Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# COMPARISON OF THE EUCLIDEAN AND LINEAR DISCRIMINANT FUNCTIONS IN STATISTICAL DISCRIMINANT ANALYSIS

A thesis presented to Massey University
in partial fulfillment of the requirements
for the degree of Master of Science
in Operations Research at
Massey University

Tiew-Kim Lim, BSc.

1992

# Abstract

*It is known that in the problem of statistical discriminant analysis, the linear discriminant function performs poorly when the dimension of the data, p, is large. It has been demonstrated by Marco, Young and Turner (1987) that the much simpler Euclidean distance classifier may out-perform the usual linear discriminant function under certain conditions. Their conclusions were arrived at from a simulation experiment which compared the probabilities of misclassification associated with the Euclidean distance classifier with those of the linear discriminant function, under certain conditions. In this dissertation, the asymptotic expansions of the probabilities of misclassification (the expected actual and expected plug-in error rates) associated with the two discriminant functions are obtained. These error rates are then used to investigate the relative performances of the two methods.*

*Chapter 1 introduces the problem of discriminant analysis and describes the two competing procedures for discriminant analysis and some associated error rates. Then Chapter 2 reviews previous results, in the literature which show that the Euclidean distance classifier can perform better than the linear discriminant function. Chapter 3 gives the asymptotic expansions of the error rates, i.e. the expected actual error rate, and the expected plug-in error rate. The relative performances of the two methods on the basis of the asymptotic expansions are discussed in Chapter 4. The results show that in general the plug-in error rates for the*

*Euclidean distance classifier give better estimates of the actual error rates for all dimensions of* $p$ *which were considered, when compared to the linear discriminant function. Furthermore, the actual error rates for the Euclidean distance classifier also seem to give better estimates of the true error rates at large dimensions of* $p$*, when compared to the linear discriminant function. Certain situations where the linear discriminant function performs better than the Euclidean distance classifier are also identified. Final conclusions, discussions and recommendations for further work are given in Chapter 5.*

# Acknowledgements

First of all I would like to thank my supervisor, Dr. C.R.O. Lawoko for his supervision and encouragement, and for presenting the results of this dissertation at the International Symposium on Multivariate Analysis and its Applications in Hong Kong.

The department of Statistics at Massey University and the New Zealand Statistics Association sponsored my presentation of the results of this dissertation at the Annual Conference of the New Zealand Statistical Association at Victoria University of Wellington. This is hereby acknowledged.

I am also particularly grateful to Dr. S. Ganesalingam for his time and help looking through my algebra, and to Richard Rayner who helped with computing facilities.

A big thank you to Uncle K.C. and Aunty Lee Lee for their loan to make my studies possible. Thank you for your care, love, encouragement and prayers. Thanks also go to Yuin-Khai for loaning me his Smart Alec's Wally Jokes book to read when I was bored and frustrated with my work. I would also like to thank my parents and brother for their support and patience through my studies.

Finally, lots of thanks go to thank Simon, Jeanne and Ming for their friendship, care and encouragement.

# Contents

ABSTRACT ii

ACKNOWLEDGEMENTS iv

LIST OF FIGURES AND TABLES vii

CHAPTER PAGE

1 INTRODUCTION

Section 1.1 : Introduction 1
Section 1.2 : Classification Rules 4
Section 1.3 : Error Rates 7
Section 1.4 : Aim of study 11

2 REVIEW OF PREVIOUS RELATED WORK

Section 2.1 : Introduction 12
Section 2.2 : S. Raudys and V. Pikelis (1980) 12
Section 2.3 : R. Peck and J. Van Ness (1982) 13
Section 2.4 : V. R. Marco, D. M. Young and D. W. Turner (1987) 15
Section 2.5 : Motivation for this project 17

3 ASYMPTOTIC EXPANSIONS OF ERROR RATES

Section 3.1 : Introduction 19
Section 3.2 : Asymptotic expansions for the actual error rate 21
Section 3.2.1 : Case A1 22
Section 3.2.2 : Case A2 24
Section 3.2.3 : Case A3 25
Section 3.2.4 : Case A4 27

Section 3.3 : Asymptotic expansions for the plug-in error rate 29
Section 3.3.1 : Case P1 29
Section 3.3.2 : Case P2 31
Section 3.3.3 : Case P3 32
Section 3.3.4 : Case P4 34

4 COMPUTATION RESULTS AND DISCUSSION
Section 4.1 : Introduction 36
Section 4.2 : Discussion 43
Section 4.2.1 : Actual error rates 44
Section 4.2.2 : Plug-in error rates 55

5 SUMMARY AND CONCLUSION 61

BIBLIOGRAPHY 65

APPENDIX
A1.1 Asymptotic expansion of the expected actual error rate for the Euclidean distance classifier 68
A1.2 Asymptotic expansion of the expected actual error rate for the linear discriminant function 91
A2.1 Asymptotic expansion of the expected plug-in error rate for the Euclidean distance classifier 124
A2.2 Asymptotic expansion of the expected plug-in error rate for the linear discriminant function 153
A3 Computer programs 174
A4 Computational results (Table 9 to Table 23) 202
E1 Some values of $\mathbf{\Sigma}^{-1}$ to explain odd results 222

# *List of figures and tables*

PAGE

**Figure 1** : Illustration of a basic problem of statistical discriminant analysis with two populations. 2

**Table 1** : Values of $m^*$ under the case of "non-equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ (i.e.for cases A1 and P1). 38

**Table 2** : Values of $m^*$ under the case of "non-equivalence" with $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( i.e.for cases A2 and P2 ). 39

**Table 3** : Values of m under the case of "equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ ( i.e.for cases A3 and P3 ). 40

**Table 4** : Values of m under the case of "equivalence" with $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( positive $\rho$ in cases A4 and P4 ). 41

**Table 4a** : Values of m under the case of "equivalence" with $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( negative $\rho$ in cases A4 and P4 ). 42

**Table 5** : The 'true', expected actual and expected plug-in error rates of the EDC and LDF under the case of 'non - equivalence' with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$. 45

**Table 6** : The 'true', expected actual and expected plug-in error rates of the EDC and LDF under the case of 'non - equivalence' with $\mathbf{\Sigma} = \mathrm{AR}(1)$. 48

**Table 6a** : The 'true', expected actual and expected plug-in error rates of the EDC and LDF under the case of 'non-equivalence' with $\mathbf{\Sigma} = \mathrm{AR}(1)$. 49

**Table 7** : The 'true', expected actual and expected plug-in error rates of the EDC and LDF under the case of 'equivalence' with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$. 51

**Table 8** : The 'true', expected actual and expected plug-in error rates of the EDC and LDF under the case of 'equivalence' with $\mathbf{\Sigma} = \mathrm{AR}(1)$. 52

**Table 8a** : The 'true', expected actual and expected plug-in error rates of the EDC and LDF under the case of 'equivalence' with $\mathbf{\Sigma} = \mathrm{AR}(1)$. 54

**Table 9** : The expected actual error rate of the EDC under the case of "non-equivalence" $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ ( i.e. case A1 ). 202

**Table 10** : The expected actual error rate of the LDF under the case of "non-equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ ( i.e. case A1 ). 203

**Table 11** : The expected actual error rate of the EDC under the case of " non-equivalence " when $\mathbf{\Sigma} = \mathrm{AR}(1)$ (with positive ρ in case A2). 204

**Table 11a:** The expected actual error rate of the EDC under the case of "non-equivalence" when $\mathbf{\Sigma} = \mathrm{AR}(1)$ (with negative ρ in case A2). 205

**Table 12** : The expected actual error rate of the LDF under the case of "non-equivalence" with $\mathbf{\Sigma} = \mathrm{AR}(1)$ (i.e. case A2). 206

**Table 13** : The expected actual error rate of the EDC under the case of "equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ (i.e. case A3 ). 207

**Table 14** : The expected actual error rate of the EDC under the case of "equivalence" when. $\mathbf{\Sigma} = \mathrm{AR}(1)$ (with positive ρ in case A4 ). 208

**Table 14a** : The expected actual error rate of the EDC under the case of "equivalence" when $\mathbf{\Sigma} = \mathrm{AR}(1)$ (with negative ρ in case A4 ). 209

**Table 15** : The expected plug-in and the expected actual error rates of the EDC under the case of "non-equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ (i.e. case P1, n1=n2=50). 210

**Table 15a** : The expected plug-in and the expected actual error rates of the EDC under the case of "non-equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ (i.e. case P1, n1=n2=100). 211

**Table 16** : The expected plug-in and expected actual error rates of the LDF under the case of "non-equivalence " with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ (i.e. case P1 ). 212

**Table 17** : The expected plug-in and the expected actual error rates of the EDC under the case of "non-equivalence" when $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( with positive $\rho$ in case P2 ). 213

**Table 17a** : The expected plug-in and the expected actual error rates of the EDC under the case of "non-equivalence" when $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( with negative $\rho$ in case P2 ). 214

**Table 18** : The expected plug-in and expected actual error rates of the LDF under the case of "non-equivalence " when $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( with positive $\rho$ in case P2 ).

**Table 18a** : The expected plug-in and expected actual error rates of the LDF under the case of " non-equivalence " when $\mathbf{\Sigma} = \mathrm{AR}(1)$ ( with negative $\rho$ in case P2 ). 216

**Table 19** : The expected plug-in and the expected actual error rate of the EDC under the case of "equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ ( i.e. case P3 ). 217

**Table 20** : The expected plug-in and the expected actual error rates of the EDC under the case of "equivalence" with $\mathbf{\Sigma} = \mathrm{AR}(1)$ (i.e. case P4). 218

**Table 21** : The expected plug-in error rate of the LDF under the case of "equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ (i.e. case P3). 219

**Table 22** : The expected plug-in error rate of the LDF under the case of "equivalence" with $\mathbf{\Sigma} = \mathrm{AR}(1)$ (i.e. case P4). 220

**Table 23** : The expected plug-in error rate of the LDF under the case of "non-equivalence" with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ or $\mathbf{\Sigma} = \mathrm{AR}(1)$, (i.e. case P1 or P2 ). 221