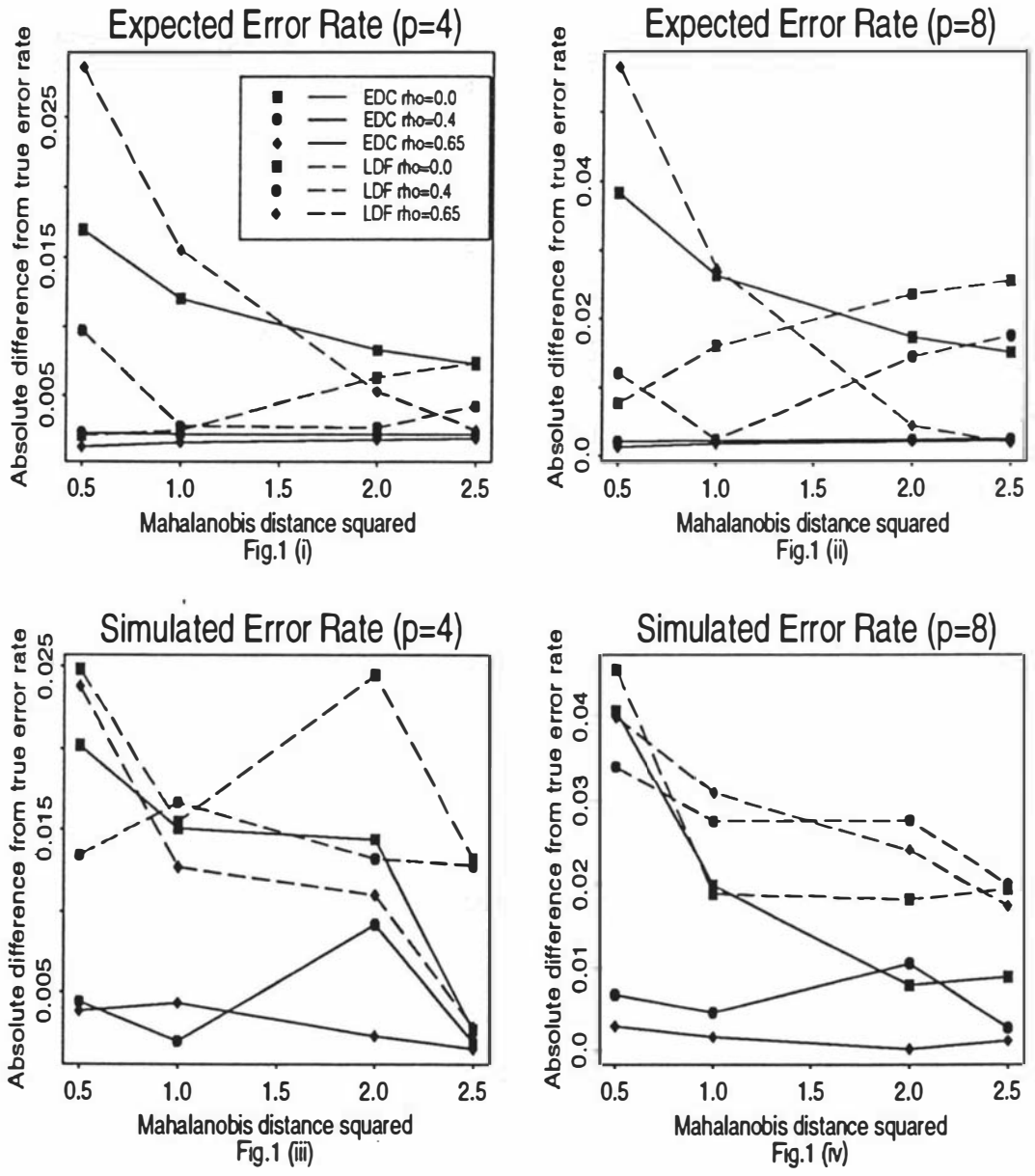


FIG. 1 Displays showing the absolute difference between the true error rate and a) the expected actual error rate (i.e. the evaluated asymptotic expansions) (graphs (i) and (ii)); and b) the simulated error rates (graphs (iii) and (iv)) for $\Sigma = \Sigma_A$, dimension p=(4,8), and various Mahalanobis distance squared (Δ^2) and ρ.

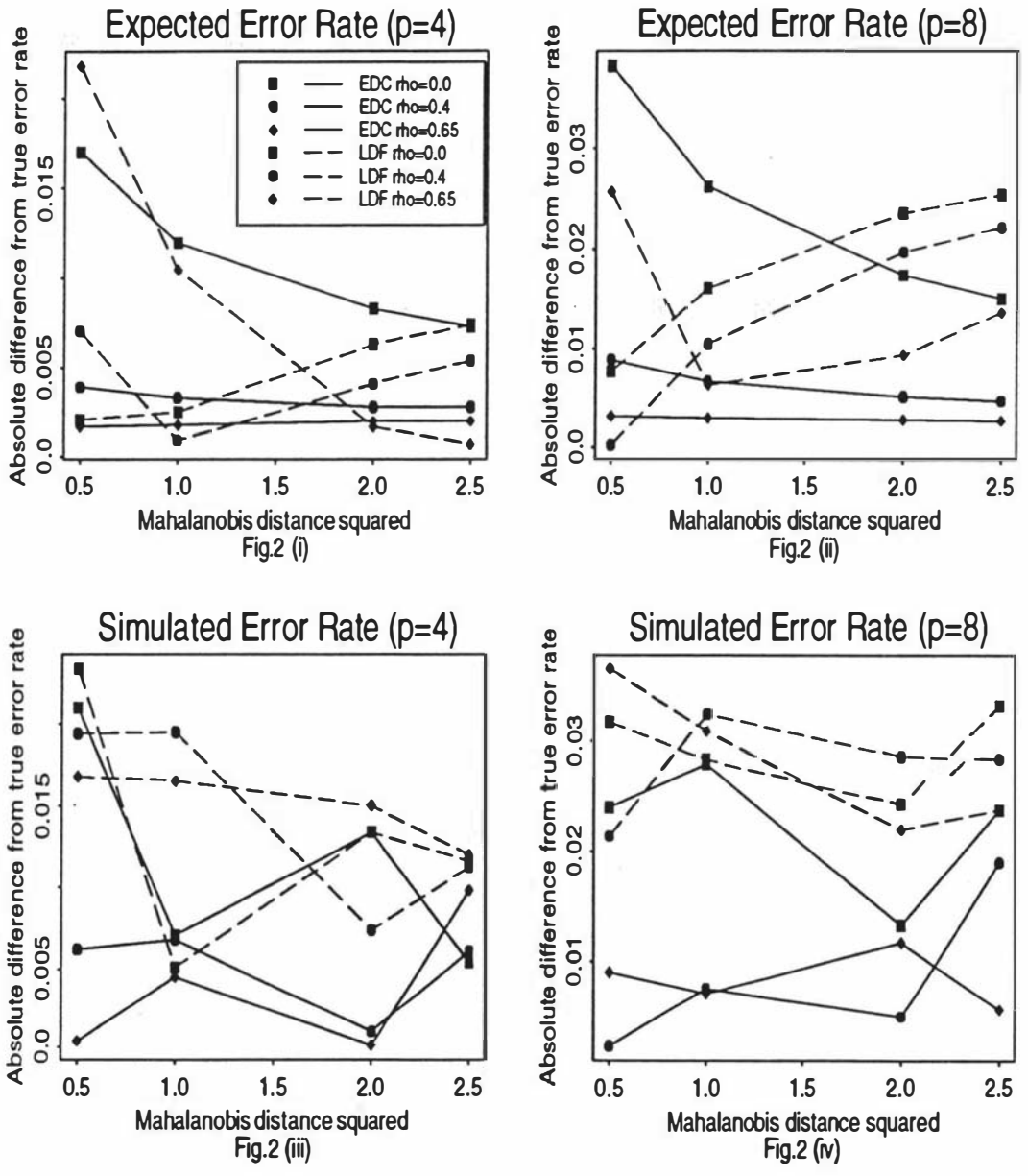


FIG. 2 Displays showing the absolute difference between the true error rate and a) the expected actual error rate (i.e. the evaluated asymptotic expansions) (graphs (i) and (ii)); and b) the simulated error rates (graphs (iii) and (iv)) for $\Sigma = \Sigma_B$, dimension p=(4,8), and various Mahalanobis distance squared (Δ^2) and positive ρ.

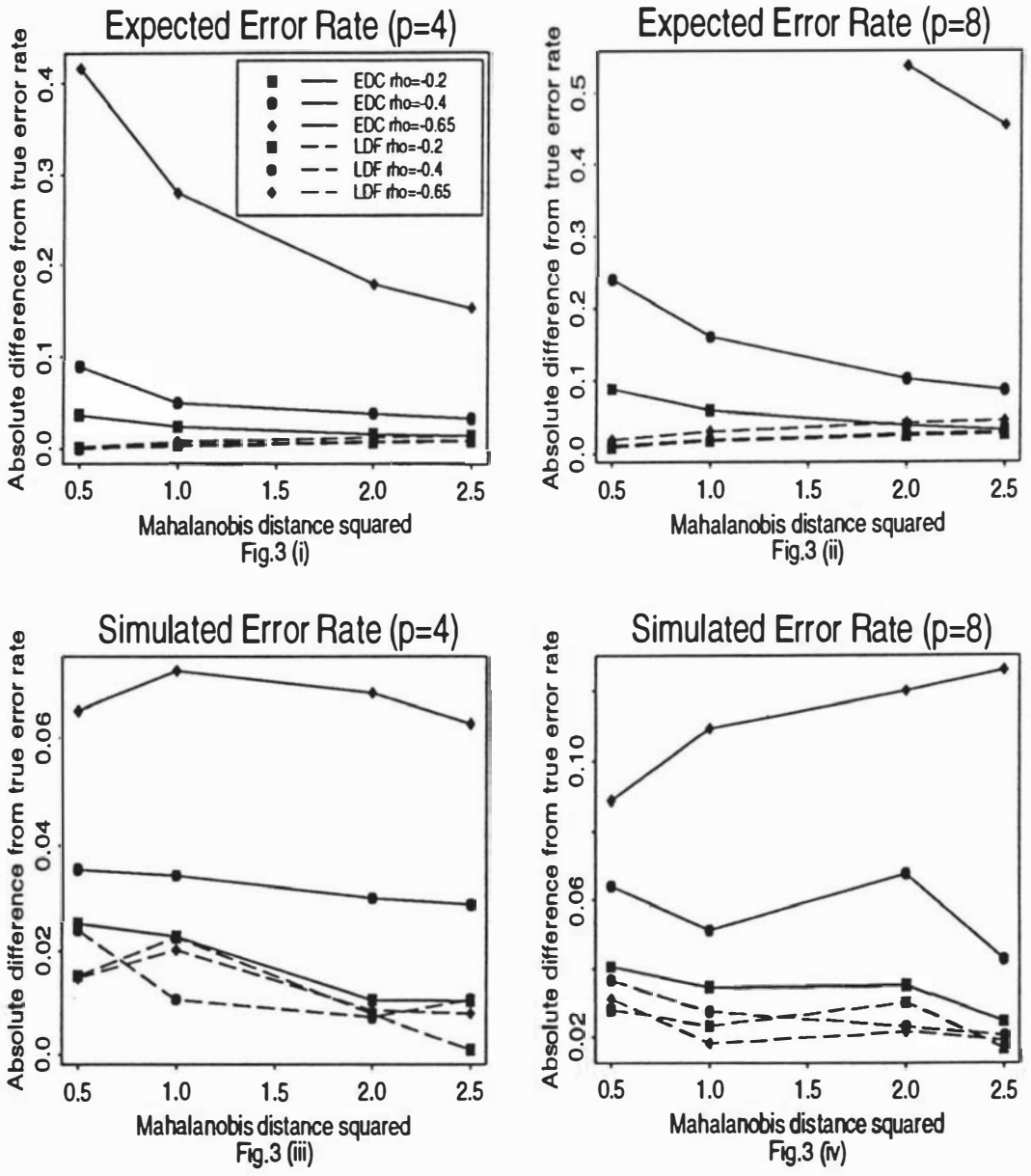


FIG. 3 Displays showing the absolute difference between the true error rate and a) the expected actual error rate (i.e. the evaluated asymptotic expansions) (graphs (i) and (ii)); and b) the simulated error rates (graphs (iii) and (iv)) for $\Sigma = \Sigma_B$, dimension p=(4,8), and various Mahalanobis distance squared (Δ^2) and negative ρ.

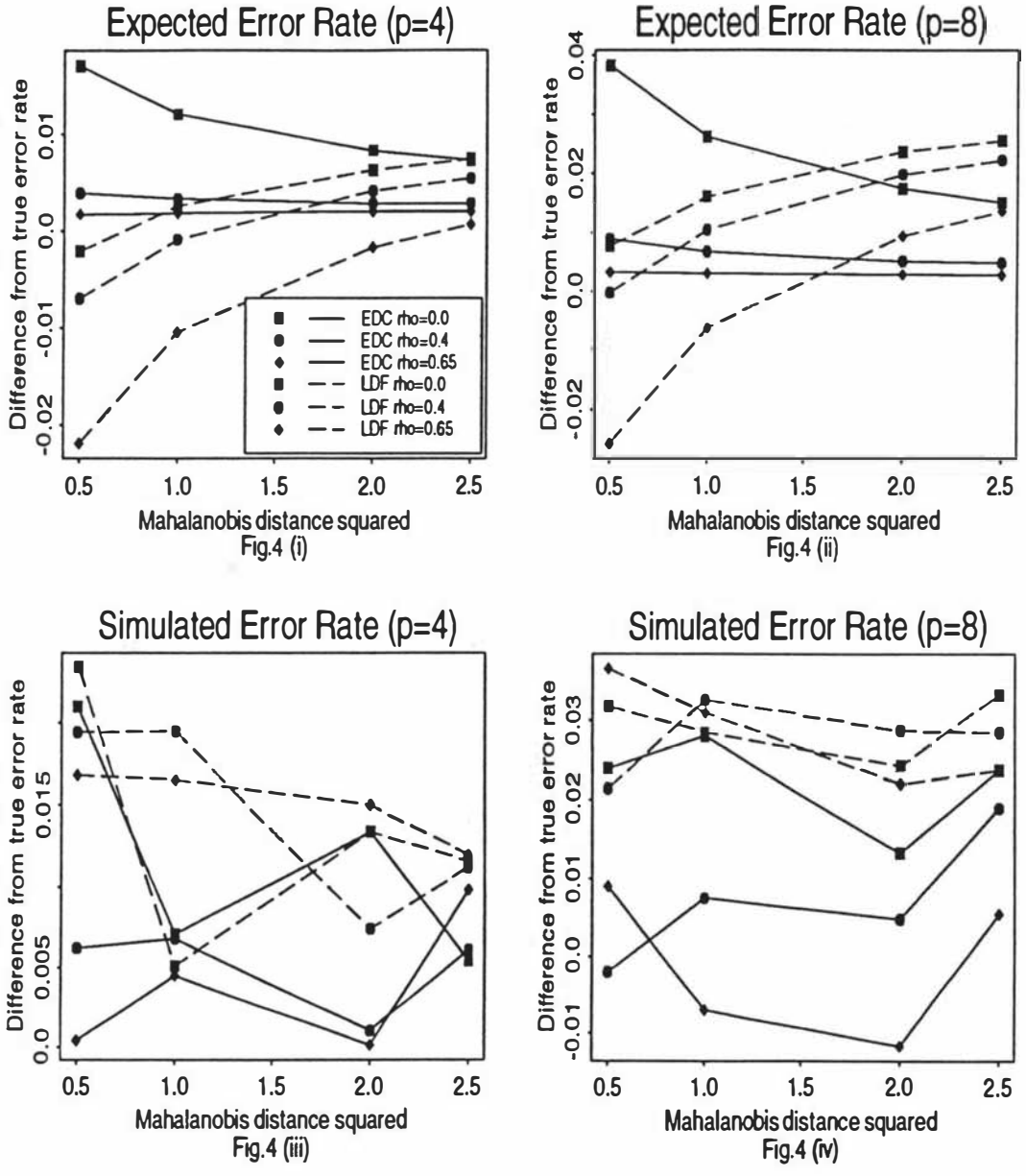


FIG. 4 Displays showing the difference between the true error rate and a) the expected actual error rate (i.e. the evaluated asymptotic expansions) (graphs (i) and (ii)); and b) the simulated error rates (graphs (iii) and (iv)) for $\Sigma = \Sigma_B$, dimension p=(4,8), and various Mahalanobis distance squared (Δ^2) and positive ρ.