DEPARTMENT OF COMPUTING
IMPERIAL COLLEGE LONDON
UNIVERSITY OF LONDON

Maximum Entropy Covariance Estimate for Statistical Pattern Recognition

Carlos Eduardo Thomaz

A thesis submitted for the degree of
Doctor of Philosophy (Ph.D.) of the University of London
and for the Diploma of the Imperial College (D.I.C.)

2004

To my wife and my parents.

My lovely son Lucas and
My little one who is coming soon
Welcome Life !

Abstract

In classification problems, when the number of examples per class is less than (or comparable to) the dimension of the feature space, the performance of statistical pattern recognition techniques tends to deteriorate. This problem, called the “limited sample size problem”, is indeed quite common nowadays, especially in image recognition applications. For instance, in face recognition each individual or class is defined by a small number of pictures but the features used for recognition may be hundreds of pre-processed image attributes.

In the statistical approach, the region of the feature space occupied by each class is generally determined by the probability distribution of the observations belonging to each class, which must be either specified or learned. Many of these probability distribution estimations are based on Gaussian kernels that involve the inverse of the true covariance matrix of each class. The usual choice for estimating the true covariance matrices is the maximum likelihood estimator defined by the corresponding sample group covariance matrices. However, these matrices are either poorly estimated or cannot be inverted when the group sample sizes are smaller than the number of features or parameters.

In this thesis, a new covariance estimate called Maximum Entropy Covariance Selection (MECS) is proposed. This estimate is based on combining covariance matrices to take into account the principle of maximum uncertainty. When limited information is provided, we show that the problem of estimating covariance matrices for classification is affected not only by the way this information is optimised but also by its reliability. The new covariance method does not require an iterative optimisation procedure and, hence, its estimation, differently from others, is not exclusive to the parametric Bayesian classifier. In fact, we demonstrate that MECS can be used in the parametric as well as non-parametric Bayesian classifiers whenever the sample group covariance matrices are ill-posed or poorly estimated.

The singularity and instability of covariance matrices is a critical issue not only for Bayesian classifiers but also other statistical covariance-based analysis, such as the Linear Discriminant Analysis (LDA). We also show that the novel method of combining covariance matrices in limited sample size problems improves the LDA classification performance, with or without an intermediate dimensionality reduction step and using few linear discriminant features.

Acknowledgements

Since I came here, approximately three years ago, I have been very, very pleased to receive all kind of support from my supervisor, Dr. Duncan Fyfe Gillies. I would like to take this opportunity to say that Duncan is an admirable supervisor. In fact, he is a human being that I very much respect for not only his scientific skills but also his standards of behaviour. Duncan, thanks a lot for all you have done for me and my family.

I have been very fortunate to receive the support of Dr. Raul Queiroz Feitosa from Pontifical Catholic University of Rio de Janeiro, PUC-RJ, Brazil. Raul was my MSc. supervisor in Brazil some five years ago and the person who introduced me to the research fields of Computer Vision and Pattern Recognition. Some of the ideas presented here were topics of very stimulating electronic discussions that happened in the first two years of this study.

Also, I would like to express my gratitude for the support of Professor Guang-Zhong Yang, Dr. Daniel Rueckert, and all the past and current Visual Information Processing members, especially Philip, Dave, and Alex. It has been an honour and a pleasure to belong to such a motivating team.

The sponsorship of the Brazilian Government agency CAPES, which covered my Imperial College fees and my family living expenses, made this work possible. In addition, the financial support of the Department of Computing, which approved my funding requests for six international conferences, was definitely a real boost for my research development.

I would like to say a huge thank you to my brothers Fernando and Paulinho; my friends Luis, Stellet, and Enedir; and my family-in-law: Seu Hamilton (in memoriam), Dona Neuza, Seu Adalberto, Dona Marly, Tia Sely, Elaine, Ricardo, Lucia, and Max; for your support and friendship. Moreover, I would like to thank some old and new Brazilian friends that I have met here in London. Torres, Artur, Simone, Ronaldo, Sonja, Kiko, Karina, Julinha, Roberto, Lamb, and Lucio, are only some of the names to be mentioned.

Most importantly, my gratitude and love go to my wife Sheila, my babies Lucas and the one who is coming soon, and my parents Antonio and Ludi. This thesis is theirs as well as the great happiness behind it. You have given so much to me and I would like to dedicate this work to you. Thank God for that.

Contents

Abstract. iii

Acknowledgements. iv

List of Figures. viii

List of Tables. ix

1 Introduction **1**

1.1 Motivation 1

1.2 Objective 3

1.3 Outline of the Thesis 5

2 Mathematical Foundations **8**

2.1 Vectors and Matrices 8

2.2 Eigenvectors and Eigenvalues 10

2.3 Entropy and Information 11

2.4 A Typical Statistical Pattern Recognition System 13

2.5 Principal Component Analysis 14

2.6 The Statistical Decision Making Process 16

2.7 The Bayes Decision Rule 17

2.8 Parametric and Non-Parametric Statistical Methods 19

2.9 Notation 20

3 The Bayes Plug-in Classifier **21**

3.1 The Conventional Bayes Plug-in Classifier 21

3.2 Limited Sample Size Problem 23

3.3 Linear Discriminant Classifier 26

3.4 Unconventional Bayes Plug-in Classifiers 27

3.4.1 Regularised Discriminant Analysis Method 27

3.4.2 Empirical Bayesian Method. 30

3.4.3 Leave-One-Out Likelihood Method 33

3.4.4 Bayesian Leave-One-Out Likelihood Method 36

3.4.5 Simplified Quadratic Discriminant Function Method 38

3.5 Summary and Conclusions. 39

4 The Sample Size Problem in Image Recognition **41**

4.1 Mixing Sample Group and Pooled Covariance Matrices 41

4.1.1 Definition. 42

4.1.2 The Mixture Parameter. 43
4.1.3 Experiments. 46
4.1.3.1 ORL Face Database. 46
4.1.3.2 Facial Expression Database. 46
4.1.3.3 Implementation. 47
4.1.4 Results. 48
4.1.5 Discussion. 49
4.2 The Loss of Covariance Information. 50
4.2.1 Definition. 50
4.2.2 Covariance Projection Ordering Method 53
4.2.3 Experiments. 54
4.2.4 Results. 54
4.2.5 Discussion. 57
4.3 Summary and Conclusions 57
5 Covariance Matrix Estimation 59
5.1 The Maximum Entropy Covariance Selection Method. 59
5.1.1 Definition. 60
5.1.2 Algorithm. 62
5.2 Visual Analysis of Covariance Matrix Estimates. 63
5.2.1 Mixture Covariance Approaches. 63
5.2.2 Experiments. 64
5.2.2.1 FERET Face Database. 64
5.2.2.2 Implementation. 65
5.2.3 Results. 66
5.2.4 Discussion. 72
5.3 Synthetic Data Analysis. 72
5.3.1 Estimation of the Bayes Classification Error. 72
5.3.2 Experiments. 73
5.3.3 Results. 74
5.3.3.1 Equal Spherical Covariance Matrices. 75
5.3.3.2 Equal Ellipsoidal Covariance Matrices. 77
5.3.3.3 Unequal Ellipsoidal Covariance Matrices. 80
5.3.4 Discussion. 82
5.4 Image Data Analysis. 83
5.4.1 Experiments. 83
5.4.1.1 Fingerprint Database. 83
5.4.1.2 OCR Database. 84

5.4.1.3 Implementation. 85

5.4.2 Results. 86

5.4.3 Discussion. 89

5.5 Summary and Conclusions. 91

6 The Parzen Window Classifier **93**

6.1 The Conventional Parzen Window Classifier. 94

6.2 Unconventional Parzen Window Classifiers. 95

6.2.1 Van Ness Covariance Method 95

6.2.2 Toeplitz Covariance Method. 96

6.3 Synthetic Data Analysis. 98

6.3.1 Experiments. 98

6.3.2 Results. 99

6.3.3 Discussion. 103

6.4 Image Data Analysis. 104

6.4.1 Experiments. 105

6.4.2 Results. 106

6.5 Summary and Conclusions. 109

7 Fisher Discriminant Analysis **111**

7.1 Linear Discriminant Analysis (LDA). 112

7.2 LDA Limited Sample Size Approaches. 114

7.2.1 Fisherfaces Method. 114

7.2.2 Chen et al.'s Method. 115

7.2.3 Yu and Yang's Method. 116

7.2.4 Yang and Yang's Method. 117

7.3 The Maximum Uncertainty LDA-based Approach. 118

7.3.1 Related Methods. 118

7.3.2 The Proposed Method. 119

7.4 Experiments. 122

7.5 Results. 123

7.6 Memory Issues. 125

7.7 Summary and Conclusions. 127

8 Conclusions and Future Work **129**

8.1 Conclusions. 129

8.2 Future Work. 131

Bibliography **137**

List of Figures

Figure 1.1. Geometric idea of covariance matrices for parametric classifiers. 3

Figure 1.2. Geometric idea of covariance matrices for non-parametric classifiers. 4

Figure 2.1. A typical statistical pattern recognition system. 13

Figure 2.2. An example of concatenating rows of an image to form a vector. 15

Figure 4.1. Geometric idea of the mixture covariance matrix. 43

Figure 4.2. Ten images' set of one individual from the ORL Face Database. 46

Figure 4.3. Anger, disgust, fear, …, sadness, and surprise (top left – bottom 47

Figure 4.4. Geometric idea of a hypothetical "loss of covariance information". . . 52

Figure 5.1. Some example images from the FERET Database. 65

Figure 5.2a. ORL visual analysis. The top row shows … entropy mixture. 68

Figure 5.2b. ORL visual analysis. The top row shows … entropy mixture. 69

Figure 5.3a. FERET visual analysis. The top row shows … entropy mixture. 70

Figure 5.3b. FERET visual analysis. The top row shows … entropy mixture. . . . 71

Figure 5.4. Equal spherical covariance matrices – QDF classification results. . . . 76

Figure 5.5. Equal spherical covariance matrices – RDA and LOOC parameters. . 76

Figure 5.6. Equal ellipsoidal covariance matrices – QDF classification results. . . 79

Figure 5.7. Equal ellipsoidal covariance matrices – RDA and LOOC parameters. 79

Figure 5.8. Unequal ellipsoidal covariance matrices – QDF classification results. 81

Figure 5.9. Unequal ellipsoidal covariance matrices – RDA and LOOC parame- 81

Figure 5.10. (a) Arch, (b) left loop, …, and (e) whorl images. 84

Figure 5.11. FERET face database recognition error for Bayes plug-in classifiers. 86

Figure 5.12. Tohoku facial expression recog. error for Bayes plug-in classifiers. . 87

Figure 5.13. NIST-4 fingerprint error for Bayes plug-in classifiers. 88

Figure 5.14. NIST-3/7 OCR databases error for Bayes plug-in classifiers. 89

Figure 6.1. FERET face database recognition error for Parzen classifiers. 106

Figure 6.2. Tohoku facial expression recognition error for Parzen classifiers. . . . 107

Figure 6.3. NIST-4 fingerprint recognition error for Parzen classifiers. 108

Figure 7.1. Geometric idea of the new LDA-based method. 122

Figure 8.1. Scatter plot of four hypothetical observations belonging to a class. . . 135

Figure 8.2. Scatter plot of other four observations belonging to the class. 136

List of Tables

Table 2.1. Notation used throughout this work. 20

Table 4.1. ORL face recognition results (Smix). 48

Table 4.2. Tohoku facial expression recognition results (Smix). 49

Table 4.3. The average (with std deviations) of optimised mixture parameters. . . 49

Table 4.4. ORL face recognition results (COPO). 55

Table 4.5. Tohoku facial expression recognition results (COPO). 56

Table 5.1. ORL classification results 66

Table 5.2. FERET classification results 66

Table 5.3. Computational time (in seconds) for the quadratic classifiers. 90

Table 6.1. Equal spherical covariance matrices – EUC and Parzen results. 100

Table 6.2. Equal ellipsoidal covariance matrices – EUC and Parzen results. 101

Table 6.3. Unequal ellipsoidal covariance matrices – EUC and Parzen results. . . 102

Table 6.4. Lower classif. bounds of non-parametric and parametric classifiers. . . 104

Table 6.5. Computational time (in seconds) for the Parzen classifiers. 109

Table 7.1. ORL (32x32 pixels) LDA classification results. 124

Table 7.2. FERET (16x16 pixels) LDA classification results. 124

Table 7.3. ORL (64x64 pixels) and FERET (96x64 pixels) LDA classif. results. . 126

Chapter 1

Introduction

Pattern recognition is known as an important area of research in Computer Science and Mathematics. More specifically, statistical pattern recognition is a well-established field of study that has been successfully applied in several domains, such as Engineering, Biology, Psychology, and Medicine.

Although the concept of statistical pattern recognition reflects theoretical approaches to problems of clustering and feature extraction or selection, its primary goal is classification [JDM00]. In other words, the main idea of a statistical pattern recognition system is to formulate a decision-making process where a pattern is assigned to one of a finite number of pre-specified classes characterised by their respective probability density functions. This is essentially our main topic of research here.

In the next two sections of this first chapter, we explain our motivation for this work and provide the details of the objective to be achieved. An outline of the dissertation is described in the last section.

1.1 Motivation

In many pattern recognition applications nowadays there are often a large number of features available, but the number of training patterns per class or group may be significantly less than the dimension of the feature space. For instance, in image recognition problems, each group is commonly defined by a small number of pictures but the number of features used for recognition may be thousands of pixels or even hundreds of pre-processed image attributes. This implies that the performance of classical statistical pattern recognition techniques, which have been used successfully to design several recognition systems, deteriorate in such limited sample size settings.

In the statistical approach, a pattern is represented by a set of features or parameters and the region of the feature space occupied by each class is determined by the probability distribution of its corresponding patterns, which must be either specified (parametric approach) or learned (non-parametric approach). There are a number of classification rules available to define appropriate statistical decision-making boundaries. The well-known Bayes' decision rule that assigns a pattern to the class with the highest posterior probability is the one that achieves minimal misclassification risk among all possible rules (see, e.g., [And84]).

The idea behind Bayes' rule is that the main information available about class member-ship is contained in the set of conditional probability densities. Among the various density functions that have been investigated, none has received more attention than the multivariate normal or Gaussian density [DHS01]. The main reason for choosing a Gaussian kernel to model a class-conditional probability density is not only its relatively simple analytical properties, which can be completely specified by its corresponding true mean and covariance matrix, but also because Gaussian distributions are frequently found experimentally in natural systems.

The class membership similarities given by Gaussian kernel functions involve the inverse of the true covariance matrix of each class. As in real-world problems, the true covariance matrices are seldom known and estimates must be computed based on the patterns available in a training set. The usual choice for estimating the true covariance matrices is the maximum likelihood estimate defined by the corresponding sample group covariance matrices. However, in limited sample size applications the inverse of sample group covariance matrices is either poorly estimated or cannot be calculated when the number of training patterns per class is smaller than the number of features. Therefore, the most common Bayesian classification methods, which are based on the inverse of the sample group covariance matrices, cannot be used.

In the last 25 years, several researchers have proposed different modifications to the sample group covariance matrices that increase their stability when limited information is available [VNe80, Fri89, Fuk90, GR89, GR91, HFT96, HL96, OSA00]. However, as we will present in this work, most of these covariance estimate approaches rely on optimisation techniques that are time consuming and exclusive to the type of Bayesian classifier used. Therefore, to some extent, the tractability of Gaussian kernels have been lost because the sample group covariance matrices have been replaced with covariance estimates that are mathematically invertible, but not necessarily computationally feasible when the number of classes is large. It would be certainly rewarding if we could restore the simplicity and tractability of the Gaussian distribution in such limited sample size and high dimensional problems.

1.2 Objective

The objective of this thesis is to investigate and develop a new covariance approximation for the sample group covariance matrices used in Bayesian classifiers. Our main concern here is with pattern recognition problems composed of limited training sets, a large number of features, and several groups. Biometric image recognition applications, such as face recognition and fingerprint recognition, are examples of the applications studied.

The geometric idea of sample group covariance matrices for parametric Bayesian classifiers can be illustrated as follows. Let there be a two-dimensional feature space containing three hypothetical samples drawn randomly from three distinct classes normally distributed, as shown in Figure 1.1.

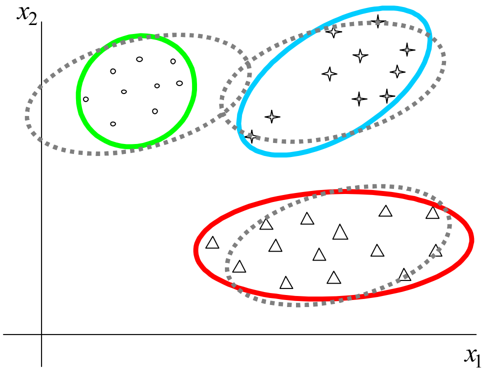


Figure 1.1. Geometric idea of covariance matrices for parametric classifiers.

The coloured ellipses correspond to contours of constant probability densities governed by the sample group covariance matrices centred at the corresponding mean vectors. That is, the principal directions of the ellipses are aligned with the eigenvectors of the respective sample group covariance matrices and the magnitude of these eigenvectors are given by the square root of each corresponding eigenvalue [Fuk90]. However, in sparse and high dimensional problems, these ellipses become ellipsoids that cannot be represented because some of those eigenvalues are zero.

One way to overcome this problem in parametric Bayesian classifiers is to assume that all groups have equal covariance matrices and to use as their estimates the weighted average of each sample group covariance matrix, given by the pooled covariance matrix calculated from the whole training set. The pooled covariance matrix is shown as dotted grey ellipses in Figure 1.1. As we can see, the pooled covariance estimate is not a convenient approximation for all the three hypothetical sample group covariance matrices presented.

Considering the same three hypothetical samples described previously, Figure 1.2 illustrates the geometric idea of the sample group covariance matrices for non-parametric Bayesian classifiers. Since non-parametric Bayesian classifiers are not restricted to unimodal distributions, each class can be described by more than one Gaussian kernel. However, analogously, in sparse and high dimensional problems the ellipses become ellipsoids that cannot be represented because some of the sample group covariance eigenvalues are zero.

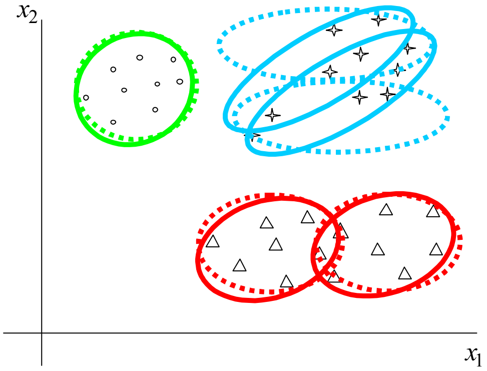


Figure 1.2. Geometric idea of covariance matrices for non-parametric classifiers.

The usual way to overcome this limited sample size problem in non-parametric Bayesian classifiers is to assume that each sample group covariance matrix has a diagonal form. As a consequence, all the ellipses described in Figure 1.2 would have the same principal directions, as shown by the corresponding coloured dotted ellipses. Again, we can see that this approximation for the sample group covariance matrices would not necessarily suit all the three hypothetical classes shown.

Therefore, in order to propose a new covariance estimate for sample group covariance matrices, we need to address the following question: Is it possible to develop an approximation for singular or poorly estimated covariance matrices that (1) improves the classification accuracy in limited sample size settings; (2) is computationally feasible if several classes are considered; (3) does not have a unique or restrictive form; and (4) is valid for parametric and non-parametric Bayesian classifiers and any other statistical covariance-based analysis ? We believe that this thesis will show that the answer to this question is: “Yes! It is possible.”

1.3 Outline of the Thesis

This thesis can be summarised as follows.

In Chapter 2, we present some mathematical concepts that have been used throughout this work. These concepts are essentially fundamental topics from linear algebra, information theory, and probability theory that are important within multivariate statistical analysis. The reader will be referred to the detailed background material for a comprehensive exposition of these topics. The notations for the most commonly occurring quantities used in this dissertation are shown in the last section of this chapter.

In Chapter 3, we present the main idea of the parametric Bayes plug-in classifier and review its most important non-conventional implementations, such as the Regularised Discriminant Analysis (RDA) [Fri89] and the Leave-One-Out Likelihood (LOOC) method [HL96], which address the difficulties caused by limited sample size problems. Experiments carried out by a number of researchers have shown that choosing a non-conventional Bayes plug-in classifier improves the classification accuracy in settings for which sample sizes are limited and the number of features is large. These ideas have proved to be true in cases where no more than 20 groups are required, but have not been verified for a large number of groups.

Chapter 4 analyses the performances of several non-conventional Bayes plug-in covariance estimators, reviewed previously, in pre-processed image recognition problems that consist of small and moderate training sets, a large number of features, and a moderate number of groups. Experiments carried out on face and facial expression recognition confirm the findings of other researchers that choosing a non-conventional Bayes plug-in classifier between the linear and quadratic ones improves the classification accuracy in settings for which sample sizes are small and the number of features is large. However, in those well-framed applications where the sources of variation are the same from group to group and consequently a similar covariance shape might be assumed for all groups, linear combinations of the sample group covariance matrices and the pooled covariance matrix may lead to a loss of covariance information. An initial approach to understand this problem will be presented and shows the importance of taking into account the distinct information provided by the sample group covariance matrix and the pooled covariance matrix in the whole high-dimensional feature space.

In Chapter 5, a new non-conventional Bayes Plug-in classifier is proposed. This classifier is based on a new covariance matrix estimate, called Maximum Entropy Covariance Selection (MECS) method, which combines covariance matrices under the principle of maximum uncertainty. The main idea of the MECS approach is to expand in a straightforward way the smaller and consequently less reliable eigenvalues of the sample group covariance matrix while trying to keep most of its larger eigenvalues unchanged. The results indicate that in image recognition applications where the sources of variation are commonly the same from group to group, limited training samples sizes are considered, and high concerns about computation costs exist, the MECS approach is preferable to RDA and LOOC non-conventional quadratic classifiers.

In Chapter 6, we initially present the basic concepts of the Parzen Window classifier and review its most relevant non-conventional approaches for limited sample size problems. Since the MECS approach is a direct procedure that is not exclusive to the parametric Bayes Plug-in classifier, we then investigate the MECS performance as a new kernel covariance estimator for the non-parametric Parzen Window classifier. The experimental results carried out on synthetic and image data indicate that the less restricted MECS covariance estimate improves the classification performance of the Parzen Window classifier with Gaussian kernels, especially when the sample size is small and the data parameters are highly correlated.

In Chapter 7, a new Fisher-based method of linear discriminant analysis (LDA) is proposed. Analogously to the procedure described in the chapter 5, the new LDA-based method is a straightforward approach that considers the issue of stabilising the ill posed or poorly estimated within-class scatter matrix with a multiple of the identity matrix. In order to evaluate its effectiveness, experiments on face recognition using benchmark databases are carried out and compared with other LDA-based methods. The results indicate that our method improves the LDA classification performance when the within-class scatter matrix is singular as well as poorly estimated, with or without a Principal Component Analysis intermediate step and using fewer linear discriminant features.

In Chapter 8, we conclude the thesis and discuss some issues that have emerged from this work, such as the feasibility of quantifying the assumption that the covariance matrix estimations share some similarities and the association of the MECS approach with other regularisation methods of combining singular and non-singular covariance estimates.

Chapter 2

Mathematical Foundations

In this chapter, we present a number of basic mathematical concepts from linear algebra, information theory, and probability theory that are important within statistics and have been used throughout this work.

The chapter is organised in three parts. The first part, consisting of Sections 2.1 and 2.2, provides some definitions and results of linear algebra, more specifically matrix algebra, that have been used in the study of multivariate statistical analysis. A comprehensive exposition of these topics can be found in [Sea66, Str88]. The second part of the chapter, consisting of Section 2.3, discusses briefly the idea of entropy as a quantitative measure of information, which leads to the maximum entropy statistical inference used in this thesis. For a broad treatment of information theory, the reader is referred to the book by Cover and Thomas [CT91]. In the final and third part, consisting of Sections 2.4 through 2.8, we describe some relevant topics in one of the most general frameworks to formulate solutions to pattern recognition problems, the statistical pattern recognition approach. The chapter concludes with the notation used throughout this dissertation, which is presented in Section 2.9.

2.1 Vectors and Matrices

An array x of n real numbers $x_1, x_2, ..., x_n$ is called a vector, and it is written as

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{or} \quad x^T = [x_1, x_2, \ldots, x_n], \tag{2.1}$$

where x^T is its transpose.

A vector can be represented geometrically as a point in an n-dimensional space or a line in n dimensions with component x_1 along the first axis, x_2 along the second axis, …, and x_n along the nth axis [JW98]. In statistics, vectors are often referred as patterns and are used to represent the measurements of a number of variables.

A matrix is defined as a rectangular array of real numbers arranged into rows and columns. It is said to be square if it has as many rows as it has columns. A particular square and diagonal matrix is the *identity* matrix I whose on-diagonal elements are 1's and all off-diagonal elements 0. The matrix I plays the same role in matrix multiplication as the number 1 does in ordinary multiplication [Jam85]. In other words, the following equation

$$IA = AI = A \qquad \textbf{(2.2)}$$

is valid for any matrix A of the appropriate size so that the multiplications can be performed.

There are two other particular square matrices that are of special importance in multivariate statistical analysis: the *symmetric* and *orthogonal* matrices. A square $n \text{x} n$ matrix A is called *symmetric* if

$$A = A^T. \qquad \textbf{(2.3)}$$

For example, covariance matrices are symmetric matrices. A real square matrix is said to be *orthogonal* if

$$AA^T = A^T A = I \quad \text{or} \quad A^{-1} = A^T \qquad \textbf{(2.4)}$$

and its columns are *orthonormal.* That is, for $A = [a_1, a_2, \ldots, a_n]$

$$a_i^T a_j = \begin{cases} 1 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j. \end{cases} \qquad \textbf{(2.5)}$$

The eigenvector matrix of a covariance matrix is, for instance, an orthogonal matrix. It is important to mention that although a rectangular matrix can still have the property that $AA^T = I$ or $A^T A = I$, it cannot have both and, consequently, is said to be a semi-orthogonal matrix [MN99].

2.2 Eigenvectors and Eigenvalues

One of the most important results of matrix algebra that finds application within statistics is the topic of eigenvectors and eigenvalues. We can describe the main idea of this linear transformation as follows.

Let A be an $n \mathrm{x} n$ square matrix. The eigenvalues of A are defined as the roots of the following equation

$$\det(A - \lambda I) = |A - \lambda I| = 0\,, \tag{2.6}$$

where I is the $n \mathrm{x} n$ identity matrix. Equation (2.6), which is called the characteristic equation [MN99], has n roots. These roots can be complex numbers. Let λ be an eigenvalue of A. Then there exists a vector x such that

$$Ax = \lambda x\,. \tag{2.7}$$

The vector x is called an eigenvector of A associated with the eigenvalue λ. Ordinarily, we normalise x so that it has length one, that is, $x^T x = 1$.

In general, the vector Ax defined in equation (2.7) is a new vector that will not be simply related to x. That is, x changes direction when multiplied by A, so that Ax is not a multiple of x. This means that only certain special numbers λ are eigenvalues, and only certain special vectors x are eigenvectors. However, as pointed out by Strang [Str88], if A were a multiple of the identity matrix, then no vector would change direction and all vectors would be eigenvectors.

Although eigenvalues are in general complex, the eigenvalues of a real symmetric matrix are always real [MN99]. This is a fundamental and remarkable result for the covariance matrices used in multivariate statistical analysis, because not only do the eigenvectors exist but also there exists a complete set of n eigenvectors and their corresponding eigenvalues. In other words, there exist an orthogonal $n \mathrm{x} n$ matrix Φ whose columns are eigenvectors of the covariance matrix Σ_x and a diagonal matrix Λ whose diagonal elements are the eigenvalues of Σ_x, such that

$$\Phi^T \Sigma_x \Phi = \Lambda\,. \tag{2.8}$$

Therefore, the linear transformation given by the eigenvectors matrix Φ diagonalises the covariance matrix Σ_x in the new coordinate system, creating a set of new variables

$$y = \Phi x \tag{2.9}$$

that are uncorrelated. In fact, as we will describe later in this chapter, this linear transformation essentially finds the principal components of the covariance structure.

2.3 Entropy and Information

In 1948, Claude Shannon introduced the mathematical foundations of information theory and the remarkable concept of entropy as an information measure in statistics [Sha48]. At that time, Shannon's original work on information theory was in direct response to the need for electrical engineers to design communication systems that are both efficient and reliable [Hay99].

Despite its practical origin, information theory as it is known nowadays is not only a deep mathematical theory concerned with the very essence of the communication process, but also a framework of study that provides a constructive criterion for setting up probability distributions on the basis of partial knowledge or limited information [Jay82]. This is essentially our main context of study here and, hence, the purpose of this section is to discuss the idea of entropy as a quantitative measure of information, which leads to the type of statistical inference used in this work, the maximum entropy estimate [Jay57].

Let an event X have N possible values, that is, X is capable of assuming the discrete values x_j ($j = 1,2,\ldots,N$). Each one of these values x_j has $p(x_j)$ probability of occurrence with the two fundamental requirements that

$$0 \le p(x_j) \le 1 \quad \text{and} \quad \sum_{j=1}^{N} p(x_j) = 1 . \tag{2.10}$$

The amount of information gained after observing the event $X = x_j$ with probability $p(x_j)$ is defined by the following equation [Hay99]

$$I(x_j) = \ln\left(\frac{1}{p(x_j)}\right) = -\ln p(x_j) . \tag{2.11}$$

Equation (2.11) states basically that the amount of information described by the value x_j is related to the inverse of its probability of occurrence. In other words, if the N possible values for the event X occur with different probabilities and, in particular, the probability $p(x_j)$ is low, then there is more surprise and, consequently, more information when X takes the value x_j rather than another value x_k ($k = 1,2,\ldots,N$) with higher probability [Hay99].

The entropy $H(X)$ of the event X is defined as the expected value, or mean, of the information described in equation (2.11), such that

$$H(X) = E\{I(x_j)\} = \sum_{j=1}^{N} p(x_j) I(x_j) = -\sum_{j=1}^{N} p(x_j) \ln p(x_j) \,. \qquad \textbf{(2.12)}$$

In case any of the probabilities vanish, that is, $p(x_j) = 0$ for any $0 \le j \le N$, we use the fact that $\lim_{p(x)\to 0} p(x) \ln p(x) = 0$ to take $0 \ln 0$ to be 0 [DHS01]. Analogously, for a continuous n-dimensional random vector X_i, the entropy $h(X_i)$ is given by [Hay99]

$$h(X_i) = -\int_{-\infty}^{\infty} p(x) \ln p(x) \, \mathrm{dx} = -E\{\ln p(x)\}, \qquad \textbf{(2.13)}$$

where $x \equiv [x_1, x_2, \ldots, x_n]^T$ and $p(x)$ is the probability density function of X_i.

Equation (2.12), or equivalently equation (2.13), describes a quantity that increases with increasing uncertainty. As pointed out by Jaynes [Jay57], this is an impressive result because not only is the entropy a unique and unambiguous criterion for the amount of uncertainty inherent in a discrete or continuous event, but also it agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one. In other words, the higher the entropy, the higher the uncertainty as to which possibility of the analysed event will occur.

Therefore, when we are making inferences on the basis of limited information it is natural to think that we should use that probability distribution which has the maximum uncertainty or entropy. In fact, as stated by Jaynes [Jay57], this is the only unbiased assignment we can make because any other would amount to an arbitrary assumption of information, which by definition we do not have.

In the discrete case, $H(X)$ is strictly non-negative and will be maximised when the distribution is uniform, i.e. all outcomes are equally likely. However, in the multivariate continuous case, the entropy $h(X_i)$ may be negative and its maximum value, among all continuous probability density functions having a given mean and covariance matrix for the random vector X_i, will be attained by the multivariate Gaussian distribution [DHS01].

2.4 A Typical Statistical Pattern Recognition System

Although a pattern recognition investigation may consist of several steps, a fairly typical statistical recognition system is commonly partitioned into components such as the ones shown in Figure 2.1.

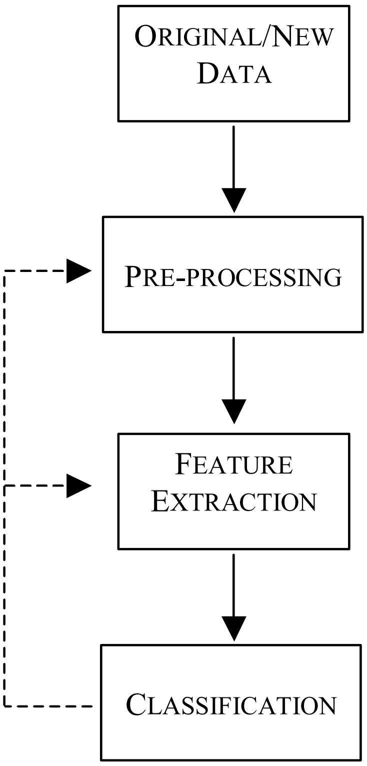


Figure 2.1. A typical statistical pattern recognition system.

The pre-processing and feature extraction (or selection) stages operate on the original (or new) samples (or data) in a way that normalises the pattern of interest, segmenting it from the background, diminishing noise, and removing redundant or irrelevant information. These stages attempt to transform data in a way that input vectors belonging to distinct classes should occupy as compact and disjoint regions in a lower dimensional feature space as possible for a subsequent classification.

The task of the classification stage or the classifier component, which is our main concern in this work, is to use the feature vectors provided by the previous stages to assign the object to a specific class or group. In the formal setting, an object is assumed to be a member of one, and only one, class and an error with an associated cost or loss is incurred if it is assigned to a different class [Fri89].

In order to improve the classification results of both the original and new data, those pre-stages might often be revisited (feedback) during the decision-making process. This feedback procedure can essentially use the output of the classifier to recommend further adjustments on the statistical recognition system to improve its classification performance.

2.5 Principal Component Analysis

Principal Component Analysis, or simply PCA, is a feature extraction procedure concerned with explaining the covariance structure of a set of variables through a small number of linear combinations of these variables. It is a well-known statistical technique that has been used in several image recognition problems, especially for dimensionality reduction. A comprehensive description of this multivariate statistical analysis method can be found in [Fuk90, JW98].

Let us consider the face recognition problem as an example to illustrate the main idea of the PCA. In any image recognition, and particularly in face recognition, an input image with n pixels can be treated as a point in an n-dimensional space called the image space [KS90, TP91]. The coordinates of this point represent the values of each pixel of the image and form a vector $x^T = [x_1, x_2, \ldots, x_n]$ obtained by concatenating the rows (or columns) of the image matrix.

Figure 2.2 shows an example of concatenating the rows of a 64x64 (or 4096) pixels image to represent a feature vector in the 4096-dimensional space. For this representation to make sense in classification problems, we are assuming implicitly that two images that look like one another correspond to two close points in the high dimensional image space.

$$= \begin{bmatrix} 150 & 152 & \dots & 151 \\ 153 & 154 & \dots & 155 \\ \vdots & \vdots & \vdots & \vdots \\ 254 & 255 & \dots & 252 \end{bmatrix}_{64 \text{ x } 64 \text{ pixels}}$$

$$\Big\downarrow$$

$$\begin{bmatrix} 150 & 152 & \dots & 151 & 153 & 154 & \dots & 155 & \dots & 254 & 255 & \dots & 252 \end{bmatrix}_{4096 \text{ variables}}$$

Figure 2.2. An example of concatenating the rows of an image matrix to form a vector.

It is well-known that well-framed face images are highly redundant not only owing to the fact that the image intensities of adjacent pixels are often correlated but also because every individual has one mouth, one nose, two eyes, etc. As a consequence, an input image with n pixels can be projected onto a lower dimensional space without significant loss of information.

Let an $N \text{ x } n$ training set matrix X be composed of N input face images with n pixels. This means that each column of matrix X represents the values of a particular pixel observed all over the N images. Let this data matrix X have covariance matrix Σ_x with respectively Φ and Λ eigenvector and eigenvalue matrices, as described in equation (2.8). It is a proven result that the set of m ($m \leq n$) eigenvectors of Σ_x, which corresponds to the m largest eigenvalues, minimises the mean square reconstruction error over all choices of m orthonormal basis vectors [Fuk90]. Such a set of eigenvectors that defines a new uncorrelated coordinate system for the training set matrix X is known as the principal components. In the context of face recognition, those components are frequently called eigenfaces [TP91].

Therefore, although n variables are required to reproduce the total variability (or information) of the sample X, much of this variability can be accounted for by a smaller number m of principal components [JW98]. That is, the m principal components can then replace the initial n variables and the original data set, consisting of N measurements on n variables, is reduced to a data set consisting of N measurements on m principal components.

There is always the question of how many m principal components to retain in order to reduce the dimensionality of the original sample X. Unfortunately, there is no definitive answer to this question. As pointed out by Fukunaga [Fuk90], one useful property of such a linear transformation to consider is the effectiveness of each principal component. In terms of representing the total information of X, this effectiveness is determined by the magnitude of its corresponding eigenvalue [Fuk90].

Although the absolute value of the eigenvalue does not give adequate information for selection, the ratio of the eigenvalue to the summation of all eigenvalues expresses the percentage of the mean square reconstruction error introduced by eliminating the corresponding eigenvector or principal component [Fuk90]. Thus, it is possible to use as a criterion for eliminating or retaining the *i*th principal component the following ratio

$$r_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} = \frac{\lambda_i}{\mathrm{tr}(\Sigma_x)} \le t \, , \tag{2.14}$$

where t is a threshold value such that $0 \le t \le 1$, and the notation ‘tr’ denotes the trace of a matrix. For example, if we choose $t = 0.1$ then every eigenvalue which explains 10% or less of the total variance is eliminated.

Since the criterion described in equation (2.14) for determining an appropriate number of principal components would give different results for different pattern recognition problems, we have carried out our image recognition tests by considering several numbers of principal components and selecting the best number of principal components experimentally, based on the classification accuracy.

2.6 The Statistical Decision Making Process

In the statistical pattern recognition approach, the decision-making process consists of assigning a given pattern with n pre-processed feature values $x = [x_1, x_2, ..., x_n]^T$ to one of g groups or classes $\pi_1, \pi_2, ..., \pi_g$ on the basis of a set of measurements or observations obtained for each pattern.

The measurements associated with the population of patterns belonging to the π_i class are assumed to have a distribution of values with probability conditioned on the pattern class (or probability density function). That is, a pattern vector x belonging to class π_i is viewed as an observation drawn randomly from the class-conditional probability function $p(x \mid \pi_i)$ [JDM00].

The misclassification risk is the usual measure of the decision-making process (or classifier) performance. This measure can be defined as the expected misclassification loss over the sample to be classified [Fri89]. In most pattern recognition problems, it is customary to consider as the loss function the 0/1 or symmetrical function, which assigns no loss to a correct decision and assigns a unit loss to any error [DHS01]. Therefore, we are assuming basically that all errors are equally costly and, consequently, the misclassification risk is then just the classification error rate, that is, the percentage of new patterns that are assigned to the wrong class. Thus, it is common to evaluate the decision-making process by seeking maximum classification accuracy or, equivalently, minimum error rate classification.

There are a number of decision rules available to define appropriate statistical decision-making boundaries. However, it is a proven result [And84] that the Bayes decision rule that assigns a pattern to the group with the highest conditional probability is the one that achieves minimal misclassification risk among all possible rules.

2.7 The Bayes Decision Rule

The idea behind the Bayes decision rule is that all of the information available about group membership is contained in the set of conditional (or posterior) probabilities.

In the case of a 0/1 or symmetrical loss function, the Bayes decision rule for minimising the risk can be formally stated as follows: Assign input pattern x to class π_i if

$$P(\pi_i \mid x) > P(\pi_j \mid x)\,, \qquad \textbf{(2.15)}$$

for all $j \neq i$ and $i, j = 1,2,\ldots,g$ groups. If there is more than one group with the largest conditional probability then the tie may be broken by allocating the object randomly to one of the tied groups [Jam85].

Although quantities such as $P(\pi_i \mid x)$ are difficult to find by standard methods of estimation, this is not the case, however, for quantities such as $p(x \mid \pi_i)$. The probability of getting a particular set of measurements x given that the object comes from class π_i, that is the class-conditional probability $p(x \mid \pi_i)$ or likelihood information, can be estimated simply by taking a sample of patterns from class π_i.

Fortunately there is a connection between $P(\pi_i \mid x)$ and $p(x \mid \pi_i)$ known as the Bayes theorem [Jam85]:

$$P(\pi_i \mid x) = \frac{p(x \mid \pi_i) p(\pi_i)}{\sum_{\text{all k}} p(x \mid \pi_k) p(\pi_k)}, \tag{2.16}$$

where $k = 1,2,\ldots,g$ groups. As we can note, all the items on the right-hand side of the equation (2.16) are measurable quantities and so can be found by sampling. The item $p(\pi_i)$ defined as the prior probability is simply the probability that the pattern comes from class π_i in the absence of any information, i.e. it is the proportion of class π_i in the population.

Use of Bayes theorem as described in equation (2.16) with the previous Bayes rule described in (2.15), gives the following decision rule: Assign input pattern x to class π_i if

$$\frac{p(x \mid \pi_i) p(\pi_i)}{\sum_{\text{all k}} p(x \mid \pi_k) p(\pi_k)} > \frac{p(x \mid \pi_j) p(\pi_j)}{\sum_{\text{all k}} p(x \mid \pi_k) p(\pi_k)}, \tag{2.17}$$

for all $j \neq i$ and $k = 1,2,\ldots,g$. As on both sides of the inequality the denominators are equal, Bayes rule can be conveniently written as follows: Assign pattern x to class π_i if

$$p(x \mid \pi_i) p(\pi_i) = \max_{1 \le j \le g} p(x \mid \pi_j) p(\pi_j). \tag{2.18}$$

The classification rule defined in (2.18) is the final practical form of the optimal Bayes decision rule. This practical form of Bayes decision rule is also called the maximum a posteriori rule.

2.8 Parametric and Non-Parametric Statistical Methods

Several methods have been utilized to design a statistical pattern recognition classifier. Strategies for choosing the most appropriate method basically depend on the type and the amount of information available about the class-conditional probability densities.

The optimal Bayes rule discussed in the previous section can be used to design a classifier when all of the class-conditional densities are specified. In practice, however, the true class-conditional densities are typically not known and must be estimated from the available samples or training sets. If at least the form of the class-conditional densities is known (e.g. multivariate Gaussian distributions) but some of the parameters of these densities (e.g. mean vectors and covariance matrices) are unknown, then this problem is defined as a parametric decision problem. A common strategy to tackle this problem is to replace the unknown parameters in the density functions by their respective estimated values calculated from the training sets. This strategy is often referred to as the Bayes plug-in classifier, which will be described in the next chapter.

When the form of the class-conditional densities is either not known or assumed, non-parametric models have to be considered. In non-parametric problems, either the density functions must be estimated by using kernel functions or the class decision boundary has to be directly constructed based on the available training samples. These ideas form respectively the bases of the two most common non-parametric models: the Parzen Window classifier and the *k*-nearest neighbour (k-NN) classifier. The Parzen Window classifier will be the topic of discussion in Chapter 6.

As mentioned previously, another subtle point in choosing a convenient statistical pattern method is related to the amount of information available. Although intuitively it could be expected that an increase in the number of features or dimensionality (more details) of a particular problem would lead to a better classification performance, it has been found in practice that beyond a certain point and when no additional information (more samples) is available, exactly the opposite occurs. This apparently paradoxical and well-known behaviour is commonly called the curse of dimensionality or peaking phenomena [Fuk90, Bis97, JDM00]. For fixed and limited sample sizes, as the number of features increases, the reliability of the parameter estimates decreases, degrading the corresponding classifier performance.

In addition, when a classifier is designed using a finite number of training samples, the expected probability of error is greater than if an infinite number of training samples were available. It is reasonable to expect that the probability of error decreases as more training samples are added and this behaviour obviously depends on the complexity of the classifier used. Raudys and Jain [RJ91] found that the additional error due to finite training sample size decreases more quickly for parametric classifiers than for non-parametric ones. As a result, non-parametric methods rely on more densely populated feature spaces for reliable classification than parametric ones.

2.9 Notation

Table 2.1 presents the notations for the most commonly occurring quantities used in this work. Other variables or functions, which are not listed below, have their specific definitions and usages clarified in the text.

Symbol	Description
n	Number of features or dimensionality
g	Number of total classes or groups
N	Number of total samples or observations
N_i	Number of class i samples
π_i	Class i
$x = [x_1, x_2, ..., x_n]^T$	Vector of features
μ_i	True mean vector of π_i
Σ_i	True covariance matrix of π_i
$p(\pi_i)$	A priori probability of π_i
$p(x \mid \pi_i)$	A class conditional probability of x
$P(\pi_i \mid x)$	A posteriori probability of π_i given x

Table 2.1. Notation used throughout this work.

Chapter 3

The Bayes Plug-in Classifier

The Bayes plug-in classifier is one of the most common parametric methods applied to statistical pattern recognition systems. This classifier is based on similarity measures that involve the inverse of the true covariance matrix of each class or group. Since in practical cases these matrices are not known, estimates must be computed based on patterns available in a training set. The usual choice for estimating the true covariance matrix is the maximum likelihood estimator defined by its corresponding sample group covariance matrix. However, in limited sample size applications the sample group covariance estimates become highly variable or even not invertible. Thus, a considerable amount of effort has been devoted to the design of other non-conventional Bayes plug-in classifiers, for use in limited sample and high dimensional problems. This chapter presents the basic concepts of the Bayes plug-in classifier and reviews its most important implementations.

3.1 The Conventional Bayes Plug-in Classifier

The Bayes plug-in classifier, also called the Gaussian maximum likelihood classifier, is based on the n-multivariate normal or Gaussian class-conditional probability densities

$$p(x \mid \pi_i) \equiv f_i(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{n/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right], \qquad \textbf{(3.1)}$$

where μ_i and Σ_i are the true class π_i population mean vector and covariance matrix, and n is the dimension of the pattern vector x. The notation "| |" denotes the determinant of a matrix. The class-conditional probability densities $f_i(x \mid \mu_i, \Sigma_i)$ defined in (3.1) are also known as the likelihood density functions.

Substituting equation (3.1) into the maximum a posteriori rule (2.18) defined in the previous chapter, leads to the following Bayes classification rule form: Assign pattern x to class π_i if

$$f_i(x \mid \mu_i, \Sigma_i) p(\pi_i) = \max_{1 \le j \le g} f_j(x \mid \mu_j, \Sigma_j) p(\pi_j), \tag{3.2}$$

where, as a reminder, $p(\pi_i)$ is the prior probability associated with the *i*th group and g is the number of groups or classes. Another way to specify equation (3.2) is to take the natural logarithms of the quantities involved, such as

$$\begin{aligned} d_i(x) &= \ln\left[f_i(x \mid \mu_i, \Sigma_i) p(\pi_i)\right] \\ &= \ln\left[\frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] p(\pi_i)\right] \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln(p(\pi_i)) \end{aligned} \tag{3.3}$$

where $d_i(x)$ is often called the quadratic discriminant score for the *i*th class. Since the constant $(n/2)\ln(2\pi)$ is the same for all classes, it can be ignored. Therefore, the optimal Bayes classification rule defined in equation (3.2) may be simplified even further to: Assign pattern x to class π_i if

$$\begin{aligned} d_i(x) &= \max_{1 \le j \le g}[-\frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) + \ln(p(\pi_j))] \\ &= \min_{1 \le j \le g}[\ln|\Sigma_j| + (x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) - 2\ln(p(\pi_j))] \\ &= \min_{1 \le j \le g} d_j^*(x) \end{aligned} \tag{3.4}$$

The Bayes classification rule specified in (3.4) is known as the quadratic discriminant rule (QD). Also the measure $d_j^*(x)$ without the prior probability information $p(\pi_j)$ is sometimes referred to as the generalized distance between x and μ_j. The first term is related to the generalized variance of the *j*th group and the second term is the Mahalanobis distance between x and the mean vector for the *j*th group.

In practice, however, the true values of the mean and covariance matrix, i.e. μ_i and Σ_i, are seldom known and must be replaced by their respective estimates calculated from the training samples available- this is where the term “plug-in” of the Bayes plug-in classifier takes its name. The mean is estimated by the usual sample mean $\bar{x}_i$ which is the maximum likelihood estimator of μ_i [And84], that is

$$\mu_i \equiv \bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{i,j} \ , \qquad \textbf{(3.5)}$$

where $x_{i,j}$ is observation *j* from class π_i, and N_i is the number of training observations from class π_i. The covariance matrix is commonly estimated by the sample group covariance matrix S_i which is the unbiased maximum likelihood estimator of Σ_i [And84], that is

$$\Sigma_i \equiv S_i = \frac{1}{(N_i - 1)}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T \ . \qquad \textbf{(3.6)}$$

In addition, on the assumption that the data is not only normally distributed but also statistically independent, the sample mean and the sample group covariance matrix estimates have the property of maximising the joint likelihood of the training observations [JW98]. In other words, the maximum likelihood estimates maximise the product of the marginal normal density functions:

$$(\bar{x}_i, S_i) = \underset{\mu_i,\Sigma_i}{\arg\max} \prod_{j=1}^{N_i} f_i(x_{i,j} \mid \mu_i, \Sigma_i) \ . \qquad \textbf{(3.7)}$$

After replacing (“pluging-in”) the true values of the mean and covariance matrix in (3.4) by their respective estimates, the Bayes rule can be rewritten as: Assign pattern *x* to class *i* that *minimises*:

$$d_i(x) = \ln|S_i| + (x - \bar{x}_i)^T S_i^{-1}(x - \bar{x}_i) - 2\ln(p(\pi_i)) \ . \qquad \textbf{(3.8)}$$

The rule described in equation (3.8) is also known as the standard or conventional quadratic discriminant function (QDF) classifier.

3.2 Limited Sample Size Problem

According to the previous section, the similarity measures of the Bayes plug-in classifier use the inverse of the true covariance matrices. Since in practice these matrices are not known, estimates must be computed based on patterns or observations available in a training set.

Although $\bar{x}_i$ and S_i are maximum likelihood estimators of μ_i and Σ_i, the misclassification rate calculated from equation (3.8) approaches the optimal rate obtained by equation (3.4) only when the sample sizes in the training set approach infinity [And84]. In fact, the performance of (3.8) can be seriously degraded if there are only limited samples owing to the instability of $\bar{x}_i$ and most significantly of S_i. For instance, the use of S_i is especially problematic if for *n*-dimensional patterns less than n + 1 training observations from each class π_i are available. Since the sample group covariance matrix is a function of $(N_i - 1)$ or fewer linearly independent vectors (see equation (3.6)), its rank is $(N_i - 1)$ or less. Therefore, S_i is a singular matrix if N_i is less than the dimension of the feature space. As a general guideline, Jain and Chandrasekaran [JC82] have suggested that the class sample sizes N_i should be at least five to ten times the dimension of the feature space n.

The effect of the sample group covariance instability on the conventional QDF classifier can be explicitly seen by representing those matrices in their spectral decomposition forms [Fuk90]

$$S_i = \Phi_i \Lambda_i {\Phi_i}^T = \sum_{k=1}^{n} \lambda_{ik} \phi_{ik} \phi_{ik}^T \, , \qquad \textbf{(3.9)}$$

where λ_{ik} is the *k*th eigenvalue of S_i and ϕ_{ik} is the corresponding eigenvector. According to this representation, the inverse of the sample group covariance matrix is

$$\begin{aligned} S_i^{-1} &= (\Phi_i \Lambda_i \Phi_i^T)^{-1} \\ &= (\Phi_i^T)^{-1} (\Lambda_i)^{-1} (\Phi_i)^{-1} \\ &= \Phi_i \Lambda_i^{-1} \Phi_i^T \\ &= \sum_{k=1}^{n} \frac{\phi_{ik} \phi_{ik}^T}{\lambda_{ik}} , \end{aligned} \qquad \textbf{(3.10)}$$

where the following property described in the previous chapter for orthogonal matrices is used: $\Phi^{-1} = \Phi^T$. Since the determinant of a matrix is equal to the product of all its eigenvalues [Str88], substituting equation (3.10) into equation (3.8) gives the spectral decomposition form of the conventional Bayes plug-in classifier, as follows:

$$\begin{aligned} d_i(x) &= \ln|S_i| + (x-\bar{x}_i)^T S_i^{-1}(x-\bar{x}_i) - 2\ln(p(\pi_i)) \\ &= \ln(\prod_{k=1}^{n}\lambda_{ik}) + (x-\bar{x}_i)^T \left[\sum_{k=1}^{n}\frac{\phi_{ik}\phi_{ik}^T}{\lambda_{ik}}\right](x-\bar{x}_i) - 2\ln(p(\pi_i)) \\ &= \sum_{k=1}^{n}\ln\lambda_{ik} + \sum_{k=1}^{n}\frac{[(x-\bar{x}_i)^T\phi_{ik}][\phi_{ik}^T(x-\bar{x}_i)]}{\lambda_{ik}} - 2\ln(p(\pi_i)) \\ &= \sum_{k=1}^{n}\ln\lambda_{ik} + \sum_{k=1}^{n}\frac{[\phi_{ik}^T(x-\bar{x}_i)]^2}{\lambda_{ik}} - 2\ln(p(\pi_i)), \end{aligned} \qquad \textbf{(3.11)}$$

where the following relationship is used: for any product of a 1 by n vector y_1^T (the row vector $(x-\bar{x}_i)^T$) with a n by 1 vector y_2 (the column vector ϕ_{ik}) $y_1^T y_2 = y_2^T y_1$ [Str88].

As can be observed, the discriminant score in equation (3.11) is heavily weighted by the smallest eigenvalues and the directions associated with their eigenvectors [Fri89]. Therefore, a poor or unreliable estimation of the sample group covariance matrices tends to exaggerate the importance associated with the low-variance information and consequently distorts the quadratic discriminant analysis.

Another problem related to the first two terms of equation (3.11) is the upward bias of the large sample group covariance eigenvalues and downward bias of the smaller ones. When the sample size decreases the estimates based on the maximum likelihood equation (3.6) produce biased estimates of the corresponding eigenvalues, that is, the largest eigenvalues are larger than the eigenvalues of the true covariance and the smallest ones are biased toward lower values. This effect is most pronounced when the true eigenvalues tend to be equal rather than highly different [Fri89]. In fact, when the sample covariance matrix is singular the smallest ($n - N_i + 1$) eigenvalues are estimated to be 0 and the corresponding eigenvectors are arbitrary, though constrained by the orthogonality assumption.

In the past, several investigators [Haf79, Haf80, DS85, LP85] demonstrated that Stein-like biased estimators, which basically shrink or expand the sample eigenvalues depending on their magnitude, dominate the sample covariance matrix under a variety of loss functions. Moreover, in [EM76] an estimate for the inverse of the sample covariance matrix that shrinks the eigenvalues of the sample group covariance matrix toward a common value was developed. In these works, the problem of estimating the covariance matrix S_i was based on its distribution, often called the Wishart distribution[1] [And84]. Since the probability density function of a Wishart distribution does not exist unless the sample size N_i is greater than the number of parameters n [JW98], in all of these eigenvalues shrinkage methods quoted the sample covariance matrix S_i must be non-singular. This constraint has been shown quite restrictive in practice.

3.3 Linear Discriminant Classifier

One straightforward method routinely applied to overcome the limited sample size problem and consequently deal with the singularity and instability of the sample group covariance matrices S_i is to employ the so-called linear discriminant function (LDF) classifier.

The LDF classifier can be obtained by replacing the S_i in (3.8) with the pooled sample covariance matrix defined as

$$S_p = \frac{1}{N-g}\sum_{i=1}^{g}(N_i - 1)S_i = \frac{(N_1-1)S_1 + (N_2-1)S_2 + \cdots + (N_g-1)S_g}{N-g}, \qquad \textbf{(3.12)}$$

where $N = N_1 + N_2 + \cdots + N_g$. Since more observations are taken to calculate the pooled covariance matrix, S_p is indeed a weighted average of all the S_i, S_p will potentially have a higher rank than S_i and would normally be a full rank matrix.

Although, theoretically, S_p is a consistent estimator of the true covariance matrices Σ_i only when $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$, studies have shown that the LDF classifier is not only simple and easy to use but also works particularly well in small sample size situations [MD74, WK77]. In such situations, the LDF classifier can outperform the conventional QDF even though the true covariance matrix of each group is known to differ.

In the classification context, as long as the covariance matrices are not quite dissimilar and small sample sizes are available, the ellipsoidal symmetry associated with the normal distribution seems to be the relevant aspect to consider rather than its detailed shape [Lac75, Jam85]. In fact, it has been argued [Jam85] that in regions where the groups are most represented, the linear discriminant functions would be close to the quadratic ones, giving similar results for observations that are located near the sample mean of each

[1] The Wishart distribution is a multivariate generalization of the gamma distribution. It is originally derived as the sampling distribution of the sample group covariance matrices.

group. These regions are obviously where the majority of the cases to be classified supposedly occur. Furthermore, it is known that the QDF classifier often requires larger samples than those based on the LDF classifier and seems to be more sensitive to the violation of the normal distribution assumption. These practical advantages have made the LDF classifier one of the most popular methods of classification.

3.4 Unconventional Bayes Plug-in Classifiers

As discussed previously in this chapter, a critical issue for the Bayes plug-in classifier is the instability and singularity of the sample group covariance estimates. Hence, a considerable amount of effort has been devoted to the design of other unconventional Bayes plug-in classifiers, for use in limited sample and high dimensional problems. By "unconventional Bayes Plug-in classifiers" we mean any quadratic classifier that is not based solely on the data via the sample group covariance estimate. In the following subsections, most of these approaches that essentially bias the sample group covariance estimates towards non-singular matrices are described.

3.4.1 Regularised Discriminant Analysis Method

Regularisation methods have been used successfully in solving poorly and ill-posed inverse problems [OSu86].

An estimation problem can be defined as a poorly posed problem when the number of patterns or observations available (training set) is not considerably larger than the number of parameters (dimension of the feature space) to be estimated and ill-posed if this number of parameters exceeds the training sample size. As a result, such parameter estimates become highly variable owing to limited training set size.

Regularization methods attempt to reduce the variability of poorly and ill-posed estimates by biasing them toward values that are considered to be more "physically plausible" [Fri89]. The idea behind the term "regularization" is to decrease the variance associated with the limited sample based estimate at the expense of potentially increased bias. The extent of this variance-bias trade-off is controlled by one or more regularization parameters [Fri89].

Friedman [Fri89] has proposed one of the most important regularization procedures for QDF classifiers called "Regularised Discriminant Analysis" (RDA) classifier. RDA is an alternative to the usual Bayes plug-in classifier and can be viewed as an intermediate classifier between the LDF and QDF classifiers.

Friedman's RDA approach is basically a two-dimensional optimisation method that shrinks both the S_i towards S_p and also the eigenvalues of the S_i towards equality by blending the first shrinkage with multiples of the identity matrix. In this context, the sample covariance matrices S_i of the discriminant rule defined in (3.8) are replaced by the following $S_i^{rda}(\lambda,\gamma)$

$$\begin{aligned} S_i^{rda}(\lambda,\gamma) &= (1-\gamma)S_i^{rda}(\lambda) + \gamma\left(\frac{tr(S_i^{rda}(\lambda))}{n}\right)I, \\ S_i^{rda}(\lambda) &= \frac{(1-\lambda)(N_i-1)S_i + \lambda(N-g)S_p}{(1-\lambda)N_i + \lambda N} \end{aligned} \quad , \qquad \textbf{(3.13)}$$

where the notation "tr" denotes the trace of a matrix, that is, the sum of all its eigenvalues [Str88]. Thus the regularization parameter λ controls the degree of shrinkage of the sample group covariance matrix estimates toward the pooled covariance estimate, while the parameter γ controls the shrinkage toward a multiple of the identity matrix. Since the multiplier $tr(S_i^{rda}(\lambda))/n$ is just the average eigenvalue of $S_i^{rda}(\lambda)$, the shrinkage parameter γ has the effect of decreasing the larger eigenvalues and increasing the smaller ones [Fri89]. This effect counteracts the aforementioned upward and downward biases of the sample group covariance estimates and favours true covariance matrices that are some multiples of the identity matrix. In fact, the RDA method provides a number of regularization alternatives. Holding the mixing parameter γ at 0 and varying λ yields classification models between LDF and QDF classifiers. Holding λ at 0 and increasing γ attempts to unbias the sample-based eigenvalue estimates while holding λ at 1 and varying γ gives rise to ridge-like estimates of the pooled covariance matrix [DPi77, Cam80].

The mixing parameters λ and γ are restricted to the range 0 to 1 (optimisation grid) and are selected to maximise the leave-one-out classification accuracy based on the discriminant rule defined in (3.8). In other words, the following classification rule is developed on the $N-1$ training observations exclusive of a particular observation $x_{i,v}$ and then used to classify $x_{i,v}$: Choose class *k* such that

$$d_k(x_{i,v}) = \min_{1 \le j \le g} d_j(x_{i,v}) \text{, with}$$

$$d_j(x_{i,v}) = \ln\left|S_{j/v}^{rda}(\lambda,\gamma)\right| + (x_{i,v} - \bar{x}_{j/v})^T \left(S_{j/v}^{rda}(\lambda,\gamma)\right)^{-1} (x_{i,v} - \bar{x}_{j/v}) - 2\ln(p(\pi_j)) \tag{3.14}$$

where the notation */v* represents the corresponding quantity with observation $x_{i,v}$ removed. Each of the training observations is in turn held out and then classified in this manner. The resulting misclassification loss, i.e. the number of cases in which the observation left out is allocated to the wrong class, averaged over all the training observations is then used to choose the best grid-pair (λ,γ).

Although Friedman's RDA method is theoretically a well-established approach and has the benefit of being directly related to classification accuracy, it has practical drawbacks. RDA is indeed a computationally intensive method. For each point on the two-dimensional optimisation grid, RDA requires the evaluation of the proposed estimates of every class. In situations where the optimisation has to be done over a fine grid and a large number of g groups is considered, for instance g is a number of order 10^2, the RDA becomes unfeasible. Also, despite the substantial amount of computation saved by taking advantage of matrix updating formulas based on the Sherman-Morrison-Woodbury formula [GL89], which we will discuss in detail in chapter 4, RDA requires the computation of the eigenvalues and eigenvectors for a (n by n) matrix for each value of the mixing parameter λ.

In addition to the computational limitation, Greene and Rayens [GR91] have observed that RDA has the disadvantage of partially ignoring information from a considerable portion of the data in the selection of the mixing parameters λ and γ - the same error rates could take place over a wide range of parameter values - and the optimal values of the grid-pair (λ,γ) are not unique. Therefore, a tie-breaking method needs to be applied. Finally, as RDA maximises the classification accuracy calculating all covariance estimates simultaneously, it is restricted to using the same value of the mixing parameters for all the classes. These same values may not be optimal for all classes.

3.4.2 Empirical Bayesian Method

The Bayes plug-in classifier is completely specified by both the true mean μ_i and the true covariance Σ_i. In this sense, the statistical approach can be viewed as a problem of estimating properly these parameters on the basis of a sample of patterns or observations.

Several authors [EM76, Haf79, Haf80, And84] have observed that the method of unbiased maximum likelihood estimation for the true covariance matrices may be improved by other considerations, such as empirical Bayes estimators, which are better with respect to certain squared-error loss functions.

Following this idea, Greene and Rayens [GR89] have developed the empirical Bayes covariance estimator which provides a theoretical basis for controlling the shrinkage of the S_i in such a way that the amount of shrinkage is related to the number of training samples N_i and the estimated concentration of the true covariance matrices Σ_i.

According to Anderson [And84], the assumption of normal *n*-dimensional observations implies that, conditionally on the Σ_i, the sample group covariance matrices S_i are mutually independent with Wishart distribution:

$$f_i S_i \sim W_n(\Sigma_i, f_i), \qquad \textbf{(3.15)}$$

where $f_i = N_i - 1$, and $W_n(\Sigma_i, f_i)$ denotes the central Wishart distribution with f_i degrees of freedom and parameter matrix Σ_i. The family of inverted Wishart distributions for Σ_i is conjugate to the family of Wishart distributions, i.e. when there is a sufficient statistic (sample mean or sample covariance matrix) there will exist a family of prior distributions for the estimated parameter such that the posterior distribution is a member of this family. Then the family of inverted Wishart distributions provides a convenient family of prior distributions for true covariance Σ_i [And84].

Assuming that Σ_i are mutually independent, Greene and Rayens [GR89] have proposed that Σ_i have the following prior distribution

$$\Sigma_i \sim W_n^{-1}((m-n-1)\Psi, m), \qquad \textbf{(3.16)}$$

where $m > n+1$, and W_n^{-1} is the inverted Wishart distribution with parameters Ψ and m. The parameter Ψ represents the central location of the prior distribution of Σ_i, i.e. Ψ is the expected value or prior mean of Σ_i, and *m* represents the degree of concentration of the Σ_i around Ψ. Under equations (3.15) and (3.16), and following the inverted Wishart distribution's properties [And84], the posterior distribution of Σ_i given $S_1, S_2, \ldots, S_g$ is

$$W_n^{-1}(f_i S_i + (m-n-1)\Psi, f_i + m), \qquad \textbf{(3.17)}$$

and the Bayes estimator of Σ_i (posterior mean), under squared error loss, is

$$\hat{\Sigma}_i(\Psi, m) = \frac{f_i}{f_i + m - n - 1} S_i + \frac{m-n-1}{f_i + m - n - 1}\Psi \qquad \textbf{(3.18)}$$

Equation (3.18) shows that $\hat{\Sigma}_i(\Psi, m)$ approaches S_i as the degrees of freedom $f_i \to \infty$ and approaches Ψ as the concentration parameter $m \to \infty$. Also the Bayes estimator $\hat{\Sigma}_i(\Psi, m)$ can be simply written as a weighted average between S_i and Ψ given by

$$\hat{\Sigma}_i(\Psi, w_i) = (1 - w_i) S_i + w_i \Psi, \qquad \text{where } w_i = \frac{m-n-1}{f_i + m - n - 1}. \qquad \textbf{(3.19)}$$

As pointed out by Greene and Rayens [GR89], it is important to observe that equation (3.19) is intuitively plausible regarding the similarity of the Σ_i. For instance, when there are large training samples and/or large variation among the Σ_i, that is, f_i is high and/or $m \to n+1$, the shrinking parameter w_i will be small and consequently $\hat{\Sigma}_i(\Psi, w_i) \approx S_i$. Analogously, when there are small training samples and/or similar Σ_i, i.e. f_i is low and/or $m >> n+1$, w_i will be large and so $\hat{\Sigma}_i(\Psi, w_i) \approx \Psi$.

In order to estimate the Bayesian parameters Ψ and m of equation (3.18), Greene and Rayens have employed the following empirical approaches. For a given m, it is possible to compute the generalised least squares estimator of Ψ, designated as $S_p^*(m)$, as:

$$S_p^*(m) = \left[\sum_{i=1}^{g} \frac{f_i}{f_i + m - n - 1}\right]^{-1} \sum_{i=1}^{g} \frac{f_i}{f_i + m - n - 1} S_i. \qquad \textbf{(3.20)}$$

Note that when the number of training observations in each class is equal, that is $f_1 = f_2 = \cdots = f_g$, the least squares estimator of Ψ does not depend on the concentration parameter m and $S_p^*(m) = S_p = S$, where

$$S = \frac{1}{g}\sum_{i=1}^{g} S_i \tag{3.21}$$

is called the unweighted common covariance estimate of the sample group covariance matrices. Substituting $S_p^*(m)$ for Ψ in equation (3.18) gives the following empirical Bayes estimate of Σ_i for known m:

$$\hat{\Sigma}_i^*(m) = \frac{f_i}{f_i + m - n - 1} S_i + \frac{m - n - 1}{f_i + m - n - 1} S_p^*(m) . \tag{3.22}$$

For the remaining concentration parameter m, Greene and Rayens have suggested three different estimators, where two of them depend on probabilistic assumptions concerning Σ_i and the third one is based on the generalized distances between the training set observations and the mean vectors of the corresponding classes. In this latter criterion the parameter m has been expressed in terms of the shrinking parameter w_i as

$$m = \frac{w_i(f_i - n - 1) + n + 1}{1 - w_i} , \tag{3.23}$$

and a grid of values for m is given by $0 < w_i < 1$. Analogous to Friedman's RDA approach, the empirical Bayes estimate of Σ_i has been again modified by shrinking the eigenvalues of $\Sigma_i^*(m)$ towards equality to form the estimator

$$\hat{\Sigma}_i^*(m,\gamma) = (1-\gamma)\hat{\Sigma}_i^*(m) + \gamma\left(\frac{tr(\hat{\Sigma}_i^*(m))}{n}\right)I, \tag{3.24}$$

where $0 \le \gamma \le 1$. The two mixing parameters m and γ should be chosen to minimise over a grid of candidate values (m,γ) the following leave-one-out generalized distance

$$\hat{D}(m,\gamma) = \sum_{i=1}^{g}\sum_{v=1}^{N_i}\left[\ln\left|\hat{\Sigma}_{i/v}^*(m,\gamma)\right| + (x_{i,v} - \bar{x}_{i/v})^T(\hat{\Sigma}_{i/v}^*(m,\gamma))^{-1}(x_{i,v} - \bar{x}_{i/v})\right], \tag{3.25}$$

where again the notation /v represents the corresponding quantity with observation $x_{i,v}$ left out. According to Greene and Rayens results, this leave-one-out generalized distance criterion performs at least as well as the first two other probabilistic estimators of m while requiring fewer distribution assumptions for their justification.

In their research, Greene and Rayens have also enhanced the important result related to the optimisation of the leave-one-out discriminant distance. They have observed that minimising the leave-one-out generalized distance of each training observations group is indeed equivalent to minimising the Kullback-Leibler (KL) distance measure of each class [Sil86, GR91]. The KL distance is often called *relative entropy* [Hay99] and measures basically the divergence (distance) between two density functions. In this case, it can be defined as the following:

$$KL^i(m) = \int f_i(x) \log\left(\frac{f_i(x)}{\hat{f}_i^m(x)} \right) dx \qquad \mathbf{(3.26)}$$

where $\hat{f}_i^m(\cdot)$ denotes the *n*-multivariate normal density function with mean vector $\bar{x}_i$ and covariance matrix $\hat{\Sigma}_i^*(m,\gamma)$, and $f_i(\cdot)$ denotes the true density of the *i*th group which may or may not be a multivariate normal.

Although Greene and Rayens method of biasing S_i towards the pooled covariance and identity matrix is similar to Friedman's RDA approach, their respective parameter optimisations are different. The empirical Bayes estimator does not optimise its parameters with respect to classification accuracy, but to a loss function based on the generalized distance between the training set observations and their respective group means. Therefore, the Bayes estimator is able to circumvent the RDA non-unique "optimal" problem related to ignoring a considerable portion of the data in the selection of the corresponding mixing parameters. However, although Greene and Rayens have also proposed a convenient rank-one updating algorithm based on the Sherman-Morrison-Woodbury formula [GL89], the computational issues involving the empirical Bayes covariance approach are as severe as in Friedman's RDA approach. In addition, both methods are restricted to using the same values of mixing parameters for all classes, which may not be optimal especially for Greene and Rayens method, which is based on the optimisation of a maximum likelihood generalized distance.

3.4.3 Leave-One-Out Likelihood Method

The RDA [Fri89] and the empirical Bayes [GR89, GR91] methods described in the previous sub-sections use the leave-one-out procedure to optimise their respective mixing parameters under different loss functions. Since both loss functions depend on calculating all covariance estimates simultaneously, Friedman's as well as Greene and Rayens' approaches must employ the same mixing parameters for all classes. In practice, however, it is common to have classes with different forms and, consequently, different covariance matrices. In such situations, it seems appropriate to allow these covariance matrices to be estimated by distinct mixing parameters.

Hoffbeck [Hof95] has proposed a leave-one-out covariance estimator (LOOC) that depends only on covariance estimates of single classes. In LOOC each covariance estimate is optimised independently and a separate mixing parameter is computed for each class based on the corresponding likelihood information. The idea is to examine pair-wise mixtures of the sample group covariance estimates S_i and the unweighted common covariance estimate S (defined in equation (3.21)), together with their diagonal forms.

The LOOC estimator has the following form:

$$S_i^{looc}(\alpha_i) = \begin{cases} (1-\alpha_i)\text{diag}(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\text{diag}(S) & 2 < \alpha_i \le 3 \end{cases}, \qquad \textbf{(3.27)}$$

where the mixing or shrinkage parameter α_i determines which covariance estimate or mixture of covariance estimates is selected. That is: if $\alpha_i = 0$ then the diagonal of sample covariance is used; if $\alpha_i = 1$ the sample covariance is returned; if $\alpha_i = 2$ the common covariance is selected; and if $\alpha_i = 3$ the diagonal form of the common covariance is considered. Other values of α_i lead to mixtures of two of the aforementioned estimates [Hof95].

In order to select the appropriate mixing parameter α_i, the leave-one-out likelihood (LOOL) parameter estimation has been considered. In the LOOL technique [Fuk90], one training observation of the *i*th class training set is removed and the sample mean and sample group covariance are estimated from the remaining $N_i - 1$ samples. Then the likelihood of the excluded sample is calculated given the previous mean and covariance estimates. This operation is repeated $N_i - 1$ times and the average log likelihood is computed over all the N_i observations. Hoffbeck's strategy is to evaluate several values of α_i over the optimisation grid $0 \le \alpha_i \le 3$, and then choose the α_i that maximizes the average log likelihood of the corresponding *n*-multivariate normal density function, computed as follows:

$$
\begin{aligned}
LOOL_i(\alpha_i) &= \frac{1}{N_i}\sum_{v=1}^{N_i}\left[f\left(x_{i,v}\,|\,\bar{x}_{i/v}, S_{i/v}^{looc}(\alpha_i)\right)\right] \\
&= \frac{1}{N_i}\sum_{v=1}^{N_i}\left[-\ln\left|S_{i/v}^{looc}(\alpha_i)\right| - \frac{1}{2}(x_{i,v}-\bar{x}_{i/v})^T\left(S_{i/v}^{looc}(\alpha_i)\right)^{-1}(x_{i,v}-\bar{x}_{i/v})\right],
\end{aligned}
\qquad \mathbf{(3.28)}
$$

where the notation */v* represents the corresponding quantity with observation $x_{i,v}$ left out. Once the mixture parameter α_i is selected, the corresponding leave-one-out covariance estimate $S_i^{looc}(\alpha_i)$ is calculated using all the N_i training observations and substituted for S_i into the Bayes discriminant rule defined in (3.8) [Hof95].

As can be seen, the computation of the LOOC estimate requires only one density function be evaluated for each point on the α_i optimisation grid, but also involves calculating the inverse and determinant of the (n by n) matrix $S_i^{looc}(\alpha_i)$ for each training observation belonging to the *i*th class. Although this is a one-dimensional optimisation procedure for each sample group and consequently requires less computation, for instance, than the two-dimensional RDA estimator, LOOC is still computationally expensive. Hoffbeck has reduced significantly the LOOC required computation by considering valid approximations of the covariance estimates and using the Sherman-Morrison-Woodbury formula [GL89] to write the estimates in a form that allows the determinant and inverse of each corresponding class to be computed only once, followed by a relatively simple computation for each left out observation. The final form of the LOOC requires less computation than the RDA estimator.

LOOC differs from the similar covariance methods already described in the mixtures it considers and the optimisation index utilised to select the best estimator. Although both RDA and the empirical Bayes estimator make use of the sample covariance matrix, pooled covariance matrix and the identity matrix multiplied by a scalar, LOOC employs the sample covariance matrix, unweighted common covariance matrix and the diagonal forms of these matrices. In LOOC the optimisation search is one-dimensional and limited to pair-wise mixtures, while in RDA and the empirical Bayes estimator more general two-dimensional mixtures are considered. Moreover, the optimisation index maximised in LOOC is the leave-one-out group likelihood that allows a separate mixing parameter to be computed for each class. On the other hand, RDA and the empirical Bayes estimator use leave-one-out optimisation procedures based on all the training observations of all classes and are restricted to using the same mixing parameters for all classes.

Hoffbeck and Landgrebe have carried out several experiments with computer generated and remote sensing data to compare LOOC and RDA performances [Hof95, HL96]. In about half of these experiments, LOOC has led to higher classification accuracy than RDA and required less computation.

3.4.4 Bayesian Leave-One-Out Likelihood Method

In 1998, Tadjudin has proposed another covariance estimation method [Tad98] called Bayesian leave-one-out covariance estimation (bLOOC). This method is essentially an extension of the previous works RDA [Fri89], empirical Bayesian approach [GR89, GR91], and LOOC [Hof95, HL96].

Basically, Tadjudin has developed two covariance estimators. The first one (bLOOC1) intends to represent a wide variety of covariance matrices, including the RDA identity matrix multiplied by a scalar for estimating spherical structures, and has the following form [Tad98]:

$$S_i^{blooc1}(\alpha_i) = \begin{cases} (1-\alpha_i)\dfrac{tr(S_i)}{p}I + \alpha_i S_i & 0 \le \alpha_i < 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(m) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\dfrac{tr(S)}{n}I & 2 \le \alpha_i \le 3 \end{cases}, \qquad \textbf{(3.29)}$$

where S_i is the sample group covariance matrix, $S_p^*(m)$ is the Bayesian generalised least squares estimator defined in (3.20), and S is the unweighted common covariance estimate defined in (3.21). In addition, in order to consider group covariance matrices that are highly ellipsoidal, Tadjudin has proposed a second estimator (bLOOC2) defined as:

$$S_i^{blooc2}(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(S_i) + \alpha_i S_i & 0 \le \alpha_i < 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(m) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\mathrm{diag}(S) & 2 \le \alpha_i \le 3 \end{cases}. \qquad \textbf{(3.30)}$$

Note that bLOOC2 is quite similar to Hoffbeck LOOC estimator described in (3.27). In fact, when all classes have equal number of training samples, bLOOC2 has the same form as LOOC.

Analogously to Hoffbeck approach, Tadjudin has used the leave-one-out average likelihood to select the appropriate mixing parameter for both bLOOC1 and bLOOC2 estimators. In terms of computational costs, Tadjudin has shown that an efficient implementation of the two methods can be achieved again by using rank-one down-data based on the Sherman-Morrison-Woodbury formula [GL89]. The computational issues involved in bLOOC1 are much more severe than LOOC, but not as severe as Friedman's and Greene and Rayens' approaches. The bLOOC2 second estimator requires almost the same computation as LOOC and, consequently, less computation than RDA and the empirical Bayesian estimator.

In [Tad98, TL99], Tadjudin has presented various experimental results from computer generated and remote sensing data. These results essentially compared both bLOOC1 and bLOOC2 with LOOC. In addition, the effects of substituting the Bayesian covariance estimations for the pooled covariance matrix on the linear discriminant analysis (LDA) feature extraction technique have been discussed. Situations where the sample sizes are unequal and the training set size is large enough to reflect the true covariance matrices favour the estimators derived under the Bayesian framework. Moreover, the first Bayesian estimator bLOOC1 combined with LDA can achieve better performance when the total number of training samples N is less than the dimension of the feature space n. These results confirm that the ridge estimator gives rise to a better pooled covariance estimate by counteracting the upward and downward biases described previously. On the other hand, when the pooled covariance matrix is non-singular, the other estimator bLOOC2 should be used. Under these conditions, the proposed estimators perform better than the LOOC , LDF and QDF classifiers. No comparison results with Friedman's RDA or Greene and Rayens' empirical Bayesian approaches have been provided.

3.4.5 Simplified Quadratic Discriminant Function Method

The Simplified Quadratic Discriminant Function (SQDF) classifier has been proposed by Omachi et al. [OSA00] and can be viewed as an approximation method of the standard or conventional quadratic discriminant function (QDF) classifier.

The main idea of the SQDF approach is to divide the *n*-dimensional feature space into two subspaces: a primary subspace of dimension s ($s < n$) containing the largest eigenvalues of the sample group covariance matrices and a complementary subspace of dimension $n - s$ containing the smallest ones. All the largest eigenvalues (associated with the primary subspace) are estimated by using the actual eigenvalues calculated from the sample group covariance matrices, while the smallest eigenvalues (associated with the complementary subspace) are replaced by a constant determined by maximum likelihood estimation. Similar approaches of splitting the quadratic discriminant feature space into two subspaces were investigated by other researchers [Wol76, FF89].

In order to partition the *n*-dimensional feature space, the SQDF method approximates the spectral decomposition form of the QDF classifier described in (3.11) by the following function:

$$d_i(x) = \sum_{k=1}^{s} \ln \lambda_{ik} + \sum_{k=s+1}^{n} \ln \lambda + \sum_{k=1}^{s} \frac{[\phi_{ik}^T (x - \bar{x}_i)]^2}{\lambda_{ik}} + \sum_{k=s+1}^{n} \frac{[\phi_{ik}^T (x - \bar{x}_i)]^2}{\lambda} - 2\ln(p(\pi_i)) \qquad \textbf{(3.31)}$$

where $(\lambda_{ik}, \phi_{ik})$ are the *k*-th eigenvalue-eigenvector pair of S_i, λ is the simplification constant and $s < n$. In the case of $s = n$, SQDF is exactly the conventional QDF classifier described in (3.11).

Performing the maximum likelihood estimation on the $n - s$ complementary subspace, Omachi et al. [OSA00] have determined that the constant value λ of each group is defined as the mean value of the small $n - s$ eigenvalues of the corresponding sample group covariance matrix, that is

$$\lambda = \frac{1}{n - s} \sum_{k=s+1}^{n} \lambda_{ik} \,, \qquad \textbf{(3.32)}$$

where λ_{ik} correspond to the smallest eigenvalues of S_i. However, the procedure to determine the parameter s, that is, the number of reliable eigenvalues to preserve, is not straightforward. The parameter s should be defined arbitrarily, experimentally or by minimising information criteria such as Akaike's Information Criteria or Minimum Description Length [OSA00].

In [OSA00] Omachi et al. have presented experimental results from some computer generated data and character recognition of digits '0' and '1'. These results were compared only with the conventional QDF classifier and showed that the SQDF classifier reduces the computation cost and improves the classification accuracy in small sample settings. However, as SQDF approximates the standard QDF classification rule considering solely the information provided by each sample group covariance matrix, it seems to be more sensitive to poor sample covariance estimation than other unconventional QDF classifiers. In addition, SQDF addresses the conventional QDF problem when the sample group covariance matrices S_i are singular, but does not avoid the covariance estimation instability of S_i when these matrices are non-singular but poorly estimated.

3.5 Summary and Conclusions

In this chapter, the conventional Bayes Plug-in classifier, Linear Discriminant Function classifier and a number of non-conventional Bayes plug-in classifiers available in statistical pattern recognition have been reviewed with regard to the difficulties caused by limited sample size problems.

Several simulation experiments carried out by various researchers have shown that choosing an unconventional Bayes plug-in classifier, mostly either the linear and quadratic ones, improves the classification accuracy in settings for which sample sizes are limited and the number of parameters or features is large. In situations, however, where the class sample sizes are all very large compared with the number of features, and consequently the maximum likelihood estimators of the true covariance matrices are not ill-posed or poorly estimated, all aforementioned approaches obtain little benefit from the unconventional methods and sometimes display a small degradation in classification performance.

From computer-generated data, comparisons between classifiers like RDA [Fri89], the Empirical Bayes method [GR89], LOOC [Hof95] and the both bLOOC1 and bLOOC2 [Tad98] have been provided by previous studies of several researchers. In these comparisons, standard small samples considering equal/unequal spherical covariance matrices, equal/unequal highly ellipsoidal covariance matrices, as well as equal/unequal training set sizes have been analysed. In short, problems with spherical true covariance matrices frequently favour estimators that shrink the sample group covariance matrices towards the identity matrix, such as RDA, the Empirical Bayes method and bLOOC1. On the other hand, in cases where highly ellipsoidal forms for the true covariance matrices are used, LOOC and bLOOC2 commonly outperform the other estimators. Therefore, it has been generally accepted that different Bayes plug-in classifiers should be optimal depending not only on the true covariance statistics of each class, but also on the number of training observations, the dimension of the feature space and even the ellipsoidal symmetry associated with the normal multivariate distributions.

In the context of real data, the aforementioned unconventional classifiers have been evaluated in applications that involve mostly remote sensing data. In fact, several related results have been produced in this field during the last eight years [Hof95, HL96, Tad98, TL99, BL00]. According to these remote sensing experiments, one important point can be drawn. All the compared mixing parameters of RDA, LOOC, bLOOC1 and bLOOC2 have been optimised inside the range that requires solely blending the sample group covariance matrices towards the pooled estimate for almost all classes. In the case of recognition applications where the sources of variation are often the same from group to group, this suggests that when no information about the true covariance forms are available and when the total number of the training observations N is larger than the number of features n (so the pooled covariance matrix is non-singular), the sample group covariance matrices and the pooled covariance matrix may be reasonable covariance extremes for intermediate estimations, especially when concerns about computational costs exist.

Finally, the unconventional Bayes plug-in classifiers have been shown to improve the classification accuracy for limited training set recognition problems and small number of groups. These ideas have proved to be true in cases where no more than 20 groups are required, but have not been verified for a large number of groups. In this way, biometric image recognition problems, such as face recognition, which involve extremely small training sets, a large number of features and a large number of groups, are examples of promising applications for further research.

Chapter 4

The Sample Size Problem in Image Recognition

In image recognition applications, patterns are frequently composed of thousands of pixels or even hundreds of pre-processed image features, but the number of training examples per class is limited. In such situations, the conventional Bayes Plug-in or Quadratic Discriminant Function (QDF) classifier based on maximum likelihood covariance estimation either performs poorly or cannot be calculated when the group sample sizes are smaller than the number of features or parameters.

As described in the previous chapter, other unconventional QDF classifiers have been proposed in order to overcome the difficulties of estimating reliable covariance matrices in limited sample, high dimensional classification problems. However, most of these quadratic approaches rely on optimisation techniques that are time consuming and do not necessarily lead to the highest classification accuracy for all circumstances.

This chapter analyses the performance of the aforementioned Bayes Plug-in covariance estimators in image recognition problems that consider limited training sets, large number of features and a number of groups. Biometric image recognition problems, such as face recognition, have been chosen as the applications to study.

4.1 Mixing Sample Group and Pooled Covariance Matrices

The maximum likelihood covariance estimate (or sample group covariance matrix) and the pooled covariance estimate represent two possible estimates for the true covariance matrices of Bayes Plug-in classifiers. In this section, we carry out experiments on face and facial expression recognition applications in order to investigate in practice the combination of these two covariance matrices, called mixture covariance matrices [TFV00, TGF01a, TGF03a].

4.1.1 Definition

We define the mixture covariance matrix as being simply a linear or convex combination between the sample group covariance matrix S_i (defined in equation (3.6)) and the pooled sample covariance matrix S_p (defined in equation (3.12)). It is given by

$$S_i^{mix}(w_i) = w_i S_p + (1 - w_i) S_i \,, \qquad \textbf{(4.1)}$$

where the mixture parameter w_i takes on values $0 < w_i \leq 1$ and is different for each class. This parameter controls the degree of shrinkage of the sample group covariance estimates toward the pooled one.

The motivation of combining solely the sample group and pooled covariance matrices comes from one of the conclusions of the previous chapter that the sample group covariance matrices and the pooled covariance matrix may be reasonable extremes for intermediate estimations, especially when concerns about computational costs exist and the total number of real data training patterns is larger than the number of features.

We can visualise the geometric idea of the mixture covariance matrix as follows. Let a two-dimensional feature space contain three hypothetical normally distributed classes, as illustrated in Figure 4.1. The constant probability densities contours of S_i and S_p are represented by the dashed and dotted grey ellipses, respectively, and defined as surfaces of ellipsoids centred at the corresponding sample mean vectors $\bar{x}_i$ [JW98], that is

$$\begin{aligned} (x - \bar{x}_i)^T S_i^{-1} (x - \bar{x}_i) &= c_i^2 \\ (x - \bar{x}_i)^T S_p^{-1} (x - \bar{x}_i) &= t_i^2 \,, \end{aligned} \qquad \textbf{(4.2)}$$

where c_i and t_i are constants. The mixture covariance estimates assume geometrically that the ellipses corresponding to the true covariance matrices are placed somewhere in between S_i and S_p contours, as shown by the solid black ellipses [TFV00, TGF01a, TGF03a].

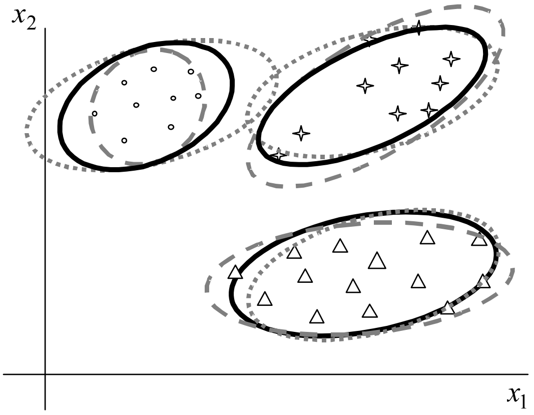


Figure 4.1. Geometric idea of the mixture covariance matrix.

Each mixture covariance matrix S_i^{mix} defined in equation (4.1) has the important property of admitting an inverse if the pooled estimate S_p does so [MN99]. This implies that if the pooled estimate is non-singular and the mixture parameter takes on values $w_i > 0$, then S_i^{mix} will be non-singular.

Therefore the important question is [TGF01a, TGF03a]: what is the value of w_i that gives a relevant linear mixture between the pooled and sample covariance estimates ? A method that determines an appropriate value of the mixture parameter, which is based on the Hoffbeck approach [Hof95, HL96] discussed in the previous chapter, is described in the next sub-section.

4.1.2 The Mixture Parameter

According to Hoffbeck and Landgrebe [Hof95, HL96], the value of the mixture parameter w_i can be appropriately selected so that a best fit to the training samples is achieved. Their technique is based on the leave-one-out likelihood parameter estimation and allows different mixing parameters for each class without increasing highly the computational burden.

In this leave-one-out likelihood method [Fuk90], one observation of the class π_i training set is removed and the mean and covariance matrix from the remaining $N_i - 1$ examples is estimated. Afterwards the likelihood of the excluded sample is calculated given the previous mean and covariance matrix estimates. This operation is repeated a further $N_i - 1$ times and the average log likelihood is computed over all the N_i observations.

The strategy is to evaluate several different values of w_i in the range $0 < w_i \leq 1$, and then choose the w_i that maximizes the average log likelihood.

The sample mean of class π_i (defined in equation (3.5)) without observation r may be computed as

$$\bar{x}_{i\backslash r} = \frac{1}{(N_i - 1)} \left[\left(\sum_{j=1}^{N_i} x_{i,j} \right) - x_{i,r} \right], \tag{4.3}$$

where $x_{i,j}$ is the *n*-dimensional observation *j* from class π_i and, as a reminder, N_i is the number of training observations from class π_i. The notation $\backslash r$ conforms to the Hoffbeck and Landgrebe [Hof95, HL96] works. It indicates that the corresponding quantity is calculated with the *r*-th observation from class π_i removed.

Following the same idea, the sample group covariance matrix of class π_i (defined in equation (3.6)) without observation r is

$$S_{i\backslash r} = \frac{1}{(N_i - 2)} \left[\left(\sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_{i\backslash r})(x_{i,j} - \bar{x}_{i\backslash r})^T \right) - (x_{i,r} - \bar{x}_{i\backslash r})(x_{i,r} - \bar{x}_{i\backslash r})^T \right]. \tag{4.4}$$

On the assumption that all classes have the same number of training observations, the pooled covariance matrix (defined in equation (3.12)) without observation r is

$$S_{p_{i\backslash r}} = \frac{1}{g} \left[\left(\sum_{j=1}^{g} S_j \right) - S_i + S_{i\backslash r} \right], \tag{4.5}$$

where, as a reminder, g is the number of groups or classes. Thus the average log likelihood with the excluded observations can be calculated as follows:

$$\bar{L}_i(w_i) = \frac{1}{N_i} \left[\sum_{r=1}^{N_i} \ln\left[f_i\left(x_{i,r} \mid \bar{x}_{i\backslash r}, S_{i\backslash r}^{mix}(w_i) \right) \right] \right], \tag{4.6}$$

where $f_i\left(x_{i,r} \mid \bar{x}_{i\backslash r}, S_{i\backslash r}^{mix}(w_i) \right)$ is the Gaussian class-conditional probability function (defined in equation (3.1)) with $\bar{x}_{i\backslash r}$ mean vector and $S_{i\backslash r}^{mix}(w_i)$ covariance matrix given by

$$S_{i\backslash r}^{mix}(w_i) = w_i S_{p_{i\backslash r}} + (1 - w_i) S_{i\backslash r}. \tag{4.7}$$

As Hoffbeck and Landgrebe pointed out, this approach, if implemented in a straightforward way, would require computing the inverse and determinant of the $S_{i\backslash r}^{mix}(w_i)$ for each training observation. Since the $S_{i\backslash r}^{mix}(w_i)$ is an n by n matrix and n is typically a large number, this computation would be quite expensive [Hof95, HL96]. However, they showed that it is possible to significantly reduce the required computation by using the Sherman-Morrison-Woodbury formula [GL89, p. 51] given by

$$\left(A+uu^T\right)^{-1} = A^{-1} - \frac{A^{-1}uu^T A^{-1}}{1+u^T A^{-1}u}, \tag{4.8}$$

where A is a n by n matrix and u is a n by 1 vector. This allowed them to write the log likelihood of the excluded samples in an analogous form as follows:

$$\ln\left[f_i\left(x_{i,r} \mid \bar{x}_{i\backslash r}, S_{i\backslash r}^{mix}(w_i)\right)\right] = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left[|Q|(1-vd)\right] - \frac{1}{2}\left(\frac{N_i}{N_i-1}\right)^2\left[\frac{d}{1-vd}\right], \tag{4.9}$$

where

$$Q = \left[(1-w_i)\frac{(N_i-1)}{(N_i-2)} + w_i\frac{1}{g(N_i-2)}\right]S_i + w_i S_p, \tag{4.10}$$

$$v = \frac{N_i}{(N_i-1)(N_i-2)}\left[1 - w_i\frac{(g-1)}{g}\right], \tag{4.11}$$

$$d = (x_{i,r} - \bar{x}_i)^T Q^{-1}(x_{i,r} - \bar{x}_i). \tag{4.12}$$

As can be seen from equations (4.10) and (4.11), the matrix Q and the value v do not depend on the removed training observation. Therefore, the determinant and inverse of matrix Q calculated in equations (4.9) and (4.12) respectively, as well as the value v calculated by equation (4.11), can be computed only once for each mixing parameter w_i, reducing significantly the computational burden.

Finally, when the parameter w_i is selected, the mixture covariance matrix estimate defined in equation (4.1) is calculated using all the training examples and placed into the quadratic discriminant rule defined in equation (3.8).

4.1.3 Experiments

As mentioned before, experiments on face and facial expression recognition applications were carried out in order to investigate the mixture of sample group and pooled covariance matrices on the QDF classifier.

4.1.3.1 ORL Face Database

In the face recognition experiments the Olivetti Face Database[1] (ORL) was used. This database contains a set of face images taken between April 1992 and April 1994 at the Olivetti Research Laboratory in Cambridge, U.K, with ten images for each of 40 individuals, a total of 400 images. All images were taken against a dark homogeneous background with the person in an upright frontal position, with tolerance for some tilting and rotation of up to about 20 degrees. Scale varies about 10%. The original size of each image is 92x112 pixels, with 256 grey levels per pixel. Figure 4.2 shows as an example of the set of 10 images of one individual cropped to the size of 64x64.

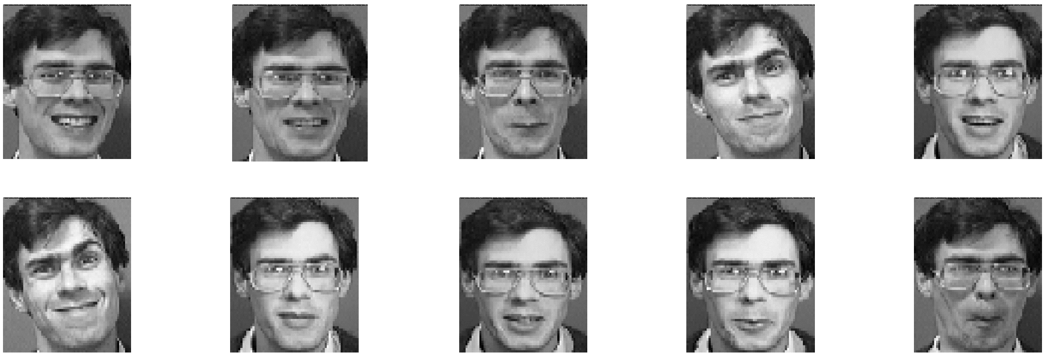


Figure 4.2. A set of ten images of one individual from the ORL Face Database.

4.1.3.2 Facial Expression Database

Tohoku University has made available a database which was used for the facial expression experiment. This database is composed of 193 images of expressions posed by nine Japanese females [LBA99]. Each person posed three or four examples of each of six fundamental facial expression: anger, disgust, fear, happiness, sadness and surprise. The

[1] The ORL database has been available online on the website http://www.cam-orl.co.uk/facedatabase.html

database has at least 29 images for each fundamental facial expression. Figure 4.3 illustrates some examples of each one of the six fundamental facial expression images (from top left to bottom right) of the Tohoku Facial Expression database, cropped to the size of 64x64 pixels.

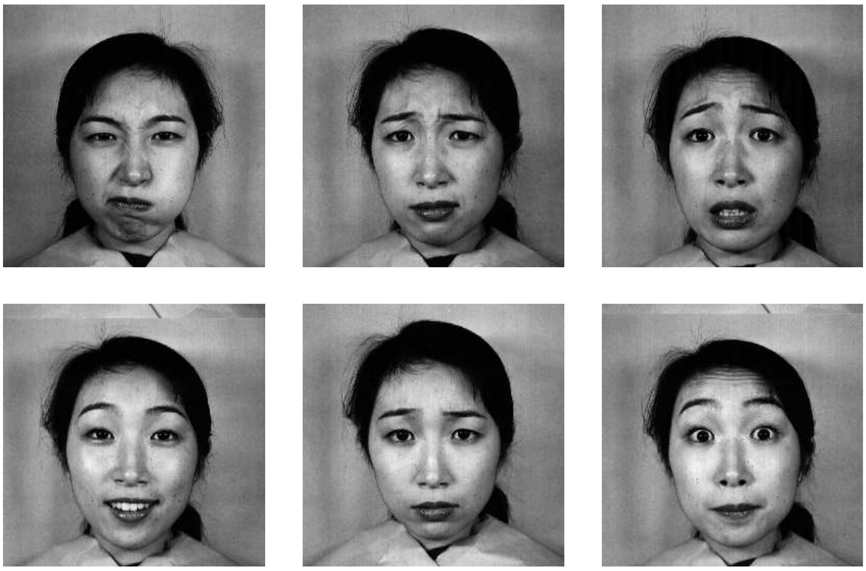


Figure 4.3. Anger, disgust, fear, happiness, sadness, and surprise (top left - bottom right).

4.1.3.3 Implementation

Instead of analysing the unconventional QDF classifier directly on the face or facial expression images, the standard and lower dimensional image representation [TP91] using Principal Component Analysis (PCA), described in the previous chapter, was applied first to provide dimensionality reduction.

Thus, the experiments were carried out as follows. First PCA reduces the dimensionality of the original images (which were resized to 64x64 pixels for implementation convenience) and secondly the discriminant quadratic rule (defined in equation (3.8)) was applied using each one of the three following covariance estimates: S_i (or Sgroup), S_p (or Spooled), and S_i^{mix} (or Smix). Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate were calculated.

The face recognition classification was computed using for each individual 5 images to train and 5 images to test. In the facial expression recognition, the training and test sets were respectively composed of 20 and 9 images. The size of the mixture parameter ($0 < w_i \leq 1$) optimisation range was taken to be 20, that is $w_i = [0.05, 0.10, 0.15, \ldots, 1]$.

4.1.4 Results

The training and test average recognition rates (with standard deviations) of the face and facial expression databases, respectively, over the different PCA dimensions are presented in Tables 4.1 and 4.2.

Since only 5 images of each individual were used to form the face recognition training set, the results relative to the sample group covariance estimate were limited to 4 PCA components. Table 4.1 shows that in all but one experiment the S_i^{mix} (or Smix) estimate led to higher accuracy than did both the pooled covariance and sample group covariance matrices. In terms of how sensitive the mixture covariance results were to the choice of the training and test sets, it is fair to say that the S_i^{mix} standard deviations were similar to the pooled estimate.

PCA	**Sgroup**		**Spooled**		**Smix**	
Components	Training	Test	Training	Test	Training	Test
4	99.5 (0.4)	51.6 (4.4)	73.3 (3.1)	59.5 (3.0)	90.1 (2.1)	70.8 (3.2)
10			96.6 (1.2)	88.4 (1.4)	99.4 (0.5)	92.0 (1.5)
20			99.2 (0.6)	91.8 (1.8)	100.0 (0.1)	94.5 (1.7)
30			99.9 (0.2)	94.7 (1.7)	100.0 (0.0)	95.9 (1.5)
40			100.0 (0.0)	95.4 (1.5)	100.0 (0.0)	96.2 (1.6)
50			100.0 (0.0)	95.7 (1.2)	100.0 (0.0)	96.4 (1.5)
60			100.0 (0.0)	95.0 (1.6)	100.0 (0.0)	95.8 (1.6)
70			100.0 (0.0)	94.9 (1.6)	100.0 (0.0)	95.4 (1.6)

Table 4.1. ORL face recognition results (Smix)

Table 4.2 shows the results of the facial expression recognition. For more than 20 components when the sample group covariance estimate became singular, the mixture covariance estimate reached higher recognition rates than the pooled covariance estimate. Again, regarding the computed standard deviations, the S_i^{mix} estimate was shown to be as sensitive to the choice of the training and test sets as the other two estimates.

PCA	Sgroup		Spooled		Smix	
Components	Training	Test	Training	Test	Training	Test
5	41.5 (4.2)	20.6 (3.9)	32.3 (3.0)	21.6 (3.8)	34.9 (3.3)	21.3 (4.1)
10	76.3 (3.6)	38.8 (5.6)	49.6 (3.9)	26.5 (6.8)	58.5 (3.7)	27.9 (5.6)
15	99.7 (0.5)	64.3 (6.4)	69.1 (3.6)	44.4 (5.3)	82.9 (2.9)	49.7 (7.7)
20			81.2 (2.6)	55.9 (7.7)	91.4 (2.8)	61.3 (7.1)
25			86.9 (2.8)	64.9 (6.9)	94.8 (2.2)	68.3 (5.1)
30			91.9 (1.7)	70.1 (7.8)	96.8 (1.3)	72.3 (6.2)
35			94.3 (1.7)	72.0 (7.4)	97.7 (1.1)	75.6 (5.5)
40			95.9 (1.4)	75.6 (7.1)	98.3 (1.1)	77.2 (5.7)
45			96.7 (1.3)	78.4 (6.5)	98.6 (0.8)	79.1 (5.4)
50			97.6 (1.0)	79.4 (5.8)	99.2 (0.7)	81.0 (6.6)
55			98.5 (0.9)	81.6 (6.6)	99.5 (0.6)	82.8 (6.3)
60			99.1 (0.8)	82.1 (5.9)	99.6 (0.6)	83.6 (7.2)
65			99.5 (0.6)	83.3 (5.5)	99.8 (0.4)	84.5 (6.2)

Table 4.2. Tohoku facial expression recognition results (Smix).

4.1.5 Discussion

An important result revealed by these experiments is related to the mixture parameters w_i optimised by the leave-one-out likelihood procedure described in the sub-section 4.1.2.

Table 4.3 shows the average (with standard deviations) of the optimised mixture parameter w_i over the common face and facial expression PCA components.

PCA	Linear Mixture Parameter	
Components	Face	Facial Expression
10	0.58 (0.25)	0.76 (0.19)
20	0.65 (0.21)	0.49 (0.15)
30	0.71 (0.18)	0.56 (0.15)
40	0.77 (0.16)	0.67 (0.15)
50	0.82 (0.13)	0.77 (0.11)
60	0.85 (0.11)	0.85 (0.09)

Table 4.3. The average (with standard deviations) of the optimised mixture parameters.

It can be seen from Table 4.3 that as the dimension of the feature space increases, the average and standard deviation of the mixture parameter w_i in all but one experiment increases and decreases respectively, making the mixture covariance of each class S_i^{mix} more similar to the pooled covariance S_p than the sample group one S_i.

Although this behaviour depends on the applications considered, it suggests that in both well-framed and pre-processed image classification tasks the sparseness of the sample group covariance matrix could influence its linear combination to the pooled covariance matrix. In other words, it seems that when the group sample sizes are small compared with the dimension of the feature space, the pooled information is more reliable than that provided sparsely by each group.

4.2 The Loss of Covariance Information

Motivated by the results presented in the preceding section, we have observed that in situations where the sample group covariance matrices S_i are singular, linear combinations of S_i and, for instance, the pooled covariance matrix S_p may lead to a "loss of covariance information".

In the following sub-sections, we define the problem of "loss of covariance information" and describe an initial approach to understand this concept in practice, called the Covariance Projection Ordering (COPO) method. Experiments on face and facial expression databases are shown and compared with the RDA and LOOC methods described in the previous chapter [TG01, TGF01b].

4.2.1 Definition

The theoretical interpretation of the "loss of covariance information" can be described as follows. Let a matrix S_i^{mix} be given by the following linear combination:

$$S_i^{mix} = aS_i + bS_p, \tag{4.13}$$

where the mixing parameters a and b are positive constants that sum to 1, and the pooled covariance matrix S_p is a non-singular (or full-rank) matrix.

The S_i^{mix} eigenvectors and eigenvalues are given by the matrices Φ_i^{mix} and Λ_i^{mix}, respectively. From the covariance spectral decomposition formula (defined in equation (3.9)), it is possible to write

$$(\Phi_i^{mix})^T S_i^{mix} \Phi_i^{mix} = \Lambda_i^{mix} = \begin{bmatrix} \lambda_1^{mix} & & & 0 \\ & \lambda_2^{mix} & & \\ & & \ddots & \\ 0 & & & \lambda_n^{mix} \end{bmatrix} = diag[\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_n^{mix}], \qquad \textbf{(4.14)}$$

where $\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_n^{mix}$ are the S_i^{mix} eigenvalues and n is the dimension of the measurement space considered. Using the information provided by equation (4.13), equation (4.14) can be rewritten as:

$$\begin{aligned} \Lambda_i^{mix} &= diag[\lambda_1^{mix}, \lambda_2^{mix}, ..., \lambda_n^{mix}] \\ &= (\Phi_i^{mix})^T \left[aS_i + bS_p \right] \Phi_i^{mix} \\ &= a(\Phi_i^{mix})^T S_i \Phi_i^{mix} + b(\Phi_i^{mix})^T S_p \Phi_i^{mix} \\ &= a\mathrm{Z}^i + b\mathrm{Z}^p. \end{aligned} \qquad \textbf{(4.15a)}$$

The matrices Z^i and Z^p are not diagonal matrices because Φ_i^{mix} does not necessarily diagonalises both matrices simultaneously. However, as Φ_i^{mix} is the eigenvector matrix of the linear combination of S_i and S_p, the off-diagonal elements of Z^i and Z^p necessarily cancel each other in order to generate the eigenvalues matrix Λ_i^{mix}. Therefore, the string of equalities in (4.15a) can be extended to

$$\begin{aligned} \Lambda_i^{mix} &= a\mathrm{Z}^i + b\mathrm{Z}^p \\ &= diag[a\zeta_1^i, a\zeta_2^i, ..., a\zeta_n^i] + diag[b\zeta_1^p, b\zeta_2^p, ..., b\zeta_n^p] \\ &= diag[a\zeta_1^i + b\zeta_1^p, a\zeta_2^i + b\zeta_2^p, ..., a\zeta_n^i + b\zeta_n^p] \end{aligned} \qquad \textbf{(4.15b)}$$

where $\zeta_1^i, \zeta_2^i, ..., \zeta_n^i$ and $\zeta_1^p, \zeta_2^p, ..., \zeta_n^p$ are respectively the variances of the sample and pooled covariance matrices spanned by the S_i^{mix} eigenvectors matrix Φ_i^{mix}. Then, the spectral decomposition form of the conventional QDF (defined in equation (3.11)) becomes:

$$\begin{aligned} d_i(x) &= \sum_{k=1}^{n} \ln \lambda_k^{mix} + \sum_{k=1}^{n} \frac{[(\phi_{ik}^{mix})^T (x - \bar{x}_i)]^2}{\lambda_k^{mix}} - 2\ln(p(\pi_i)) \\ &= \sum_{k=1}^{n} \ln\left(a\zeta_k^i + b\zeta_k^p\right) + \sum_{k=1}^{n} \frac{[(\phi_{ik}^{mix})^T (x - \bar{x}_i)]^2}{a\zeta_k^i + b\zeta_k^p} - 2\ln(p(\pi_i)) \end{aligned} \qquad \textbf{(4.16)}$$

where ϕ_{ik}^{mix} is the corresponding k-th eigenvector of the matrix S_i^{mix} and, as a reminder, $p(\pi_i)$ is the prior probability associated with the ith group.

The discriminant score described in equation (4.16) considers the dispersions of sample group covariance matrices spanned by all the S_i^{mix} eigenvectors. However, when the group sample sizes N_i are small or not large enough compared with the dimension of the feature space n, the corresponding lower dispersion values are often estimated to be 0 or approximately 0, implying that these values are not reliable. Therefore, a linear combination of S_i and S_p that uses the same parameters a and b as defined in (4.16) for the whole feature space fritters away some pooled covariance information.

The geometric idea of a hypothetical "loss of covariance information" on a three-dimensional feature space is illustrated in Figure 4.4. The constant probability density contour of S_i and S_p are represented by the two-dimensional (x_1, x_2) dark grey ellipse and three-dimensional (x_1, x_2, x_3) light grey ellipsoid, respectively.

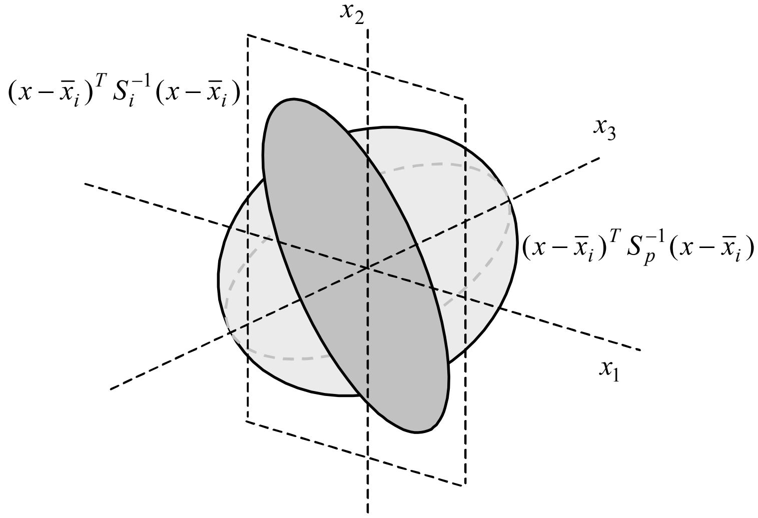

x_2
$(x-\bar{x}_i)^T S_i^{-1}(x-\bar{x}_i)$
x_3
$(x-\bar{x}_i)^T S_p^{-1}(x-\bar{x}_i)$
x_1

Figure 4.4. Geometric idea of a hypothetical "loss of covariance information".

As can be seen, S_i is well defined on the plane (x_1, x_2) but not defined at all on (x_1, x_2, x_3). In fact, there is no information from S_i on the x_3 axis. As a consequence, a linear combination of S_i and S_p that shrinks or expands both matrices equally all over the feature space simply ignores this evidence from S_p. Other covariance estimators have not addressed this problem.

4.2.2 Covariance Projection Ordering Method

The Covariance Projection Ordering (COPO) method is an intuitive, initial, approach to understand in practice the problem of loss of covariance information [TG01, TGF01b]. It assumes that all groups have similar covariance shapes and has the property of having the same rank as the pooled estimate.

The COPO idea is, basically, to use all the sample group covariance information available whenever possible and the pooled covariance information otherwise. Looking at equations (4.13) and (4.15) and writing the covariance matrix on its spectral decomposition form (defined in equation (3.9)), this idea can be derived as follows:

$$\begin{aligned} S_i^{copo} &= \sum_{k=1}^{n} \zeta_k^{copo} \phi_{ik}^{copo} (\phi_{ik}^{copo})^T, \\ \zeta_k^{copo} &= \begin{cases} \zeta_k^i & \text{if } 1 \le k \le rank(S_i) \\ \zeta_k^p & \text{otherwise,} \end{cases} \end{aligned} \tag{4.17}$$

where ϕ_{ik}^{copo} is the corresponding *k*-th eigenvector of the matrix given by $S_i + S_p$ ordered in sample group ζ_k^i variance decreasing values. Thus, the discriminant score described in (4.16) becomes:

$$d_i(x) = \sum_{k=1}^{r} \ln \zeta_k^i + \sum_{k=r+1}^{n} \ln \zeta_k^p + \sum_{k=1}^{r} \frac{[(\phi_{ik}^{copo})^T (x - \bar{x}_i)]^2}{\zeta_k^i} + \sum_{k=r+1}^{n} \frac{[(\phi_{ik}^{copo})^T (x - \bar{x}_i)]^2}{\zeta_k^p}, \tag{4.18}$$

where $r = rank(S_i)$.

The Covariance Projection approach provides another combination of the sample group covariance matrices S_i and the pooled covariance matrix S_p in such a way that this combination is strongly related to the rank of S_i or, equivalently, to the number of training samples N_i. It can be viewed as an *n*-dimensional non-singular approximation of an *r*-dimensional singular matrix that considers explicitly the sample group singularity effects.

The COPO requires an eigenvector-eigenvalue ordering process to select information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise. Therefore, the COPO method is not restricted to use the same covariance combination for all classes, allowing a better understanding of the loss of covariance information issue.

4.2.3 Experiments

In order to evaluate the COPO method, experiments on face and facial expression recognition applications, using the same ORL Face and Tohoku Facial Expression databases described in the sub-sections 4.1.3.1 and 4.1.3.2, respectively, were carried out.

Following the same procedure of the section 4.1 experiments, first PCA reduced the dimensionality of the original images (which were resized to 64x64 pixels for implementation convenience) and secondly the discriminant quadratic rule (defined in equation (3.8)) was applied using each one of the following five covariance estimators: 1) Sample group covariance matrix (Sgroup) defined in equation (3.6); 2) Pooled covariance matrix (Spooled) defined in equation (3.12); 3) Covariance projection ordering matrix (Scopo) defined in equation (4.17); 4) Friedman's RDA matrix (Srda) defined in equation (3.13); 5) Hoffbeck's covariance matrix (Slooc) defined in equation (3.27). Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate were calculated.

The face recognition classification was computed using for each individual 5 images to train and 5 images to test. In the facial expression recognition, the training and test sets were respectively composed of 20 and 9 images. The RDA optimisation grid was taken to be the outer product of $\lambda = [0, 0.125, 0.354, 0.650, 1]$ and $\gamma = [0, 0.25, 0.5, 0.75, 1]$, identically to that in Friedman's work [Fri89]. Analogously, the size of the LOOC mixture parameter [HL96] was $\alpha_i = [0, 0.25, 0.5, ..., 2.75, 3.0]$.

4.2.4 Results

Tables 4.4 and 4.5 present the training and test average recognition rates (with standard deviations) of the ORL and Tohoku face and facial expression databases, respectively, over the different PCA dimensions. Also presented are the mean of the optimised RDA and LOOC parameters. For the ORL face database, only 6 out of the 40 LOOC parameters corresponding to the subjects 1, 5, 10, 20, 30 and 40 are shown. The notation "-" in the Sgroup rows indicates that the sample group covariance was singular and could not be used to classify the samples.

As a reminder, the RDA parameter λ controls the degree of shrinkage of the sample group covariance matrix estimates toward the pooled covariance one, whereas the parameter γ controls the shrinkage toward a multiple of the identity matrix. Analogously, the LOOC parameter α_i in between 0 and 1 leads to mixtures of the diagonal of the sample group covariance and sample group covariance itself, in between 1 and 2 it leads to mixtures of the sample group covariance and the common covariance (which in this case is equal to the pooled one), and in between 2 and 3 to mixtures of common covariance and the matrix of its diagonal elements.

	PCAs				
	4	10	20	40	60
Training					
Sgroup	99.5(0.4)	-	-	-	-
Spooled	73.3(3.1)	96.6(1.2)	99.2(0.6)	100.0(0.0)	100.0(0.0)
Scopo	97.0(1.1)	99.9(0.2)	100.0(0.0)	100.0(0.0)	100.0(0.0)
Srda	81.2(2.8)	99.5(0.7)	99.9(0.2)	100.0(0.0)	100.0(0.0)
Slooc	89.4(1.9)	98.9(0.7)	99.6(0.4)	99.8(0.3)	99.9(0.2)
Test					
Sgroup	51.6(4.4)	-	-	-	-
Spooled	59.5(3.0)	88.4(1.4)	91.8(1.8)	95.4(1.5)	95.0(1.6)
Scopo	69.8(3.4)	90.2(2.5)	94.0(1.9)	96.4(1.6)	95.9(1.5)
Srda	64.7(3.9)	92.4(1.9)	94.0(1.4)	96.0(1.7)	95.6(1.6)
Slooc	70.1(3.1)	90.8(2.2)	93.5(2.2)	93.0(1.8)	92.0(1.8)
RDA					
λ	0.1	0.1	0.2	0.3	0.3
γ	0.0	0.0	0.1	0.2	0.3
LOOC					
α1	1.6	2.0	2.6	2.9	3.0
α5	1.3	1.6	1.6	1.7	1.8
α10	2.3	2.4	2.2	2.8	2.9
α20	1.6	1.6	1.9	2.3	2.7
α30	1.5	1.5	1.6	1.6	1.8
α40	1.4	1.6	1.8	2.1	2.5

Table 4.4. ORL face recognition results (COPO).

Table 4.4 shows that on the training set and for less than 20 PCA components the Scopo estimator led to higher face recognition classification accuracy than the linear covariance estimator (Spooled) and both optimised quadratic discriminant estimators (Srda and Slooc). For the test samples, the Srda and Slooc estimators often outperformed the Scopo in lower dimensional space, but these performances deteriorated when the dimensionality increased, particularly the Slooc ones. It seems that in higher dimensional space, when the Sgroup estimate became extremely poorly represented, the RDA and LOOC parameters, despite the optimisation of the classification accuracy and likelihood indexes respectively, did not counteract the Sgroup mixing singularity effect. The Scopo estimator achieved the best recognition rate – 96.4% – for all PCA components considered. In terms of how sensitive the covariance results were to the choice of training and test sets, the covariance estimators had similar performances, particularly in high dimensional space.

	PCAs				
	10	30	50	70	100
Training					
Sgroup	76.3(3.6)	-	-	-	-
Spooled	49.6(3.9)	91.9(1.7)	97.6(1.0)	99.6(0.5)	100.0(0.0)
Scopo	66.6(3.2)	95.8(1.6)	99.2(0.8)	100.0(0.0)	100.0(0.0)
Srda	75.0(6.7)	96.7(2.9)	98.5(1.0)	99.2(1.0)	99.9(0.2)
Slooc	51.4(4.9)	91.0(4.1)	95.8(2.0)	98.8(1.3)	99.9(0.3)
Test					
Sgroup	38.8(5.6)	-	-	-	-
Spooled	26.5(6.8)	70.1(7.8)	79.4(5.8)	83.9(7.0)	84.4(6.5)
Scopo	31.5(5.8)	68.3(5.5)	79.5(5.8)	85.0(7.0)	84.1(6.0)
Srda	37.8(5.9)	73.0(7.4)	80.1(6.2)	79.9(8.7)	81.3(6.7)
Slooc	26.3(5.3)	65.2(5.6)	71.2(8.2)	79.9(8.7)	87.2(5.8)
RDA					
λ	0.0	0.4	0.8	0.7	0.7
γ	0.0	0.0	0.0	0.1	0.3
LOOC					
$\alpha 1$	2.3	0.6	0.9	2.9	3.0
$\alpha 2$	2.4	1.4	2.3	2.9	2.9
$\alpha 3$	2.8	1.0	1.7	2.8	3.0
$\alpha 4$	2.8	2.3	2.1	3.0	3.0
$\alpha 5$	2.8	0.6	0.9	2.3	3.0
$\alpha 6$	2.6	0.5	1.1	2.8	3.0

Table 4.5. Tohoku facial expression recognition results (COPO).

The results of the Tohoku facial expression recognition are presented in table 4.5. For more than 50 PCA components on the training set, the Scopo estimator performed as well or better than all the other covariance estimators considered. Regarding the test samples, however, there is no overall dominance of any covariance estimator. In lower dimension spaces, Srda led to higher classification accuracies, followed by Scopo, Spooled and Slooc. On the other hand, when the dimensionality increased and the true covariance matrices became apparently equal and highly ellipsoidal, Srda performed poorly while Scopo, Spooled and Slooc improved. In the highest dimensional space the LOOC optimisation, which considers the diagonal elements of the pooled estimate, took advantage of the equal-ellipsoidal behaviour (for more than 70 PCAs all α_i parameters are close to the value 3) achieving the best recognition rate – 87.2% – for all PCA components calculated. In this recognition application, all the computed covariance estimators were quite sensitive to the choice of the training and test sets.

4.2.5 Discussion

This section described the problem of loss of covariance information when singular sample group and non-singular covariance matrices, such as the pooled estimation, are linearly combined using the same parameters all over the feature space.

In limited sample size and well framed image recognition applications, like the face and facial expression classification problems presented, an eigenvector-eigenvalue ordering procedure that selects information from the projected sample group covariance matrices whenever possible and the pooled covariance otherwise showed the significance of the loss of covariance information problem when blending singular and non-singular covariance matrices. It seems that when limited information is provided, the problem of estimating covariance matrices for classification is affected not only by the way that information is optimised but also by its reliability.

Although the COPO method can be viewed as a valid *n*-dimensional non-singular approximation of an *r*-dimensional singular matrix that considers explicitly the sample group singularity effects, it is still strongly related to the reliable information provided by the sample group covariance matrices. In image recognition applications, where the sample group covariance matrices would be eventually full rank but not accurately estimated, this initial approach may not perform well.

4.3 Summary and Conclusions

In this chapter, the performances of several unconventional Bayes plug-in covariance estimators were evaluated in pre-processed image recognition problems that used small and moderate training sets, a large number of features, and a moderate number of groups.

Experiments carried out on face and facial expression recognition confirmed the findings of other researchers that choosing an unconventional Bayes plug-in classifier between the linear and quadratic ones improves the classification accuracy in settings for which sample sizes are small and the number of features is large. In those well-framed applications, however, where the sources of variation were the same from group to group and consequently a similar covariance shape might be assumed for all groups, linear combinations of the sample group covariance matrices and the pooled covariance matrix led to a loss of covariance information.

An initial and intuitive approach to understanding this problem was developed and showed the importance of taking into account the distinct information provided by the sample group covariance matrix and the pooled covariance matrix in the whole high-dimensional feature space. When limited information is provided, the problem of estimating covariance matrices for classification is affected not only by the way that information is optimised but also by its reliability.

Chapter 5

Covariance Matrix Estimation

The previous chapter's investigations demonstrated that linear combinations of singular and non-singular covariance matrices may lead to a loss of reliable covariance information in small sample size problems. In this chapter, a new unconventional Bayes Plug-in or Quadratic Discriminant Function (QDF) classifier is proposed. This classifier is based on a covariance matrix estimation that combines covariance matrices under the principle of maximum entropy. It assumes that the sources of variation are similar from group to group and consequently a similar covariance shape may be expected for all classes. This has often been the case for pre-processed image recognition applications. The new covariance matrix estimation not only deals with the singularity and instability of the maximum likelihood covariance estimator, but also is computed directly without requiring a time-consuming optimisation procedure.

5.1 The Maximum Entropy Covariance Selection Method

The Maximum Entropy Covariance Selection (MECS) method considers the issue of combining the sample group covariance matrices and the pooled covariance matrix based on the maximum entropy (ME) principle, stated briefly as:

> *"When we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have."* [Jay82]

In the problem of estimating covariance matrices for Gaussian classifiers, it is known that different covariance estimators might be optimal depending not only on the true covariance statistics of each class, but also on the number of training observations, the dimension of the feature space and even the ellipsoidal symmetry associated with the normal distribution [Jam85, Fri89, Hof95]. In fact, such covariance optimisation can be viewed as a problem of estimating the parameters of Gaussian probability distributions under uncertainty. Therefore, the ME criterion that maximises the uncertainty under an incomplete information context should be a promising solution [TGF02, TGF03b].

5.1.1 Definition

The Maximum Entropy Covariance Selection (MECS) method assumes that the sources of variation are similar from group to group and consequently a similar covariance shape may be expected for all classes.

Let an n-dimensional sample X_i of class π_i be normally distributed with true mean μ_i and true covariance matrix Σ_i, i.e. $X_i \sim N_n(\mu_i, \Sigma_i)$. As described in chapter 2, the entropy $h(X_i)$ of such a sample X_i is defined as the expected value of the natural logarithm of the inverse of the probability density function of X_i, which in this case can be written as (e.g., [Fuk90]):

$$\begin{aligned} h(X_i) &= -E\{\ln[p(x \mid \pi_i)]\} \\ &= -E\left\{\ln\left[\frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right]\right]\right\} \\ &= -E\left\{-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \\ &= -E\left\{-\frac{n}{2}\ln(2\pi)\right\} - E\left\{-\frac{1}{2}\ln|\Sigma_i|\right\} - E\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \\ &= \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma_i| + \frac{n}{2}. \end{aligned} \tag{5.1}$$

As can be seen from equation (5.1), the first $(n/2)\ln 2\pi$ and third $(n/2)$ terms are constants and can be ignored. Consequently, the entropy $h(X_i)$ is simply a function of the determinant of Σ_i, which is invariant under any orthonormal transformation. Thus, when Φ_i consists of n eigenvectors of Σ_i it is possible to write [Fuk90]

$$\ln\left|\Phi_i^T \Sigma_i \Phi_i\right| = \ln|\Lambda_i| = \sum_{k=1}^{n} \ln \lambda_k \ , \tag{5.2}$$

where Λ_i is the diagonal λ_k eigenvalues matrix of Σ_i. In order to maximise (5.2) or equivalently (5.1), we must select the covariance estimation of Σ_i that gives the largest eigenvalues.

If we consider linear combinations S_i^{mix} between the sample group S_i and pooled S_p covariance matrices, equation (5.2) can be rewritten (by using equations (4.15)) as

$$\begin{aligned}\ln\left|(\Phi_i^{mix})^T\left(aS_i+bS_p\right)\Phi_i^{mix}\right| &= \ln\left|a(\Phi_i^{mix})^T S_i\Phi_i^{mix}+b(\Phi_i^{mix})^T S_p\Phi_i^{mix}\right| \\ &= \ln\left|diag[a\zeta_1^i,a\zeta_2^i,\ldots,a\zeta_n^i]+diag[b\zeta_1^p,b\zeta_2^p,\ldots,b\zeta_n^p]\right| \\ &= \ln\left|diag[a\zeta_1^i+b\zeta_1^p,a\zeta_2^i+b\zeta_2^p,\ldots,a\zeta_n^i+b\zeta_n^p]\right| \\ &= \ln(\prod_{k=1}^{n} a\zeta_k^i+b\zeta_k^p) \\ &= \sum_{k=1}^{n}\ln(a\zeta_k^i+b\zeta_k^p),\end{aligned} \qquad \textbf{(5.3)}$$

where $\zeta_1^i,\zeta_2^i,\ldots,\zeta_n^i$ and $\zeta_1^p,\zeta_2^p,\ldots,\zeta_n^p$ are respectively the variances of the sample and pooled covariance matrices spanned by the S_i^{mix} eigenvector matrix Φ_i^{mix}. The parameters a and b are nonnegative and sum to 1.

Since the natural logarithm is a monotonic increasing function, the maximum of the function "$\ln(a\zeta_k^i+b\zeta_k^p)$" is at the same point in the space as the maximum of "$a\zeta_k^i+b\zeta_k^p$", we do not change the problem if instead of maximising equation (5.3) we maximise

$$\sum_{k=1}^{n}(a\zeta_k^i+b\zeta_k^p)\,. \qquad \textbf{(5.4)}$$

However, $a\zeta_k^i+b\zeta_k^p$ is a convex combination of two real numbers and the following inequality is valid [HJ85]

$$a\zeta_k^i+b\zeta_k^p \le \max(\zeta_k^i,\zeta_k^p)\,, \qquad \textbf{(5.5)}$$

for any $1 \le k \le n$ and convex parameters a and b that are nonnegative and sum to 1.

Equation (5.5) shows that the maximum of $a\zeta_k^i+b\zeta_k^p$ depends on k and is attained at the extreme values of the convex parameters, that is, either $a=1$ and $b=0$ or $a=0$ and $b=1$ [TGF03b]. Therefore, in order to maximise equation (5.4) we shall not choose the same parameters a and b for the whole feature space as previously, but select appropriately the maximum variances of the corresponding matrices.

5.1.2 Algorithm

One possible way to maximise equation (5.4) and consequently the entropy given by the convex combination of S_i and S_p is to select the maximum variances of the sample and pooled covariance matrices given by an orthonormal projection basis that diagonalises an unbiased ($a = b$) linear mixture of the corresponding matrices [TGF03b].

If we recall the assumption made that all classes have similar covariance shapes, it is reasonable to expect that the dominant eigenvectors (i.e., the eigenvectors with largest eigenvalues) of this unbiased mixture would be mostly orientated by the eigenvectors of the covariance matrix with largest eigenvalues. The choice of sample group or pooled information is then made purely on the size of the eigenvalue, which reflects the reliability of the poor information available. Since any unbiased combination of S_i and S_p gives the same set of eigenvectors, an orthonormal basis that would avoid the loss of covariance information is the one composed of the eigenvectors of the covariance matrix given by $S_i + S_p$.

Therefore, the MECS estimator S_i^{mecs} can be calculated by the following procedure:

i. Find the eigenvectors Φ_i^{me} of the covariance given by $S_i + S_p$.

ii. Calculate the variance contribution of both S_i and S_p on the Φ_i^{me} basis, i.e.

$$\begin{aligned} diag[(\Phi_i^{me})^T S_i \Phi_i^{me}] &= [\zeta_1^i, \zeta_2^i, \ldots, \zeta_n^i] \\ diag[(\Phi_i^{me})^T S_p \Phi_i^{me}] &= [\zeta_1^p, \zeta_2^p, \ldots, \zeta_n^p] \end{aligned} \qquad \textbf{(5.6a)}$$

iii. Form a new variance matrix based on the largest values, that is

$$Z_i^{me} = diag[\max(\zeta_1^i, \zeta_1^p), \max(\zeta_2^i, \zeta_2^p), \ldots, \max(\zeta_n^i, \zeta_n^p)] \qquad \textbf{(5.6b)}$$

iv. Form the MECS estimator

$$S_i^{mecs} = \Phi_i^{me} Z_i^{me} (\Phi_i^{me})^T . \qquad \textbf{(5.6c)}$$

The new unconventional quadratic classifier is constructed by substituting S_i^{mecs} for S_i in the Bayes discriminant rule defined in equation (3.8).

The main idea of the MECS approach is to expand in a straightforward way the smaller and consequently less reliable eigenvalues of S_i while trying to keep most of its larger eigenvalues unchanged. In fact, MECS estimation will select the most reliable linear features of a n-dimensional sample X_i described by a combination between its sample group covariance matrix S_i and the weighted average of all the sample group covariance matrices considered. It is a direct procedure that not only deals with the singularity and instability of S_i but also with the loss of information when similar covariance matrices are linearly combined. Furthermore, as MECS does not require an iterative optimisation procedure, its computational cost is much less severe than, for instance, the RDA and LOOC methods described in chapter 3.

5.2 Visual Analysis of Covariance Matrix Estimates

Before evaluating the MECS effectiveness on synthetic and real image recognition data, we first visually analyse the mixture covariance matrix estimates defined by the linear combination of the sample group and pooled covariance information. Given the high-dimensionality of the limited sample size problems, little attention has been paid to understanding what has happened to the final shape of such covariance matrix estimates in the recognition space [TG03a].

In this section, we describe an image analysis of the eigenvectors and eigenvalues of the mixture covariance matrices in the commonly used principal components (or eigenfaces) space [KS90, TP91]. This analysis is particularly helpful because by using the characterization of human faces we can distinguish clearly the hyper ellipsoids formed by the different mixture covariance matrix approaches in the sparse and high dimensional classification space.

5.2.1 Mixture Covariance Approaches

A number of optimisation approaches can be used to determine the appropriate parameter w_i for mixing the sample group covariance matrices S_i and the pooled covariance S_p. As a reminder (equation (4.1)), the mixture covariance matrices can be defined as

$$S_i^{mix}(w_i) = w_i S_p + (1 - w_i) S_i , \qquad \textbf{(5.7)}$$

where the mixture parameter w_i takes on values $0 < w_i \leq 1$ and could be different for each class depending on the optimisation technique used to solve the problem.

In light of the methods described in this chapter and the previous ones, these mixture covariance matrices can be calculated by using basically the following three approaches: maximum likelihood, maximum classification accuracy, and maximum entropy.

The maximum likelihood (ML) strategy, adopted in the section 4.1 of the previous chapter, is essentially a simplification of the Hoffbeck and Landgrebe approach [Hof95, HL96] described in chapter 3 (equation (3.28)). It consists of evaluating several values of w_i over the optimisation grid $0 < w_i \leq 1$, and then selecting w_i so that a best leave-one-out fit to the training patterns is achieved. Analogously, the maximum classification (MC) strategy is a simplification of the Friedman work [Fri89]. In this approach, all the mixture parameters w_i of each class are equal and selected to maximise the leave-one-out classification accuracy over all the training patterns of all classes (equation (3.14)). The maximum entropy (ME) approach has been described in the preceding section and maximises the information contained in the combined sample group and pooled covariance matrices. It basically selects accordingly the maximum variances of the corresponding covariance matrices (equations (5.6)).

5.2.2 Experiments

In order to investigate and visualise the different mixture covariance approaches considered, two experiments using the well known ORL (described in the subsection 4.1.3.1) and FERET face databases were performed. The FERET images pose an alternative analysis where the faces are better framed than the ORL ones.

5.2.2.1 FERET Face Database

The FERET database is the United States Army Face Recognition Technology facial database that has become the standard data set for benchmark studies [PWH98]. Sets containing 4 “frontal b series” images for each of 200 total subjects were considered. Each image set is composed of a regular facial expression (referred as “ba” images in the FERET database), an alternative expression (“bj” images), and two symmetric images (“be” and “bf” images) taken with the intention of investigating the effects of 15 degrees of pose angle variation. Figure 5.1 illustrates some example images from the FERET database cropped to the size of 96x64 pixels.

Figure 5.1. Some example images from the FERET Database.

5.2.2.2 Implementation

The experiments were implemented as follows. First the face images from the original vector space are projected to a lower dimensional space (face subspace) using Principal Component Analysis (PCA) [KS90, TP91] and then classified using the pooled covariance matrix and the three mixture covariance approaches described in the previous sub-section 5.2.1. Each experiment was repeated 25 times using several eigenfaces. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate, as well as the mean of the likelihood and classification accuracy mixture parameters, were calculated. Then, based on the best classification accuracy of the several PCA features used, the number of eigenfaces to visualise and calculate the covariance eigenvectors and eigenvalues on the face subspace was determined.

The ORL face experiments were computed using for each individual 5 images to train and 5 images to test. The FERET training and test sets were composed of 3 and 1 images respectively. Since in all applications the same number of training examples per subject was considered, the prior probabilities were assumed equal for all classes and recognition tasks. For implementation convenience, all ORL and FERET images were first resized to

64x64 and 96x64 pixels. The mixture parameter range was taken to be [0.1,0.2,...,1.0] for both w_i likelihood and w classification accuracy optimisations.

As can be seen from tables 5.1 and 5.2, the best classification results were obtained by using 40 and 50 eigenfaces (which we call most effective eigenfaces) for the ORL and FERET databases. These most effective eigenfaces correspond to approximately 82% and 83% of the total sample variance explained by the principal components transformation matrices of ORL and FERET face images respectively.

Eigenfaces	Pooled	Smix - ML	Smix - MC	Smix - ME
10	88.4% (1.4%)	91.9% (1.6%)	93.8% (1.7%)	93.5% (1.5%)
20	91.8% (1.8%)	94.4% (1.7%)	94.7% (1.4%)	95.2% (1.8%)
40	95.4% (1.5%)	96.2% (1.5%)	96.5% (1.6%)	96.7% (1.5%)
60	95.0% (1.6%)	95.7% (1.5%)	95.4% (1.6%)	95.9% (1.6%)
80	94.6% (1.9%)	94.9% (1.7%)	94.7% (1.9%)	94.8% (1.7%)

Table 5.1. ORL classification results

Eigenfaces	Pooled	Smix - ML	Smix - MC	Smix - ME
10	94.9% (1.1%)	94.7% (1.4%)	95.3% (1.1%)	95.3% (1.2%)
30	96.8% (0.8%)	96.6% (1.1%)	97.0% (0.9%)	97.2% (1.0%)
50	96.9% (0.8%)	96.7% (1.1%)	97.3% (1.0%)	97.8% (0.9%)
70	96.7% (0.9%)	96.5% (0.9%)	96.9% (0.9%)	97.3% (0.9%)

Table 5.2. FERET classification results

5.2.3 Results

Figures 5.2 and 5.3 present the visual analysis of two examples of each ORL and FERET covariance blending experiments using the most effective eigenfaces of each. These examples were chosen based on the closeness of the likelihood and classification mixture parameters to their respective mean values.

The visual results of Figures 5.2a-5.2b, and 5.3a-5.3b, can be described as follows. The first image row corresponds to the training images of a specific subject, and the second and third following rows correspond to the eigenvectors (in descending ordering of eigenvalues, from left to right) of the respective sample group and pooled covariance matrices transformed back to the image space by using the corresponding principal components transformation matrix. Accordingly, the fourth, fifth and sixth image rows correspond to the eigenvectors of the maximum likelihood, maximum classification accuracy, and maximum entropy mixture covariance matrices. The numbers below each image row describe the magnitude of the eigenvalue of each covariance eigenvector with its corresponding percentage of total variance shown in parentheses.

Since only 5 images of each individual were used to form the ORL training set, the results of Figures 5.2a and 5.2b relative to the sample group covariance estimates were limited, in terms of total variation within the subject's images, to the first 4 eigenvectors. The remaining eigenvectors (only 4 more shown) are arbitrary, apart from being constrained by the orthogonality assumption on the face subspace, and should be replaced or modified using the pooled information. In Figures 5.2a and 5.2b, the likelihood mixture parameters are respectively 0.9 (with mean value 0.92) and 1.0 (0.83), whereas the classification accuracy mixture parameter is 0.8 (0.61).

As can be seen on Figures 5.2a and 5.2b, the mixture covariance matrices that preserve as much of the sample group covariance information as possible were the ones blended using the maximum entropy approach. It is important to note that although the percentage of total variation of each eigenvalue was different due to the use of the pooled information, the first eigenvectors and eigenvalues of the maximum entropy covariance matrices are quite similar to those of the respective sample group covariance matrix.

Figures 5.3a and 5.3b show the results of the FERET experiments. Analogously to the ORL experiments, the sample group covariance information became limited to the first 2 eigenvectors. The remaining eigenvectors (only 3 more shown) represent no subject variation at all and are arbitrary, apart from being constrained by the orthogonality assumption on the face subspace. In these FERET figures, the likelihood mixture parameters are respectively 0.5 (with mean value 0.62) and 0.9 (0.95), whereas the classification accuracy mixture parameter is 0.8 (0.85).

As can be observed in Figures 5.3a and 5.3b, the visual results of the mixture covariance estimates seem to be more related to the pooled information than the sample group one. Again, the maximum entropy approach was the one that preserved as much of the sample group covariance information in the covariance matrix blending as possible. However, in this application where the face images are well framed thus favouring the pooled covariance matrix, there is no significant visual improvement in using mixture covariance matrices.

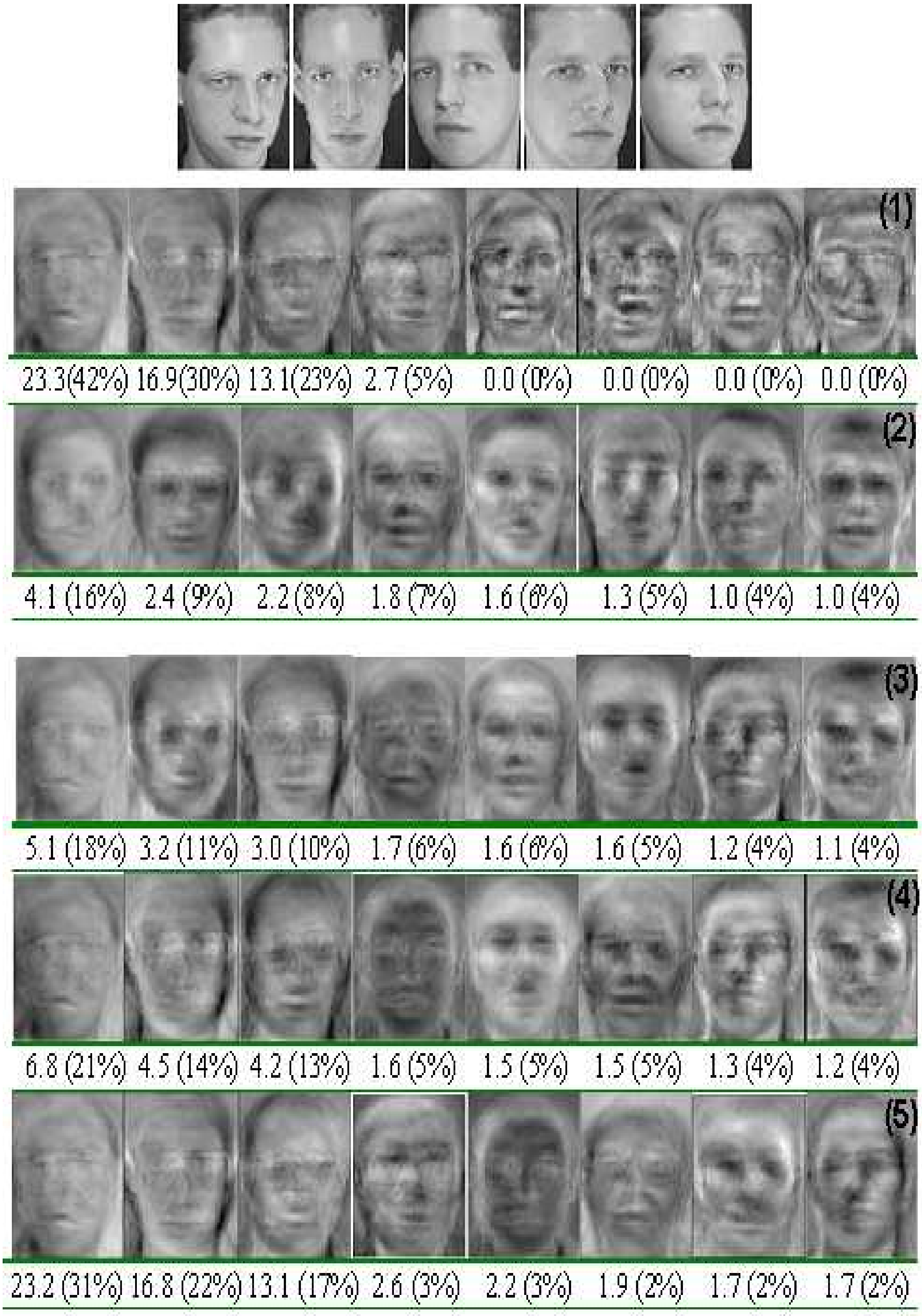


Figure 5.2a. ORL visual analysis. The top row shows the 5 image training examples of a subject and the subsequent rows show the image eigenvectors (with the corresponding eigenvalues below) of the following covariance matrices: (1) sample group; (2) pooled; (3) maximum likelihood mixture; (4) maximum classification mixture; and (5) maximum entropy mixture.

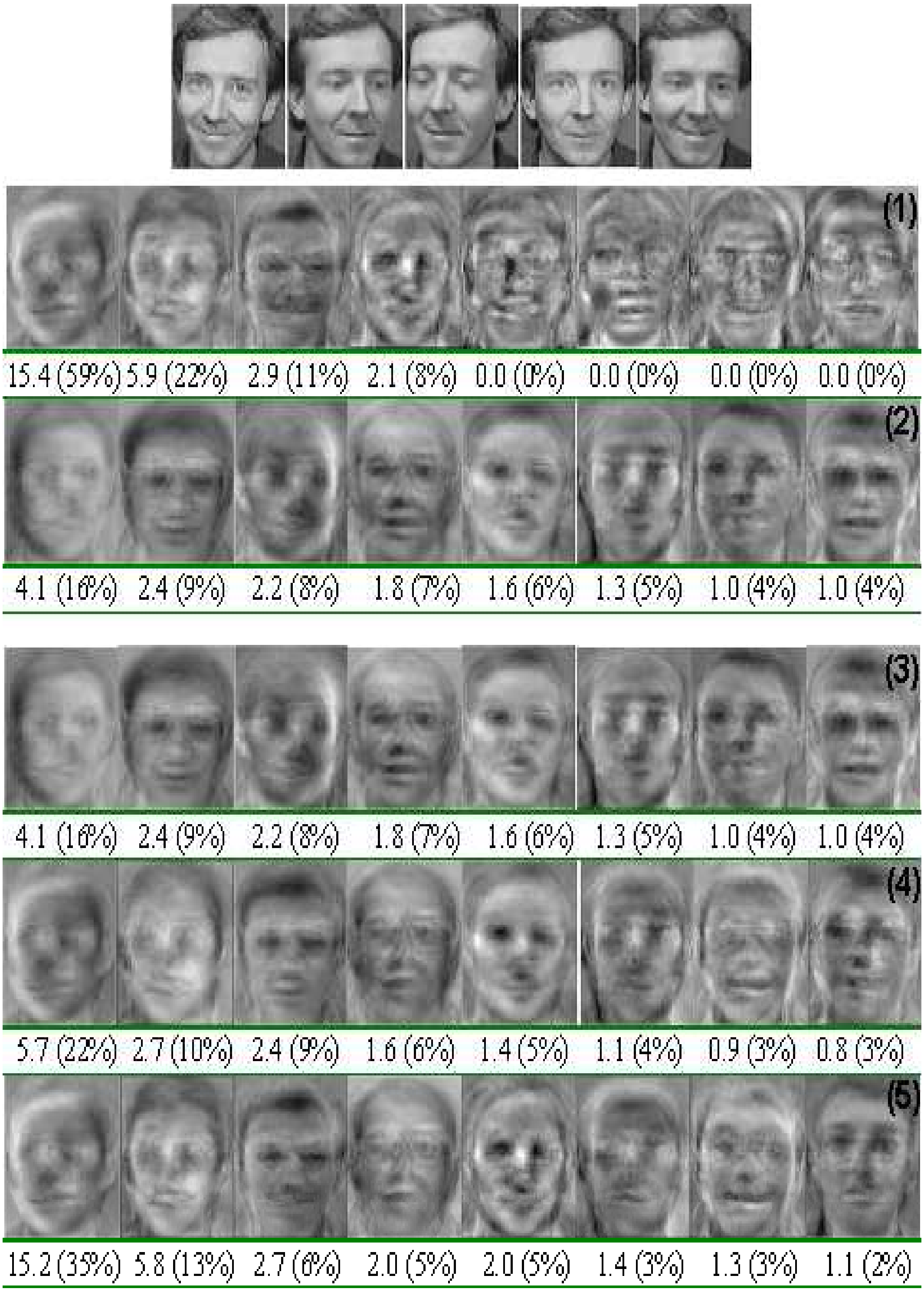


Figure 5.2b. ORL visual analysis. The top row shows the 5 image training examples of a subject and the subsequent rows show the image eigenvectors (with the corresponding eigenvalues below) of the following covariance matrices: (1) sample group; (2) pooled; (3) maximum likelihood mixture; (4) maximum classification mixture; and (5) maximum entropy mixture.

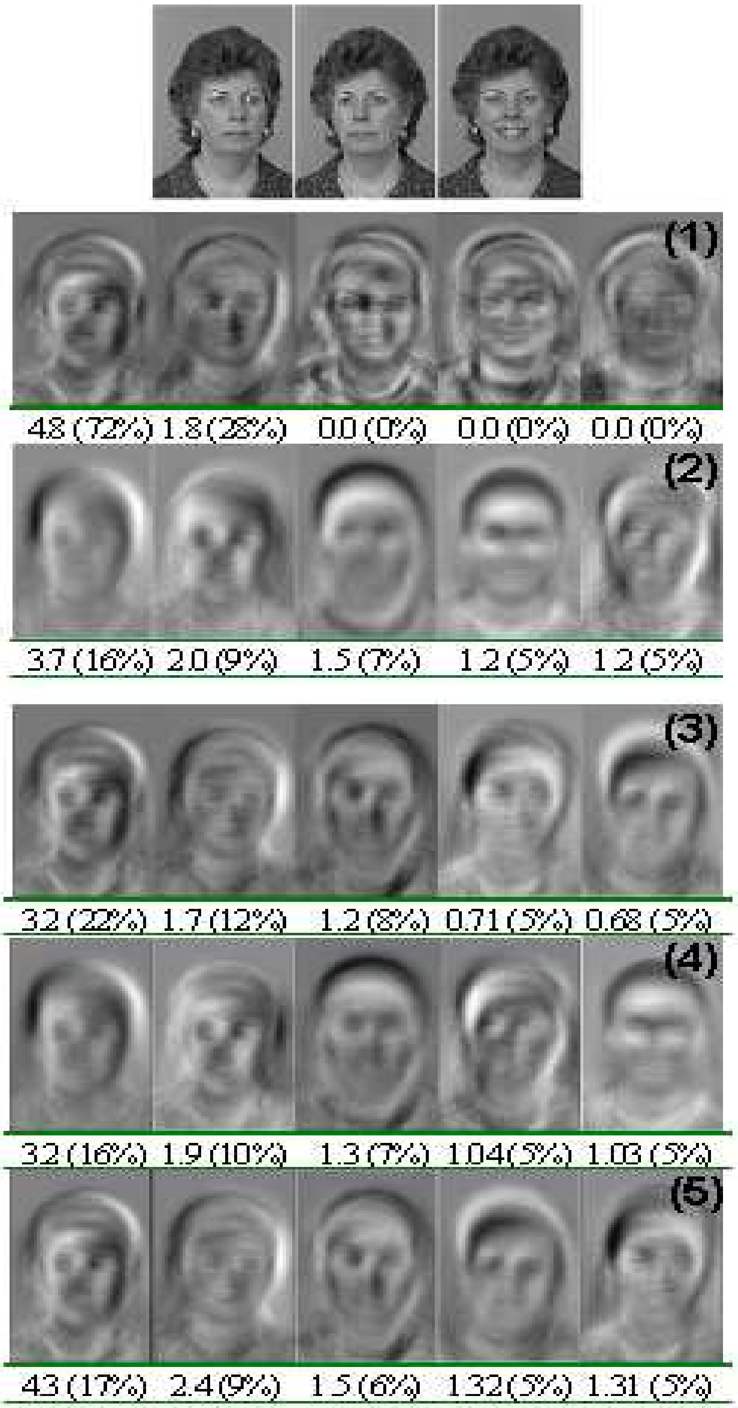


Figure 5.3a. FERET visual analysis. The top row shows the 3 image training examples of a subject and the subsequent rows show the image eigenvectors (with the corresponding eigenvalues below) of the following covariance matrices: (1) sample group; (2) pooled; (3) maximum likelihood mixture; (4) maximum classification mixture; and (5) maximum entropy mixture.

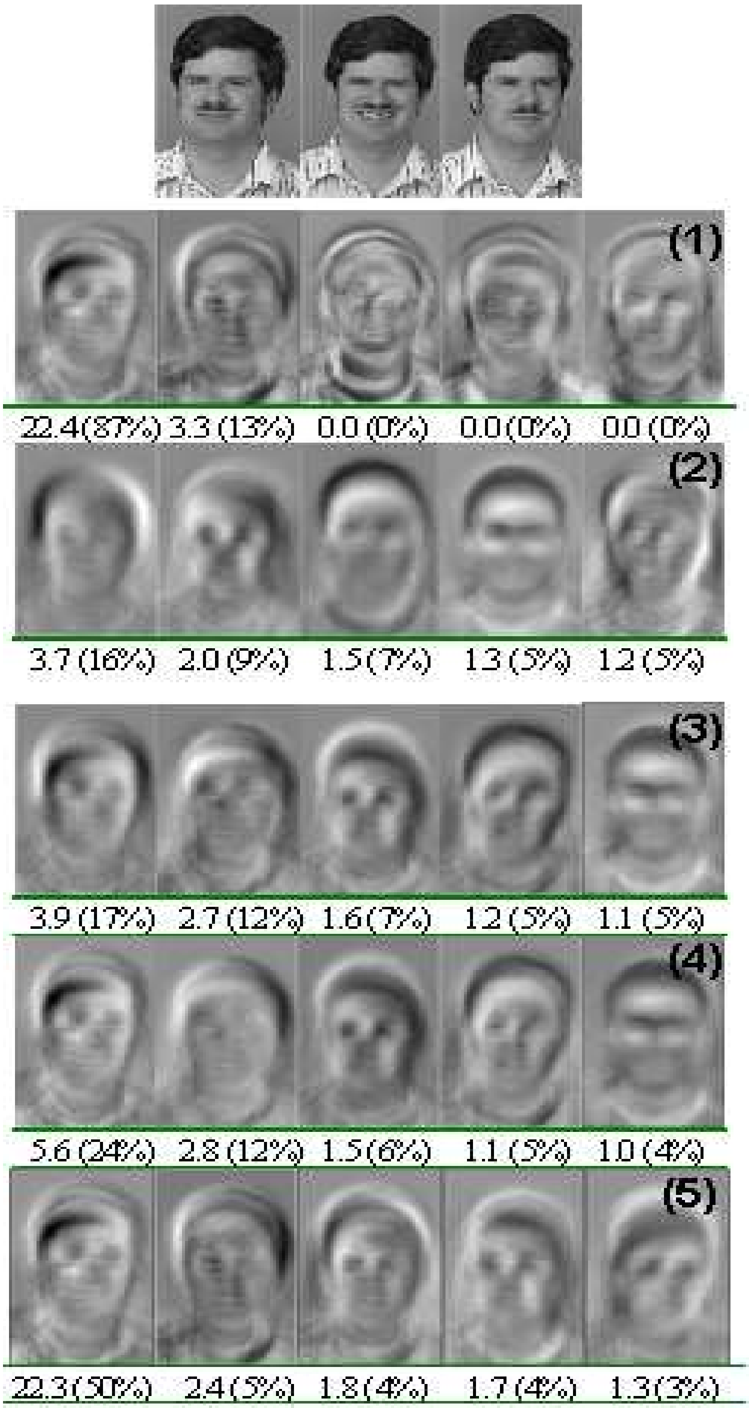


Figure 5.3b. FERET visual analysis. The top row shows the 3 image training examples of a subject and the subsequent rows show the image eigenvectors (with the corresponding eigenvalues below) of the following covariance matrices: (1) sample group; (2) pooled; (3) maximum likelihood mixture; (4) maximum classification mixture; and (5) maximum entropy mixture.

5.2.4 Discussion

A visual study of three mixture covariance matrix approaches for the QDF classifier has been undertaken in the context of characterising human faces. This analysis allows a better understanding of not only the final shape of such covariance matrix estimates, but also the importance and applicability of blending the sample group and the pooled covariance matrices in small sample size, high-dimensional problems.

The experiments performed in this section show that the maximum entropy approach preserves as well of the sample group covariance information as possible, achieving a more intuitive visual performance. This behaviour was especially identified in the ORL face experiments where moderate changes in facial expressions, pose, and scale, occurred.

In order to explore and understand the full scope of the MECS approach in limited sample and high dimensional problems, experiments on synthetic and other real image data are described in the following sections.

5.3 Synthetic Data Analysis

The main idea of the synthetic data analysis is to evaluate the effectiveness of the maximum entropy quadratic classifier and possibly predict situations where we might expect improvement with MECS compared to the other similar approaches aforementioned.

We have used the classification error as the evaluation criterion to compare the performance of the competing covariance estimators. This, also called "error-counting procedure" [Fuk90], is the only feasible possibility when a finite number of samples is available in practice [MYT87, PJY88, Fri89, Fuk90, HL96].

5.3.1 Estimation of the Bayes Classification Error

According to Fukunaga [Fuk90], the Bayes classification error of a sample-based estimate can be determined by a function of two sets of data, that is $\varepsilon(\Omega_D,\Omega_T)$ where Ω_D and Ω_T are respectively the design (or training) and test sets.

Fukunaga [Fuk90] has shown that the upper and lower expected values of the Bayes classification error of the true population Ω can be calculated by the following estimated bounds

$$E\left\{\varepsilon(\hat{\Omega}_D,\hat{\Omega}_D)\right\} \le \varepsilon(\Omega,\Omega) \le E_{\hat{\Omega}_T}\left\{\varepsilon(\hat{\Omega}_D,\hat{\Omega}_T)\right\}, \tag{5.8}$$

where $\hat{\Omega}_T$ is a sample generated from Ω independently of $\hat{\Omega}_D$.

The formula (5.8) expresses that on the one hand, the rightmost term $\varepsilon(\hat{\Omega}_D,\hat{\Omega}_T)$ of equation (5.8) is calculated by generating two different sample sets, $\hat{\Omega}_D$ and $\hat{\Omega}_T$, and using $\hat{\Omega}_D$ for training and $\hat{\Omega}_T$ for testing. This procedure is called the holdout method and its expected value gives the upper bound of the Bayes classification error. On the other hand, the leftmost term $\varepsilon(\hat{\Omega}_D,\hat{\Omega}_D)$ is obtained by using the same sample data for training and testing and its expected value gives the lower bound of the Bayes error. This procedure is called the re-substitution method [Fuk90].

It is common practice to replace the expectation values of the formula (5.8) by the average over the available samples randomly replicated.

5.3.2 Experiments

We have used the holdout and re-substitution methods to estimate respectively the upper and lower bounds of the classification errors of Bayes Plug-in classifiers.

Analogously to the synthetic analyses developed by other researchers [MYT87, PJY88, Fri89, HL96], we have implemented simulation experiments conducted as an n-multivariate normal repeated-measures design. Our attention is focused on evaluating covariance estimators while varying the dimension of the space n, the degree of similarity of the covariance matrices, and, particularly, the intra-class correlation ρ of all the groups considered. We believe that the correlation between the n parameters can play an important role when blending covariance-matrix estimates in limited sample and high dimensional problems. Other simulation experiments for combining the sample group and pooled covariance matrices, such as RDA [Fri89] and LOOC [Hof95, HL96] synthetic analyses, have not addressed this problem.

In all simulation experiments there are 9 groups or classes. We have chosen four values for the dimension of the space n (5, 10, 20, 40) and three values for the intra-class correlation factor ρ (0.0, 0.1, 0.9). Since the training sample sizes of all classes for all experiments are fixed at $N_1 = N_2 = \cdots = N_9 = 20$, those four n values would allow us to analyse the Bayes Plug-in classifiers in situations where the sample group covariance matrices are non-singular (n= 5 or 10) and singular (n= 20 or 40). The three correlation factors represent the situations where the intra-class data parameters are not correlated (ρ = 0.0), slightly correlated (ρ = 0.1), and highly correlated (ρ = 0.9).

For each pair of (n,ρ) values, the simulation consists of 25 replications of the following procedure. For each multivariate normal population of the 9 classes, we have generated 20 n-dimensional training observations and 50 n-dimensional test observations. From these distinct training and test samples randomly drawn, we have calculated the lower and upper classification rate of the following six Bayes Plug-in classifiers: Euclidean distance classifier or simply EUC (i.e., covariance matrices equal to the identity matrix), QDF (covariance matrices defined in equation (3.6)), LDF (equation (3.12)), RDA (equation 3.13)), LOOC (equation (3.27)), and MECS (equations (5.6)).

Since the same number of training examples per class is considered, the prior probabilities are assumed equal for all classes and experiments. Different forms for the 9 true mean vectors μ are selected, namely: $\mu_1 = [0,0,\ldots,0]^T$, $\mu_2 = [1,0,\ldots,\mathrm{rem}(n,2)]^T$, $\mu_3 = [0,1,\ldots,\mathrm{rem}((n+1),2)]^T$, $\mu_4 = [1,1,\ldots,1]^T$, $\mu_5 = [-1,1,\ldots,(-1)^n]^T$, $\mu_6 = -\mu_2$, $\mu_7 = -\mu_3$, $\mu_8 = -\mu_4$, and $\mu_9 = -\mu_5$, where "$\mathrm{rem}(y_1, y_2)$" is simply the remainder after the division of the value y_1 by the value y_2. These values simulate practical recognition problems where the mean differences are not restricted to a specific subspace but increase as more features are used. Moreover, the RDA optimisation grid has been taken to be the outer product of $\lambda = [0,0.125,0.354,0.65,1.0]$ and $\gamma = [0,0.25,0.5,0.75,1.0]$, as suggested by Friedman's work [Fri89]. Analogously, the size of the LOOC parameter was $\alpha_i = [0,0.25,0.5,\ldots,2.75,3.0]$, as given by [Hof95, HL96].

5.3.3 Results

The results of the different covariance structures used to evaluate the six Bayes Plug-in classifiers are presented in the following sub-sections.

All the corresponding figures can be briefly described as follows. The high-low charts present the mean of the training (upper bound) and test (lower bound) classification rates over the 25 replications regarding the four values used for the dimension parameter n (5, 10, 20, 40) and the three values for the intra-class correlation factor ρ (0.0, 0.1, 0.9). The mark "-" in between the high-low bars is illustrative and represents simply the average value between the upper and lower Bayes classification bounds.

Also presented in separate figures are the mean of the selected RDA and LOOC mixing parameters for each covariance structure with correlation factor ρ equals 0.1 and 0.9. As a reminder, RDA parameter lambda λ controls the degree of shrinkage of the sample group covariance matrix estimates toward the pooled covariance one, whereas the parameter gamma γ controls the shrinkage toward a multiple of the identity matrix. Analogously, the LOOC parameter alpha α_i in between 0 and 1 leads to mixtures of the diagonal of the sample group covariance and sample group covariance itself, in between 1 and 2 to mixtures of the sample group covariance and the common covariance (which in this case is equal to the pooled one), and in between 2 and 3 to mixtures of the common covariance and the matrix of its diagonal elements.

5.3.3.1 Equal Spherical Covariance Matrices

In this simulation, each of the $i = 1,2,\ldots,9$ classes was generated from a population with the true covariance matrix chosen to have the form of a pure intra-class correlation matrix. That is, all the true covariance matrices Σ_i are given by

$$\Sigma_i(\rho) = (1-\rho)I + \rho\mathbf{1}, \tag{5.9}$$

where I is a $n \text{x} n$ identity matrix and **1** is $n \text{x} n$ matrix of 1's [MYT87, PJY88]. The classification results of this simulation are shown in Figure 5.4. Figure 5.5 displays the means of the RDA and LOOC mixing parameters selected for this covariance structure.

As can be seen in the classification charts of Figure 5.4, the differences between the covariance estimators become more apparent as the dimension parameter n increases relative to the sample size. Since there were only 20 training observations for each class, the sample group covariance matrices were singular for n= 20 or 40 and the QDF classification rate could not be computed. The success of the EUC when the correlation factor was 0.0 is theoretically expected because in this situation the true covariance matrices of all the classes were equal to the identity matrix. Not surprisingly, when the intra-class data parameters became slightly correlated (ρ = 0.1), and, more significantly, highly correlated (ρ = 0.9), the EUC classification results deteriorated.

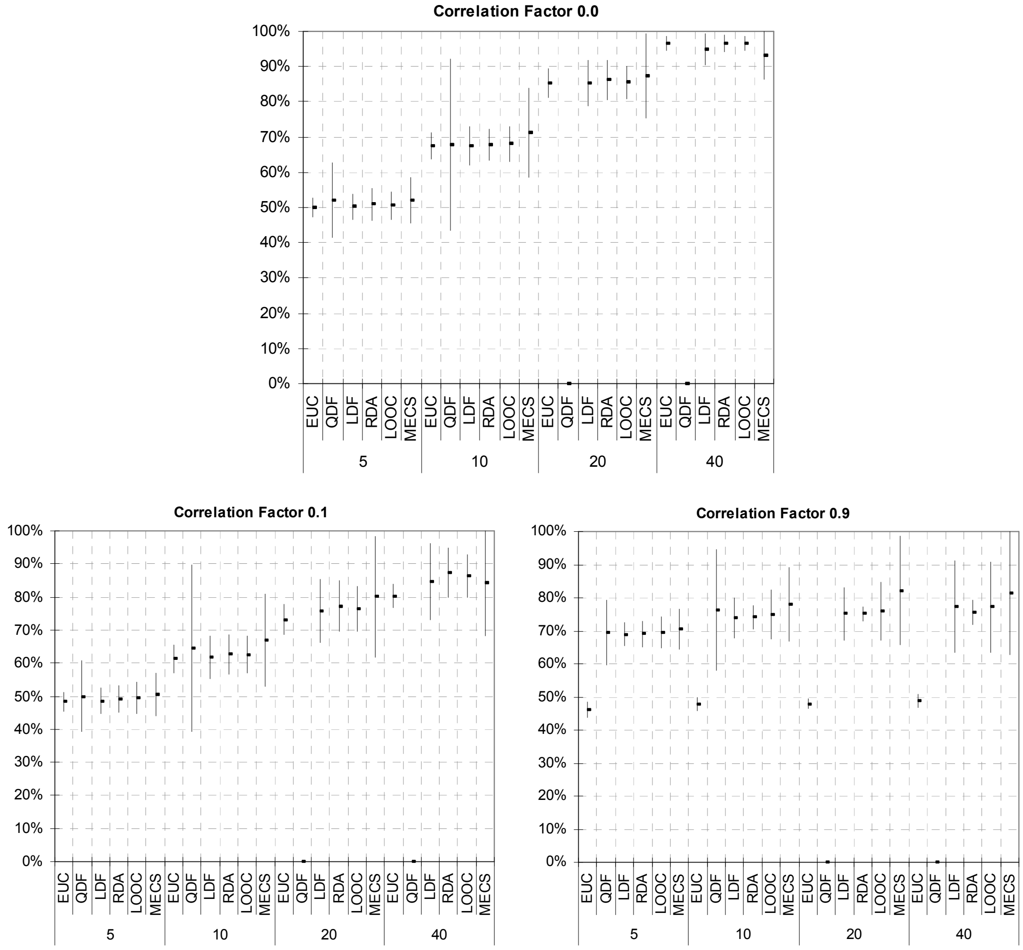


Figure 5.4. Equal spherical covariance matrices – QDF classification results.

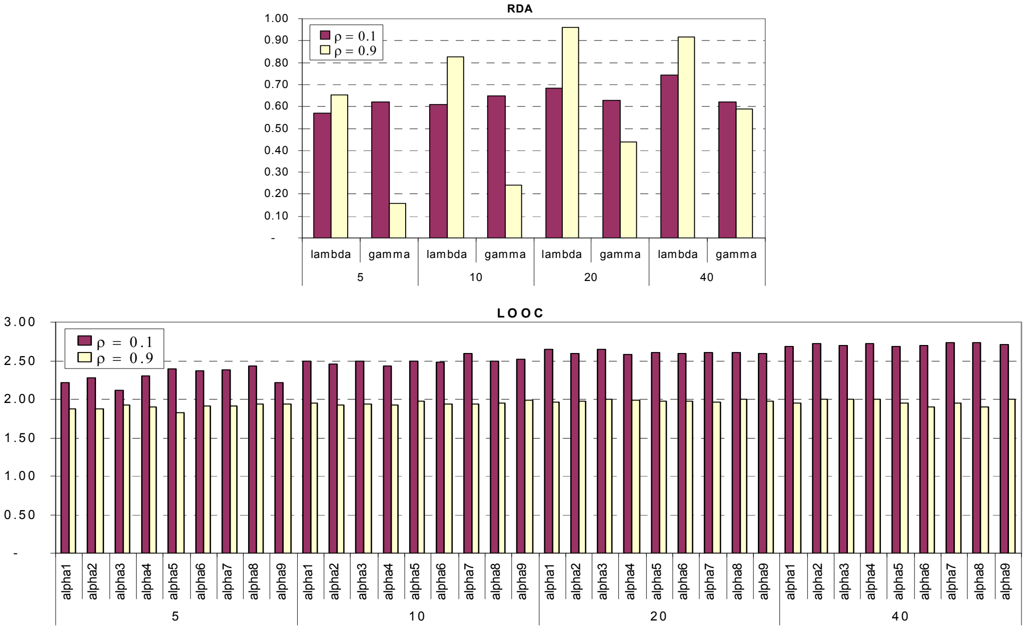


Figure 5.5. Equal spherical covariance matrices – RDA and LOOC parameters.

On the training set (or upper part of the classification bounds) the sample group (when invertible) and the MECS covariance estimators led to higher recognition accuracies than the pooled (LDF) and both optimised RDA and LOOC covariance estimators. For the test samples (or lower part of the classification bounds), MECS performed slightly worse than the LDF, RDA, and LOOC in the lowest dimensional space, but its relative performance deteriorated when the dimensionality increased, particularly in situations where the data parameters were slightly correlated (ρ = 0.1) or not correlated at all (ρ = 0.0). This result suggests that MECS depends on the relative size of the dimension of the space to the total training sample size when the true covariance matrices have essentially a diagonal form.

When data parameters were highly correlated (ρ = 0.9) and consequently the estimation of the off-diagonal elements (co-variances) of the true covariance matrices becomes as important as the estimation of the diagonal elements (variances), MECS performed equivalently in the higher dimensional spaces to LDF and LOOC. In this situation, where a high degree of regularisation for the RDA identity shrinkage parameter gamma was selected (see top chart of Figure 5.5), RDA achieved the best classification rates for the test samples. In contrast, RDA recognition accuracies of the training samples were much lower than the other covariance estimators.

Figure 5.5 illustrates the mean of the RDA and LOOC mixing parameters selected for the equal spherical covariance structures. As we should expect, when the data parameters were slightly correlated (ρ = 0.1), the RDA and LOOC methods selected their appropriate and available diagonal forms to better estimate the on-diagonal (variances) elements of the true covariance matrices. However, this strategy is undermined when the true covariance matrices are calculated based on highly correlated class-sample data.

5.3.3.2 Equal Ellipsoidal Covariance Matrices

In order to simulate equal ellipsoidal covariance matrices, each of the $i = 1,2,\ldots,9$ classes was generated from a population with the true covariance matrix chosen to have the following form:

$$\Sigma_i(\rho) = D^{1/2} R(\rho) D^{1/2} , \qquad \textbf{(5.10)}$$

where the matrix $R(\rho)$ is the $n \text{x} n$ intra-class correlation matrix given by

$$R(\rho) = (1-\rho)I + \rho\mathbf{1}\,, \tag{5.11}$$

and D is the diagonal $n \text{x} n$ sample variance matrix with exponential decrease values defined as

$$D = \begin{bmatrix} \exp\left(1/1\right) & 0 & \cdots & 0 \\ 0 & \exp\left(1/2\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp\left(1/n\right) \end{bmatrix}. \tag{5.12}$$

The exponential decrease described in equation (5.12) simulates practical situations where a statistical multivariate technique, such as PCA, is applied first to reduce the dimensionality of the original data and improve the recognition rate of the Bayes Plug-in classifiers in limited sample size problems.

Figure 5.6 displays the classification results of this simulation. The equal ellipsoidal covariance results were similar to the equal spherical ones described in the previous sub-section. In the same manner, when the data parameters were not correlated (ρ = 0.0), the EUC classifier performed well owing to the rapid convergence of the diagonal elements of the true covariance matrices to 1 in higher dimensional spaces.

Regarding the training set (or upper part of the classification bounds), the QDF (when computable) and the MECS covariance estimators led, analogously, to higher recognition accuracies than the other ones. For the test samples (or lower part of the classification bounds), the classification performance of MECS compared to the other covariance estimators was slightly better than the one obtained when simulating equal spherical covariance matrices in lower dimensional spaces. However, likewise, MECS relative performance deteriorated when the dimensionality increased, particularly in the situations where the true covariance matrices have essentially a diagonal form (ρ = 0.0 or 0.1).

When data parameters were highly correlated (ρ = 0.9), RDA's previous superior performance on classifying test samples of equal spherical covariance structures was minimised because the blending toward a multiple of the identity matrix was not the most appropriate in this ellipsoidal case (see top chart of Figure 5.7).

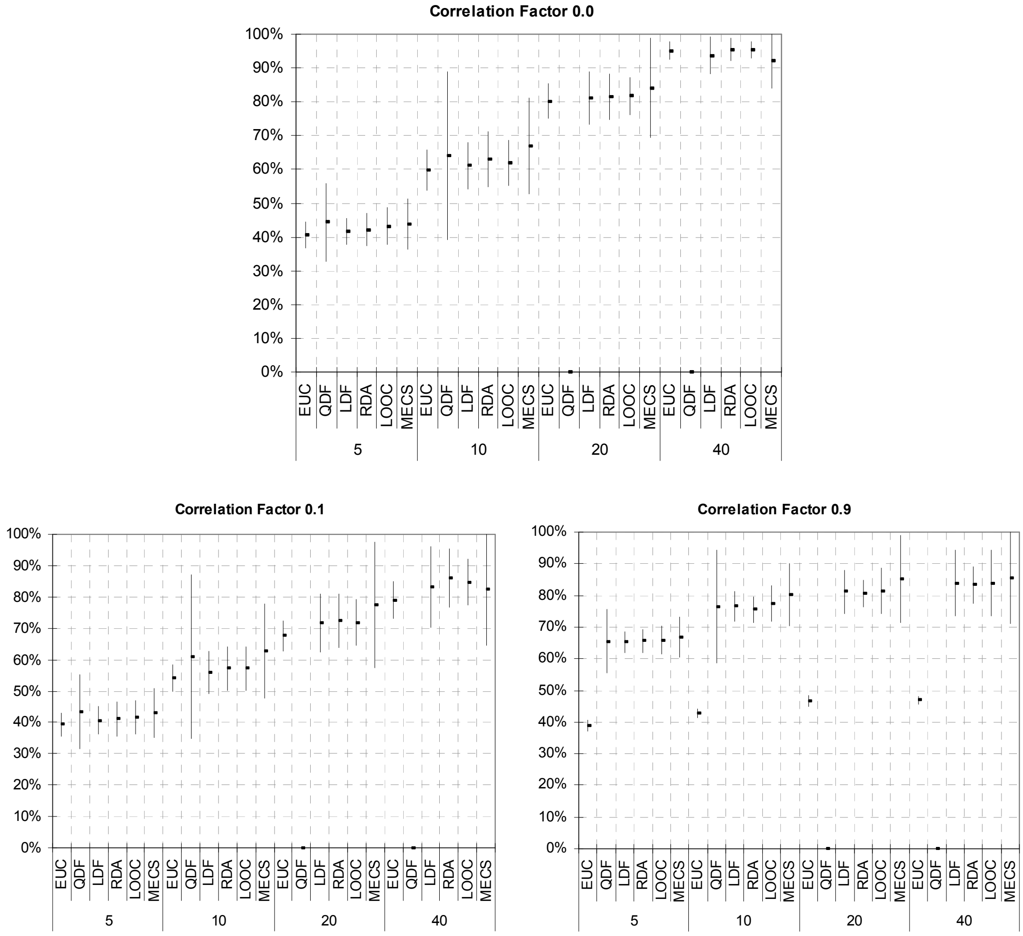


Figure 5.6. Equal ellipsoidal covariance matrices – QDF classification results.

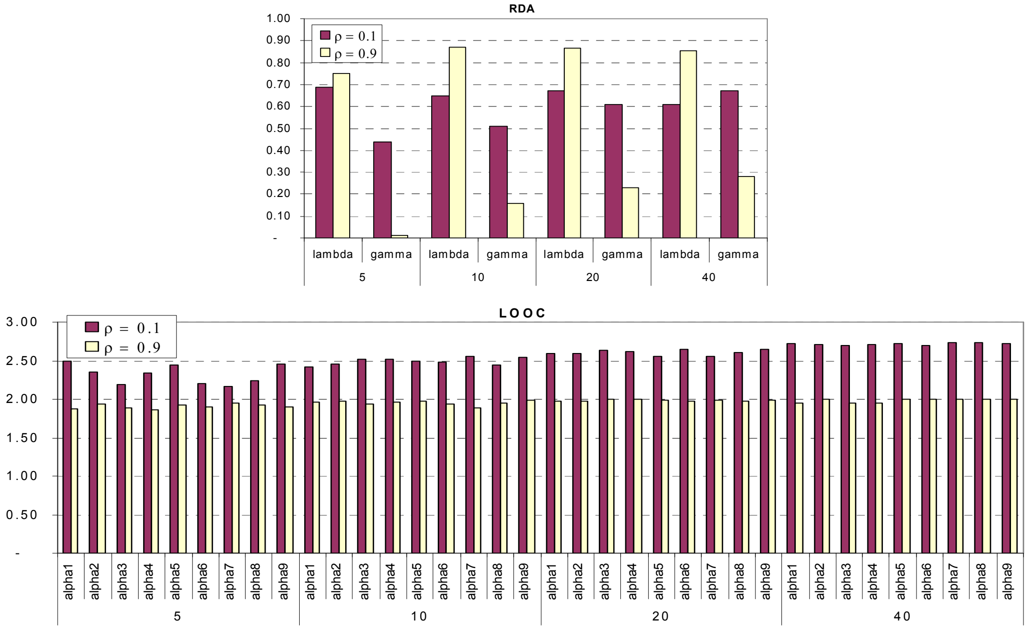


Figure 5.7. Equal ellipsoidal covariance matrices – RDA and LOOC parameters.

According to the charts of Figure 5.7, RDA and LOOC approaches depended more on the sample and pooled information when the data were highly correlated (ρ = 0.9) rather than slightly correlated (ρ = 0.1). This result was not surprising because when the intra-class data parameters were highly correlated the estimation of the off-diagonal elements (co-variances) of the true covariance matrices becomes as important as the estimation of the diagonal elements (variances), favouring mixtures of covariance estimations that have both types of elements.

5.3.3.3 Unequal Ellipsoidal Covariance Matrices

In this synthetic experiment, each of the $i = 1,2,\ldots,9$ classes was generated from a population with the true covariance matrix chosen to have the following unequal ellipsoidal covariance structure:

$$\Sigma_i(\rho) = \left[\frac{i}{3}D\right]^{1/2} R(\rho) \left[\frac{i}{3}D\right]^{1/2}, \qquad \textbf{(5.13)}$$

where $R(\rho)$ and D matrices are respectively the $n \,\mathrm{x}\, n$ intra-class correlation and diagonal matrices defined in the previous equations (5.11) and (5.12). This is a situation that ought to prove difficult for MECS because the true covariance matrices are significantly different from each other.

Figure 5.8 displays the classification results of this synthetic design. As can be seen, the differences between the covariance estimators are apparent not only in the high dimensional space, like the previous simulations, but also in the lower n dimensions.

When the data parameters were not correlated (ρ = 0.0) or slightly correlated (ρ = 0.1), and the dimension of the space was lower (n = 5 or 10), MECS performed on the test samples better than the QDF and LDF classifiers and worse than the RDA and LOOC optimised approaches. In the higher dimensional spaces (n = 20 or 40) and at the same correlation settings that favour diagonal forms for the covariance estimates (ρ = 0.0 or 0.1), although MECS classification rates of the training samples were the highest ones, its relative performance on the test samples deteriorated compared to the RDA and LOOC approaches.

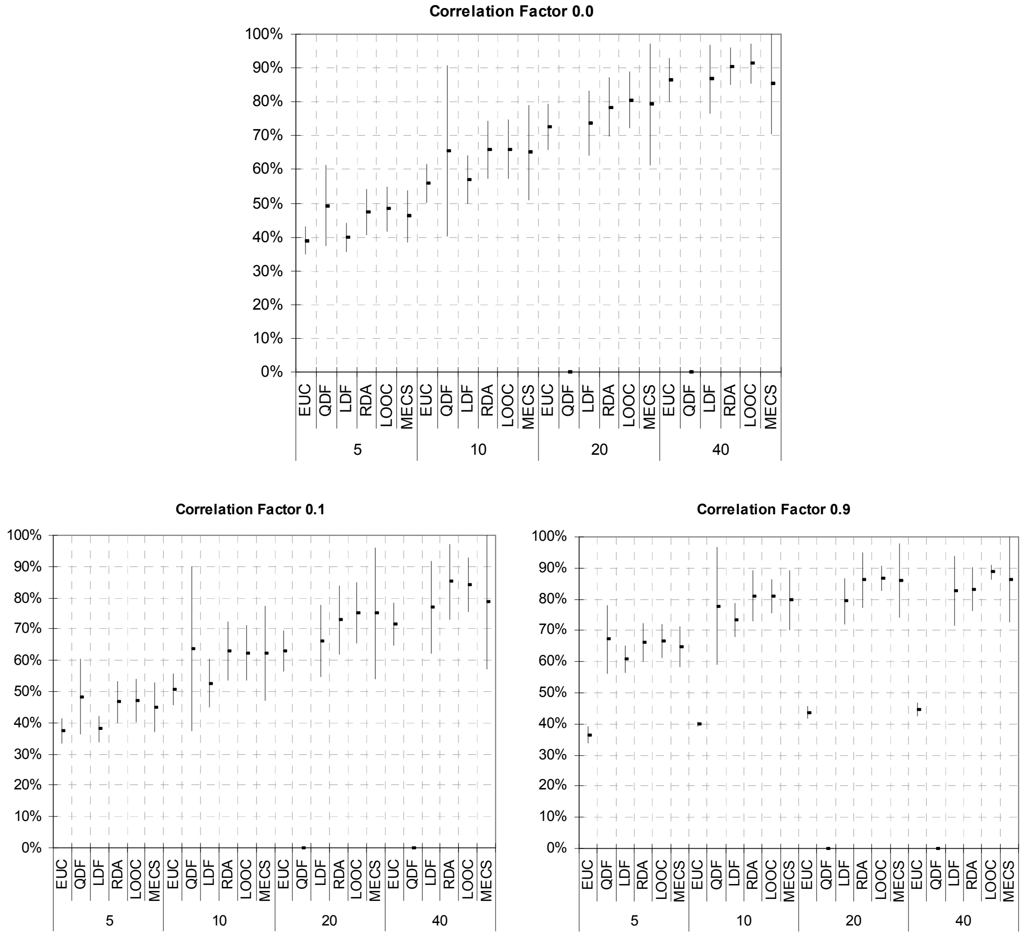


Figure 5.8. Unequal ellipsoidal covariance matrices – QDF classification results.

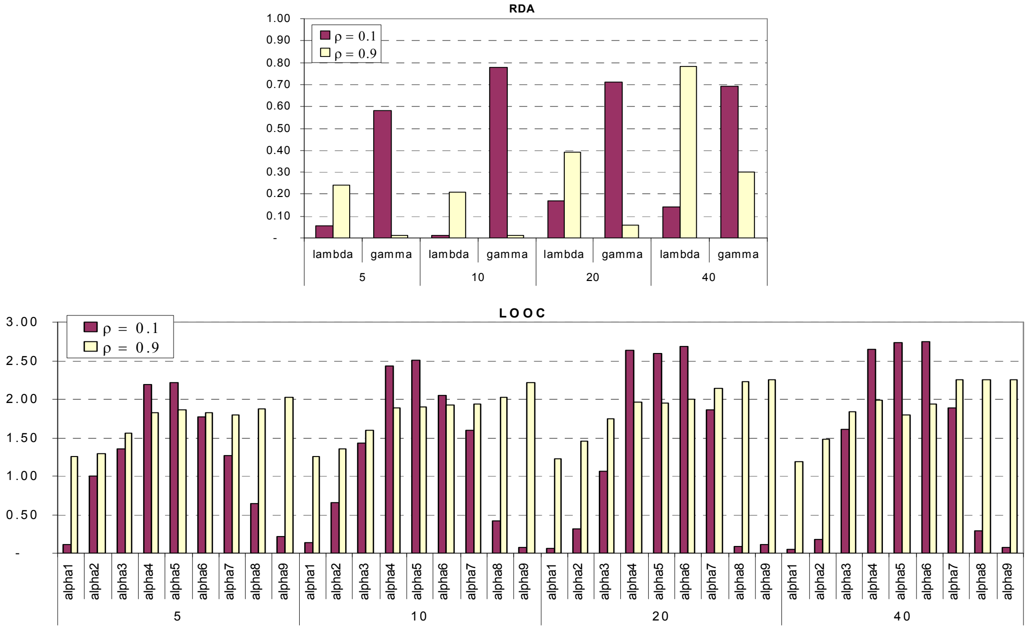


Figure 5.9. Unequal ellipsoidal covariance matrices – RDA and LOOC parameters.

In the situation where the intra-class data parameters were highly correlated (ρ = 0.9), the gap between the performance of MECS and the optimised approaches on classifying test samples was smaller than the ones obtained when the correlation factor was 0.0 or 0.1. In such situation of high data correlation (ρ = 0.9), LOOC achieved the best classification rates for the test samples. However, LOOC recognition accuracies of the training samples were lower than the other covariance estimators.

Figure 5.9 displays the mean of the RDA and LOOC mixing parameters selected for this covariance structure estimation. As can be seen, more drastically than the previous simulations, even when the true covariance matrices are unequal and ellipsoidal, but the intra-class data parameters are highly correlated, RDA and LOOC optimisation approaches relied mainly on the mixture between solely the sample and pooled covariance matrices.

5.3.4 Discussion

The simulation experiments presented in this section suggest that when the data parameters are highly correlated the MECS estimator performs as well as or better than the QDF (when computable) and the widely used LDF estimator in limited sample and high dimensional problems.

In the high correlation case, the MECS relative performance compared to the RDA and LOOC optimised approaches depends on the choice of the training samples. This is not a surprising result because the MECS procedure of blending the sample and pooled covariance matrices does not involve an optimisation procedure, but simply a selection of the most reliable information available. When the data parameters are not correlated or slightly correlated, there is little benefit to be derived from MECS and, as the experiments suggest, there is deterioration in performance when the true covariance matrices have essentially a diagonal form and the relative size of the total training sample size to the dimension of the space is small.

It would be impossible to analyse the performance of the competing estimators over all possible covariance structures. Analogously to the results obtained by other researchers [MYT87, PJY88, Fri89, HL96], these simulation experiments indicate that there is no overall optimality of any covariance estimator for all possible configurations. However, it is worthwhile to mention that the gain of computational simplicity of MECS compared to the RDA and LOOC covariance estimators is dramatic when several classes have to be discriminated. This property could be of great importance, especially in well-framed or pre-processed image recognition applications where only one limited set of highly correlated data is available.

5.4 Image Data Analysis

In order to investigate with real data the performance of MECS compared with QDF, LDF, RDA, and LOOC classifiers, four image based classification applications were considered: face recognition, facial expression recognition, fingerprint classification and optical character recognition (OCR).

In the face and facial expression recognition applications, the training sample sizes were chosen to be extremely small and small, respectively, compared to the dimension of the feature space. In contrast, moderate and large relative training sample sizes were considered for the fingerprint and OCR problems. All applications were analysed using publicly released databases.

5.4.1 Experiments

We used the same FERET and Facial Expression benchmark databases, previously described in the subsections 5.2.2.1 and 4.1.3.2, for the face and facial expression recognition experiments. The training and test feature files extracted from the NIST (US National Institute of Standards and Technology) special datasets were used in the fingerprint and OCR classification experiments. A brief description of these two feature datasets is provided in the following sub-sections.

5.4.1.1 Fingerprint Database

The fingerprint classification was performed utilising the training and test feature vectors extracted from the grey scale images of the standard NIST Special Database 4 [WCG92]. Each feature vector consists of 112 floating point numbers, made by a feature selection procedure that ends with the PCA transform.

The fingerprints were classified into one of five categories (arch, left loop, right loop, tented arch, and whorl) with an equal number of prints from each class (400). There are 2000 first-rolling fingerprint feature vectors for training and 2000 corresponding second-rolling ones for testing. Figure 5.10 illustrates some example images taken from the fingerprint database that have been displayed on the NIST Special Database 4 web site (http://www.nist.gov/srd/nistsd4.htm).

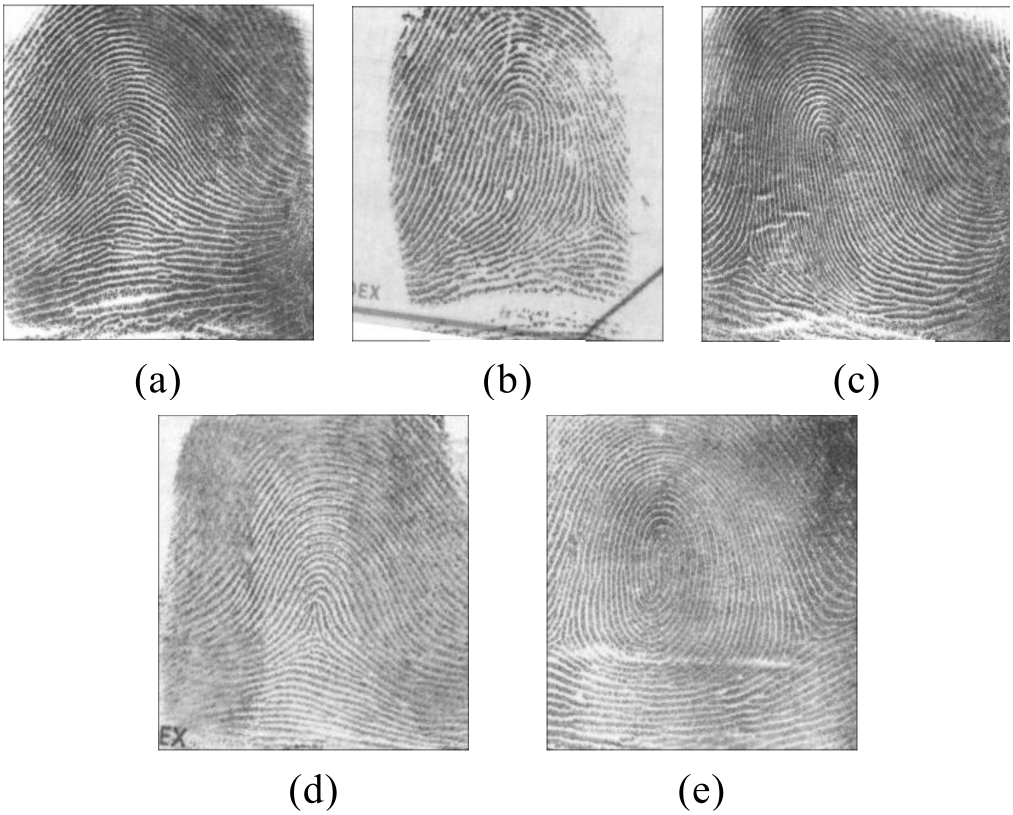


Figure 5.10. (a) Arch, (b) left loop, (c) right loop, (d) tented arch, and (e) whorl images.

5.4.1.2 OCR Database

In the OCR experiments we used the training and test feature files extracted from NIST Special Database 3 and NIST Special Database 7 respectively [BCG94]. Each feature vector consists of 96 floating point numbers made by a PCA transform on the normalized character image. Each original image is a 32 pixel square binary raster containing a hand printed numerical digit extracted from a document.

The characters were classified into one of the ten digits “0” - “9”. There are in total 7400 first-writing character feature vectors for training and 23140 second-writing ones for testing. Both training and test files contain equal numbers of prints for each digit.

5.4.1.3 Implementation

For implementation convenience, all FERET face images were first resized to 96x64 pixels and transformed into eigenfeature vectors using PCA [TP91]. Each experiment was repeated 25 times using several of those eigenfeatures. Distinct training and test samples were randomly drawn, and the mean of the recognition rate was calculated. Since the LOOC computation requires at least three examples in each class [Hof95, HL96], the recognition rate was computed utilising for each subject 3 images to train and 1 image to test.

Analogously to the face recognition experiments, first PCA reduces the dimensionality of the original Tohoku facial expression images (which were resized to 64x64 pixels for implementation convenience) and secondly the discriminant Bayes’ rule using the covariance estimators aforementioned were applied. Each experiment was repeated 25 times using several PCA dimensions. Distinct training and test sets were randomly drawn, and the mean of the recognition rate was calculated. The training and test sets were respectively composed of 20 and 9 images.

The fingerprint and OCR classifications were performed utilising the feature vector files extracted from the images of the corresponding NIST Special Databases [WCG92, BCG94]. The training/test files of the fingerprint and OCR were composed respectively of 400/400 and 740/2314 feature vectors with different number of floating numbers.

Since in all applications the same number of training examples per class was considered, the prior probabilities were assumed equal for all classes and recognition tasks. Again, the RDA optimisation grid has been taken to be the outer product of $\lambda = [0, 0.125, 0.354, 0.65, 1.0]$ and $\gamma = [0, 0.25, 0.5, 0.75, 1.0]$, as suggested by Friedman’s work [Fri89], and the size of the LOOC parameter was $\alpha_i = [0, 0.25, 0.5, \ldots, 2.75, 3.0]$, as given by Hoffbeck and Landgrebe [Hof95, HL96].

5.4.2 Results

Figure 5.11 presents the test average recognition error of the FERET face database. Since only 3 face images were used to train the classifiers, the sample group covariance matrices S_i were singular and the QDF could not be calculated. Instead, the recognition rate of the Euclidean distance classifier (EUC) that corresponds to the classical eigenfaces method proposed by Turk and Pentland [TP91] are displayed. Figure 5.11 shows that for all the feature components considered the MECS quadratic classifier performed as well or better than the other classifiers. The MECS quadratic classifier achieved the lowest classification error – 2.2% – on 50 eigenfeatures. In this application where each S_i seems to be quite similar, favouring the LDF performance, the MECS classifier did better by using the singular and pooled covariance information.

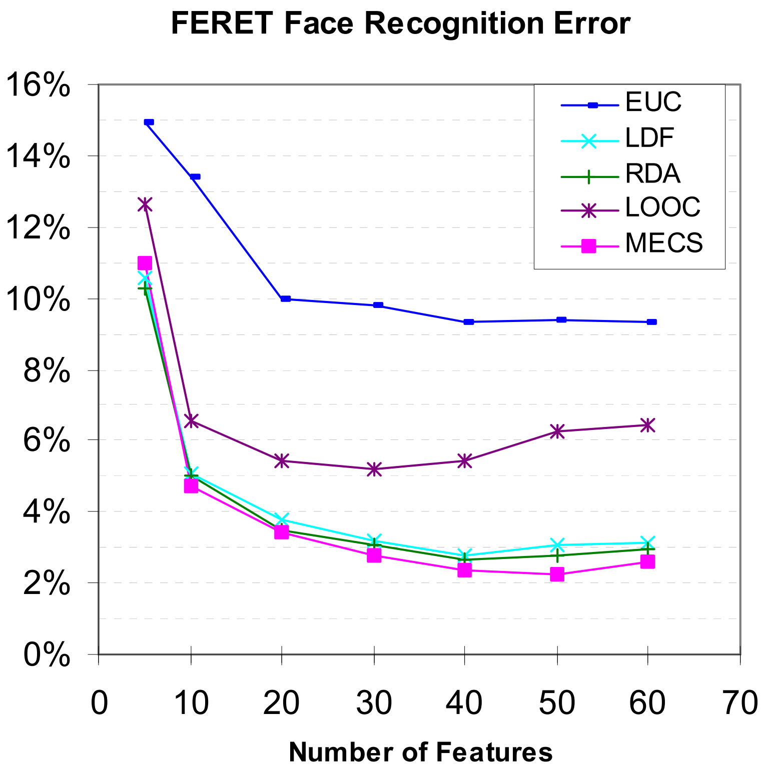


Figure 5.11. FERET face database recognition error for Bayes plug-in classifiers.

The results of the Tohoku facial expression recognition are presented in Figure 5.12. Owing to the fact that 20 images were used to form the training set of each class, the sample group covariance estimate (QDF) results were limited to 19 PCA features. These results and the EUC results were much less accurate than the others and were not plotted on Figure 5.12. As can be seen, there is no overall dominance of any covariance estimator. In lower dimension space, RDA led to lower classification error rates, followed by MECS, LDF and LOOC. When the dimensionality increased and the true covariance matrices became apparently equal and highly ellipsoidal, RDA performed poorly while MECS, LDF and LOOC improved. In the highest dimensional space the LOOC optimisation, which considers the diagonal elements of the pooled estimate, took advantage of the equal-ellipsoidal behaviour (for more than 70 PCAs all α_i parameters are closest to the value 3) achieving the lowest recognition error rate – 12.8% – for all PCA components calculated. In this recognition application, all the computed covariance estimators were quite sensitive to the choice of the training and test sets.

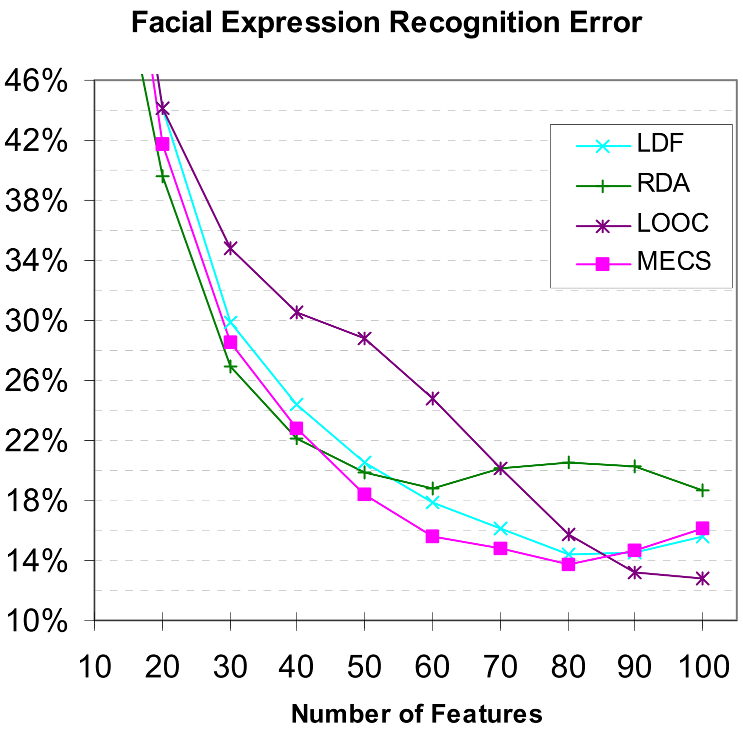


Figure 5.12. Tohoku facial expression recognition error for Bayes plug-in classifiers.

The recognition results of the NIST-4 fingerprint database are presented in Figure 5.13. The EUC results were much less accurate than the others and were not plotted on the figure. In the lowest and highest dimension spaces (28 and 112 features), RDA led to marginally lower classification error than MECS estimator. However, for 56 and 84 features the MECS performed better than the other classifiers. Although in this application the ratio of the training sample size to the number of features is moderate and large, favouring the QDF, RDA and LOOC classifiers, the MECS estimator achieved the lowest classification error – 12.5% – on 84 components. Putting this result in perspective, a classification error of 12% but with 10% rejection of the fingerprints was reported on the same training and test sets [WCG92].

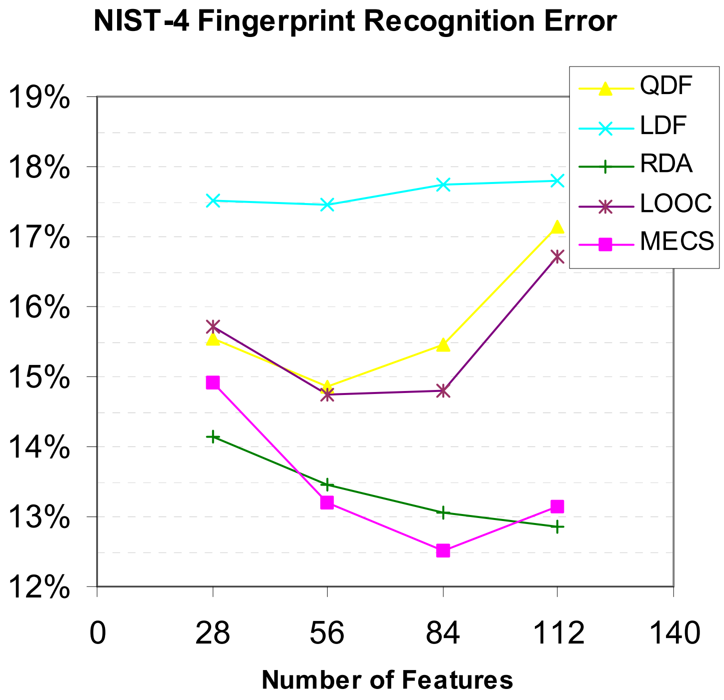


Figure 5.13. NIST-4 fingerprint error for Bayes plug-in classifiers.

Figure 5.14 presents the test recognition error of the NIST-3/7 OCR databases. In the lower dimensional spaces, where the ratio of the training sample size to the number of features were quite large and consequently the sample group covariance matrices were well-estimated, both LOOC optimised likelihood and RDA optimised classification classifiers shrank their parameters towards the sample group covariance matrices, performing identically to the conventional QDF maximum likelihood classifier. As the dimension of the feature space increased and the sample group covariance matrices became not so well posed, the sample likelihood QDF and LOOC methods slightly deteriorated while MECS and RDA improved. At the 72-dimensional space, RDA, taking advantage of blending mixture of covariance matrices with multiples of the identity matrix, achieved the lowest classification error (2.6%), followed by MECS

(3.3%), LOOC (3.6%) and QDF (3.6%). A classification error of 2.5% was reported on the same NIST OCR feature files, which used probabilistic neural networks [BCG94].

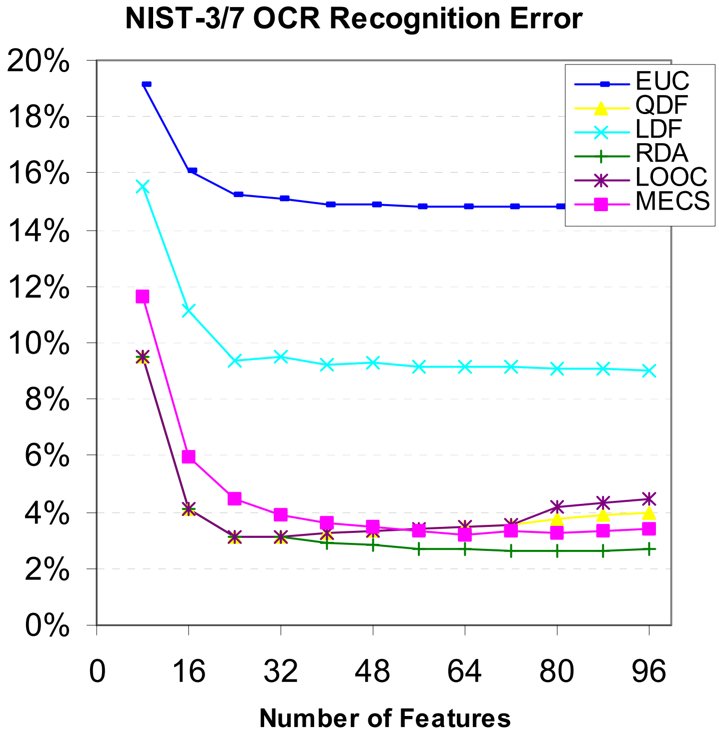


Figure 5.14. NIST-3/7 OCR databases error for Bayes plug-in classifiers.

5.4.3 Discussion

The effectiveness of the MECS method compared with other covariance estimators for Bayes plug-in classifiers (QDF, LDF, RDA and LOOC) was evaluated on four image recognition applications: face recognition, facial expression recognition, fingerprint classification, and optical character recognition (OCR).

In the face and fingerprint recognition problems, where the training sample sizes were chosen to be respectively small and moderate compared to the dimensionality of the feature space, MECS quadratic classifier achieved the lowest classification error. When the training sample sizes were small and the classes were very similar to each other, as in the facial expression application, the optimised LOOC classifier that considered the diagonal elements of the pooled covariance estimate achieved the best performance on the highest dimensional spaces, followed by MECS, LDF, and RDA. For the OCR problem, however, where the training sample sizes were not limited and quite large compared to the dimensionality of the feature space, the standard QDF was the most parsimonious classifier in terms of the recognition error and number of features required.

These results indicate that the MECS covariance estimator does increase the classification accuracy in image recognition applications where the sources of variation are frequently the same from group to group. Also, those real data experiments confirm the findings of several researchers that choosing an unconventional Bayesian parametric classifier between the linear and quadratic ones improves the classification accuracy in settings for which sample sizes are small and the number of features is large [Fri89, GR89, GR91, HF96, TGF03a].

However, the MECS new covariance approach shows that in high-dimensional and highly correlated classification problems where limited training sample sizes are provided, the problem of estimating covariance matrices for unconventional quadratic classifiers is essentially an issue of combining the reliable information available rather than optimising classification or likelihood indexes of well-behaved samples [TGF03b].

Application Features	RDA	LOOC	MECS
Face			
10	1392.42	2.95	0.04
20	1860.55	7.62	0.14
30	5549.83	23.38	0.56
40	8488.75	36.08	0.98
50	10999.77	57.08	1.75
60	14644.63	73.05	2.47
Facial Expression			
10	13.02	0.73	0.01
30	20.48	3.11	0.03
50	44.99	8.37	0.06
70	87.07	17.95	0.13
90	148.73	32.29	0.24
Fingerprint			
28	247.24	38.00	0.02
56	953.39	188.47	0.04
84	2106.20	452.56	0.13
112	4251.45	934.06	0.34
OCR			
16	1275.96	63.85	0.01
32	2977.54	243.29	0.03
64	8984.08	926.34	0.14
96	20520.84	2333.49	0.40

Table 5.3. Computational time (in seconds) for the quadratic classifiers.

Table 5.3 illustrates the CPU times for our RDA, LOOC and MECS implementations on a 1GHz desktop using a Windows based mathematical package, given as inputs the corresponding identity, sample, and pooled covariance matrices. As can be seen, the computational time spent in calculating the MECS estimate is remarkably lower than the RDA and LOOC ones. This can be explained by the fact that both RDA and LOOC covariance estimation methods require time-consuming searching processes in order to find the best linear mixture parameter regarding their respective maximum classification and maximum likelihood optimisation indices. In MECS, there is no optimisation search, but simply a selection process that maximises the inherent uncertainty when incomplete information is available.

Therefore, when concerns about the computation costs exist, MECS should be preferable to the other aforementioned unconventional quadratic classifiers, particularly when correlation is high and the sample size is limited.

5.5 Summary and Conclusions

In this chapter, the new Maximum Entropy Covariance Selection (MECS) method for the Bayes Plug-in classifier was introduced. It explored the issue of combining the sample and pooled covariance matrices under the principle of maximum entropy. The main idea of the MECS approach is to expand in a straightforward way the smaller and consequently less reliable eigenvalues of the sample group covariance matrix while trying to keep most of its larger eigenvalues unchanged.

Before evaluating the effectiveness of the MECS approach on synthetic and real image recognition data, a visual study of the maximum likelihood, maximum classification accuracy, and maximum entropy mixture covariance approaches was undertaken in the context of characterising human faces. This analysis indicated that the maximum entropy approach preserved as well of the sample group covariance information as possible, especially in the face experiments where moderate changes in facial expressions, pose, and scale, occurred.

In the synthetic data analysis, the simulation experiments suggested that when the data parameters are highly correlated the MECS estimator performs as well as, or better than, the QDF (when computable) and the widely used LDF estimator in limited sample and high dimensional problems. In such highly correlated data, MECS relative performance to the RDA and LOOC optimised approaches depends on the choice of the training samples. This is not a surprising result because the MECS procedure of blending the sample and pooled covariance matrices does not involve an iterative optimisation procedure, but simply a selection of the most reliable information available. When the data parameters are not correlated or slightly correlated and, consequently, the true covariance matrices have essentially a diagonal form, the synthetic results indicate that there is little benefit to be derived from MECS.

The performance of MECS was also compared with QDF, LDF, RDA, and LOOC classifiers in the context of real data. The following four image classification applications were studied: face recognition, facial expression recognition, fingerprint classification, and optical character recognition. The results indicated that in image recognition applications where the sources of variation are commonly the same from group to group, limited training samples sizes are considered, and concerns about high computation costs exist, the MECS approach is preferable to RDA and LOOC unconventional quadratic classifiers.

Finally, the MECS estimation, in contrast to the RDA and LOOC ones, is not exclusive to the Bayes Plug-in classifier. In fact, MECS can be used in the parametric quadratic classifier as well as in non-parametric Gaussian classifiers whenever the sample group covariance matrices are ill-posed or poorly estimated. Therefore, in the next chapter, we investigate the MECS approach as a new kernel covariance estimator for the non-parametric Parzen Window classifier.

Chapter 6

The Parzen Window Classifier

The Parzen Window Classifier is a popular non-parametric Bayesian classifier. In this classifier, the class-conditional probability densities are estimated locally by using kernel functions and a number of group neighbouring patterns. In practice, most of these probability densities are based on Gaussian kernel functions that involve the inverse of the true covariance matrix of each class.

As we have seen, the usual choice for estimating the true covariance matrices is the maximum likelihood estimator defined by the corresponding sample group covariance matrices. However, as described previously in the Bayes Plug-in classifier chapter, it is well known that in limited sample size applications the inverse of a sample group covariance matrix is either poorly estimated or cannot be calculated when the number of training patterns per class is smaller than the number of features. Thus, a significant amount of research has also been developed to design other covariance estimators for targeting limited sample and high dimensional problems in non-parametric Bayesian classifiers.

In this chapter we initially present the basic concepts of the Parzen Window classifier and review the most relevant unconventional approaches to estimating it for limited sample size problems. Since the MECS approach is a direct procedure that is not exclusive to the parametric Bayes Plug-in classifier, we then investigate the performance of using the MECS approach as a new kernel covariance estimator for the non-parametric Parzen Window classifier. The experimental results carried out on synthetic and image data indicate that the less restricted MECS covariance estimate improves the classification performance of the Parzen Window classifier with Gaussian kernels, especially when the sample size is small and the data parameters are highly correlated.

6.1 The Conventional Parzen Window Classifier

The Parzen Window classifier is a non-parametric Bayesian classifier based on class-conditional probability densities that are estimated locally by using kernel functions and a number of group neighbouring patterns [Fuk90].

If we recall the symmetrical or zero-one loss function described in the chapter 2, that is if we assume that all errors are equally costly [DHS01], the Bayes classification rule stipulates that an unknown pattern x should be assigned to the class π_i with the highest posterior probability. However, in the context of classification this rule becomes equivalent to assigning pattern x to the class π_i that has the maximum value obtained, among all the classes considered, by multiplying the prior probability by the corresponding Parzen likelihood density estimate.

In the standard or conventional Parzen classifiers with Gaussian kernels, the class-conditional Parzen likelihood density estimate is given by

$$p(x \mid \pi_i) \equiv q_i(x) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left[\frac{1}{(2\pi)^{n/2} \left|S_i\right|^{1/2}} \frac{1}{h_i^n} \exp\left[-\frac{1}{2h_i^2} (x - x_{i,j})^T S_i^{-1} (x - x_{i,j}) \right] \right], \quad \textbf{(6.1)}$$

where, as a reminder, n is the dimension of the feature space, $x_{i,j}$ is the pattern j from class π_i, N_i is the number of training patterns from class π_i, and S_i is the conventional sample group covariance matrix defined in equation (3.6). The parameter h_i is the window-width of class π_i and controls the kernel function "spread" or size.

Since the Gaussian function is symmetric, continuous, and unimodal, equation (6.1) essentially describes a multimodal probability density estimation where, according to the corresponding window-width parameter h_i, patterns that fall close to x contribute more to the estimate of this probability density at the point x than patterns that are far away from x [JR88].

The choice of the window-width parameter h_i is essentially a trade-off between reducing the bias of the estimate and increasing the variance [FH87]. In other words, as pointed out by Jain and Ramaswami [JR88], small values of h_i would give spiky or very biased estimates of $p(x \mid \pi_i)$ with each spike corresponding to the kernel itself at the training patterns. In contrast, when h_i are very large each training pattern provides basically the same contribution towards density estimation at every point x and the result is an over-smoothed estimate of $p(x \mid \pi_i)$ with almost no bias [JR88] but huge variance.

Although the choice of the window-width parameter h_i is important to the design of the Parzen classifiers, there has been no clear evidence that an optimal value for h_i can be theoretically determined, particularly in limited sample size and high dimensional problems [FH87, RJ91, HFT96]. In order to address this issue properly, since our attention here is focused on estimating the covariance matrices rather than the optimisation of h_i, we follow some authors' recommendations [FH87, RJ91] of selecting the best h_i of the Parzen Window classifier experimentally, for each particular recognition application considered.

6.2 Unconventional Parzen Window Classifiers

Owing to the limited sample size problem, several researchers have imposed, analogously to the Bayes Plug-in or QDF classifiers, some structures on the sample group covariance matrices for use in Gaussian Parzen classifiers [VNe80, JR88, Fuk90, HFT96]. Two approaches commonly employed for overcoming these estimation singularities and instabilities are described in the next sub-sections.

6.2.1 Van Ness Covariance Method

In the late 1970's, Van Ness described a number of studies [VS76, VNe80] on discriminant analysis of high dimensional Gaussian data and proposed a flexible diagonal form for the true covariance matrices of Gaussian Parzen classifiers. This diagonal form for the covariance matrix of each class is based solely on the estimation of the variances of each variable [VNe80].

In the Van Ness covariance estimation approach, the sample group covariance matrices S_i of the Parzen density estimate defined in equation (6.1) are replaced with the following matrices:

$$S_i^{ness}(\alpha) = \alpha U_i, \tag{6.2}$$

where α is a smooth or scale parameter and U_i is the diagonal $n \text{ x } n$ sample variance matrix of each class π_i. The matrix U_i is given by

$$U_i = diag(S_i) = \begin{bmatrix} \sigma_{i1}^2 & & & 0 \\ & \sigma_{i2}^2 & & \\ & & \ddots & \\ 0 & & & \sigma_{in}^2 \end{bmatrix}, \qquad \textbf{(6.3)}$$

where σ_{ij}^2 is the sample variance for the *j*th variable calculated from the training data of each class π_i. The smooth parameter α is selected to maximise the leave-one-out classification accuracy over all classes.

Since only the sample variance of each variable has to be calculated from the training patterns of each class, S_i^{ness} would be non-singular as long as there are at least two linearly independent patterns available per class. However, as equations (6.2) and (6.3) describe, the Van Ness covariance estimate relies solely on the information provided by the on-diagonal elements of each sample group covariance matrix, disregarding any information available about its off-diagonal elements.

Therefore, although S_i^{ness} could characterise different hyper-ellipsoidal shapes for each sample group, it considers that all hyper-ellipsoids have the same orientation. In image recognition applications where there are several classes to discriminate and the parameters are highly correlated, and consequently the estimation of the off-diagonal elements (co-variances) of the true covariance matrices becomes as important as the estimation of the diagonal elements (variances), the use of Van Ness covariance estimate seems to be restrictive and likely to undermine the potential recognition accuracy of the Gaussian Parzen Window classifier.

6.2.2 Toeplitz Covariance Method

Another possible structure for the Parzen Window covariance matrices is the Toeplitz approximation, based on the stationary assumption [Fuk90]. The basic idea of the Toeplitz covariance method is to allow each individual variable to have its own variance, whereas all covariance elements along any diagonal are multiplied by the same correlation factor.

The Toeplitz approximation of each group covariance matrix can be calculated as follows:

$$S_i^{toep} = \left[diag(S_i)\right]^{1/2} R_i \left[diag(S_i)\right]^{1/2}, \tag{6.4}$$

where the $diag(S_i)$ can be calculated as described in equation (6.3) and

$$R_i = \begin{bmatrix} 1 & \rho_i & \cdots & \rho_i^{n-1} \\ \rho_i & 1 & & \vdots \\ \vdots & & \ddots & \rho_i \\ \rho_i^{n-1} & \cdots & \rho_i & 1 \end{bmatrix}. \tag{6.5}$$

The correlation factor ρ_i is given by the average of the sample correlation $\rho_{k,k+1}$ over $k = 1,\ldots,n-1$ variables [Fuk90], that is

$$\rho_i = \frac{1}{n-1}\sum_{k=1}^{n-1}\rho_{k,k+1} = \frac{1}{n-1}\sum_{k=1}^{n-1}\frac{\sigma_{k,k+1}^2}{\sigma_k \sigma_{k+1}}, \tag{6.6}$$

where $\sigma_{k,k+1}^2$ is the sample co-variance between the *k*-th and (*k*+1)-th variables calculated from the training data of each class π_i, and the values σ_k and σ_{k+1} are their respective square roots or sample standard deviations. The Parzen Window classifier is then designed by substituting the sample group covariance matrices S_i for the Toeplitz covariance estimates described in equation (6.4).

As pointed out by Fukunaga [Fuk90], in the Toeplitz approach only ($n+1$) parameters, that is σ_{ik} ($k = 1,\ldots,n$) and ρ_i, are used to estimate the covariance matrix of each class π_i. Since these calculations do not require an optimisation process, its computational cost is much less severe than the Van Ness covariance method described in the previous sub-section.

Although we would not expect the Toeplitz covariance estimate to be well suited to many pattern recognition applications, Hamamoto, Fujimoto, and Tomita [HFT96] have shown, based on experiments carried out on artificial data sets, that the Toeplitz estimator can be preferable to the Van Ness [Vne80] and the orthogonal expansion [KTT87] estimators, particularly in small training sample size problems where concern about computational costs exists.

6.3 Synthetic Data Analysis

Analogously to the procedure explained in the chapter 5, more specifically in section 5.3, this section describes synthetic data experiments carried out to evaluate the MECS classification performance compared to the aforementioned kernel covariance estimators for Parzen Window classifiers.

We have adopted the same re-substitution (R) and holdout (H) methods to estimate respectively the lower and upper bounds of the classification errors of the conventional and unconventional Parzen classifiers with Gaussian kernels. As a reminder, in the R method the same samples are used to design and test the classifier. In contrast, two different sample sets are generated by the H method and one is used for training and the other for testing.

6.3.1 Experiments

Following the synthetic analyses developed by other researchers [VNe80, Fuk90, HFT96] and similar to the procedure adopted in the previous chapter, we have implemented simulation experiments conducted as an n-multivariate normal repeated-measures design.

Our attention is again concentrated on evaluating covariance estimators while varying the dimension of the space n, the degree of similarity of the covariance matrices, and the intra-class correlation ρ of all the groups considered. In the non-parametric Parzen Window classifier, however, the class-conditional probability densities are estimated by using not the centre or mean of each class as a reference value for closeness but a number of local group neighbouring patterns. Therefore, we would like to investigate whether the correlation between the n parameters would play the same important role observed in the covariance estimation for the parametric Bayes plug-in classifier in limited sample and high dimensional problems.

In all simulation experiments, we have considered the same 9 classes with the following mean vectors: $\mu_1 = [0,0,...,0]^T$, $\mu_2 = [1,0,...,\text{rem}(n,2)]^T$, $\mu_3 = [0,1,...,\text{rem}((n+1),2)]^T$, $\mu_4 = [1,1,...,1]^T$, $\mu_5 = [-1,1,...,(-1)^n]^T$, $\mu_6 = -\mu_2$, $\mu_7 = -\mu_3$, $\mu_8 = -\mu_4$, and $\mu_9 = -\mu_5$. Equivalently, we have chosen four values for the dimension of the space n (5, 10, 20, 40) and three values for the intra-class correlation factor ρ (0.0, 0.1, 0.9).

As previously, for each pair of (n, ρ) values the simulation consists of 25 replications of the following procedure. For each multivariate normal population of the 9 classes, we have generated 20 n-dimensional training observations and 50 n-dimensional test observations for each class. From these distinct training and test samples randomly drawn, we have calculated the lower and upper classification rates of the Parzen Window classifier defined in equation (6.1) using the following covariance kernels: SG (standard sample group covariance matrices as in equation (6.1)), TOEP (equation 6.4)), VNESS (equation (6.2)), and MECS (equations (5.6)). For comparison purpose, we have also calculated the lower and upper classification rates of the Euclidean distance classifier (EUC).

In order to simplify the non-parametric estimation, the Parzen window parameter h_i is assumed equal for all classes in all applications, and its optimum value was determined using the following set of ten values: 0.001, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 100, 1000. According to Raudys and Jain [RJ91], this set is usually sufficient to empirically determine the best h_i for a particular recognition application. Since the same number of training examples per class is considered, the prior probabilities are assumed equal for all classes in all experiments. As suggested by Van Ness' work [VNe80], the best smooth parameter α is found considering the following optimisation grid $\alpha = [0.2, 0.4, ..., 1.6, 1.8]$.

6.3.2 Results

Tables 6.1, 6.2, and 6.3 present the mean of the training (R method) and test (H method) classification rates over the 25 replications regarding the four values for the dimension parameter n (5, 10, 20, 40), the three values for the intra-class correlation factor ρ (0.0, 0.1, 0.9), and different covariance structures. Analogously to the synthetic parametric results, we considered the three following structures for the true covariance matrices: equal spherical covariance matrices, equal ellipsoidal covariance matrices, and unequal ellipsoidal covariance matrices.

As the training set of each synthetic class contains only 20 examples per class, the notation "-" found in a couple of rows of each classification table indicates that the sample group covariance (SG) matrices were singular. Therefore, in such cases, the standard Parzen Window classifier could not be used in order to classify the samples.

Equal Spherical Covariance Matrices

Table 6.1 presents the simulation results of the equal spherical covariance matrices defined in the previous chapter 5 (equation (5.9)).

As theoretically expected, when the correlation factor ρ was 0 the EUC classifier achieved clearly the best classification performance among all the classifiers considered. This result is due to fact that in this situation the true covariance matrices of all classes were equal to the identity matrix. However, differently from the parametric results described in the previous chapter, the EUC performance on the test sets was significantly superior to the non-parametric classifiers even when the data features became slightly correlated ($\rho = 0.1$).

		Equal Spherical Covariance Matrices					
n	Classifier	ρ = 0		ρ = 0.1		ρ = 0.9	
		R	H	R	H	R	H
5	EUC	52.5%	47.3%	51.9%	45.6%	46.2%	44.4%
	SG	78.0%	36.9%	79.9%	37.2%	90.9%	58.0%
	TOEP	77.5%	38.9%	77.8%	39.5%	85.8%	61.3%
	VNESS	66.0%	40.2%	66.3%	40.2%	85.5%	55.5%
	MECS	72.6%	40.9%	74.5%	40.4%	87.8%	61.9%
10	EUC	73.3%	63.4%	65.9%	57.0%	48.1%	44.7%
	SG	97.0%	42.8%	98.7%	37.9%	94.0%	57.7%
	TOEP	98.3%	49.5%	99.6%	46.0%	95.1%	63.6%
	VNESS	94.4%	52.8%	92.7%	48.4%	93.4%	61.5%
	MECS	96.7%	51.9%	94.8%	46.0%	94.7%	63.0%
20	EUC	90.7%	80.7%	77.8%	67.6%	49.9%	46.9%
	SG	-	-	-	-	-	-
	TOEP	100.0%	62.9%	100.0%	53.7%	96.7%	64.9%
	VNESS	99.8%	65.8%	99.2%	55.6%	98.4%	62.7%
	MECS	99.8%	65.5%	95.5%	53.3%	97.8%	63.0%
40	EUC	98.4%	94.0%	85.1%	76.3%	50.8%	47.6%
	SG	-	-	-	-	-	-
	TOEP	100.0%	77.1%	99.9%	64.1%	98.5%	64.7%
	VNESS	100.0%	79.3%	100.0%	64.3%	96.9%	63.5%
	MECS	100.0%	80.1%	99.9%	60.9%	99.6%	61.3%

Table 6.1. Equal spherical covariance matrices – EUC and Parzen classification results.

When the data features were highly correlated ($\rho = 0.9$), the MECS and Toeplitz covariance kernels performed similarly to each other and better than the other classifiers when the dimension of the space was lower (n = 5 or 10), particularly on the test sets. In this high correlated situation, the classification achievement of the MECS compared not only to the Toeplitz but also to the Van Ness kernel deteriorated when the dimensionality increased, especially when $n = 40$. This result suggests that not only MECS depends more on the relative size of the dimension of space to the total training sample size, as observed in the results of the previous chapter, but also that the specific shape and orientation described by the Parzen covariance estimate do not have the same significance as the one observed for the equal spherical parametric results.

Equal Ellipsoidal Covariance Matrices

The simulation results of the equal ellipsoidal covariance matrices defined in the previous chapter (equation (5.10)) are shown in Table 6.2.

		Equal Ellipsoidal Covariance Matrices					
n	Classifier	$\rho = 0$		$\rho = 0.1$		$\rho = 0.9$	
		R	H	R	H	R	H
5	EUC	45.6%	37.7%	42.4%	36.1%	39.1%	36.8%
	SG	76.1%	31.5%	77.0%	30.6%	88.4%	54.9%
	TOEP	74.6%	32.8%	72.1%	32.4%	83.2%	55.8%
	VNESS	64.1%	33.1%	55.0%	33.1%	86.2%	49.1%
	MECS	72.4%	34.5%	69.4%	33.1%	82.5%	58.2%
10	EUC	64.2%	53.4%	59.0%	49.4%	45.6%	41.7%
	SG	97.9%	37.5%	97.5%	32.0%	98.1%	57.3%
	TOEP	99.8%	43.0%	95.4%	38.8%	99.6%	61.4%
	VNESS	93.6%	45.5%	87.5%	41.1%	96.8%	58.6%
	MECS	94.4%	44.9%	89.8%	38.7%	96.6%	65.2%
20	EUC	85.6%	74.6%	72.4%	61.8%	47.3%	43.3%
	SG	-	-	-	-	-	-
	TOEP	100.0%	57.5%	100.0%	49.1%	98.1%	63.7%
	VNESS	100.0%	60.2%	99.0%	50.6%	99.4%	62.9%
	MECS	99.1%	59.5%	96.7%	47.4%	97.7%	67.6%
40	EUC	97.6%	91.3%	83.2%	74.2%	48.8%	46.2%
	SG	-	-	-	-	-	-
	TOEP	100.0%	73.7%	99.9%	60.9%	98.7%	64.3%
	VNESS	100.0%	76.0%	100.0%	61.2%	99.1%	64.9%
	MECS	100.0%	76.8%	99.9%	58.0%	99.9%	68.5%

Table 6.2. Equal ellipsoidal covariance matrices – EUC and Parzen classification results.

As can be seen, analogously to the aforementioned non-parametric results, when the data parameters were not correlated ($\rho = 0$) or slightly correlated ($\rho = 0.1$) the EUC classifier achieved the best classification performance on the test sets. However, in this equal ellipsoidal covariance matrices case, the superiority of the MECS estimate on highly correlated data ($\rho = 0.9$) compared to the other approaches was clear. In all dimensions with $\rho = 0.9$, MECS led to higher testing recognition accuracies than the Van Ness and Toeplitz covariance kernels.

Unequal Ellipsoidal Covariance Matrices

Table 6.3 presents the simulation results of the unequal ellipsoidal covariance matrices defined in the equation (5.13) of the previous chapter.

n	Classifier	Unequal Ellipsoidal Covariance Matrices					
		ρ = 0		ρ = 0.1		ρ = 0.9	
		R	H	R	H	R	H
5	EUC	43.7%	36.7%	41.3%	34.3%	37.4%	34.2%
	SG	66.7%	35.0%	66.1%	33.5%	80.2%	56.2%
	TOEP	62.4%	35.9%	61.6%	35.2%	72.4%	56.6%
	VNESS	70.1%	35.9%	67.8%	35.6%	84.1%	51.4%
	MECS	69.8%	33.4%	71.0%	32.8%	78.7%	54.3%
10	EUC	62.5%	49.6%	55.4%	45.9%	42.0%	37.7%
	SG	98.1%	41.1%	97.8%	38.0%	98.8%	61.2%
	TOEP	94.1%	47.5%	91.9%	45.3%	91.0%	68.6%
	VNESS	95.2%	47.5%	93.7%	44.7%	95.2%	62.2%
	MECS	98.7%	42.3%	98.5%	38.5%	99.5%	64.7%
20	EUC	79.9%	65.3%	71.2%	56.2%	44.0%	41.2%
	SG	-	-	-	-	-	-
	TOEP	99.9%	61.1%	99.6%	56.9%	99.4%	80.0%
	VNESS	100.0%	60.3%	99.5%	55.7%	99.2%	71.1%
	MECS	100.0%	54.2%	100.0%	47.0%	100.0%	70.7%
40	EUC	93.6%	80.5%	79.8%	65.7%	45.6%	42.2%
	SG	-	-	-	-	-	-
	TOEP	100.0%	73.8%	100.0%	67.5%	100.0%	88.0%
	VNESS	100.0%	73.5%	100.0%	65.6%	99.9%	75.3%
	MECS	100.0%	66.2%	100.0%	51.9%	100.0%	71.5%

Table 6.3. Unequal ellipsoidal covariance matrices – EUC and Parzen classif. results

Although MECS achieved the best classification performance on the training sets when the data parameters were not correlated ($\rho = 0$) or slightly correlated ($\rho = 0.1$) and the dimension of the space was lower (n = 5 or 10), its performance on the test sets deteriorated compared to the Van Ness and Toeplitz approaches. However, both Van

Ness and Toeplitz testing classification results were worse than the EUC classifier when $\rho = 0$, and similar to, or slightly better than, the EUC classifier when the intra-class correlation factor was $\rho = 0.1$.

In the situation where the intra-class data parameters were highly correlated ($\rho = 0.9$), the Toeplitz performance on the training sets was worse than the Van Ness and MECS covariance estimators when the dimension was lower, that is, when n = 5 or 10. On the test sets, however, the Toeplitz covariance kernel achieved the best classification results in all dimensions considered.

The overall good performance of the Parzen classifier with the Toeplitz covariance kernel is explained by the fact that in all covariance simulations we assumed the same correlation values for all the features considered. This simulation assumption is similar to the basic idea of the Toeplitz covariance matrix that all covariance elements along any diagonal are multiplied by the same correlation factor.

6.3.3 Discussion

Since we have used the same synthetic classes and carried out the same n-multivariate normal simulations, it is possible to compare the classification results of the non-parametric Gaussian Parzen Window (PZW) classifiers described in this chapter with the parametric Bayes Plug-in (QDF) classifiers presented in the previous chapter (section 5.3). As the Euclidean classifier achieved good classification performance in those situations where the data parameters were not correlated or slightly correlated, we restrict our analysis to the case where the intra-class data parameter was highly correlated ($\rho = 0.9$).

Table 6.4 shows the lower classification bounds (based on the holdout method) over different features of the Toeplitz, Van Ness, and MECS PZW classifiers, and also the parametric RDA, LOOC, and MECS QDF classifiers on the three covariance structures previously described: equal spherical covariance matrices, equal ellipsoidal covariance matrices, and unequal ellipsoidal covariance matrices. As can be seen, the superiority of the QDF classifiers compared to the PZW classifiers is clear in all but one of the n-multivariate normal simulations considered. Moreover, in all experiments, the parametric QDF classifier using the MECS estimate performed better than its non-parametric PZW version.

Covariance Structure		**PZW**			**QDF**	
Features	Toeplitz	Van Ness	MECS	RDA	LOOC	MECS
Equal Spherical						
5	61.3%	55.5%	61.9%	65.2%	64.8%	64.4%
10	63.6%	61.5%	63.0%	70.6%	67.4%	66.7%
20	64.9%	62.7%	63.0%	73.0%	67.2%	65.6%
40	64.7%	63.5%	61.3%	71.8%	63.5%	62.7%
Equal Ellipsoidal						
5	55.8%	49.1%	58.2%	61.7%	61.5%	60.3%
10	61.4%	58.6%	65.2%	71.5%	71.7%	70.4%
20	63.7%	62.9%	67.6%	76.2%	74.0%	71.4%
40	64.3%	64.9%	68.5%	77.5%	73.3%	71.1%
Unequal Ellipsoidal						
5	56.6%	51.4%	54.3%	59.9%	61.0%	58.4%
10	68.6%	62.2%	64.7%	72.9%	75.4%	70.2%
20	80.0%	71.1%	70.7%	77.2%	82.8%	74.1%
40	88.0%	75.3%	71.5%	76.1%	86.3%	72.5%

Table 6.4. Lower classification bounds of non-parametric and parametric classifiers.

It has been suggested that even when the underlying data are ideal for the QDF classifier, that is, the sample data of each group are unimodal with a local maximum, the non-parametric classifiers formed by unconventional covariance estimators, such as the Van Ness approach [VNe80], could achieve superior classification accuracy in limited samples and high dimensional problems.

The results described in Table 6.4 suggest that what has been behind these surprising findings is an unfair comparison between a poor covariance estimate given by the conventional maximum likelihood approach (or sample group covariance matrices) used in the QDF classifier and a more reliable unconventional covariance estimate used in the non-parametric PZW classifier. When the less restricted MECS covariance estimate is used in both parametric and non-parametric classifiers, our results indicate that the parametric classifier is the better choice. Furthermore, parametric classifiers are simpler and faster to compute.

6.4 Image Data Analysis

For the purpose of investigating in practice the performance of MECS as a covariance kernel for Gaussian Parzen classifiers, the following three image based classification applications were examined: face recognition, facial expression recognition, and fingerprint classification. These experiments were designed to evaluate the effectiveness of the MECS approach compared with the sample group (when possible), Van Ness, and Toeplitz covariance estimation approaches over a comprehensive range of sample and feature sizes.

6.4.1 Experiments

For the face and facial expression recognition experiments, we used the previously described FERET (sub-section 5.2.2.1) and Tohoku facial expression (sub-section 4.1.3.2) benchmark databases. In these applications the training sample sizes were chosen to be extremely small and small respectively compared to the dimension of the feature space. The training and test feature files extracted from the NIST Special Database 4 (sub-section 5.4.1.1) were used in the fingerprint classification. This dataset represented an alternative sample size setting where moderate and large training sets compared with the number of features were considered.

As in the parametric experiments described in the previous chapter with the same datasets (section 5.4), PCA [TP91] was first used to reduce the dimensionality of the original face and facial expression images (resized to 96x64 and 64x64 pixels respectively) and then the discriminant Bayes' rule using the non-parametric approaches described previously in this chapter was applied. Each experiment for both the face and facial expression applications was repeated 25 times using a number of PCA features. Distinct training and test sets were randomly drawn, and the mean of the recognition error rate was calculated.

The face recognition error rate was computed by utilising for each subject 3 images to train and 1 image to test. The training and test sets of the facial expression experiments were respectively composed of 20 and 9 images. Analogously, in the fingerprint non-parametric classification, we used the same 112 floating point training and test feature vectors used in the parametric experiments presented in the previous chapter. The fingerprints were classified into one of the five categories using from each class 400 prints to train and 400 prints to test.

The Parzen window parameter h_i was assumed equal for all classes in all applications, and its optimum value was determined using the following set of ten values: 0.001, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 100, 1000 [RJ91]. Moreover, the prior probabilities were again supposed equal for all classes and recognition tasks, and the Van Ness smoothing parameter was $\alpha = [0.2, 0.4, 0.6, \ldots, 1.6, 1.8]$, as suggested by [VNe80].

6.4.2 Results

The results of the Gaussian Parzen Window classifiers using the sample group (SG), Van Ness, Toeplitz, and MECS covariance estimates are presented in Figures 6.1, 6.2, and 6.3.

Figure 6.1 illustrates the test average recognition error of the FERET face database. Since only 3 face images were used to train the classifiers, the SG covariance matrices were singular and the standard Gaussian Parzen Window classifier could not be calculated.

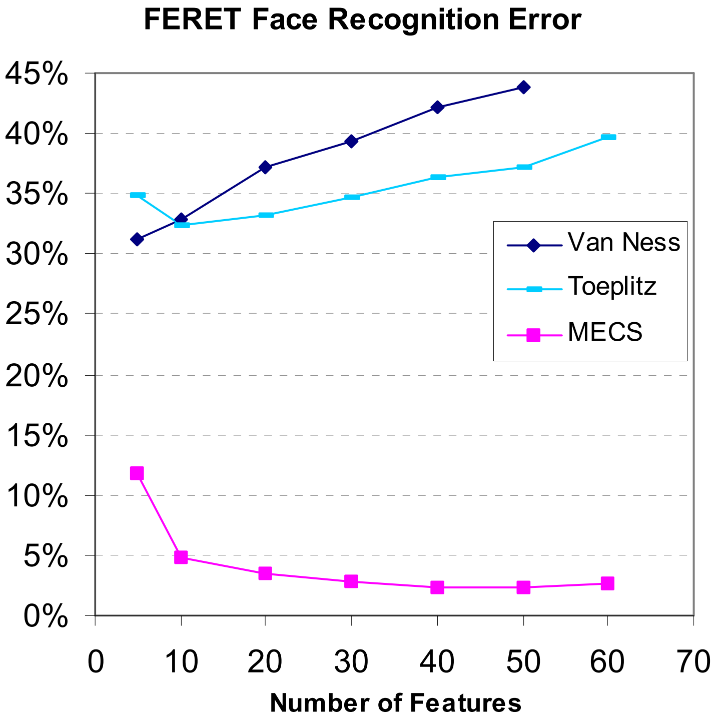


Figure 6.1. FERET face database recognition error for Parzen classifiers.

As can be seen from Figure 6.1, the MECS estimator improved significantly the face classification accuracy of the Parzen classifier compared with the other estimation approaches. The MECS Parzen classifier achieved the lowest classification error – 2.3% – on 50 eigenfaces. In this application where the off-diagonal elements (co-variances) of the covariance matrix of each class seem to be as important as the on-diagonal elements (variances), the Toeplitz approach did better than the Van Ness estimate in all but one experiment.

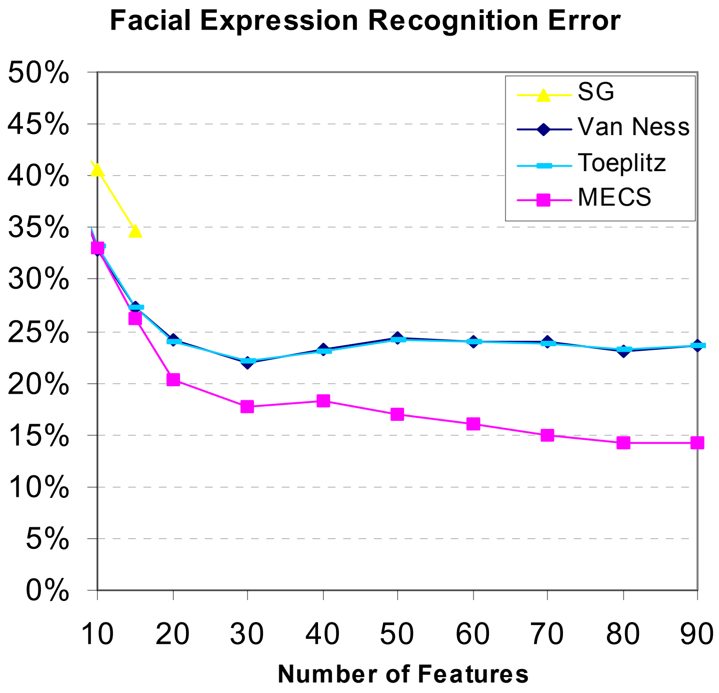


Figure 6.2. Tohoku facial expression recognition error for Parzen classifiers.

An analogous performance of the MECS estimator is shown in Figure 6.2. It presents the test average recognition error of the Tohoku facial expression database. Since 20 images were used to compose the training set of each facial expression, the results relative to the Gaussian Parzen Window classifier with the sample group covariance estimate (SG) were limited. Although there is no clear dominance of any unconventional covariance estimator in the lowest dimension spaces (10 and 15 features), when the dimensionality increased and the ratio of the training sample size to the number of features became small, MECS performed clearly better than the Van Ness and Toeplitz estimators.

In this facial expression application where the true covariance matrices seem to describe equal-ellipsoidal shapes in the higher dimensional spaces, the restrictive forms of the Van Ness and Toeplitz approaches undermined the potential recognition accuracy of the Gaussian Parzen Window classifier and they achieved approximately the same classification performance as each other.

Figure 6.3 presents the recognition error results of the NIST-4 fingerprint experiments. As expected, because in this application the sample group covariance matrices are well posed and estimated from 400 training patterns per class, the standard Parzen Window Classifier (SG) in all but one experiment led to lower recognition error than did the MECS, Van Ness and Toeplitz covariance estimators. According to the number of features considered, MECS achieved its best recognition error result – 15.05% – when the dimensionality of the patterns was reduced to 28 components. This result was slightly worse than the SG best result – 15.00% of recognition error using, however, 56 preprocessed features. On comparing the MECS classification performance with both the Van Ness and Toeplitz estimators, MECS showed again its superiority on discriminating well-framed image patterns.

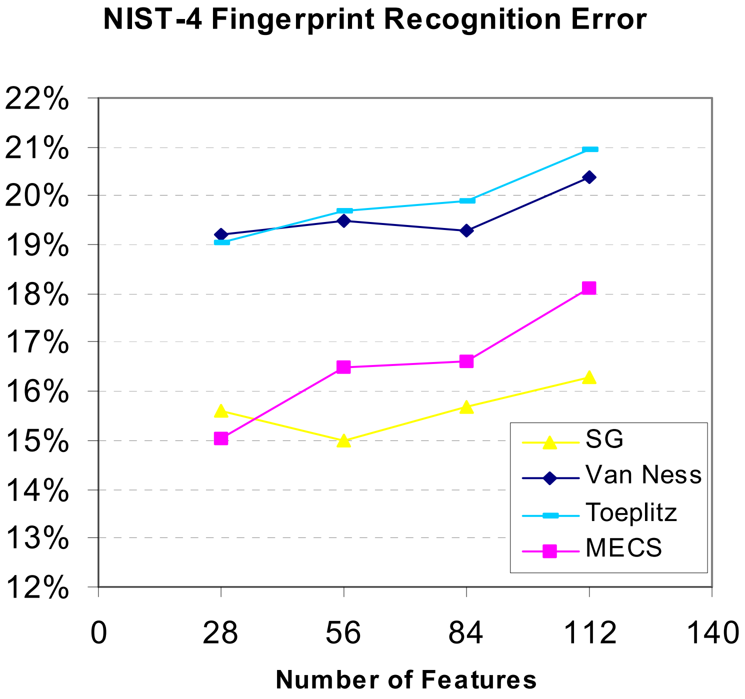


Figure 6.3. NIST-4 fingerprint recognition error for Parzen classifiers.

The computational time of the unconventional covariance kernels for Gaussian Parzen Window classifiers is shown in Table 6.5. As can be seen, MECS and Toeplitz computational times are similar to each other and much less severe than the Van Ness covariance estimation time. This result can be explained by the fact that both MECS and Toeplitz approaches do not require a time-consuming optimisation method to estimate their respective covariance matrices.

Application Features	Van Ness	Toeplitz	MECS
Face			
10	943.66	0.13	0.07
20	1370.39	0.27	0.22
30	2020.80	0.39	0.51
40	2893.75	0.54	1.04
50	4016.77	0.83	1.85
60	5428.72	1.19	2.88
Facial Expression			
10	15.56	0.01	0.01
30	54.53	0.01	0.02
50	127.81	0.03	0.05
70	236.55	0.06	0.12
90	383.28	0.08	0.24
Fingerprint			
28	12754.39	0.02	0.01
56	42181.72	0.08	0.07
84	91019.83	0.22	0.17
112	161273.64	0.68	0.40

Table 6.5. Computational time (in seconds) for the Parzen classifiers.

6.5 Summary and Conclusions

In this chapter, the non-parametric Gaussian Parzen Window classifier and its non-conventional Van Ness and Toeplitz covariance estimation approaches for solving the singularity and instability of the sample group covariance matrices have been reviewed with regard to the difficulties caused by limited sample size in high dimensional problems.

Since the MECS estimate described in the previous chapter can be used in multidimensional Gaussian classifiers whenever the sample group covariance matrices are ill posed or poorly estimated, we then evaluated the MECS effectiveness as a new kernel covariance estimator for the non-parametric Bayesian classifier. Biometric image recognition applications, such as face recognition and fingerprint classification, which involved small and limited training sets compared to the number of features, were used to compare the MECS classification accuracy with the standard and non-conventional aforementioned Gaussian Parzen Window classifiers.

The real data results indicated that when the ratio of the training sample size to the number of features was small and the inverse of the sample group covariance matrices could not be calculated, MECS achieved clearly much lower recognition error rates than did the commonly used Van Ness and Toeplitz estimators. In highly correlated and well-framed classification problems where the estimation of the off-diagonal elements (co-variances) of the covariance matrix of each class is as important as its on-diagonal elements (variances), the restrictive Van Ness and Toeplitz covariance approaches tend to undermine the potential recognition accuracy of the Gaussian Parzen Window classifier.

Furthermore, as we carried out in this chapter the same synthetic n-multivariate normal simulations previously described in the chapter 5, it was possible to compare the classification results of the non-parametric Gaussian Parzen Window classifiers with the parametric Bayes Plug-in classifiers. Our synthetic results suggest that when the less restricted MECS covariance estimate is used in both parametric and non-parametric classifiers, the parametric approach is the best choice because it is simpler to calculate requiring much less computer time, especially for testing the classifiers.

Finally, the singularity and instability of covariance matrices is a critical issue not only for parametric and non-parametric Bayesian classifiers, but also other statistical covariance-based analysis that requires the inverse of such covariance matrices. Hence, in the next chapter, we investigate the maximum entropy covariance selection method of combining singular and non-singular covariance matrices for the Linear (or Fisher) Discriminant Analysis in situations where the total number of training samples is comparable to the number of features.

Chapter 7

Fisher Discriminant Analysis

Fisher Discriminant Analysis, also called Linear Discriminant Analysis (LDA), has been used successfully as a statistical feature extraction technique in several classification problems.

Analogously to the Bayesian classifiers described in the preceding chapters, a critical issue in using LDA is, however, the singularity and instability of the within-class scatter matrix. In practice, particularly in image recognition applications such as face recognition, there are often a large number of pixels or pre-processed features available, but the total number of training patterns is limited and commonly less than the dimension of the feature space. This implies that the within-class scatter matrix either will be singular if its rank is less than the number of features or might be unstable (or poorly estimated) if the total number of training patterns is not significantly larger than the dimension of the feature space. Hence, a considerable amount of research has also been devoted to the design of other Fisher-based methods, for targeting limited sample and high dimensional problems.

In this chapter a new Fisher-based method is proposed. It is based on the straightforward maximum entropy covariance selection approach that overcomes the singularity and instability of the within-class scatter matrix when LDA is applied in limited sample and high dimensional problems. In order to evaluate its effectiveness, experiments on face recognition using the previously described ORL and FERET face databases were carried out and compared with other LDA-based methods. The results indicate that our method improves the LDA classification performance when the within-class scatter matrix is singular as well as poorly estimated, with or without a PCA intermediate step and using fewer linear discriminant features.

7.1 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a well-known feature extraction technique that has been used successfully in many statistical pattern recognition problems. It has its origin in the late 1930's when Fisher proposed a number of studies on separating or discriminating populations [Fis36, Fis38]. For this reason, Linear Discriminant Analysis is often called Fisher Discriminant Analysis.

The primary purpose of Linear Discriminant Analysis is to separate samples of distinct groups by maximising their between-class separability while minimising their within-class variability. Although LDA does not assume that the populations of the distinct groups are normally distributed, it assumes implicitly that the true covariance matrices of each class are equal because the same within-class scatter matrix is used for all the classes considered [JW98].

Let the between-class scatter matrix S_b be defined as

$$S_b = \sum_{i=1}^{g} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{7.1}$$

and the within-class scatter matrix S_w be defined as

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T \, , \tag{7.2}$$

where, as a reminder, $x_{i,j}$ is the *n*-dimensional pattern j from class π_i, N_i is the number of training patterns from class π_i, and g is the total number of classes or groups. The vector $\bar{x}_i$ and matrix S_i are respectively the sample mean and sample covariance matrix of class π_i previously defined in equations (3.5) and (3.6). The grand mean vector $\bar{x}$ is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \bar{x}_i = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j} \, , \tag{7.3}$$

where N is the total number of samples, that is, $N = N_1 + N_2 + \cdots + N_g$. It is important to note that the within-class scatter matrix S_w defined in equation (7.2) is essentially the pooled covariance matrix (previously defined in equation (3.12)) multiplied by the scalar $(N-g)$, that is

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = (N-g) S_p . \tag{7.4}$$

The main objective of LDA is to find a projection matrix P_{lda} that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix (Fisher's criterion), that is

$$P_{lda} = \arg\max_{P} \frac{\left| P^T S_b P \right|}{\left| P^T S_w P \right|} . \tag{7.5}$$

Devijver and Kittler [DK82] have shown that P_{lda} is in fact the solution of the following eigensystem problem:

$$S_b P - S_w P \Lambda = 0 . \tag{7.6}$$

Multiplying both sides by S_w^{-1}, equation (7.6) can be rewritten as

$$\begin{aligned} & S_w^{-1} S_b P - S_w^{-1} S_w P \Lambda = 0 \\ & S_w^{-1} S_b P - P \Lambda = 0 \\ & (S_w^{-1} S_b) P = P \Lambda \end{aligned} \tag{7.7}$$

where P and Λ are respectively the eigenvectors and eigenvalues of $S_w^{-1} S_b$. In other words, equation (7.7) states that if S_w is a non-singular matrix then the Fisher's criterion described in equation (7.5) is maximised when the projection matrix P_{lda} is composed of the eigenvectors of $S_w^{-1} S_b$ with at most $(g-1)$ nonzero corresponding eigenvalues. This is the standard LDA procedure.

The performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations N compared to the dimension of the feature space n. Since the within-class scatter matrix S_w is a function of $(N-g)$ or fewer linearly independent vectors, its rank is $(N-g)$ or less. Therefore, S_w is a singular matrix if N is less than $(n+g)$, or, analogously, might be unstable if N is not at least five to ten times $(n+g)$ [JC82].

In the next section, recent LDA-based methods proposed for targeting limited sample and high dimensional problems are described. A novel method of combining singular and non-singular covariance matrices for solving the singularity and instability of the within-class scatter matrix is proposed in section 7.3.

7.2 LDA Limited Sample Size Approaches

As discussed previously in this chapter, a critical issue for the standard LDA feature extraction technique is the singularity and instability of the within-class scatter matrix. Thus, a considerable amount of research has also been devoted to the design of other LDA-based methods, for overcoming the limited number of samples compared to the number of features. In the following sub-sections, recent LDA-based methods with application to face recognition are described. Since the face recognition problem involves small training sets, a large number of features, and a large number of groups, it has become the most used application to evaluate such limited sample size approaches [SW96, BHK97, ZCK98, CLK00, YY01a ,YY01b, YY03, TG03b].

7.2.1 Fisherfaces Method

The Fisherfaces [BHK97, ZCK98] method is one of the most successful feature extraction approaches for solving limited sample size problems in face recognition. It is also called the Most Discriminant Features (MDF) method [SW96].

The Fisherfaces or MDF method is essentially a two-stage dimensionality reduction technique. First the face images from the original vector space are projected to a lower dimensional space using Principal Component Analysis (PCA) [TP91] and then LDA is applied next to find the best linear discriminant features on that PCA subspace.

More specifically, the MDF projection matrix P_{mdf} can be calculated as

$$P_{mdf}^T = P_{lda}^T * P_{pca}^T, \qquad \textbf{(7.8)}$$

where P_{pca} is the projection matrix from the original image space to the PCA subspace, and P_{lda} is the projection matrix from the PCA subspace to the LDA subspace obtained by maximising the ratio

$$P_{lda} = \arg\max_{P} \frac{\left|P^T P_{pca}^T S_b P_{pca} P\right|}{\left|P^T P_{pca}^T S_w P_{pca} P\right|} . \qquad \textbf{(7.9)}$$

As described in the previous section, equation (7.9) analogously states that if $P_{pca}^T S_w P_{pca}$ is a non-singular matrix then the Fisher's criterion is maximised when the projection matrix P_{lda} is composed of the eigenvectors of $(P_{pca}^T S_w P_{pca})^{-1}(P_{pca}^T S_b P_{pca})$ with at most $(g-1)$ nonzero corresponding eigenvalues.

The singularity problem of the within-class scatter matrix S_w is then overcome if the number of retained principal components varies from at least g to at most $N-g$ PCA features [SW96, BHK97, ZCK98].

7.2.2 Chen et al.'s Method (CLDA)

Chen et al. [CLK00] have proposed another LDA-based method, here called CLDA, that overcomes the singularity problems related to the direct use of LDA in small sample size applications, particularly in face recognition.

The main idea of their approach is to use either the discriminative information of the null space of the within-class scatter matrix to maximise the between-class scatter matrix whenever S_w is singular, or the eigenvectors corresponding to the set of the largest eigenvalues of matrix $(S_b + S_w)^{-1} S_b$ whenever S_w is non-singular.

The CLDA algorithm for calculating the projection matrix P_{clda} can be summarised as follows [CLK00]:

i. Calculate the rank r of the within-class scatter matrix S_w;

ii. If S_w is non-singular, that is $r = n$, then P_{clda} is composed of the eigenvectors corresponding to the largest eigenvalues of $(S_b + S_w)^{-1} S_b$;

iii. Otherwise, calculate the eigenvectors matrix $V = [v_1, ..., v_r, v_{r+1}, ..., v_n]$ of the singular within-class scatter matrix S_w. Let Q be the matrix that spans the S_w null space, that is $Q = [v_{r+1}, v_{r+2}, ..., v_n]$;

iv. The projection matrix P_{clda} is then composed of the eigenvectors corresponding to the largest eigenvalues of $QQ^T S_b (QQ^T)^T$.

Although their experimental results have shown that CLDA improves the performance of a face recognition system compared with Liu et al.'s approach [LCY93] and the standard template matching procedure [JDM00], Chen et al.'s approach will select the same linear discriminant features as the standard LDA when S_w is non-singular [Fuk90] but poorly estimated.

7.2.3 Yu and Yang's Method (DLDA)

Yu and Yang [YY01b] have developed a direct LDA algorithm (DLDA) for high dimensional data with application to face recognition.

The key idea of their method is to discard the null space of S_b rather than discarding the null space of S_w by diagonalising S_b first and then diagonalising S_w. This diagonalisation process avoids the singularity problems related to the use of the pure LDA in high dimensional data where the within-class scatter matrix S_w is likely to be singular. Also, differently from Chen et al.'s algorithm previously described, DLDA uses all the within-class scatter matrix information, i.e. both within and outside information of S_w's null space [YY01b].

The DLDA algorithm for calculating the projection matrix P_{dlda} can be described as follows [YY01b]:

i. Diagonalise S_b, that is calculate the eigenvector matrix V such that $V^T S_b V = \Lambda$;

ii. Let Y be the first m columns of V corresponding to the S_b largest eigenvalues, where $m \leq rank(S_b)$. Calculate $D_b = Y^T S_b Y$;

iii. Let Z be a whitening transformation of S_b that also reduces its dimensionality from n to m, that is calculate $Z = Y D_b^{-1/2}$;

iv. Diagonalise $Z^T S_w Z$, that is compute U and D_w such that $U^T (Z^T S_w Z) U = D_w$;

v. Calculate the projection matrix P_{dlda} given by $P_{dlda} = D_w^{-1/2} U^T Z^T$.

Using computational techniques to handle large scatter matrices, Yu and Yang's [YY01b] experimental results have shown that DLDA can be applied on the original vector space of face images without any explicit intermediate dimensionality reduction step. However, they pointed out [YY01b] that by replacing the between-class scatter matrix S_b with the total scatter matrix S_T, given by $S_T = S_b + S_w$, the first two steps of their algorithm becomes exactly the PCA dimensionality reduction technique.

7.2.4 Yang and Yang's Method (YLDA)

More recently, Yang and Yang [YY03] have proposed a linear feature extraction method, here called YLDA, which is capable of deriving discriminatory information of the LDA criterion in singular cases.

Analogous to the Fisherfaces method described previously in the subsection 7.2.1, the YLDA is explicitly a two-stage dimensionality reduction technique. That is, PCA [TP91] is used firstly to reduce the dimensionality of the original space and then LDA, using a particular Fisher-based linear algorithm called Optimal Fisher Linear Discriminant (OFLD) [YY01a], is applied next to find the best linear discriminant features on that PCA subspace.

The OFLD algorithm [YY01a] can be described as follows:

i. In the m-dimensional PCA transformed space, calculate the within-class and between-class scatter matrices S_w and S_b;

ii. Calculate the eigenvectors matrix $V = [v_1, v_2, ..., v_m]$ of S_w. Suppose the first q eigenvectors of S_w correspond to its positive eigenvalues;

iii. Let a projection matrix be $P_1 = [v_{q+1}, v_{q+2}, ..., v_m]$. Form the transformation matrix Z_1 composed of the eigenvectors of $P_1^T S_b P_1$. The first k_1 YLDA discriminant vectors are given by $P_{ylda}^1 = P_1 Z_1$, where generally $k_1 = g - 1$;

iv. Let a second projection matrix be $P_2 = [v_1, v_2, ..., v_q]$. Form the transformation matrix Z_2 composed of the eigenvectors corresponding to the k_2 largest eigenvalues of $(P_2^T S_w P_2)^{-1} (P_2^T S_b P_2)$. The remaining k_2 YLDA discriminant vectors are given by $P_{ylda}^2 = P_2 Z_2$, where k_2 is an input parameter that can extend the final number of LDA features beyond the $(g - 1)$ nonzero S_b eigenvalues;

v. Form the projection matrix P_{ylda} given by the concatenation of P_{ylda}^1 and P_{ylda}^2.

Yang and Yang [YY03] have proved that the number m of principal components to retain for a best LDA performance should be equal to the rank of the total scatter matrix

S_T, given, as a reminder, by $S_T = S_b + S_w$ and calculated on the original space [YY03]. However, no procedure has been shown to determine the optimal value for the parameter k_2. This parameter is context dependent and consequently can vary according to the application studied. Moreover, although YLDA addresses the PCA+LDA problems when the total scatter matrix S_T is singular, such PCA strategy does not avoid the within-class scatter instability when S_T is non-singular but poorly estimated.

7.3 The Maximum Uncertainty LDA-based Approach

In order to avoid the singularity and instability critical issues of the within-class scatter matrix S_w when LDA is used in limited sample and high dimensional problems, we propose a new LDA-based approach based on a straightforward covariance selection method for the S_w matrix.

7.3.1 Related Methods

In the past, a number of researchers [DPi79, Cam80, PN82, Ray90] have proposed a modification in LDA that makes the problem mathematically feasible and increases the LDA stability when the within-class scatter matrix S_w has small or zero eigenvalues.

The idea is to replace the pooled covariance matrix S_p of the scatter matrix S_w (equation (7.4)) with a ridge-like covariance estimate of the form

$$\widehat{S}_p(k) = S_p + kI, \qquad \textbf{(7.10)}$$

where I is the *n* by *n* identity matrix and $k \geq 0$. DiPillo [DPi79] attempted to determine analytically the optimal choice for the value k. However, such solution has been shown intractable in practice and several researchers have performed simulation studies to choose the best value for k [DPi79, PN82, Ray90].

According to Rayens [Ray90], a reasonable grid of potential simulation values for the optimal k could be

$$\lambda_{\min} \leq k \leq \lambda_{\max}, \qquad \textbf{(7.11)}$$

where the values $\lambda_{\min}$ and $\lambda_{\max}$ are respectively the non-zero smallest and largest eigenvalues of the pooled covariance matrix S_p. Rayens [Ray90] has suggested that a more productive searching process should be based on values near $\lambda_{\min}$ rather than $\lambda_{\max}$. However, this reasoning is context-dependent and a time-consuming leave-one-out optimisation process is necessary to determine the best multiplier for the identity matrix.

As described in chapter 3, other researchers have imposed regularisation methods to overcome the singularity and instability in sample based covariance estimation, especially to improve the Bayes Plug-in or QDF classification performance [Fri89, GR91, Tad98]. Most of these works have used shrinkage parameters that combine linearly a singular or unstable covariance matrix, such as S_p, to a multiple of the identity matrix.

According to these regularisation methods, the ill posed or poorly estimated S_p could be replaced with a convex combination matrix $\widehat{S}_p(\gamma)$ of the form

$$\widehat{S}_p(\gamma) = (1-\gamma)S_p + (\gamma)\overline{\lambda}I \,, \tag{7.12}$$

where the shrinkage parameter γ takes on values $0 \le \gamma \le 1$ and could be selected to maximise the leave-one-out classification accuracy. The identity matrix multiplier k would be given by the average eigenvalue $\overline{\lambda}$ of S_p calculated as

$$\overline{\lambda} = \frac{1}{n}\sum_{j=1}^{n} \lambda_j = \frac{tr(S_p)}{n} \,, \tag{7.13}$$

where the notation “tr” denotes the trace of a matrix.

The regularisation idea described in equation (7.12) would have the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in sample-based estimation of eigenvalues [Fri89].

7.3.2 The Proposed Method

The proposed method considers the issue of stabilising the S_p estimate with a multiple of the identity matrix by selecting the largest dispersions regarding the S_p average eigenvalue. It is based on the maximum entropy covariance selection idea described in chapter 5 to improve quadratic classification performance on limited sample size problems [TG03b, TGF03b, TG04].

Following equation (7.10), the eigen-decomposition of a combination of the covariance matrix S_p and the *n* by *n* identity matrix I can be written as [Mar87]

$$\begin{aligned} \widehat{S}_p(k) &= S_p + kI \\ &= \sum_{j=1}^{r} \lambda_j \phi_j (\phi_j)^T + k \sum_{j=1}^{n} \phi_j (\phi_j)^T \\ &= \sum_{j=1}^{r} (\lambda_j + k) \phi_j (\phi_j)^T + \sum_{j=r+1}^{n} k \phi_j (\phi_j)^T \end{aligned} \quad \textbf{(7.14)}$$

where r is the rank of S_p ($r \le n$), λ_j is the *j*th non-zero eigenvalue of S_p, ϕ_j is the corresponding eigenvector, and k is an identity matrix multiplier. In equation (7.14), the following alternative representation of the identity matrix in terms of any set of orthonormal eigenvectors is used [Mar87]

$$I = \sum_{j=1}^{n} \phi_j (\phi_j)^T . \quad \textbf{(7.15)}$$

As can be seen from equation (7.14), a combination of S_p and a multiple of the identity matrix I as described in equation (7.10) expands all the S_p eigenvalues, independently of whether these eigenvalues are null, small, or even large.

A possible regularisation method for LDA could be the one that decreases the larger eigenvalues and increases the smaller ones, as briefly described by equation (7.12) of the previous sub-section. According to this idea, the eigen-decomposition of a convex combination of S_p and the *n* by *n* identity matrix I can be written as

$$\begin{aligned} \widehat{S}_p(k) &= (1-\gamma) S_p + \gamma \bar{\lambda} I \\ &= (1-\gamma) \sum_{j=1}^{r} \lambda_j \phi_j (\phi_j)^T + \gamma \sum_{j=1}^{n} \bar{\lambda} \phi_j (\phi_j)^T \end{aligned} \quad \textbf{(7.16)}$$

where the mixing parameter γ takes on values $0 \le \gamma \le 1$ and $\bar{\lambda}$ is the average eigenvalue of S_p.

Despite the substantial amount of computation saved by taking advantage of matrix updating formulas [Fri89, Ray90, Tad98], the regularisation method described in equation (7.16) would require the computation of the eigenvalues and eigenvectors of an *n* by *n* matrix for each training observation of all the classes in order to find the best mixing parameter γ. In recognition applications where several classes and a large total number of training observations are considered, such as face recognition, this regularisation method might be unfeasible.

Yet, equation (7.16) describes essentially a convex combination between a singular or poorly estimated covariance matrix, the pooled covariance matrix S_p, and a non-singular or well-estimated covariance matrix: the identity matrix I. Therefore, the same idea described in chapter 5 of selecting the most reliable linear features when blending such covariance matrices can be used.

Since the estimation errors of the non-dominant or small eigenvalues are much greater than those of the dominant or large eigenvalues [Fuk90], we propose the following selection algorithm in order to expand only the smaller and consequently less reliable eigenvalues of S_p, and keep most of its larger eigenvalues unchanged:

i. Find the Φ eigenvectors and Λ eigenvalues of S_p, where $S_p = S_w/[N-g]$;

ii. Calculate the S_p average eigenvalue $\bar{\lambda}$ using equation (7.13);

iii. Form a new matrix of eigenvalues based on the following largest dispersion values

$$\Lambda^* = diag[\max(\lambda_1,\bar{\lambda}),\max(\lambda_2,\bar{\lambda}),...,\max(\lambda_n,\bar{\lambda})]; \qquad \textbf{(7.17a)}$$

iv. Form the modified within-class scatter matrix

$$S_w^* = S_p^*(N-g) = (\Phi\Lambda^*(\Phi)^T)(N-g). \qquad \textbf{(7.17b)}$$

The new LDA (NLDA) is constructed by replacing S_w with S_w^* in Fisher's criterion formula described in equation (7.5). It is a straightforward method that overcomes both the singularity and instability of the within-class scatter matrix S_w when LDA is applied directly in limited sample and high dimensional problems. NLDA also avoids the computational costs inherent to the aforementioned shrinkage processes.

Figure 7.1 illustrates the geometric idea of the new LDA modification on a two-dimensional feature space. The constant probability density contours of S_p and S_p^* for two hypothetical "Gaussian-like" sample classes are represented respectively by the grey and black ellipses respectively. As can be seen, the new LDA expands S_p and might increase slightly the two classes overlap. However, the same optimum linear mapping v would be found by a Fisher's criterion based on the within-class variability given by S_p or S_p^*.

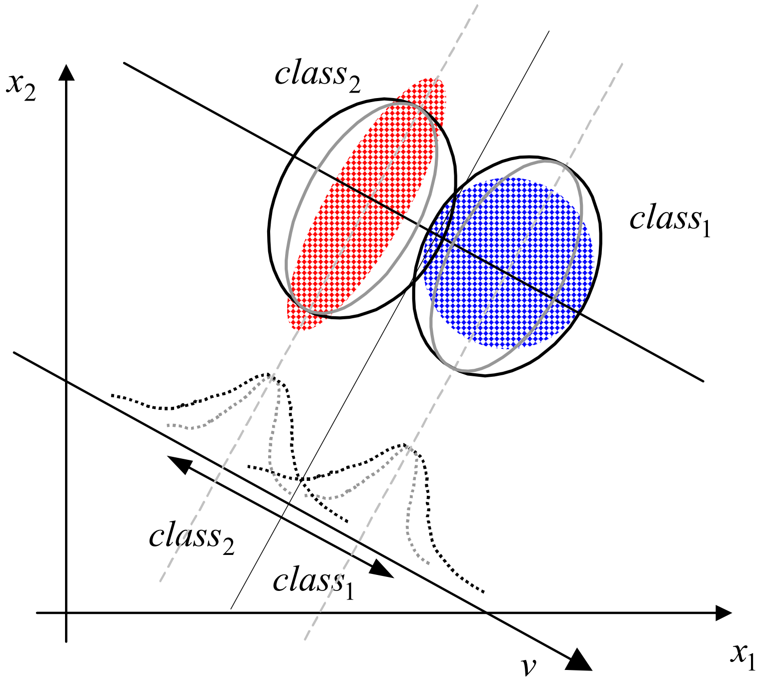


Figure 7.1. Geometric idea of the new LDA-based method.

Therefore, the main idea of the proposed LDA-based method can be summarised as follows. In limited sample size and high dimensional problems where the within-class scatter matrix is singular or poorly estimated, it is reasonable to expect that the Fisher's linear basis found by minimizing a more difficult "inflated" within-class S_p^* estimate would also minimize a less reliable "shrivelled" within-class S_p estimate.

7.4 Experiments

In order to evaluate the effectiveness of the new LDA-based method (NLDA) on face recognition, comparisons with the standard LDA (when possible), Fisherfaces, CLDA, DLDA, and YLDA, were performed using the ORL and FERET face databases previously described in the sub-sections (4.1.3.1) and (5.2.2.1).

A simple Euclidean distance classifier was used to perform classification in the projective feature space, analogously to the other approaches we investigated. Each experiment was repeated 25 times using several features. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate were calculated. The classification of the 40 subjects in the ORL database was computed using for each individual 5 images to train and 5 images to test. In the FERET database with 200 subjects, the training and test sets were respectively composed of 3 and 1 images.

For implementation convenience, the ORL face images were resized to 32x32 pixels, representing a recognition problem where the within-class scatter matrix is singular, that is the total number of training observations was $N = 200$ and the dimensionality of the original images was $n = 1024$. The FERET images were resized to 16x16 pixels in order to pose an alternative pattern recognition problem where the within-class scatter matrix is non-singular but poorly estimated, i.e. $N = 600$ and $n = 256$.

To determine the number of principal components to be retained in the intermediate step of Fisherfaces, experimental analyses were carried out based on the best classification accuracy of several PCA features in the corresponding interval $(g, N-g)$. The best results were obtained when the ORL and FERET original images were first reduced respectively to 60 and 200 PCA features.

For the purpose of establishing the number of the YLDA best discriminant vectors derived from the eigenspace of the within-scatter matrix, we used for the ORL database the eigenvectors corresponding to the remaining 10 largest eigenvalues, as suggested by Yang and Yang's work [YY03]. For the FERET database, the eigenvectors corresponding to the remaining 20 largest eigenvectors were sufficient to determine the respective YLDA best discriminant vectors. We assumed that an eigenvalue λ is positive if $round(\lambda) > 0$.

7.5 Results

Tables 7.1 and 7.2 present the maximum test average recognition rates (with standard deviations) of the ORL and FERET databases over the corresponding number of PCA (when applicable) and LDA features.

Since the ORL face database contains only 40 subjects to be discriminated, the LDA features of the Fisherfaces, CLDA, DLDA, and NLDA were limited to 39 components. Using the remaining 10 largest eigenvalues, the number of YLDA discriminant vectors could be extended from 39 to 49 LDA features. Also, the notation "-" in the standard LDA (LDA) row of Table 7.1 indicates that the within-class scatter matrix was singular and consequently the standard LDA could not be calculated.

Table 7.1 shows that the new LDA (NLDA) led to higher classification accuracies than the other one-stage approaches. The overall best classification result was reached by Yang and Yang’s approach (YLDA) – 96.1% (1.4%) – which was not significantly greater than the NLDA one – 95.8% (1.6%). However, the YLDA used a much larger two-stage linear transformation matrix compared to the one-stage methods. In terms of how sensitive the NLDA results were to the choice of the training and test sets, it is fair to say that the new LDA standard deviations were similar to the other methods.

	Features		
Method	PCA	LDA	Recognition Rate
Fisherfaces	60	39	94.9% (1.9%)
YLDA	199	45	96.1% (1.4%)
LDA	-	-	-
CLDA		39	95.4% (1.5%)
DLDA		39	94.9% (1.6%)
NLDA		39	95.8% (1.6%)

Table 7.1. ORL (32x32 pixels) LDA classification results.

Table 7.2 presents the results of the FERET database. In this application, the within-class scatter was non-singular but poorly estimated and the standard LDA (LDA) could be applied directly on the face images.

	Features		
Method	PCA	LDA	Recognition Rate
Fisherfaces	200	20	91.5% (1.9%)
YLDA	256	92	94.7% (1.4%)
LDA		20	86.2% (1.9%)
CLDA		20	86.2% (1.9%)
DLDA		20	94.5% (1.3%)
NLDA		10	95.4% (1.4%)

Table 7.2. FERET (16x16 pixels) LDA classification results.

As can be seen from Table 7.2, the overall best classification result was achieved by NLDA – 95.4% (1.4%) – using remarkably only 10 features. Again, regarding the standard deviations, NLDA was shown to be as sensitive to the choice of the training and test sets as the other approaches investigated.

7.6 Memory Issues

According to Samal and Iyengar [SI92], images with 32x32 pixels and at least 4 bits per pixel are sufficient for face identification problems. However, it is possible that memory computation problems would arise when scatter matrices larger than 1024x1024 elements are used directly in the optimisation of the Fisher's criterion described in equation (7.5).

In fact, the PCA intermediate step that has been applied to project images from the original space into the face subspace has made not only some of the aforementioned LDA-based approaches mathematically feasible in limited sample size and high-dimensional classification problems, but also has allowed the within-class S_w and between-class S_b scatter matrices to be calculable on computers with a normal memory size [LKM99].

In the experiments described in the previous sections, our attention was focused on evaluating the new LDA-based performance in situations where the within-class scatter matrix was either singular or poorly estimated, without a PCA intermediate step of dimensionality reduction. However, it would be important to assess the proposed method in higher resolution images where the PCA intermediate step is made necessary to avoid such memory computation difficulties.

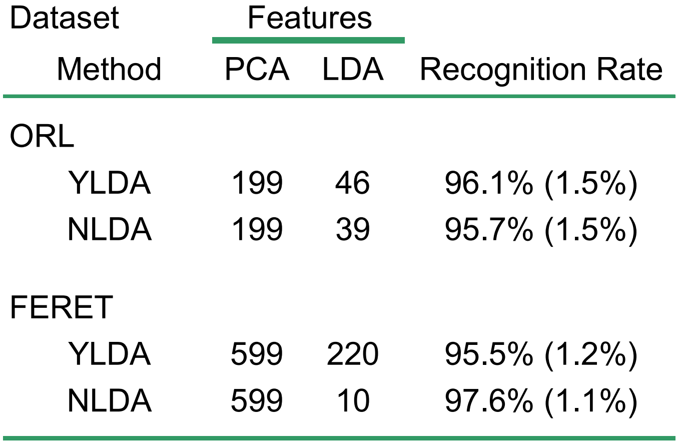
Thus, we discuss here further experimental results that evaluate the previous top 2 NLDA and YLDA approaches when the standard resolutions of 64x64 pixels and 96x64 pixels were used respectively for the ORL and FERET face images. Analogous to the previous experiments, the classification of the ORL 40 subjects was computed using in total 200 examples for training (5 images per subject) and the remaining 200 examples (5 images per subject) for testing. In the FERET database with 200 subjects, the total number of training and test sets were respectively composed of 600 (3 images per subject) and 200 (1 image per subject) images. Following Yang and Yang's work [YY03], we used again the eigenvectors corresponding to the remaining 10 largest eigenvalues to extend the number of YLDA discriminant vectors. For the FERET database, the eigenvectors corresponding to the remaining 25 largest eigenvalues were sufficient to determine the respective YLDA best discriminant vectors.

As described previously, the total number of principal components to retain for a best LDA performance should be equal to the rank of the total scatter matrix $S_T = S_w + S_b$ [YY03]. When the total number of training examples N is less than the dimension of the original feature space n, the rank of S_T can be calculated as [MN99]

$$\begin{aligned} rank(S_T) &\leq rank(S_w) + rank(S_b) \\ &\leq (N-g) + (g-1) \\ &\leq N-1. \end{aligned} \tag{7.18}$$

In order to avoid the high memory rank computation of such large scatter matrices and because both NLDA and YLDA deal with the singularity of the within-class scatter matrix, we used equation (7.18) to assume that the rank of S_T in both applications was $N-1$. Therefore, we first projected the original ORL and FERET images into the corresponding 199 and 599 largest principal components and secondly we applied the NLDA and YLDA feature classification methods.

Table 7.3 shows the maximum test average recognition rates (with standard deviations) of the ORL and FERET datasets over the corresponding number of PCA and LDA features.



Dataset	Features		
Method	PCA	LDA	Recognition Rate
ORL			
YLDA	199	46	96.1% (1.5%)
NLDA	199	39	95.7% (1.5%)
FERET			
YLDA	599	220	95.5% (1.2%)
NLDA	599	10	97.6% (1.1%)

Table 7.3. ORL (64x64 pixels) and FERET (96x64 pixels) LDA classification results.

As can be seen, as in the previous experiments, the best classification results for the ORL dataset was achieved by the Yang and Yang's approach (YLDA), which was slightly better than the NLDA result. However, the YLDA used a larger two-stage linear transformation matrix. In the FERET application, where the higher resolution images improved the classification results of both YLDA and NLDA approaches, the NLDA achieved clearly the best classification performance, using impressively only 10 LDA features again after the PCA dimensionality reduction.

7.7 Summary and Conclusions

In this chapter, we extended the idea of the maximum entropy selection method used in Bayesian classifiers to overcome the singularity and instability of the LDA within-class scatter matrix in limited sample, high dimensional problems.

Analogously to the procedure described in the previous chapter 5, the new LDA-based method is a straightforward approach that considers the issue of stabilising the ill posed or poorly estimated within-class scatter matrix with a multiple of the identity matrix. Although such modification has been used before, our method is based on selecting the largest and consequently most informative dispersions. Therefore, it avoids the computational costs inherent in the commonly used optimisation processes, resulting in a simple and efficient implementation for the maximisation of Fisher's criterion.

Experiments were carried out to evaluate this approach on face recognition, using the well-known ORL and FERET databases. Comparisons with similar methods, such as Fisherfaces [BHK97, ZCK98], Chen et al.'s [CLK00], Yu and Yang's [YY01b], and Yang and Yang's [YY01a, YY03] LDA-based methods, were made. In both databases, our method improved the LDA classification performance with or without a PCA intermediate step and using fewer linear discriminant features. Regarding the sensitivity to the choice of the training and test sets, the new LDA gave a similar performance to the compared approaches.

We have shown that, in limited sample size and high dimensional problems where the within-class scatter matrix is singular or poorly estimated, Fisher's linear basis found by minimising a more difficult but appropriate "inflated" within-class scatter matrix would also minimise a less reliable "shrivelled" within-class estimate. We believe that such LDA modification might be suitable for solving not only the singularity and instability issues of the linear Fisher methods, but also the Fisher discriminant analysis with kernels

[MRW99] where the non-linear mapping of the original space to a higher dimensional feature space would commonly lead to a ill-posed within class scatter matrix.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis, we have studied the importance and relevance of estimating reliable and computationally feasible covariance matrices for sparse and high dimensional statistical pattern recognition problems.

We frequently associate a covariance matrix with simple descriptive statistics that allow us to calculate the ellipsoidal dispersion of some data. In other words, we know that the covariance matrix is precisely calculated by a standard sample group covariance formula and its on-diagonal elements describe the spread along the main directions of an ellipsoidal shape, whereas its off-diagonal terms explain the orientation of this ellipsoid. However, the use of the covariance matrix in statistical classifiers is far more complex than its simple mathematical calculation or interpretation suggests.

As described in this thesis, the reason for that apparent contradiction is due to the fact that in constructing Bayesian classifiers based on Gaussian kernels we need to use the inverse of the covariance matrices. Hence, to estimate reliable covariance matrices in high dimensional spaces we need in general a large number of observations compared to the number of features. This assumption is quite inhibitive in practice and consequently the intuitive and mathematically convenient Gaussian kernel, which uses the covariance matrix given by the standard sample group covariance formula, cannot be used in such classification problems.

In the last years, possible ways of overcoming this limited sample size issue have been proposed. We have classified these covariance estimation approaches mainly in two categories. On the one hand, we have the covariance estimates based on some pre-defined structure and calculated in a straightforward way. For instance, the Toeplitz structure has been applied to approximate the sample group covariance matrix in Bayesian classifiers, but its restrictive form has not suited many real pattern recognition applications. On the other hand, we have flexible but optimised covariance estimates based on combining the singular sample group covariance estimate with non-singular appropriate covariance estimates. For example, the RDA and LOOC procedures have been used to maximise respectively the classification accuracy and the sample-group likelihood in parametric classifiers. As another example, the Van Ness approach has been used to maximise the classification accuracy as well, but in non-parametric Bayesian classifiers. Since these flexible approaches are based on optimisation indexes that involve one or two parameters, they are not only time-consuming but also exclusive to the type of the Bayesian classifier to be used.

Thus, what all the approaches that are reviewed have failed to suggest is a way of estimating reliable covariance approximations for the sample group covariance matrix that avoids both an over-simplification of its structure as well as an over-complication of its computation. As a consequence, the classical statistical problem of estimating class-conditional probability densities, which could be either specified or learned by using Gaussian kernels, has changed to the problem of estimating covariance matrices for either parametric or non-parametric class-conditional probability densities, but not necessarily both.

This thesis demonstrates that it is possible to calculate a reliable estimate for the sample group covariance matrix that does not have a pre-defined structure and does not require an iterative computation for its parameter estimation. By using the principle of maximum uncertainty and assuming that the covariance shapes of all the classes are not equal but share some similarities, we have shown that the maximum entropy covariance estimate (MECS) approach is not exclusive to the parametric nor the non-parametric Bayesian classifiers, and can replace the sample group covariance matrix whenever it is ill posed or poorly estimated.

In addition, we have demonstrated that this maximum uncertainty idea of combining singular or unstable covariance matrices with well estimated covariance matrices can solve the singularity and instability difficulties found in other multivariate statistical analysis, such as Linear Discriminant Analysis.

Finally, since the maximum entropy covariance estimate is based on selecting the most reliable dispersions of a mixture of covariance matrices, we do not expect that it will lead to the highest classification accuracy in all circumstances. Nonetheless, we believe it will provide at least a more parsimonious method of estimating efficient covariance structures for statistical covariance-based classifiers in limited sample size problems.

8.2 Future Work

There are some issues that have emerged from this work, which we believe are out of the scope of this thesis but might lead to future topics of investigation.

The first two issues, described in the next sub-sections, are more practical and are related basically to (1) the standard possibility of incorporating a reject option or threshold in the Bayesian classifiers based on the maximum entropy covariance estimate, and (2) the feasibility of quantifying the assumption that the covariance matrix estimations share some similarities. The third and fourth topics, presented in the last sub-sections, are more fundamental and concern (3) the relationship between the maximum entropy covariance estimate approach and the statistical framework of regularisation and (4) the variance-bias dilemma in limited sample size problems.

The Reject Threshold

In all classification experiments carried out in this work, we have assumed that only patterns belonging to one of the classes still existing in the training database would be presented to the statistical classifiers. However, there are some situations in which we should consider the possibility of having a pattern from an unknown class, that is, an impostor in the biometric context.

For this case we need not only to assign a new pattern x to the class π_i that has the maximum value calculated by multiplying its prior probability $p(\pi_i)$ with its corresponding likelihood $p(x \mid \pi_i)$, but also examine the value of its posterior probability, that is, $P(\pi_i \mid x)$. We would then be able to decide not to assign a pattern x to any of the classes if $P(\pi_i \mid x)$ is less than a specific probability value, that is, a reject threshold.

If we recall the well-known Bayes theorem described in chapter 2, the posterior probability of class π_i is given by

$$P(\pi_i \mid x) = \frac{p(x \mid \pi_i) p(\pi_i)}{\sum_{\text{all j}} p(x \mid \pi_j) p(\pi_j)}, \tag{8.1}$$

where $j = 1,2,\dots,g$ groups or classes. The practical form of the optimal Bayes decision rule defined for the case of a 0/1 or symmetrical loss function (described previously in equation (2.18)) can then be modified as follows: Assign input pattern x to class π_i if not only

$$p(x \mid \pi_i) p(\pi_i) = \max_{1 \le j \le g} p(x \mid \pi_j) p(\pi_j), \tag{8.2}$$

but also

$$P(\pi_i \mid x) \ge t, \tag{8.3}$$

where t is a threshold in the range of (0,1). The larger the value of t, the fewer points will be classified.

As we can see from equation (8.3), to incorporate the rejection option in the Bayesian classifiers we need to estimate only the posterior probability of the class π_i that satisfies equation (8.2). This is actually a very easy value to compute because in order to find the class π_i we need to calculate and compare the product $p(x \mid \pi_j) p(\pi_j)$ for all $j = 1,2,\dots,g$. The normalisation factor of the posterior probability $P(\pi_i \mid x)$ represented by its denominator does not require the computation of new values, but simply the sum of all those $p(x \mid \pi_j) p(\pi_j)$ products already computed.

In general, the total classification error rate would be reduced if the cases that are assigned to a class on the basis of a low probability of class membership were rejected [Jam85]. In fact, there are some situations where those low probabilities are not associated with "problematic" cases or patterns, but to the assumption of invalid forms for the true class-conditional densities, such as the multivariate Gaussian distribution. However, tests of multivariate normality have proved difficult to construct in limited sample size and high dimensional problems [JW98]. Therefore, in such situations, the use of the maximum entropy covariance estimate would be very helpful, because we can use not only the parametric Bayes Plug-in classifier but also the non-parametric Parzen Window classifier with Gaussian kernels to discriminate a new pattern and validate its corresponding posterior probability.

The Similarity of Covariance Matrix Estimations

Since the maximum entropy covariance estimate approach is based on the assumption that the covariance shapes of all classes are not equal but share some similarities, it would be interesting to investigate methods of quantifying that assumption. This would be particularly useful in applications where the sources of variation are not similar from group to group and consequently we could not expect a similar covariance shape for all the classes.

According to Anderson [And84], we can use a likelihood ratio criterion for testing the hypothesis that a covariance matrix Σ is similar to a given matrix Σ_0, as follows. Let $x_1, x_2, \ldots, x_N$ be a sample X of N observation vectors with n features or parameters drawn from the multivariate normal distribution $N_n(\mu, \Sigma)$. The ratio criterion for testing the hypothesis $H : \Sigma = \Sigma_0$, where Σ_0 is a specified covariance matrix, is [And84, pg. 435]

$$\kappa = \left(\frac{e}{N-1} \right)^{\frac{1}{2} n(N-1)} \left| B \Sigma_0^{-1} \right|^{\frac{1}{2}(N-1)} e^{-\frac{1}{2} tr B \Sigma_0^{-1}}, \qquad \textbf{(8.4)}$$

where B is the following scatter matrix

$$B = \sum_{j=1}^{N} (x_j - \bar{x})(x_j - \bar{x})^T = (N-1)S \qquad \textbf{(8.5)}$$

and $\bar{x}$ and S are respectively the mean and covariance matrix of the sample X. As expected, we can show that $\kappa = 1$ if

$$\Sigma_0 = S = \frac{1}{(N-1)} B \qquad \textbf{(8.6)}$$

that is,

$$\begin{aligned}
\kappa &= \left(\frac{e}{N-1}\right)^{\frac{1}{2}n(N-1)} \left| B\left(\frac{B}{(N-1)}\right)^{-1} \right|^{\frac{1}{2}(N-1)} e^{-\frac{1}{2}trB\left(\frac{B}{(N-1)}\right)^{-1}} \\
&= \left(\frac{e}{N-1}\right)^{\frac{1}{2}n(N-1)} \left|(N-1)BB^{-1}\right|^{\frac{1}{2}(N-1)} e^{-\frac{1}{2}tr\left((N-1)BB^{-1}\right)} \\
&= \left(\frac{e}{N-1}\right)^{\frac{1}{2}n(N-1)} \left[(N-1)^{n}\left|BB^{-1}\right|\right]^{\frac{1}{2}(N-1)} e^{-\frac{1}{2}(N-1)tr(BB^{-1})} \\
&= \left(\frac{e}{N-1}\right)^{\frac{1}{2}n(N-1)} (N-1)^{\frac{1}{2}n(N-1)} e^{-\frac{1}{2}n(N-1)} \\
&= 1.
\end{aligned} \qquad \textbf{(8.7)}$$

Thus, by using the formula described in equation (8.4), it would be possible to calculate the average similarity of the covariance estimates of all the classes, as follows:

$$\bar{\kappa} = \frac{2}{g(g-1)} \sum_{i=2}^{g} \sum_{j=1}^{i-1} \left(\frac{e}{N_i - 1}\right)^{\frac{1}{2}n(N_i-1)} (N_i - 1)^{\frac{1}{2}n(N_i-1)} \left|S_i S_j^{-1}\right|^{\frac{1}{2}(N_i-1)} e^{-\frac{1}{2}(N_i-1)trS_iS_j^{-1}} , \qquad \textbf{(8.8)}$$

where, as a reminder, n is the dimension of the feature space, N_i is the number of training patterns from class π_i, g is the total number of classes or groups, and S_i and S_j are respectively non-singular covariance estimates for the samples drawn from group i and j.

It is reasonable to expect that if the covariance shapes of all the classes share some similarities, the average likelihood ratio criterion $\bar{\kappa}$ described in equation (8.8) would give a result close to 1. However, further investigation must be done in order to evaluate the effectiveness of this analysis in practice, especially in limited sample and high dimensional problems.

Is MECS another regularisation method ?

One of the issues that have been raised in this work is the association of the maximum entropy covariance estimate with other regularisation methods of combining singular and non-singular covariance estimates to solve ill-posed or poorly estimated classification problems.

In order to answer this question, we recall the concept of the term "regularisation", previously described in section 3.4.1 of the chapter 3. According to Friedman [Fri89], the main idea of regularisation is to decrease the variance associated with the limited sample based estimate at the expense of potentially increased bias. The attempt is to approach the variance-bias trade-off by reducing the variability of ill-posed or poor estimates while biasing them toward values that are considered to be more physically plausible [Fri89].

As we have seen in this work, the idea of the maximum entropy covariance estimate is essentially to keep the dominant variance associated with the limited sample based estimate at the expense of potentially increased bias. Our attempt is essentially to increase the reliability of ill-posed or poor covariance estimates while biasing them toward physically plausible values. Therefore, we cannot say that MECS is another regularisation method.

The Variance-Bias Dilemma

In classification problems, it is well known that low variance tends to be more important than low bias. In other words, we are not supposed to be especially concerned if our estimation is biased, as long as the corresponding variance is kept low [DHS01]. Then the following question necessarily arises: Is MECS approaching the variance-bias trade-off in the wrong way ? The answer is: No, as far as limited information is concerned.

To clarify the previous statement, let us consider the following example. Figure 8.1 illustrates a two-dimensional feature space containing four observations drawn randomly from a specific small sample group or class.

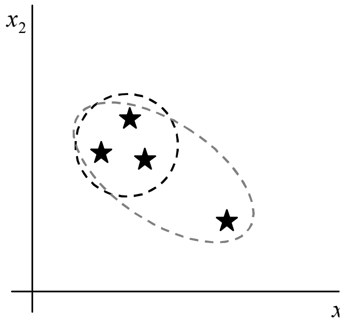


Figure 8.1. Scatter plot of four hypothetical observations belonging to a class.

In order to reduce the variance, we would tend to consider that the dashed black circle (with low variance) represents a better probability density contour for the sample than the dashed grey ellipse (with high variance) which encloses all the four observations. However there are only four observations to represent this probability density contour and consequently we cannot be certain that the fourth observation is actually an "outlier" and should be disregarded.

That is, by chance, we could have selected during the training stage the following four observations instead, illustrated in Figure 8.2. The two transparent stars are the ones shown in the previous Figure 8.1 and are displayed only for comparison. Analogously, we would tend to consider that the dashed black circle (with low variance) represents a better probability density contour for the new sample than the dashed grey ellipse (with high variance) which encloses all the four observations. However, as can be seen, the previous "problematic" fourth observation is not an "outlier" anymore.

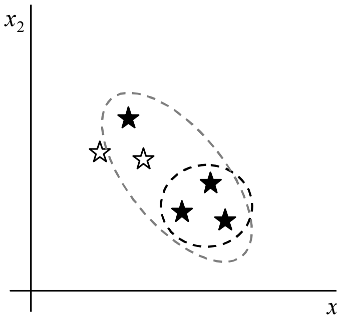


Figure 8.2. Scatter plot of other four hypothetical observations belonging to the class.

Although the example given above is simple, it illustrates our inevitable uncertainty in estimating reliable covariance structures when few examples per class are available. In such situations, we believe that the principle of maximum uncertainty should dominate the principle of minimum variance because in sparse conditions a high variance does not necessarily imply a weak match.

Bibliography

[And84] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, second edition. New York: John Wiley & Sons, 1984.

[BCG94] J. L. Blue, G. T. Candela, P. J. Grother, R. Chellappa and C. L. Wilson, "Evaluation of Pattern Classifiers for Fingerprint and OCR Applications", *Pattern Recognition*, vol. 27, pp. 485-501, 1994.

[BHK97] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[Bis97] C.M. Bishop, *Neural Networks for Pattern Recognition*, third edition. New York: Oxford University Press, 1997.

[BL00] L. Biehl and D. Landgrebe, "Effect of the Number of Samples Used in Leave-One-Out Covariance Estimator", Proceedings of SPIE International Symposium on Aerosense, Orlando, Florida, 24-28 April 2000.

[Cam80] N.A. Campbell, "Shrunken estimator in discriminant and canonical variate analysis", *Applied Statistics*, vol. 29, pp. 5-14, 1980.

[CLK00] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, 33 (10), pp. 1713-1726, 2000.

[CT91] T.M Cover and J.A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[DHS01] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, second edition. New York: John Wiley & Sons, 2001.

[DPi79] P.J. Di Pillo, "Biased Discriminant Analysis: Evaluation of the optimum probability of misclassification", *Communications in Statistics-Theory and Methods*, vol. A8, no. 14, pp. 1447-1457, 1979.

[DK82] P.A. Devijver and J. Kittler, *Pattern Classification: A Statistical Approach.* Prentice-Hall, Englewood Cliffs, N. J., 1982.

[DS85] D.K. Dey and C. Srinivasan, "Estimation of a covariance matrix under Stein's loss", *Annals of Statistics*, vol. 13, pp. 1581-1591, 1985.

[EM76] B. Efron and C. Morris, "Multivariate empirical Bayes and estimation of covariance matrices", *Annals of Statistics*, vol. 4, pp. 22-32, 1976.

[FF89] I. E. Frank and J. H. Friedman, "Classification: Oldtimers and Newcomers", *Journal of Chemometrics*, vol. 3, pp. 463-475, 1989.

[FH87] K. Fukunaga and D.M. Hummels, "Bayes error estimation using Parzen and k-NN procedures", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, pp. 634-643, 1987.

[Fis36] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[Fis38] R.A. Fisher, "The Statistical Utilization of Multiple Measurements", *Annals of Eugenics*, vol. 8, pp. 376-386, 1938.

[Fri89] J.H. Friedman, "Reguralized Discriminant Analysis", *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, March 1989.

[Fuk90] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition. Boston: Academic Press, 1990.

[GL89] G.H. Golub and C.F. Van Loan, *Matrix Computations*, second edition. Baltimore: Johns Hopkins, 1989.

[GR89] T. Greene and W.S. Rayens, "Partially pooled covariance matrix estimation in discriminant analysis", *Communications in Statistics-Theory and Methods*, vol. 18, no. 10, pp. 3679-3702, 1989.

[GR91] T. Greene and W.S. Rayens, "Covariance pooling and stabilization for classification", *Computational Statistics & Data Analysis*, vol. 11, pp. 17-42, 1991.

[Haf79] L.R. Haff, "Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity", *Annals of Statistics*, vol. 7, no. 6, pp. 1264-1276, 1979.

[Haf80] L.R. Haff, "Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix", *Annals of Statistics*, vol. 8, no. 3, pp. 586-597, 1980.

[Hay99] S. Haykin, *Neural Networks: A Comprehensive Foundation*, second edition. New Jersey: Prentice Hall, 1999.

[HFT96] Y. Hamamoto, Y. Fujimoto and S. Tomitan, "On the Estimation of a Covariance Matrix in Designing Parzen Classifiers", *Pattern Recognition*, vol. 29, no. 10, pp. 1751-1759, 1996.

[HJ85] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[HL96] J.P. Hoffbeck and D.A. Landgrebe, "Covariance Matrix Estimation and Classification With Limited Training Data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 763-767, July 1996.

[Hof95] J.P. Hoffbeck, "Classification of High Dimensional Multispectral Data", PhD thesis, Purdue University, West Lafayette, Indiana, 1995.

[Jam85] M. James, *Classification Algorithms.* London: William Collins Sons & Co. Ltd, 1985.

[Jay57] E.T. Jaynes, "Information Theory and Statistical Mechanics", *Physical Review*, vol. 106, pp. 620-630, 1957.

[Jay82] E.T. Jaynes, "On the rationale of maximum-entropy methods", *Proceedings of the IEEE*, vol. 70, pp. 939-952, 1982.

[JC82] A. K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice", *Handbook of Statistics*, P.R. Krishnaiah and L.N. Kanal Eds, vol. 2, pp. 835-855, North Holland, 1982.

[JDM00] A. K. Jain, R. P. W. Duin and J. Mao, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, January 2000.

[JR88] A. K. Jain and M. D. Ramaswami, "Classifier design with parzen windows", *Pattern Recognition and Artificial Intelligence*, E. S. Gelsema and L. N. Kanal Eds, pp. 211-228, North Holland, 1988.

[JW98] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, fourth edition. New Jersey: Prentice Hall, 1998.

[KS90] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103-108, 1990.

[KTT87] F. Kimura, K. Takashima, S. Tsuruoka and Y. Miyake, "Modified quadratic discriminant functions and the application to chinese character recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 149-153, 1987.

[Lac75] P. A. Lachenbruch, *Discriminant Analysis*. London: Hafner Press, 1975.

[LBA99] M. J. Lyons, J. Budynek, and S. Akamatsu, “Automatic Classification of Single Facial Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, December 1999.

[LCY93] K. Liu, Y. Cheng, and J. Yang, “Algebraic feature extraction for image recognition based on an optimal discriminant criterion”, *Pattern Recognition*, 26 (6), pp. 903-911, 1993.

[LKM99] Y. Li, J. Kittler, and J. Matas, “Effective Implementation of Linear Discriminant Analysis for Face Recognition and Verification”, *Computer Analysis of Images and Patterns: 8th International Conference CAIP'99*, Springer-Verlag LNCS 1689, pp. 232-242, Ljubljana, Slovenia, September 1999.

[LP85] S. P. Lin and M. D. Perlman, “A Monte Carlo Comparison of Four Estimators of a Covariance Matrix”, *Multivariate Analysis - VI*, pp. 411-429, 1985.

[Mar87] S.L. Marple, *Digital Spectral Analysis with Applications*. Englewood Cliffs, N.J: Prentice-Hall, 1987.

[MD74] S. Marks and O.J. Dunn, “Discriminant functions when the covariance matrices are unequal”, *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 555-559, June 1974.

[MN99] J.R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. Chichester: John Wiley & Sons Ltd., 1999.

[MRW99] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. –R. Muller, “Fisher discriminant analysis with kernels”, *IEEE Neural Networks for Signal Processing IX*, pp. 41-48, 1999.

[MYT87] V.R. Marco, D.M. Young, and D.W. Turner, “The Euclidean Distance Classifier: An Alternative to Linear Discriminant Function” *Communications in Statistics - Part B Simulation and Computation*, vol. 16, no. 2, pp. 485-505, 1987.

[OSA00] S. Omachi, F. Sun, and H. Aso, “A New Approximation Method of the Quadratic Discriminant Function”, *SSPR&SPR 2000,* Springer-Verlag LNCS 1876, pp. 601-610, 2000.

[OSu86] F. O’Sullivan, “A Statistical Perspective on Ill-Posed Inverse Problems”, *Statistical Science*, vol. 1, pp. 502-527, 1986.

[PJY88] R. Peck, L.W. Jennings, and D.M. Young, “A Comparison of Several Biased Estimators for Improving the Expected Error Rate of the Sample Quadratic Discriminant Function” *Journal of Statistical Computation and Simulation*, vol. 29, pp. 143-156, 1988.

[PN82] R. Peck and J. Van Ness, "The use of shrinkage estimators in linear discriminant analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 5, pp. 531-537, September 1982.

[PWH98] P. J. Phillips, H. Wechsler, J. Huang and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms", *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295-306, 1998.

[Ray90] W.S. Rayens, "A Role for Covariance Stabilization in the Construction of the Classical Mixture Surface", *Journal of Chemometrics*, vol. 4, pp. 159-169, 1990.

[RJ91] S.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252-264, 1991.

[Sea66] S. R. Searle, *Matrix Algebra for the Biological Sciences*: Wiley, 1966.

[Sha48] C. E. Shannon, "A Mathematical Theory of Communication", *The Bell System Technical Journal*, 27 (3), pp. 379-423 and 623-656, 1948.

[SI92] A. Samal and P. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey", *Pattern Recognition*, 25 (1), pp. 65-77, 1992.

[Sil86] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall, 1986.

[Str88] G. Strang, *Linear Algebra and its Applications*, third edition. Orlando: Harcourt Brace Jovanovich College Publishers, 1988.

[SW96] D. L. Swets and J. J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, 1996.

[Tad98] S. Tadjudin, "Classification of High Dimensional Data With Limited Training Samples", PhD thesis, Purdue University, West Lafayette, Indiana, 1998.

[TFV00] C.E. Thomaz, R.Q. Feitosa, A. Veiga, "Separate-Group Covariance Estimation with Insufficient Data for Object Recognition". In Proc. of *Fifth All-Ukrainian International Conference*, pp. 21-24, Ukraine, November 2000.

[TG01] C.E. Thomaz and D.F. Gillies. "'Small Sample Size': A Methodological Problem in Bayes Plug-in Classifier for Image Recognition", *Technical Report TR-2001-06*, Department of Computing, Imperial College, London, UK, June 2001.

[TG03a] C. E. Thomaz and D. F. Gillies. "Visual Analysis of the Use of Mixture Covariance Matrices in Face Recognition", in Proc. of the *4th International*

Conference of Audio- and Video-Based Biometric Person Authentication AVBPA'03, Springer-Verlag LNCS 2688, pp. 172-181, Guildford, UK, June 2003.

[TG03b] C. E. Thomaz and D. F. Gillies. "A New Fisher-based method applied to Face Recognition", in Proc. of the *10^{th} International Conference on Computer Analysis of Images and Patterns CAIP2003*, Springer-Verlag LNCS 2756, pp. 596-605, Groningen, The Netherlands, August 2003.

[TGF01a] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa, "Using Mixture Covariance Matrices to Improve Face and Facial Expression Recognitions", in Proc. of the *3^{rd} International Conference of Audio- and Video-Based Biometric Person Authentication AVBPA'01*, Springer-Verlag LNCS 2091, pp. 71-77, Halmstad, Sweden, June 2001.

[TGF01b] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa, "Small Sample Problem in Bayes Plug-in Classifier for Image Recognition", in Proc. of *Int'l Conference on Image and Vision Computing New Zealand*, pp. 295-300, Dunedin, New Zealand, November 2001.

[TGF02] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa, "A New Quadratic Classifier applied to Biometric Recognition", in Proc. of *Post-ECCV Int'l Workshop on Biometric Authentication*, Springer-Verlag LNCS 2359, pp. 186-196, Copenhagen, Denmark, June 2002.

[TGF03a] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa. "Using Mixture Covariance Matrices to Improve Face and Facial Expression Recognitions", extended version of the AVBPA'01 paper published in *Pattern Recognition Letters*, Special Issue on Biometric Person Authentication, vol. 24, no. 13, pp. 2159-2165, 2003.

[TGF03b] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa. "A New Covariance Estimate for Bayesian Classifiers in Biometric Recognition", *IEEE Transactions on Circuits and Systems for Video Technology,* Special Issue on Image- and Video-Based Biometrics, February 2004 (to appear).

[TG04] C. E. Thomaz and D. F. Gillies. "A Maximum Uncertainty LDA-based approach for Limited Sample Size problems - with application to Face Recognition, *Technical Report TR-2004-01*, Department of Computing, Imperial College, London, UK, January 2004.

[TL99] S. Tadjudin and D.A. Landgrebe, "Covariance Estimation With Limited Training Samples", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, July 1999.

[TP91] M. Turk and A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.

[VNe80] J. Van Ness, “On the Dominance of Non-Parametric Bayes Rule Discriminant Algorithms in High Dimensions”, *Pattern Recognition*, vol. 12, pp. 355-368, 1980.

[VS76] J. Van Ness and C. Simpson, “On the effects of dimension in discriminant analysis”, *Technometrics*, vol. 18, pp. 175-187, 1976.

[WCG92] C. L. Wilson, G. T. Candela, P. J. Grother, C. I. Watson, and R. A. Wilkinson, “Massively Parallel Neural Network Fingerprint Classification System”, *Technical Report NIST IR 4880, National Institute of Standards and Technology*, July 1992.

[WK77] P.W. Wahl and R.A. Kronmall, “Discriminant functions when the covariance are equal and sample sizes are moderate”, *Biometrics*, vol. 33, pp. 479-484, September 1977.

[Wol76] S. Wold, “Pattern Recognition by Means of Disjoint Principal Component Models”, *Pattern Recognition*, vol. 8, pp. 127-139, 1976.

[YY01a] J. Yang and J. Yang, “Optimal FLD algorithm for facial feature extraction”, *SPIE Proceedings of the Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision*, vol. 4572, pp. 438-444, 2001.

[YY01b] H. Yu and J. Yang, “A direct LDA algorithm for high dimensional data – with application to face recognition”, *Pattern Recognition*, vol. 34, pp. 2067-2070, 2001.

[YY03] J. Yang and J. Yang, “Why can LDA be performed in PCA transformed space? ”, *Pattern Recognition*, vol. 36, pp. 563-566, 2003.

[ZCK98] W. Zhao, R. Chellappa and A. Krishnaswamy, *“*Discriminant Analysis of Principal Components for Face Recognition*”*, in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition*, 336-341, 1998.