Dimensionality and prototype reduction techniques for pattern analysis

Qin, Kai

2007

Qin, K. (2007). Dimensionality and prototype reduction techniques for pattern analysis.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/3153>

<https://doi.org/10.32657/10356/3153>

Nanyang Technological University

Downloaded on 26 Apr 2023 22:35:12 SGT

Dimensionality and Prototype Reduction Techniques for Pattern Analysis

Qin Kai

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2007

ACKNOWLEDGEMENT

Acknowledgement

First of all, I would like to express my sincere and utmost gratitude to my supervisor, Associate Professor Ponnuthurai Nagaratnam Suganthan for his invaluable support, advices and encouragement throughout my Ph.D. study. I would like to thank Assistant Professor Marco Loog in IT University of Copenhagen, Demark for his insightful suggestions especially in the mathematical aspects. I am grateful that Nanyang Technological University (NTU) has provided me the research scholarship to support my Ph.D. study.

I owe special thanks to my colleagues, Ms. Huang Ling, Ms. Liang Jing, Mr. Tang Ke, Mr. Yemin Shi, Mr. Lin Jun and Mr. Cao runzi who have given me great encouragement and help in study, life and many other aspects.

I have received all kinds of support from a number of knowledgeable and friendly people in NTU. They have also shared great fun with me. I would like to thank all my lab colleagues to for having created a very comfortable research atmosphere.

Last but not the least, I would like to thank my parents. They have been with me every step of the way and I could not have completed anything without them.

Summary

This thesis investigates two important topics in the statistical pattern recognition field, namely dimensionality reduction for supervised classification and prototype reduction for unsupervised classification. For dimensionality reduction part, we concentrate on the Discriminative Linear Dimensionality Reduction (DLDR) techniques with feature extraction for supervised classification as the major application. For prototype reduction part, we focus on the prototype-based clustering algorithms.

A thorough literature survey is provided on two categories of DLDR techniques: (1) the Bayes Optimality-motivated DLDR (BODLDR) techniques, in which the criterion functions are related to the Bayes error formulated parametrically or non-parametrically under certain conditions and suboptimal solutions can be numerically derived at the expense of high computational costs; (2) The Scatter Matrix-based DLDR (SMDLDR) techniques, in which the criterion functions are directly formulated based on the geometric class structures of the training samples and globally optimal solutions in the closed form can be analytically derived at the expense of losing the link with the Bayes error. As their eminent unions, the Fisher Linear Discriminant Analysis (FLDA) technique as well as its heteroscedastic extension, the Chernoff Linear Discriminant Analysis (CLDA) technique proposed recently, is revisited in-depth. The Bayes optimality condition for the FLDA solutions is recapitulated. The potential deficiencies inherent in the conventional Fisher and Chernoff criteria are unveiled and discussed.

Two categories of enhanced parametric BODLDR techniques are developed under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. In the first category, a Gaussian homoscedastic direct BODLDR (DBODLDRho) technique and a class of Class-wise Weighted FLDA (CWFLDA) techniques motivated by the Bayes optimality under the Gaussian homoscedastic assumption are developed. The former one directly uses the union of the pairwise Bayes errors in the transformed subspace as the criterion function, which is analytically formulated under the Gaussian

SUMMARY

homoscedastic assumption. The gradient descent coupled with the Gram-Schmidt orthogonalization scheme is used to derive a suboptimal orthonormal transformation matrix. The later class of techniques employ the newly devised class-wise weighted Fisher criteria in quotient and subtraction forms, respectively. The discriminant vectors subject to different conjugated orthogonality constraints are extracted in the batch, sequential forward or sequential backward manner. In the second category, a Gaussian heteroscedastic direct BODLDR (DBODLDRhe) technique and a class of Class-wise Weighted CLDA (CWCLDA) techniques motivated by the Bayes optimality under the Gaussian heteroscedastic assumption are developed. The former one directly used the union of the pairwise Chernoff upper bounds of the Bayes error in the transformed subspace as the criterion function, which is analytically formulated under the Gaussian heteroscedastic assumption. A suboptimal orthonormal transformation matrix can be obtained by numerically optimizing the criterion function via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme. The later class of techniques employ the newly devised class-wise weighted Chernoff criteria in quotient and subtraction forms, respectively, with discriminant vectors subject to different conjugated orthogonality constraints being extracted in the batch, sequential forward or sequential backward manner. The class-wise weighted Chernoff criteria can be regarded as the heteroscedastic extension of the class-wise weighted Fisher criteria so as to capture the discriminatory information present in the difference between both class means and class covariance matrices. Experiments on many artificial and UCI multi-class datasets have demonstrated the promising performance of the proposed techniques in comparison with their existing archetypes.

The existing discriminative dimensionality reduction techniques capable of handling high-dimensional and undersampled problems are extensively overviewed along two main avenues: discriminative subspace analysis and discriminative manifold learning. Two classes of subspace based CWFLDA techniques, namely Complementary Subspaces-based CWFLDA (CS-CWFLDA) techniques and Regularized Subspace-based CWFLDA (RS-CWFLDA) techniques are developed, which can take advantage of the discriminatory information present in different subspaces while retaining the promising performance in well-sampled cases. Specifically, the discriminant vectors

SUMMARY

subject to different conjugated orthogonality constraints will be either extracted in the sequential forward manner from two complementary subspaces, i.e. the null and column spaces of the average within-class scatter matrix, or extracted in the batch or sequential backward fashion from a single subspace, i.e. the column space of the regularized average with-class scatter matrix, based on the quotient class-wise weighted Fisher criterion. The performance of the proposed techniques is evaluated on various types of high-dimensional data sources including handwritten digit images, facial images, text documents and microarray gene expression profiles, and compared with that of several state-of-the-art FLDA variants.

Many robust variants of the conventional prototype-based clustering algorithms and several prototype-based clustering algorithms that can automatically determine the optimal number of clusters are reviewed. A Robust Growing Neural Gas (RGNG) algorithm is developed for cluster analysis. By incorporating several robustifying strategies, such as the outlier resistant scheme, the adaptive modulation of learning rates and cluster repulsion method into the conventional Growing Neural Gas (GNG) framework, the proposed RGNG algorithm can be less sensitive to initializations, input sequence orders and the presence of outliers. Moreover, the RGNG can automatically determine the optimal number of clusters by seeking the extreme value of the Minimum Description Length (MDL) measure calculated at the completion of each growing stage. Furthermore, topology relationships among prototypes can also be established. Experimental results have demonstrated the superior performance of the proposed RGNG algorithm over the conventional GNG incorporating MDL method, named GNG-M.

Table of Contents

Acknowledgement	I
Summary	II
Table of Contents	V
List of Figures	VIII
List of Tables	X
Chapter 1: Introduction	1
1.1 Research background	1
1.2 Motivation	13
1.3 Objectives.....	17
1.4 Major contributions of the thesis.....	18
1.5 Organization of the thesis.....	21
Chapter 2: An Overview of Discriminative Linear Dimensionality Reduction	
Techniques	23
2.1 Introduction	23
2.2 Overview of BODLDR and SMDLDR techniques.....	25
2.2.1 The BODLDR techniques	25
2.2.1.1 Parametric BODLDR techniques.....	29
2.2.1.2 Non-parametric BODLDR techniques	39
2.2.2 The SMDLDR techniques.....	41
2.3 Fisher and Chernoff linear discriminant analysis.....	49
2.3.1 Fisher linear discriminant analysis (FLDA).....	49
2.3.1.1 Recapitulation of the FLDA.....	50
2.3.1.2 Bayes optimality analysis of the FLDA solutions.....	53
2.3.1.3 Conjugated orthogonality constraints on FLDA	59
2.3.1.4 Pros and cons of the Fisher criterion.....	63
2.3.1.5 The subtraction Fisher criterion	70
2.3.2 Chernoff linear discriminant analysis (CLDA).....	71
2.3.2.1 Related work	71
2.3.2.2 Recapitulation of the CLDA	72
2.3.2.3 Conjugated orthogonality constraints on CLDA.....	76
2.3.2.4 Pros and cons of the Chernoff criterion	77
2.3.2.5 The subtraction Chernoff criterion.....	79
Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems	80
3.1 Introduction	80
3.2 Gaussian homoscedastic direct BODLDR	81
3.2.1 Existing technique	82
3.2.2 Proposed technique	83

TABLE OF CONTENTS

3.3	Class-wise weighted FLDA motivated by Gaussian homoscedastic Bayes optimality	85
3.3.1	Existing techniques	85
3.3.2	Proposed techniques	89
3.4	Gaussian heteroscedastic direct BODLDR	97
3.4.1	Existing technique	98
3.4.2	Proposed technique	99
3.5	Class-wise weighted CLDA motivated by Gaussian heteroscedastic Bayes optimality	100
3.6	Experimental results	105
3.6.1	Experimental setup	106
3.6.2	Experiments on artificial data	108
3.6.3	Experiments on UCI data	117
3.7	Conclusions	139
Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems		140
4.1	Introduction	140
4.2	Review of DDR on high-dimensional and undersampled problems	142
4.2.1	Discriminative subspace analysis	143
4.2.2	Discriminative manifold learning	154
4.3	Complementary subspaces based CWFLDA techniques	158
4.4	Regularized subspace based CWFLDA techniques	165
4.5	Experiments	168
4.5.1	Experimental setup	169
4.5.2	Result analysis	172
4.6	Conclusions	183
Chapter 5: Robust Growing Neural Gas Algorithm with Application in Cluster Analysis		184
5.1	Introduction	184
5.2	Review of growing neural gas algorithm	191
5.3	Robust growing neural gas algorithm	196
5.3.1	Outlier resistant strategy	197
5.3.2	Adaptive learning rates and cluster repulsion scheme	200
5.3.3	Determination of the optimal number of clusters	202
5.3.4	Implementation of the RGNG algorithm	206
5.4	Experimental results	209
5.4.1	Experimental setup	209
5.4.2	Experiments on artificial data	210
5.4.3	Experiments on UCI data	220
5.5	Conclusions	221
Chapter 6: Conclusions and Future Work		223
6.1	Conclusions	223
6.2	Recommendations for future work	226
Author's Publication List		231

TABLE OF CONTENTS

Bibliography..... 234

List of Figures

Figure 1.1: Flow chart of a typical statistical pattern recognition system	2
Figure 2.1: Illustration of the 2D three classes with Gaussian distributions.....	67
Figure 2.2: One-dimensional class distributions after mapping the original feature space on $\mathbf{p}_1$ (left) and $\mathbf{p}_2$ (right), respectively.	67
Figure 3.1: (a) The mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier on D_1 . (b) The mean classification error versus the reduced dimension by using Chernoff criteria based techniques followed by the LDC classifier on D_1	112
Figure 3.2: (a) The mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier on D_2 . (b) The mean classification error versus the reduced dimension by using Chernoff criteria based techniques followed by the QDC classifier on D_2	114
Figure 3.3: (a) The mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier on D_3 . (b) The mean classification error versus the reduced dimension by using Chernoff criteria based techniques followed by the QDC classifier on D_3	116
Figure 3.4: The RBPI value averaged over all the experimental datasets for each experimental DLDR technique in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.....	130
Figure 3.5: The RAPI value averaged over all the experimental datasets for each of the enhanced parametric BODLDR techniques under the Gaussian homoscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.	133
Figure 3.6: The ROPI value for each of the enhanced parametric BODLDR techniques under the Gaussian homoscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.	134
Figure 3.7: The RAPI value averaged over all the experimental datasets for each of the enhanced parametric BODLDR techniques under the Gaussian heteroscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.	137

LIST OF FIGURES

Figure 3.8: The ROPI value for each of the enhanced parametric BODLDR techniques under the Gaussian heteroscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN	138
Figure 4.1: The RBPI (top), RAPI (middle) and ROPI (bottom) values of each experimental technique averaged over all the experimental datasets in terms of the k NN classifier.....	181
Figure 4.2: The RBPI (top), RAPI (middle) and ROPI (bottom) values of each experimental technique averaged over all the experimental datasets in terms of the NC classifier.....	182
Figure 5.1: (a) Plot of the clean dataset D_1 ; (b) Plot of the contaminated dataset D_1 ; (c) and (d) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the clean D_1 ; (e) and (f) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the contaminated D_1	215
Figure 5.2: (a) Plot of the clean multimodal data set D_2 ; (b) Plot of the contaminated multimodal data set D_2 ; (c) and (d) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the clean D_2 ; (e) and (f) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the contaminated D_2	219

List of Tables

Table 3.1: The 36 DLDR techniques involved in the experiments. Information is provided on technique index, technique name, dimensionality reduction manner, conjugated orthogonality constraint and technique provenance in this thesis.....	105
Table 3.2: The 10 datasets used in our experiments. Information is provided on dataset index, dataset name, number of classes (C), feature space dimensionality (n), total number of samples (N) and dimensionality after applying the PCA (PC).....	107
Table 3.3: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 36 experimental techniques as well as the mean classification error obtained in the original feature space (org) with respect to the LDC classifier. Note that each row denotes a specific experimental technique with its technique index indicated in Table 3.1, and each column represents a specific dataset with its dataset index indicated in Table 3.2.....	120
Table 3.4: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 36 experimental techniques as well as the mean classification error obtained in the original feature space (org) with respect to the QDC classifier. Note that each row denotes a specific experimental technique with its technique index indicated in Table 3.1, and each column represents a specific dataset with its dataset index indicated in Table 3.2.....	122
Table 3.5: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 36 experimental techniques as well as the mean classification error obtained in the original feature space (org) with respect to the k NN classifier. Note that each row denotes a specific experimental technique with its technique index indicated in Table 3.1, and each column represents a specific dataset with its dataset index indicated in Table 3.2.....	123
Table 3.6: The group indices ($Gidx$) and the indices of the enhanced parametric BODLDR techniques ($Eidx$) and their possibly existing archetypes techniques ($Aidx$) within each group.....	128
Table 3.7: The group indices ($Gidx$) and the indices of the enhanced parametric BODLDR techniques ($Eidx$) under the Gaussian homoscedastic assumption and their possibly existing archetypes techniques ($Aidx$) within each group.....	131
Table 3.8: The group indices ($Gidx$) and the indices of the enhanced parametric BODLDR techniques ($Eidx$) under the Gaussian heteroscedastic	

LIST OF TABLES

assumption and their possibly existing archetypes techniques ($Aidx$) within each group.	135
Table 4.1: Pseudo-code of the CS-UCWFLDA technique	161
Table 4.2: The MECS method used in the MUFLDA	167
Table 4.3: The 20 DLDR techniques evaluated in our experiments. Information is provided on technique index, technique name and their provenance in this thesis or previous literatures.....	168
Table 4.4: The 10 datasets used in our experiments. Information is provided on dataset index, dataset name, number of classes (C), feature space dimensionality (n) and total number of samples (N).	170
Table 4.5: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by the null space related FLDA (NS-FLDA) techniques with respect to the k NN classifier.	173
Table 4.6: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by the null space related FLDA (NS-FLDA) techniques with respect to the NC classifier.	174
Table 4.7: The average dimensionality (standard deviation in bracket) of $C(\mathbf{S}_T)$ and $N^d(\mathbf{S}_w)$ over 20 pairs of training and testing sets with regard to different data splits, determined by the matrix rank and matrix stability index, respectively.	174
Table 4.8: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 20 experimental techniques as well as the mean classification errors obtained in the original feature space (org) with respect to the k NN classifier.	177
Table 4.9: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 20 experimental techniques as well as the mean classification errors obtained in the original feature space (org) with respect to the NC classifier.	178
Table 5.1: Pseudo-code of the Growing Neural Gas (GNG) algorithm.....	194
Table 5.2: Pseudo-code of the Robust Growing Neural Gas (RGNG) algorithm.....	207
Table 5.3: Clustering results of GNG-M and RGNG algorithms on datasets D_1 over 10 runs (standard deviation in bracket).	215
Table 5.4: Clustering results of GNG-M and RGNG algorithms on multimodal datasets D_2 over 10 runs (standard deviation in bracket).	216

LIST OF TABLES

Table 5.5: Clustering results of GNG-M and RGNG algorithms on “Wine” datasets over 10 runs (standard deviation in bracket).....	221
--	-----

Chapter 1: Introduction

1.1 Research background

“Pattern” generally refers to information that characterizes an entity (object) existing in the specific time and space zones, which can be described by the outputs of some observing, measuring and/or recording equipments. Representations of patterns vary from the properties of the investigated objects, which may be, digit images, text documents, state vectors depicting a physical system, DNA sequences, time series signals and commercial transaction records, etc. The primary goal of a pattern analysis system is to detect any relation, regularity and structure inherent in the data source by analyzing or learning from some already available patterns sampled from the data source. Consequently, the system can acquire some generalization power of using these discovered relations, regularities and structures to make reasonable predictions on new previously unseen patterns coming from the same data source. Generally speaking, pattern analysis consists of three major tasks: (1) supervised tasks, in which each available input pattern is associated with a target output (label) and the objective is to construct a prediction function based on the available labeled input patterns so as to accurately predict the target outputs of future input patterns. (2) Unsupervised tasks, in which no target outputs are provided in advance and the objective is to extract relations, regularities and structures inherent in the data source based on the intrinsic similarities among the available unlabelled patterns. (3) Semisupervised tasks, in which target outputs are only partially known and the objective is to construct a prediction function based on the available partially labeled patterns.

Pattern recognition plays a central role in pattern analysis, which includes three major tasks, namely supervised, semisupervised and unsupervised classification of input patterns into the pre-specified distinct categories or inherently natural groups. In general, the available labeled or unlabelled patterns and the newly presented unlabeled patterns are called training and testing samples, respectively. The procedure of constructing a pattern recognizer based on the training samples so as to capture

relations, regularities or structures inherent in the data source from which the training samples came is called training (learning) process. The procedure of evaluating the established recognizer on the testing samples is called testing (predicting) process. According to different types of pattern representations, pattern recognition tasks can be further dichotomized into statistical pattern recognition [1], [2] and syntactical pattern recognition [3]. The former one deals with patterns in the vector form under specific assumptions on their statistical distributions. The later one handles structured patterns such as strings, sequences and semantics with little requirement of statistics. This thesis work will only focus on statistic pattern recognition, in which patterns are always represented as feature vectors in the feature space with dimensionality equal to the number of features.

Figure 1.1 shows the flow chart of a typical statistical pattern recognition system, which consists of 4 basic components to address different aspects of a pattern recognition task. The function of each component and where the research work in this thesis locates are described as follows.

Figure 1.1: Flow chart of a typical statistical pattern recognition system

- **Pre-processing**

Input signals to feed into a statistical pattern recognition system are first sensed and measured by some transducers, such as cameras, microphones and physical peripheral equipments, where the sensing environment and characteristics of the sensing equipments may influence the quality of the measurement. Next, the objects of interest are segmented from the background followed by noise removal, normalization and registration. The resulting patterns are represented as feature vectors.

- **Feature extraction and selection**

Ideally, the number of features collected at the pre-processing stage, i.e. the feature space dimensionality, can be as large as possible so as to comprehensively characterize the objects of interest from plentiful aspects. However, in practice, high-dimensional feature vectors may not be suitable for the subsequent recognition tasks because of the following reasons: (1) pattern recognizers established in a high-dimensional feature space may be highly complex since the number of recognizer parameters typically grows exponentially with the feature space dimensionality. Training a highly complex recognizer based on the limited number of training samples may possibly lead to overfitting [1] and consequently poor generalization performance to predict test samples. (2) Large ratio between the feature space dimensionality and the training sample size may result in the “curse of the dimensionality” problem [4], [5], where the feature space is fairly sparse and consequently parameter estimation becomes ill-posed. As a result, the probability of misclassification will increase significantly or some recognizers even cannot be established at all. (3) The computational and memory costs to train a recognizer in the high-dimensional feature space may be too burdensome to make the system practical by using common computation facilities or on the field use. (4) The costs of measuring some features can be very expensive, which require discarding some less important features to reduce costs at the smallest expense of the degeneration of the recognition accuracy. (5) Many original features may be highly noisy, correlated with each other or irrelevant to the subsequent recognition task, which may disguise the underlying relations, regularities or structures of interest.

Therefore, it is desirable to seek for parsimonious pattern representations with the smaller number of salient features to improve the efficiency and effectiveness of the subsequent recognizer. Feature extraction and selection [6] are two most commonly used techniques for this purpose.

The distinction between feature extraction and selection is that feature extraction techniques create a set of new features by transforming the original features in a linear

or nonlinear manner while feature selection techniques select some useful features out of the original features and discard the others. Relatively speaking, feature selection can retain the physical meaning of those selected features, which can be essential for certain applications, while feature extraction can generate a set of new features with stronger description or discrimination power and condensed size but may lose the physical interpretation. Moreover, feature selection techniques can save the sensing and memory costs by reducing the number of original features that must be measured and stored, while most feature extraction techniques will combine linearly or nonlinearly all the raw features to generate a new feature set and thus require all the raw features to be retained. Feature extraction and selection can be exclusively applied or used in conjunction for a given task. The order of executing feature extraction and selection is problem-dependent. In some cases, feature extraction may precede feature selection to first create a set of task-oriented features from the original features obtained at the pre-processing step, then to further select some salient features out of these new features. In other cases, feature selection may be first implemented to choose a smaller number of salient features out of the large original set, followed by feature extraction performed on these selected features to generate much more discriminative features with even condensed size to facilitate the subsequent recognition task. Note that there exists no explicit boundary between the feature extraction step and the subsequent recognition step. An ideal feature extractor may even directly map the feature vectors to their corresponding labels and thus leave a trivial task for the subsequent recognizer, or the feature extraction step can be embedded into the recognizer.

Dimensionality reduction [7], [8] attempts to find a lower-dimensional representation of the original feature space via linear or nonlinear mappings such that most useful information present in the original feature space can be retained in the transformed lower-dimensional space. Since feature extraction and selection can commonly result in the reduction of the feature space dimensionality, they are often referred to as dimensionality reduction techniques in pattern recognition literature. In fact, feature selection can be regarded as a special feature extraction task by using a linear transformation matrix having constrained binary elements. Besides pattern

recognition, dimensionality reduction can also be applied for visualization, exploratory data analysis and manifold learning, etc. During the past several decades, many dimensionality reduction techniques have been developed and applied successfully in a diverse spectrum of scientific and engineering disciplines ranging from multivariate statistical analysis, scientific visualization, multimedia retrieval and computer vision to web mining and bioinformatics. Dimensionality reduction can generally be categorized into the descriptive, discriminative and random projection techniques.

In descriptive techniques, the label information is unavailable or not taken into consideration. The general objective therein is to seek for a lower-dimensional pattern representation that maximally retains the global or local descriptive information in the original feature space. Some representative feature extraction techniques in this category include Principle Component Analysis (PCA) [9] and its variants such as local PCA [10] probabilistic PCA [11], kernel PCA [12], principle curves [13], online PCA [14], [15], [16], [17] and so on, self-organizing maps [18], generative topology maps [19], unsupervised projection pursuit [20], independent component analysis [21] and descriptive manifold learning such as multidimensional scaling [22], local linear embedding (LLE) [23], Hessian LLE [24], Isotropic Feature Mapping (ISOMAP) [25], diffusion maps [26], local tangent plane alignment [27] and Laplacian eigenmaps [28] and so on. It is worth noting that the unsupervised projection pursuit denotes a large family of methods. These methods generally start from a pre-specified linear projection matrix and continue to seek for better solutions based on the currently obtained one by numerically optimizing specific projection pursuit indices characterizing the amount of global or local descriptive information retained after the projection. For example, independent component analysis can be categorized into the unsupervised projection pursuit technique under a certain choice of the projection pursuit index. Some representative feature selection methods in this category include unsupervised feature selection based on maximum information compression index [29], feature selection for Gaussian mixture model [30] and feature subset selection using Expectation-Maximization clustering [31], etc.

In discriminative techniques, the label information will be taken into account. The ultimate goal is to map the original feature space into a lower-dimensional space with minimal loss of discriminatory information. Many representative feature extraction techniques in this category will be reviewed in Chapters 2 and 4. Feature selection techniques in this category can be generally subdivided into three groups, i.e. filter, wrapper and hybrid methods. Filter methods are independent of the recognizer to be established on the selected feature subset, which use the general characteristics of the training set as the evaluation criterion to select features. In contrast, wrapper methods will simultaneously consider feature selection and recognizer to be established on the selected feature subset. That is, the performance of a specific recognizer to be used subsequently will be directly used as the evaluation criterion to select features. Hybrid feature selection attempts to take advantage of both filter and wrapper methods by exploiting their different evaluation criteria in different stages. Representative feature selection techniques belonging to different groups have been extensively studied in many literatures [6], [32], [33], [34], [35]. It deserves noting that feature extraction techniques in this category can also be assigned to filter, wrapper and hybrid types. For example, the BODLDR techniques and SMDLDR techniques to be described later can be regarded as the wrapper and filter type techniques, respectively. In recent years, online discriminative dimensionality reduction techniques have received considerable attention [36], [37] which will not be covered in the current thesis work. However, this promising research direction has been recommended for our future work in Chapter 6.

Random projection [38] specifically refers to a category of feature extraction techniques. In this category, the original high-dimensional feature space is randomly projected into a lower-dimensional subspace, which attempt to retain the similarity relationship among the original high-dimensional feature vectors.

In comparison, descriptive and random projection based techniques are more general in the sense that they can be coupled with any kind of pattern analysis techniques for the subsequent unsupervised, supervised or semi-supervised tasks while discriminative techniques are apt to retaining the discriminatory information and thus more suitable for supervised classification tasks. Note that there exist many alternative

categorizations of dimensionality reduction techniques, e.g. linear and nonlinear, global and local and so on.

The research work in this thesis will mainly focus on the discriminative linear dimensionality reduction techniques with feature extraction for the supervised classification as the major application. Determination of the intrinsically optimal dimensionality that can be reduced to is one of the most challenging tasks in dimensionality reduction, which is out of the scope of this thesis work. For supervised classification tasks, some measures evaluating the generalization performance, such as the leave-one-out error and k -fold cross-validation error on the training samples [1] can be used to estimate the optimal dimensionality.

- **Recognition**

Feature vectors obtained from the feature extraction and selection step are ready to be used for training a pattern recognizer to perform supervised, unsupervised or semi-supervised classification tasks. This thesis will only cover supervised and unsupervised classification tasks, which will be described in more details as follows.

Existing supervised classification techniques can be summarized into four categories [39]: similarity-based approach, statistical decision theory based approach, decision boundary construction approach and decision tree approach.

- *Similarity-based approaches* include template matching such as rigid and deformable templates, minimum distance classifier such as nearest center classifier, nearest neighbor classifier and learning vector quantization. The key issues in designing this kind of classifiers are the proper choice of the similarity metric and prototype representation.
- *Statistical decision theory based approaches* are generally based on class distributions, among which the Bayesian decision rule and empirical Bayesian decision rule [1], [2] are two most popular methods to construct an optimal

classifier. In the former case, class-conditional probability density functions, class prior probabilities and loss functions are known in advance. In the later case, the class-conditional probability density functions are required to be estimated parametrically or non-parametrically based on the training samples. Representative parametric classifiers include the Linear Discriminant Classifier (LDC) and Quadratic Discriminant Classifier (QDC) under Gaussian assumptions and the logistic classifier based on the maximum likelihood of the logistic posterior probability. The Parzen window density estimation based technique is an example of the non-parametric methods.

- *Direct construction of geometrical decision boundary* by optimizing certain criterion functions relating to the classification performance leads to a popular family of supervised classification techniques, such as linear classifiers (e.g. least-square classifiers [40]), Neural Network based classifiers [41] and Support Vector Machine (SVM) [12], etc. Linear classifiers usually adjust the assumed linear decision boundary to well separate patterns belonging to distinct classes by directly minimizing the mean squared error between the classifier's outputs and the corresponding actual class labels. The complexity and characteristics of decision boundaries established by NN classifiers depend on the structure of the network, i.e. the number of hidden layers, the number of neurons per layer and the type of activation functions of each neuron. Overly simple network structure may lead to under-fitting and accordingly bad training and testing performance. However, highly complex network structures may result in over-fitting and consequently excellent training accuracy yet poor generalization performance. As a result, different regularization schemes have been introduced to find a tradeoff between the network complexity and training accuracy. It is worth noting that different network structures and the corresponding training strategies will yield different classifiers, some of which may be equivalent to the existing classifiers with other forms. A typical class of NN classifiers includes the multiplayer perception NN (e.g. feedforward, recurrent and convolution), the radial basis function NN, the auto-associative NN and a most recently developed extreme learning machine [42], etc. The SVM is another eminent state-of-the-art classifier.

The decision boundaries established in SVM have the linear form in the original feature space or in the high-dimensional kernel-induced feature space mapped via polynomial, radius basis function, sigmoid or tangent kernel. They are determined by optimizing a criterion function reflecting the trade-off between the margin and empirical training error. The resulting SVM decision boundaries will only depend on a few support vectors instead of the whole set of training samples. The increased sparseness reduces the computational complexity at the testing stage. A major advantage of SVM is that it can deal well with nonlinearly separable pattern distributions. Note that kernel tricks [12] originated from the SVM have been successfully used to extend many existing linear classification techniques into their kernel counterparts so as to enhance their abilities to handle the nonlinearly separable pattern distributions. Such kernel-based classification methods can sometimes be considered as a unique branch of classification techniques.

- *Decision tree* [43] is another popular class of classification methods, which iteratively selects salient features at each tree node to optimize certain criteria such as information gain, node purity and class separability. The main characteristic of the decision tree classifier is that it naturally embeds the feature selection step into the classifier training process and accordingly decision rules can be readily extracted at the completion of the training process. Moreover, it may facilitate the interaction with domain experts. Decision tree classifiers, such as CART and C4.5 have been successfully applied in many fields.

Traditionally, unsupervised classification techniques, namely clustering or cluster analysis [44]-[46], can be divided into the partitional clustering and hierarchical clustering. Partitional methods directly assign all available patterns, according to specific criteria such as similarity and density, into different natural groups. Hierarchical approaches sequentially build nested clusters with a graphical representation known as dendrogram. Representative algorithms for the partitional clustering include Prototype-based Clustering (PBC) [47], density-based clustering [48], [49] and graph theory based clustering [50], etc. Agglomerative [52] and divisive [53] clustering algorithms constitute two main types of hierarchical clustering. In

recent years, many novel clustering techniques have been developed such as neural network based clustering (e.g. self-organizing maps [18], Neural Gas (NG) networks [54] and adaptive resonance theory network [55], etc.), kernel clustering [56], [57], spectral clustering [59], information clustering [58], relationship based clustering [60], subspace clustering [51], evolutionary clustering [61], and online clustering [62], [63], [64], etc. These clustering algorithms can either fit into one of the two traditional categories or be regarded as a unique category of clustering techniques.

Among various supervised and unsupervised classification techniques, prototype reduction schemes [47], [65], [66] provide a fundamentally simple but very effective way to establish a class of prototype-based methods. In these methods, the training samples are condensed into a smaller set of representative prototypes that can capture the salient underlying structures relevant to the subsequent classification task. In general, the prototype can be parameterized to represent a subset of training samples. For example, a point prototype can be described by a reference vector and a proper similarity function involving several parameters to characterize the center, shape and size of a local region in the feature space. Accordingly, the training samples can be either assigned to their respective nearest prototypes or assigned to all prototypes with different degrees of memberships, based on the values of the similarity function. According to the operating characteristics, prototype reduction schemes can be dichotomized into the creative and selective methods [65]. The former one can create new previously unavailable reference vectors to construct prototypes while the later one merely selects some of the training samples to construct prototypes. Each of these two categories can be further divided into the static, growing and pruning approaches. In static approaches, the number of prototypes is fixed in advance with the associated parameters being adapted to optimize some performance measure. Growing approaches gradually increase the number of prototypes from the pre-specified minimum to maximum values. At each growing stage, some new prototypes are generated via specific rules. The optimal set of prototypes to be sought should optimize some performance measure over the whole growing process. Pruning approaches gradually decrease the number of prototypes from the pre-specified maximum to minimum values. At each pruning stage, some less useful prototypes

with respect to the given task are identified and removed. The eventually obtained optimal set of prototypes should optimize some performance measure over the whole pruning process. Note that static approaches can be readily extendible to growing or pruning approaches by introducing proper rules to increase or decrease the prototypes. The definitions of performance measures usually differ in the supervised and unsupervised classification. For supervised classification tasks, the training error and generalization performance will be simultaneously considered. For unsupervised classification tasks, the average similarity between the training samples and their respective prototypes and the average dissimilarity between the training samples assigned to different prototypes will be taken into account. In many prototype-based methods, the performance measure can be explicitly formulated as a criterion function taking the parameters of prototypes as its arguments, and accordingly the optimal set of prototypes can be derived via some numerical optimization techniques.

Many existing supervised classification techniques can be categorized into the prototype-based methods [47], such as minimum distance classifiers (prototypes characterize class centers), the Bayes classifier (prototypes characterize class distributions), the radial basis function NN (prototypes characterize radial basis function centers) and the SVM with the radial basis function kernel (prototypes characterize support vectors), etc. In these techniques, prototypes can be either used for the distribution estimation, e.g. minimum distance classifiers and radial basis function NN, or used to capture the decision boundary information between different classes, e.g. SVM. Prototype-based unsupervised classification (i.e. clustering) is a most popular class of partitional clustering techniques, which can also be generalized to the hierarchical clustering through superposition. Some typical characteristics of prototype-based clustering methods are: (1) the number of prototypes can be determined in advance or obtained automatically during the training procedure. (2) Each cluster is characterized by a prototype. According to their distances from all prototypes, calculated by certain similarity functions, input patterns can be assigned to different clusters with certain degrees of membership. (3) The optimal set of prototypes can be obtained by minimizing a criterion function defined as the average distance from the training samples to each prototype weighted by the corresponding

degrees of membership. Hard clustering and soft clustering can be regarded as two main branches of prototype-based clustering methods. Hard clustering assigns each pattern to exactly one cluster with the most “similar” prototype, while soft clustering assigns each pattern to all clusters with different degrees of membership. According to the updating strategies, prototype-based clustering methods can also be dichotomized into the batch and sequential versions.

The research work in this thesis will mainly focus on the prototype-based clustering algorithms with each prototype simply characterized by a reference vector and the Euclidean distance metric involving no additional parameters. Extensions to other distance metrics, e.g. the Mahalanobis distance involving the covariance matrix as an additional parameter, and more complex prototypes like shell prototypes [286] in which the similarity is measured with respect to specific shapes instead of a single reference vector, are straightforward and planned as the further research work. Note that the clustering algorithm proposed in this thesis is currently applied to offline clustering tasks by using the sequential learning scheme. The extension to the on-line version is recommended for the future work in Chapter 6.

- **Post-processing**

The same data source using different pattern representations or different data sources collected at various times and locations may result in distinct performance of the subsequent recognizer. Even under the same pattern representation, different recognizers or the same recognizer under different parameter settings can make distinct decisions. Therefore, it is desirable to combine individual decisions made from different recognizers based on the same pattern representation or from the same classifier built on pattern representations of different data sources to generate a consistent final decision. In recent years, ensemble learning [67], [287] in both supervised and unsupervised manner has become a very active research area in pattern analysis. Since the recognition results may further guide some actions from the system to the external environments, costs and risks associated with any recognition results are required to be considered in order to minimize the possible loss. In addition, some

contextual information can be explored from the available patterns, which may simplify some complex pattern recognition tasks. Moreover, available prior knowledge can be utilized effectively to guide the design of all components in a statistical pattern recognition system. Our research work will not cover the post-processing techniques.

In common sense, a desirable pattern recognition system must rely on good coordination among every components described above. Therefore, feedbacks from latter components to the former ones may be used to further improve the system performance.

1.2 Motivation

The research work in this thesis is motivated from two different areas in the statistical pattern recognition field: dimensionality and prototype reduction for supervised and unsupervised classification, respectively.

Since raw feature vectors obtained at the pre-processing stage may possibly contain noise, redundant or irrelevant features that may deteriorate the speed, reliability and accuracy of the subsequent classifier, it is desirable to reduce the feature space dimensionality as much as possible subject to the minimum loss of the discriminatory information. During the past few decades, a large number of dimensionality reduction techniques have been developed to seek for an optimal lower-dimensional transformed space that can maximally retain the useful descriptive or discriminative information present in the original feature space. It is generally believed that discriminative dimensionality reduction methods are more suitable for supervised classification than descriptive dimensionality reduction approaches because criterion functions in the former methods are directly related to the class distinguishability in the transformed space.

Discriminative dimensionality reduction includes discriminative feature extraction and selection techniques. In some literatures [68], it was argued that finding a

transformation to reduce dimensionality might be easier than selecting features provided an appropriate criterion function measuring the class distinguishability in the lower-dimensional transformed space. Such preference of feature extraction than selection relies on two-fold justifications: (1) feature extraction is usually a smooth process. When a criterion function is differentiable with respect to the parameters of the transformation, the optimal transformation can be derived via some numerical optimization techniques, such as the gradient descent. In contrast, feature selection is a discrete event, which generally requires resolving a prohibitively expensive combinatorial optimization problem. (2) Feature extraction can effectively combine the useful information spread across several original features and consequently result in a compact representation with the stronger discrimination power. However, in some applications, when it is crucial to discard some of the original features so as to reduce the sensing and storage costs or it is necessary to interpret the physical meaning of the selected features, the preference should be given to the feature selection techniques. This thesis will concentrate on the study of discriminative feature extraction techniques for discriminative dimensionality reduction. Therefore, in the following context, discriminative dimensionality reduction will specifically refer to discriminative feature extraction unless otherwise stated. Moreover, in the current thesis, we only consider the Discriminative Linear Dimensionality Reduction (DLDR) techniques since their nonlinear counterparts can be readily obtained via the kernel trick [12] or the functional replacement [69] concept.

In general, the performance of DLDR techniques relies on the class separability measure defined in the lower-dimensional transformed subspace. This measure evaluates the amount of discriminatory information retained after the linear transformation. Ideally, the Bayes error should be the optimal choice, which is determined by applying the Bayes decision rule to classify input patterns to one of the pre-specified distinct categories and represents the minimum probability of misclassification. It characterizes the source of error stemming from underlying class distributions, which can hardly be further reduced by any eminent classifiers. However, the Bayes errors under arbitrary class distributions and/or in the multi-class scenario usually cannot be analytically formulated since it is prohibitive to calculate

the integration of complex class distribution functions over multiple discontinuous integral regions. Moreover, even if simple class distributions in two-class cases are known in the original feature space, there is no guarantee that they can be derived in the transformed subspace, and accordingly the Bayes error in the transformed subspace still may not be computed. In the past few decades, many efforts have been devoted to the design of class separability measures that can relate to the Bayes error to certain degrees. These measures, usually taking the linear transformation matrix as a parameter, can be optimized via some numerical optimization techniques to seek for the optimal linear transformation. We named the class of techniques employing such kind of class separability measures as criterion functions BODLDR techniques since they attempt to find an optimal transformed subspace with the minimum Bayes error therein, i.e. pursuing the Bayes optimality. The common disadvantages of the BODLDR are the high computational complexity and the risk of getting stuck into local optima during the optimization. Another family of DLDR techniques directly uses the geometric class structures of the available training samples to construct the class separability measure, which can significantly reduce the computational complexity coming from the class distribution estimation and may also avoid the numerical optimization at the expense of the link with the Bayes error. Fisher Linear Discriminant Analysis (FLDA) is the most famous member in this family, which uses scatter matrices to formulate the class separability measure, called the Fisher criterion. The most attractive merit of the FLDA is that globally optimal solutions in the closed form can be derived analytically in an efficient manner. Although the optimal solution obtained by the FLDA can lead to a transformed subspace achieving the Bayes optimality under certain conditions, such intimate link with the Bayes optimality will be lost in more general cases since several potential deficiencies inherent in the conventional Fisher criterion prevent its link with the Bayes error in general sense. These facts motivate us to modify the conventional Fisher criterion so as to compensate those undesirable deficiencies.

The conventional Fisher criterion also suffers from a notorious limitation, which restricts the utility of the FLDA to handle high-dimensional and undersampled problems. Specifically, optimizing the conventional Fisher criterion with respect to the

transformation matrix will lead to solve an eigenvalue decomposition problem involving the inverse of some scatter matrix. However, in the high-dimensional and undersampled scenario, all the estimated scatter matrices will become singular, which consequently disables the applicability of the FLDA. To prevent the modified Fisher criterion suffering from such singularity problem [2], [70], we will embark on further generalizing its ability to deal with high-dimensional and undersampled problems by exploring the singularity overcoming schemes used in the existing FLDA variants exempting from the singularity problem.

Prototype-based clustering techniques, such as k -means clustering [71], fuzzy c -means clustering [72], Gaussian mixture model [1] and NG network [54], etc., have been applied widely in diverse fields and demonstrated promising performance. However, the robustness [73], [74] of most of these algorithms may not be satisfactory. By “robustness”, we mean that the performance of an algorithm should not be significantly influenced by the small deviations from the assumed model and meanwhile should not deteriorate much when the outliers exist. In term of cluster analysis, outliers can be regarded as the feature vectors far away from any clusters and consequently have low degrees of membership with respect to all clusters. Several major robustness issues associated with the prototype-based clustering can be outlined as follows.

- Sensitivity to initializations
- Sensitivity to the order of the input pattern sequence
- Sensitivity to the outliers present in the dataset to be dealt with
- Sensitivity to the subjectively pre-specified number of clusters
- Sensitivity to cluster sizes, shapes and densities

In this thesis, we will address the first four issues mentioned above based on a representative prototype-based clustering algorithm: Growing Neural Gas (GNG) [75] network, which employ the creative prototype reduction scheme to derive prototypes characterized by reference vectors and the Euclidean distance metric. The GNG actually belongs to the growing approaches and thus can enable automatic detection of

the number of underlying clusters. By enhancing its robustness, we expect that the performance of the improved algorithm can be less sensitive to initializations, input sequence orders, and the presence of outliers while the actual number of underlying clusters can be automatically discovered.

1.3 Objectives

Two major objectives of this thesis are to improve the performance of some existing DLDR techniques for the supervised classification and to enhance the robustness of some existing prototype-based clustering techniques for the unsupervised classification. More specifically, the objectives of our research work can be summarized as follows.

- To survey the existing BODLDR and SMDLDR techniques, which highlight the performance and efficiency, respectively. As a union of these two techniques, the FLDA as well as its heteroscedastic extension, the Chernoff linear discriminant analysis (CLDA) [76], [77] proposed recently, will be reviewed. The specific conditions under which the optimal linear transformation obtained by the FLDA can result in a transformed subspace achieving the Bayes optimality will be recapitulated. The potential deficiencies inherent in the conventional Fisher and Chernoff criteria, which destroy their link with the Bayes error in general sense, will be in-depth analyzed.
- To develop two categories of enhanced parametric BODLDR methods under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. In each category, two classes of techniques will be developed, which give higher priority to performance and efficiency, respectively. The first class directly formulates the upper bound of the multi-class Bayes error under the corresponding assumption as the criterion function, which can be optimized via numerical optimization methods to iteratively derive the suboptimal solutions. The second class employs the pairwise two-class Bayes error or its upper bound analytically formulated under the corresponding assumption to modify the conventional Fisher and Chernoff criteria

to the class-wise weighted Fisher and Chernoff criteria. Discriminant vectors subject to different conjugated orthogonality constraints will be extracted based on the class-wise weighted Fisher and Chernoff criteria, respectively. Experiments will be performed on many multi-class artificial and real-world datasets taken from the UCI machine learning repository [78] to evaluate the effectiveness of the proposed techniques in comparison with their archetypes. Some relative performance indices will be designed to comprehensively evaluate the performance of each experimental DLDR technique over all the experimental datasets.

- To survey the existing discriminative dimensionality reduction techniques suitable in the high-dimensional and undersampled scenario, and generalize the applicability of several representative class-wise weighted Fisher criterion based DLDR techniques to handle high-dimensional and undersampled problems. Extensive experiments will be conducted on various high-dimensional data sources to compare the performance of the generalized techniques with that of several state-of-the-art FLDA variants. Each experimental technique will be comprehensively evaluated according to the designed relative performance indices.
- To improve the robustness of the GNG for cluster analysis. The developed Robust GNG (RGNG) algorithm should be less sensitive to initializations, input sequence orders and the presence of outliers, and can automatically determine the optimal number of clusters at the completion of the growing process. The clustering performance of the proposed RGNG algorithm will be compared with that of the original GNG algorithm on some artificial and UCI datasets.

1.4 Major contributions of the thesis

The major contributions in this thesis are summarized as follows.

- We conducted a thorough overview of the existing BODLDR and SMDLDR techniques. As their eminent unions, the FLDA technique as well as its heteroscedastic extension, the CLDA technique proposed recently, was reviewed.

The Bayes optimality condition for the FLDA solutions was revisited. The potential deficiencies inherent in the conventional Fisher and Chernoff criteria, which destroy their link with the Bayes error in general sense, were unveiled and discussed in-depth.

- We developed two categories of enhanced parametric BODLDR techniques under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. *In the first category*, a Gaussian homoscedastic direct BODLDR technique and a class of 8 Class-wise Weighted FLDA (CWFLDA) techniques motivated by Bayes optimality under the Gaussian homoscedastic assumption were developed. The former one directly used the union of the pairwise Bayes errors in the transformed subspace as the criterion function, which was analytically formulated under the Gaussian homoscedastic assumption. A suboptimal orthonormal transformation matrix can be obtained by numerically optimizing the criterion function via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme. The later class of techniques employed the newly devised class-wise weighted Fisher criteria in the quotient and subtraction forms, respectively, with discriminant vectors subject to different conjugated orthogonality constraints being extracted in the batch, sequential forward or sequential backward manner. The newly proposed class-wise weighted Fisher criteria compensated the deficiencies of the conventional Fisher criterion. Specifically, the between-class weights and within-class weights were introduced into the conventional between-class and average within-class scatter matrices, respectively. The resulting weighted pairwise between-class scatter matrix can somewhat approximate the average pairwise two-class Bayes accuracy instead of representing the average pairwise distance over all class pairs. The resulting average weighted within-class scatter matrix can suppress the undesirable contributions from the outlier classes (some classes that are faraway from the remaining classes) to the estimation of the assumed homoscedastic covariance matrix. *In the second category*, a Gaussian heteroscedastic direct BODLDR technique and a class of 8 Class-wise Weighted CLDA (CWCLDA) techniques motivated by Bayes optimality under the Gaussian heteroscedastic assumption were developed. The former one directly used the union of the pairwise

Chernoff upper bounds of the Bayes error in the transformed subspace as the criterion function, which was analytically formulated under the Gaussian heteroscedastic assumption. A suboptimal orthonormal transformation matrix can be obtained by numerically optimizing the criterion function via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme. The later class of techniques employed the newly devised class-wise weighted Chernoff criteria in the quotient and subtraction forms, respectively, with discriminant vectors subject to different conjugated orthogonality constraints being extracted in the batch, sequential forward or sequential backward manner. The class-wise weighted Chernoff criteria can be regarded as the heteroscedastic extension of the class-wise weighted Fisher criteria so as to capture the discriminatory information present in the difference between both class means and class covariance matrices. We conducted experiments on several multi-class artificial and UCI datasets to evaluate the effectiveness of the proposed enhanced parametric BODLDR techniques in comparison with their archetypes¹ in terms of both optimal and average dimensionality reduction performance with respect to three classifiers, i.e. LDC, QDC, k -Nearest Neighbor (k NN) classifier. Three relative performance indices, namely relative best performance index, relative average performance index and relative overall performance index, were developed to comprehensively evaluate the performance of each experimental technique.

- We conducted a thorough overview of the existing discriminative dimensionality reduction techniques capable of handling high-dimensional and undersampled problems. To generalize the applicability of several representative CWFLDA techniques in the high-dimensional and undersampled scenario, we developed two classes of subspace based CWFLDA techniques, which can take advantage of the discriminatory information present in different subspaces while retaining the promising performance in the well-sampled cases. Specifically, the discriminant vectors subject to different conjugated orthogonality constraints were either extracted in a sequential forward manner from two complementary subspaces, i.e.

¹ We also proposed some of these archetypes based on the conventional Chernoff criterion, which extend corresponding FLDA techniques in a heteroscedastic manner.

the null and column spaces of the average within-class scatter matrix, or extracted in a batch or sequential backward fashion from a single subspace, i.e. the column space of the regularized average within-class scatter matrix, based on the class-wise weighted Fisher criterion. The effectiveness of the proposed techniques was extensively evaluated on various types of high-dimensional data sources, such as handwritten digit images, facial images, text documents and microarray gene expression profiles, and compared with that of several state-of-the-art FLDA variants.

- We developed a RGNG algorithm for cluster analysis. By incorporating several robustifying strategies, such as outlier resistant scheme, adaptive modulation of learning rates and cluster repulsion method into the conventional GNG framework, the proposed RGNG algorithm was less sensitive to initializations, input sequence orders and the presence of outliers. Moreover, the RGNG can automatically determine the optimal number of clusters by seeking the extreme value of the Minimum Description Length (MDL) measure [79] calculated at the completion of each growing stage. Furthermore, topology relationships among prototypes can also be established. Experimental results on some artificial and UCI datasets demonstrated the superior performance of the proposed RGNG algorithm over the conventional GNG incorporating MDL method, named the GNG-M.

1.5 Organization of the thesis

The remainder of this thesis is organized as follows.

In Chapter 2, the existing BODLDR and SMDLDR techniques are extensively overviewed. The FLDA and CLDA techniques are revisited. The deficiencies associated with the conventional Fisher and Chernoff criteria are analyzed in-depth.

In Chapter 3, two categories of enhanced parametric BODLDR techniques are presented, which include one direct BODLDR technique and a class of 8 CWFLDA techniques developed under the Gaussian homoscedastic assumption as well as one

direct BODLDR technique and a class of 8 CWCLDA techniques developed under the Gaussian heteroscedastic assumption. The experiments are conducted on several multi-class artificial and UCI datasets to evaluate the performance of the proposed techniques in comparison with their archetypes.

In Chapter 4, the existing dimensionality reduction techniques capable of handling high-dimensional and undersampled problems are thoroughly overviewed. To generalize the utility of some representative CWFLDA techniques proposed in Chapter 3 in the high-dimensional and undersampled scenario, two classes of subspaces based CWFLDA techniques, i.e. Complementary Subspaces-based CWFLDA (CS-CWFLDA) and Regularized Subspace-based CWFLDA (RS-CWFLDA), are proposed. The experiments are conducted on various types of high-dimensional data sources including 1 handwritten digit image dataset, 3 facial image datasets, 5 microarray gene expression profile datasets and 1 text document datasets. The performances of the proposed techniques are compared with that of several state-of-the-art FLDA variants.

In Chapter 4, many robust variants of the conventional prototype-based clustering algorithms and several prototype-based clustering algorithms that can automatically determine the optimal number of clusters are reviewed. After revisiting the GNG algorithm, a novel RGNG is presented. Experiments are conducted on several artificial and UCI datasets to evaluate the robustness of the proposed RGNG algorithm in comparison with the GNG-M algorithm.

Chapter 5 concludes the thesis with some recommended further research directions.

Chapter 2: An Overview of Discriminative Linear Dimensionality Reduction Techniques

2.1 Introduction

The general goal of Discriminative Linear Dimensionality Reduction (DLDR) is to linearly transform the original feature space into a lower-dimensional subspace with the minimal loss of the discriminatory information. Although suffering from some deficiencies in comparison with its nonlinear counterpart, the DLDR remains popular till now mainly because of its mathematical tractability, computational simplicity and reliability to explore the underlying data structures in the original feature space. Furthermore, nonlinear extension can be readily implemented via the kernel trick [12] or function replacement concept [69]. Therefore, our current research work only concerns the DLDR techniques. According to information theory [80], no downsized transformations can create new valuable information (discriminatory information in the context of the supervised pattern analysis) and only information loss will happen. As a result, minimizing the loss of the discriminatory information after transformation is equivalent to maximizing the amount of the discriminatory information in the transformed subspace, which can be evaluated by some class separability measures. In general, nonlinearly separable class boundaries can represent good class separation when applying nonlinear classifiers in the post-end although they indicate bad class separability with respect to linear classifiers. Moreover, different types of classifiers may emphasize different aspects of the class distributions, e.g. classifiers based on the statistical decision theory will depend on the whole class distributions while classifiers directly constructing geometrical decision boundaries may only rely on the partial class distributions around the decision boundaries. Therefore, it is desirable to evaluate the class separability in terms of both the class distributions and the classifiers to be used. Given certain classifiers, the dimensionality reduction step can be directly embedded into the classifier formulation with the transformation matrix serving as an additional parameter, which can be jointly optimized with the classifier parameters to

pursue the minimum classification error. However, many existing classifiers do not provide analytical formulations for evaluating the classification performance and also there exists no practical guidelines on how to choose the most suitable classifier to solve a given problem.

Considering the effectiveness and efficiency of a DLDR technique, many class separability measures have been derived to serve as the criterion functions to be optimized, which take the linear transformation matrix as an argument. When the higher priority is given to the effectiveness, the Bayes error will be the ideal choice since it represents the minimum probability of misclassification, which cannot be further reduced by any eminent classifiers. However, the Bayes error under arbitrary class distributions and/or in the multi-class scenario usually cannot be analytically formulated since it is prohibitive to calculate the integration of complex class distribution functions over multiple discontinuous integral regions. Moreover, even if simple class distributions are known in the original feature space, there is no guarantee that they can be derived in the transformed subspace, and accordingly the Bayes error in the transformed subspace still may not be computed. In the past few decades, many efforts have been dedicated to the design of analytical criterion functions that can to some degree relate to the Bayes error in the transformed subspace under certain assumptions. The resulting class of DLDR techniques is hereby named Bayes Optimality Motivated DLDR (BODLDR), which commonly resorts to the numerical optimization methods to iteratively derive the suboptimal solutions². According to whether the parametric assumptions on class distributions are required or not when designing the criterion function, the BODLDR techniques can be dichotomized into parametric and non-parametric versions. The major disadvantages of the BODLDR techniques are the high computational complexity and the risk of getting stuck into local optima during the optimization. On the other hand, when the higher priority is given to the efficiency, the geometric class structures of the training samples can be used to construct the class separability measures. It may significantly reduce the computational cost coming from the class distribution estimation and also avoid the time-consuming numerical optimization at the expense of losing the direct attempt to

² In this sense, the BODLDR can be categorized into the supervised projection pursuit techniques.

minimizing the Bayes error. Scatter matrices based criterion functions constitute most representative examples of such type of class separability measures, which can usually result in globally optimal solutions in the closed form. The resulting class of DLDR methods is hereby named Scatter Matrix-based DLDR (SMDLDR). Although highly efficient, the SMDLDR techniques generally lack the link with the Bayes error, and accordingly their effectiveness may be worse than that of the BODLDR techniques. Fisher Linear Discriminant Analysis (FLDA) can be regarded as an intersection of the BODLDR and SMDLDR techniques, which uses the between-class scatter matrix and average-within class scatter matrix to formulate the class separability measure, called the Fisher criterion. The most attractive merit of the FLDA is that its globally optimal solution in the closed form can result in the Bayes optimality under certain conditions. However, some potential deficiencies inherent in the conventional Fisher criterion prevent its link with the Bayes error in general sense.

In this chapter, we give a thorough overview of the existing BODLDR and SMDLDR techniques. As the footstone of the methods to be proposed in the next chapter, the FLDA technique as well as its heteroscedastic extension, the CLDA technique, is reviewed. The major deficiencies inherent in the conventional Fisher and Chernoff criteria, which destroy their links with the Bayes error in general cases, are explored in-depth.

The remainder of this chapter is organized as follows. In sections 2.2 and 2.3, existing BODLDR and SMDLDR techniques are extensively overviewed, respectively. Section 2.4 reviews the FLDA and CLDA techniques with their inherent deficiencies being discussed in-depth.

2.2 Overview of BODLDR and SMDLDR techniques

2.2.1 The BODLDR techniques

The ultimate and ideal goal of the BODLDR techniques is to seek for an optimal linear transformation that can result in a transformed subspace with the minimum Bayes

error. Since the Bayes error can characterize the intrinsic source of error with respect to a specific feature space, the transformed subspace with the minimum Bayes error, i.e. achieving the Bayes optimality, should be the most discriminative one and may improve the performance of any classifier established therein. Although the Bayes error is a theoretically optimal criterion function, its computational intractability prevents its practical use in the BODLDR techniques. To address this issue, many analytical criterion functions related to the Bayes error in certain sense have been derived as the surrogates. Accordingly, the BODLDR techniques using these suboptimal criterion functions can result in solutions approximating the Bayes error to some extent under certain assumptions. Before thorough description of the existing BODLDR techniques, we recapitulate the Bayes decision theory [1], [2], which is the fundamental of the BODLDR techniques.

The decision-making task in statistical pattern recognition is to assign a pattern represented by the n -dimensional feature vector $\mathbf{x} \in \mathbf{R}^n$ to one of the C distinct classes $\omega_i, i = 1, \dots, C$. The conditional risk by taking action of assigning $\mathbf{x}$ to class ω_i , i.e. $\alpha_i(\mathbf{x})$, can be defined by:

$$R(\alpha_i(\mathbf{x}) | \mathbf{x}) = \sum_{j=1}^C \lambda(\alpha_i(\mathbf{x}) | \omega_j) P(\omega_j | \mathbf{x}) \quad (2.1)$$

where $P(\omega_j | \mathbf{x})$ is the posterior probability of class ω_j given $\mathbf{x}$, which satisfies

$\sum_{j=1}^C P(\omega_j | \mathbf{x}) = 1$, and $\lambda(\alpha_i(\mathbf{x}) | \omega_j)$ is the loss incurred by assigning $\mathbf{x}$ to the class ω_i

when its actual class is ω_j . Commonly, the loss $\lambda(\alpha_i(\mathbf{x}) | \omega_j)$ of the correct decision is zero and different types of wrong decisions take different loss. The Bayes decision theory attempts to find an optimal assignment of $\mathbf{x}$ to minimize the conditional risk. By using the “0/1” loss function that assumes uniform loss “1” for all incorrect

decisions, i.e. $\lambda(\alpha_i(\mathbf{x}) | \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$, the conditional risk in Eq. (2.1) comes down to

the conditional probability of error expressed by:

$$P_e(\alpha_i(\mathbf{x}) | \mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^C P(\omega_j | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x}) \quad (2.2)$$

According to the Bayes decision theory, to achieve the minimum conditional probability of error, the feature vector $\mathbf{x}$ should be assigned to class ω_i with $i = \arg \max_{1 \leq j \leq C} P(\omega_j | \mathbf{x})$, i.e. the class with the largest posterior probability, which is called the Bayes decision rule. The minimum expected probability of error over the whole feature space achieved by applying the Bayes decision rule is called the Bayes error, which is calculated by:

$$P_e = \int \min_{1 \leq i \leq C} (1 - P(\omega_i | \mathbf{x})) p(\mathbf{x}) d\mathbf{x} = 1 - \int \max_{1 \leq i \leq C} P(\omega_i | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (2.3)$$

where $p(\mathbf{x})$ is an evidence factor denoting the unconditional probability density of the feature vector $\mathbf{x}$.

According to the Bayes formula expressed by:

$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i) P(\omega_i)}{p(\mathbf{x})} \text{ with } p(\mathbf{x}) = \sum_{i=1}^C p(\mathbf{x} | \omega_i) P(\omega_i) \quad (2.4)$$

the posterior probability of class ω_i given $\mathbf{x}$, i.e. $P(\omega_i | \mathbf{x})$, can be derived from the prior class probability $P(\omega_i)$ and the class conditional probability density of $\mathbf{x}$ given class ω_i , i.e. $p(\mathbf{x} | \omega_i)$, which is also called the likelihood of $\mathbf{x}$ that evaluates how likely that $\mathbf{x}$ comes from class ω_i if all other conditions remain same. Accordingly, the Bayes error expression in Eq. (2.4) can be reformulated to:

$$P_e = 1 - \int \max_{1 \leq i \leq C} p(\mathbf{x} | \omega_i) P(\omega_i) d\mathbf{x} \quad (2.5)$$

Therefore, when class distributions including the prior class probabilities, the class conditional probability densities and the posterior class probabilities are completely known, the Bayes error can be calculated by Eqs. (2.4) and (2.5).

In general, the Bayes classifier can be defined in terms of a set of discriminant functions, e.g.

$$\begin{aligned} g_i(\mathbf{x}) &= P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})}, i = 1, \dots, C \\ \text{or } g_i(\mathbf{x}) &= p(\mathbf{x} | \omega_i)P(\omega_i), i = 1, \dots, C \\ \text{or } g_i(\mathbf{x}) &= \ln p(\mathbf{x} | \omega_i) + \ln P(\omega_i), i = 1, \dots, C \end{aligned} \quad (2.6)$$

It assigns the feature vector $\mathbf{x}$ to the class corresponding to the maximum discriminant function value. For example, $\mathbf{x}$ will be assigned to class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$, $j = 1, \dots, C$, $j \neq i$. Theoretically speaking, for a C -class problem, those C discriminant functions have carried the enough discriminatory information to achieve the Bayes error. Furthermore, only $C-1$ discriminant functions are linearly independent due to $\sum_{i=1}^C g_i(\mathbf{x}) = C$ with C being a constant with respect to the function index i . Therefore, any $C-1$ discriminant functions can constitute an ideal set of features with the smallest size, which can result in the Bayes error. Note that the ideal feature set may be not unique since different implementations of the discriminant functions may lead to the same Bayes error. However, in practice, since the underlying class distributions are only partially or completely unknown, they need to be estimated parametrically or non-parametrically based on the training samples. Such estimation often has high bias and variance due to the limited training sample size, the influence from noise and the possibly wrongly specified parametric assumptions. Consequently, the classification error may much exceed the Bayes error. This means that the ideal feature set with the most condensed size may seldom be attained in practice, and accordingly more than $C-1$ features are needed in order to pursue the Bayes error.

According to different empirical ways to estimate the class distributions, the existing BODLDR techniques can be dichotomized into parametric and non-parametric versions. The parametric BODLDR techniques assume certain parametric models of class distributions in the transformed space and then estimate the associated model parameters via the maximum likelihood or Bayesian estimation methods. Since the number of model parameters is not proportional to the training sample size, this class of techniques bears the reasonable computational cost. The non-parametric BODLDR techniques use some non-parametric density estimation methods, such as the Parzen window and the k -nearest neighbor density estimators [1], [2], to estimate the class distributions without any parametric assumptions. This class of techniques merely relies on the training samples themselves, and consequently their computational costs may grow dramatically with the increase of the training sample size. A common characteristic in both parametric and non-parametric BODLDR techniques is the requirement of some numerical optimization methods to derive the optimal solutions. In what follows, many existing representative parametric and non-parametric BODLDR techniques are revisited.

2.2.1.1 Parametric BODLDR techniques

Gaussian distribution may probably be the most commonly used parametric assumption of class distributions in multivariate statistics due to its particularly nice theoretical properties and frequent occurrence in real-world problems according to the central limit theory [81]. Under specific Gaussian assumptions, the Bayes error or its upper bound can be analytically formulated. Accordingly, a class of parametric BODLDR techniques with the criterion functions directly related to the estimate of the Bayes error or its upper bounds under Gaussian assumptions has been developed.

Under the Gaussian homoscedastic assumption of class distributions, all C classes $\omega_i, i = 1, \dots, C$ are normally distributed in the n -dimensional feature space with class mean vectors $\mathbf{m}_i, i = 1, \dots, C$, prior probabilities $P_i, i = 1, \dots, C$ and the common

covariance matrix $\mathbf{S}$. The Bayes classifier, commonly referred to as Linear Discriminant Classifier (LDC) in such case, is defined by:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) - \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{S}| + \ln P_i, i = 1, \dots, C \quad (2.7)$$

By removing common terms with regard to all classes, Eq. (2.7) can be simplified to:

$$g_i(\mathbf{x}) = \mathbf{m}_i^T \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}_i^T \mathbf{S}^{-1} \mathbf{m}_i + \ln P_i, i = 1, \dots, C \quad (2.8)$$

In the two-class case, the LDC can be formulated to:

$$h(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x}) = (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} (\mathbf{m}_i^T \mathbf{S}^{-1} \mathbf{m}_i - \mathbf{m}_j^T \mathbf{S}^{-1} \mathbf{m}_j) + \ln \frac{P_i}{P_j} = \mathbf{A} \mathbf{x} + \mathbf{B}$$

with $\mathbf{A} = (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}^{-1}$, $\mathbf{B} = -\frac{1}{2} (\mathbf{m}_i^T \mathbf{S}^{-1} \mathbf{m}_i - \mathbf{m}_j^T \mathbf{S}^{-1} \mathbf{m}_j) + \ln \frac{P_i}{P_j}$ (2.9)

When $h(\mathbf{x}) \geq 0$, the feature vector $\mathbf{x}$ is assigned to class ω_i , otherwise it is assigned to class ω_j . The decision boundary between two classes ω_i and ω_j is a linear hyperplane with the normal vector direction being $(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}^{-1}$. It has been demonstrated [2], [82] that the Bayes error between two classes ω_i and ω_j under the homoscedastic Gaussian assumption can be explicitly expressed by:

$$P_e(i, j) = \frac{P_j}{\sqrt{2\pi}} \int_{-\infty}^{k_1} e^{-\frac{t^2}{2}} dt + \frac{P_i}{\sqrt{2\pi}} \int_{k_2}^{+\infty} e^{-\frac{t^2}{2}} dt \quad (2.10)$$

$$= (P_i + P_j) P_e^{ij}$$

with $k_1 = \left(\ln \frac{P_i}{P_j} - \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}$, $k_2 = \left(\ln \frac{P_i}{P_j} + \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}$

$$\partial_M^{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}^{-1} (\mathbf{m}_i - \mathbf{m}_j)}$$

$$P_e^{ij} = \frac{P_j}{(P_i + P_j)\sqrt{2\pi}} \int_{-\infty}^{k_1} e^{-\frac{t^2}{2}} dt + \frac{P_i}{(P_i + P_j)\sqrt{2\pi}} \int_{k_2}^{+\infty} e^{-\frac{t^2}{2}} dt$$

Here, P_e^{ij} represents the two-class Bayes error between classes ω_i and ω_j with the relative prior probabilities of classes ω_i and ω_j denoted by $P_i/(P_i + P_j)$ and $P_j/(P_i + P_j)$, respectively. It is easy to verify that $P_e(i, j) \in [0, (P_i + P_j)/2]$ and that the Bayes accuracy between two classes ω_i and ω_j equals $(P_i + P_j) - P_e(i, j)$. Obviously, the Mahalanobis distance ∂_M^{ij} and class prior probabilities P_i and P_j constitute sufficient statistics for the Bayes error between two classes ω_i and ω_j . For the two-class Bayes error, it is easy to verify that $P_e^{ij} \in [0, 0.5]$.

Under the Gaussian heteroscedastic assumption of class distributions, all C classes $\omega_i, i = 1, \dots, C$ are normally distributed in the n -dimensional feature space with class mean vectors $\mathbf{m}_i, i = 1, \dots, C$, prior probabilities $P_i, i = 1, \dots, C$ and different covariance matrices $\mathbf{S}_i$. The Bayes classifier, commonly referred to as the Quadratic Discriminant Classifier (QDC) in such case, is defined by:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x} - \mathbf{m}_i) - \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{S}_i| + \ln P_i, i = 1, \dots, C \quad (2.11)$$

By removing common terms with regard to all classes, Eq. (2.11) can be simplified to:

$$g_i(\mathbf{x}) = -\frac{1}{2} \mathbf{x}^T \mathbf{S}_i^{-1} \mathbf{x} + \mathbf{m}_i^T \mathbf{S}_i^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}_i^T \mathbf{S}_i^{-1} \mathbf{m}_i + \ln P_i - \frac{1}{2} \ln |\mathbf{S}_i|, i = 1, \dots, C \quad (2.12)$$

In the two-class case, the QDC can be further simplified to:

$$h(\mathbf{x}) = -\frac{1}{2} \mathbf{x}^T (\mathbf{S}_i^{-1} - \mathbf{S}_j^{-1}) \mathbf{x} + (\mathbf{m}_i^T \mathbf{S}_i^{-1} - \mathbf{m}_j^T \mathbf{S}_j^{-1}) \mathbf{x} - \frac{1}{2} (\mathbf{m}_i^T \mathbf{S}_i^{-1} \mathbf{m}_i - \mathbf{m}_j^T \mathbf{S}_j^{-1} \mathbf{m}_j) + \ln \frac{P_i}{P_j} - \frac{1}{2} \ln \frac{|\mathbf{S}_i|}{|\mathbf{S}_j|} \quad (2.13)$$

When $h(\mathbf{x}) \geq 0$, the feature vector $\mathbf{x}$ is assigned to class ω_i , otherwise it is assigned to class ω_j . The decision boundary between two classes ω_i and ω_j is a curved hypersurface. Under the Gaussian heteroscedastic assumption, the analytical expression of the two-class Bayes error no longer exists. Fortunately, some of its upper bounds [1], [2], [82]-[85] have been mathematically derived, which provide an effective way to approximate the Bayes error. According to Eq. (2.3), the Bayes error between two classes ω_i and ω_j can be expressed by:

$$\begin{aligned} P_e(i, j) &= \int \min(P(\omega_i | \mathbf{x}), P(\omega_j | \mathbf{x})) P(\mathbf{x}) d\mathbf{x} \\ &= \int \min(P(\mathbf{x} | \omega_i) P(\omega_i), P(\mathbf{x} | \omega_j) P(\omega_j)) d\mathbf{x} \\ &\leq \int P(\mathbf{x} | \omega_i)^\alpha P_i^\alpha P(\mathbf{x} | \omega_j)^{1-\alpha} P_j^{1-\alpha} d\mathbf{x} = P_i^\alpha P_j^{1-\alpha} \int P(\mathbf{x} | \omega_i)^\alpha P(\mathbf{x} | \omega_j)^{1-\alpha} d\mathbf{x} \quad \text{for } 0 \leq \alpha \leq 1 \end{aligned} \quad (2.14)$$

Note that the inequality in the third line is obtained according to $\min(a, b) \leq a^\alpha b^{1-\alpha}$, $a, b \geq 0, 0 \leq \alpha \leq 1$. Accordingly, an upper bound of the two-class Bayes error under the heteroscedastic Gaussian assumption, called Chernoff bound, can be formulated [1], [2], [85] as follows.

$$P_e(i, j) \leq U_c(i, j) = P_i^\alpha P_j^{1-\alpha} \int P(\mathbf{x} | \omega_i)^\alpha P(\mathbf{x} | \omega_j)^{1-\alpha} d\mathbf{x} = P_i^\alpha P_j^{1-\alpha} e^{-\partial_c^{ij}(\alpha)}, 0 \leq \alpha \leq 1 \quad (2.15)$$

$$\text{with } \partial_c^{ij}(\alpha) = \frac{\alpha(1-\alpha)}{2} (\mathbf{m}_i - \mathbf{m}_j)^T [\alpha \mathbf{S}_i + (1-\alpha) \mathbf{S}_j]^{-1} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2} \ln \left(\frac{|\alpha \mathbf{S}_i + (1-\alpha) \mathbf{S}_j|}{|\mathbf{S}_i|^\alpha |\mathbf{S}_j|^{1-\alpha}} \right)$$

where $\partial_c^{ij}(\alpha)$ and $U_c(i, j) \in [0, P_i + P_j]$ denote the Chernoff distance and the corresponding Chernoff upper bound of the Bayes error between two classes ω_i and ω_j , respectively. The two-class Chernoff bound can be easily derived as

$U_C^{ij} = \left(P_i / (P_i + P_j) \right)^\alpha \left(P_j / (P_i + P_j) \right)^{1-\alpha} e^{-\partial_C^{ij}(\alpha)}$. The Chernoff distance $\partial_C^{ij}(\alpha)$ is a function of the variable α , $0 \leq \alpha \leq 1$. When α is set to 0.5, the Chernoff distance and bound come down to the well-known Bhattacharyya distance and bound [1], [2]. Besides Chernoff and Bhattacharyya distances, many other probabilistic distances [84] have been devised to measure the affinity of two class distributions, which are also related the Bayes error via certain upper bounds. Among which, the Patrick-Fisher (PF) distance and its generalized version by Toussaint [86], [87] deserve attention because of their simple analytical expressions. The PF distance between two classes ω_i and ω_j can be defined by:

$$PF_{ij} = \left(\int_{\mathbf{x}} (p(\omega_i)p(\mathbf{x}|\omega_i) - p(\omega_j)p(\mathbf{x}|\omega_j))^2 d\mathbf{x} \right)^{1/2} = \left(\int_{\mathbf{x}} p(\mathbf{x}) (p(\omega_i|\mathbf{x}) - p(\omega_j|\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} \right)^{1/2} \quad (2.16)$$

which is related to the Bayes error via an upper bound as follows.

$$P_e(i, j) \leq \frac{1}{2} \left(1 - \left(\frac{PF_{ij}^2}{\|p(\mathbf{x})\|_r} \right)^s \right) \text{ and } \|p(\mathbf{x})\|_r = \left(\int_{\mathbf{x}} p^r(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \right)^{1/r} < \infty \quad (2.17)$$

with $1 < r \leq \infty$ and $1/r + 1/s = 1$

Lissack and Fu [83] proposed a group of class separability measures based on the L^α -distance between posterior class density functions, which is defined as follows:

$$\text{Two-class cases: } J_\alpha(\omega_1, \omega_2) = \int_{\mathbf{x}} |p(\omega_1|\mathbf{x}) - p(\omega_2|\mathbf{x})|^\alpha p(\mathbf{x}) d\mathbf{x} \quad (2.18)$$

with $\alpha \in [0, +\infty]$

$$\text{Multi-class cases: } J_\alpha(\omega_1, \dots, \omega_C) = \sum_{i=1}^{C-1} \sum_{j=i+1}^C \int_{\mathbf{x}} |p(\omega_i|\mathbf{x}) - p(\omega_j|\mathbf{x})|^\alpha p(\mathbf{x}) d\mathbf{x}, \quad C \geq 3 \quad (2.19)$$

with $\alpha \in [0, +\infty]$

Some appealing properties of such class of measures are: in two-class cases, $J_1(\omega_1, \omega_2)$ is simply related to the Bayes error by $P_e = 0.5(1 - J_1(\omega_1, \omega_2))$ while an increase (or decrease) in α from 1 loosens the upper bound of the Bayes error. When $1 \leq \alpha \leq 2$, the upper bounds of the two-class Bayes error derived from such class of measures are better than or as good as many commonly used bounds [83]. From Eqs. (2.16) and (2.18), it is also interesting to observe that $J_1(\omega_1, \omega_2)$ is very similar to PF_{ij}^2 expect for the factor $p(\mathbf{x})$ under the integral of PF_{ij}^2 , which account for PF_{ij}^2 taking values within $[0, +\infty)$.

Röhl and Weihs [88] presented a computationally intensive optimal linear dimension reduction method for multi-class problems, in which the Bayes error calculated by the exact integration under the Gaussian assumption was used as the criterion function and the simulated annealing [89] was employed to seek for the optimal linear transformation matrix. Demuyneck *et al.* [90] used the minimum classification error as the criterion function, which is defined as the Bayes error estimated on the transformed training samples instead of over the whole transformed subspace under the mixture of Gaussian assumption. Although this method can reduce the computational complexity, it may take the high risk of overfitting the limited number of transformed training samples.

In [83], it was proved that the Bayes error in the multi-class case can be bounded from above by the sum of prior probability weighted pairwise Bayes errors over all available class pairs. Therefore, it is reasonable to use the sum of prior probability weighted pairwise Bayes error or its upper bound formulated under Gaussian assumptions to approximate the multi-class Bayes error. Obviously, the Bayes error or its upper bound between two classes is the building block of such approximation. By numerically optimizing this upper error bound defined in the transformed subspace with respect to the linear transformation matrix, the optimal solution³ can be achieved. Based on this idea, some approaches have been proposed under specific Gaussian assumptions. Lotlikar *et al.* [91] developed an Adaptive Linear Dimensionality

³ Possibly suboptimal solution due to the existence of many local optima on the landscape of the criterion function

Reduction (ALDR) algorithm, in which the criterion function is formulated as the summation of the pairwise Bayes error in the transformed subspace under the Gaussian homoscedastic assumption with the equal class prior probability. An optimal orthonormal linear transformation matrix is obtained by optimizing this criterion function via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme [92]. Saon and Pahmanabhan [93] presented a DLDR approach, which uses the union of the pairwise Bhattacharyya bounds formulated in the transformed space under the Gaussian heteroscedastic assumption as the criterion function. The optimization process is initialized with the FLDA solution and performed by using the conjugate gradient descent scheme [94].

Another category of parametric BODLDR techniques does not explicitly estimate the Bayes error in the transformed subspace. Instead, they seek for an optimal linear transformation by maximizing the likelihood of the training samples in the original feature space, formulated in terms of the parametric model of class distributions assumed in the transformed subspace with the transformation matrix as one parameter. Assume that the transformed training samples are Gaussian distributed and all the class discriminatory information resides in a p -dimensional subspace of the original n -dimensional feature space. An optimal linear transformation matrix can be found by jointly estimating the parameters of the Gaussian class distributions and the linear transformation via maximizing the likelihood of training samples in the original feature space. Importantly, when the number of extracted features exceeds p , the obtained linear transformation can lead to the Bayes optimality due to the equivalence of the Bayes error in the transformed subspace and that in the original feature space. For example, in the Homoscedastic Gaussian Model (HOG) [82], all classes are normally distributed with the common covariance matrix and different class mean vectors. Assume that the class mean vectors only differ in the p -dimensional subspace of the original feature space, the FLDA solution to be discussed later, become an instance of the optimal linear transformations⁴ found by the maximum likelihood estimation. This proposition was initially made by Campbell [95] and further elucidated by Kumar and Andreou in [96], [97]. In the heteroscedastic Gaussian model

⁴ Since the maximum likelihood function is invariant to any further non-singular transformation on the top of the obtained transformation matrix, the optimal linear transformation may be not unique.

[96], [97], each class can follow different Gaussian distributions with the difference of class means and/or class covariance matrices only reside in a p -dimensional subspace of the original feature space. The theoretical framework for the maximum likelihood estimation of this heteroscedastic Gaussian model to seek for an optimal linear transformation was elucidated by Kumar and Andreou [96], [97]⁵ with the resulting Heteroscedastic Discriminant Analysis (HDA) technique hereby named HDA_KA. When the number of extracted features exceeds p , the obtained optimal linear transformation can lead to the Bayes optimality, which has been proved [98] in terms of the relation between the criterion function of HDA_KA and that of maximization of the mutual information between the linearly transformed feature vectors and their class labels. Although the Bayes optimality can be achieved when the number of extracted features exceeds $p-1$ provided that p is known in advance, this does not hold true when the number of extracted features is below p . Therefore, the integer p plays a crucial role in search of the optimal linear transformation, which, unfortunately, is hard to be pre-specified in practice. Using the trail and error scheme to seek for a suitable value of p may take effect only if the model assumptions are strictly satisfied, e.g. the $n-p$ dimensional complementary subspace does contain no discriminatory information. Otherwise, an accurate estimate of p cannot be found and accordingly the obtained optimal linear transformation may deviate much from the Bayes optimal one.

The symmetric divergence characterizing the affinity of two class distributions has long been used as the class separability measure, which, however, lacks an explicit link with the Bayes error. A class of parametric BODLDR techniques using the symmetric divergence as the criterion function has been developed. Tou and Heydorn [99] used the symmetric divergence as the DLDR criterion function for two-class problems, which was extended by Bahu and Kalra [100] to the multi-class scenario. Decell and Quirein [101] established a link between the Bayes error and the average interclass symmetric divergence, which indicates that under the Gaussian heteroscedastic assumption with the equal class prior probability, if the average interclass symmetric divergence in the transformed subspace equals that in the original

⁵ The pioneer study of HDA was conducted by Schukat-Talamazzini *et al.* [102], who developed a maximum likelihood rotation method that generalizes the FLDA to the HDA based on the maximum likelihood estimation.

feature space, the equality also holds true for the Bayes error. The idea behind the proof is that if the interclass divergences remain same in the transformed subspace as in the original feature spaces, the Bayes decision will also be same since the likelihood ratios are persevered almost everywhere. Saon and Padmanabhan [93] applied the symmetric divergence based DLDR method to solve the voice mail transcription task. Specifically, under the Gaussian heteroscedastic assumption with $p(\mathbf{x} | \omega_i) = N(\mathbf{x}; \mathbf{m}_i, \mathbf{S}_i), i = 1, \dots, C$, the average interclass symmetric divergence in the original feature space can be formulated by:

$$\begin{aligned}
 Div &= \frac{2}{C(C-1)} \sum_{1 \leq i < j \leq C} Div(i, j) \\
 Div(i, j) &= \int_{\mathbf{x}} \left(p(\mathbf{x} | \omega_i) \log \frac{p(\mathbf{x} | \omega_i)}{p(\mathbf{x} | \omega_j)} + p(\mathbf{x} | \omega_j) \log \frac{p(\mathbf{x} | \omega_j)}{p(\mathbf{x} | \omega_i)} \right) d\mathbf{x} \\
 &= \frac{1}{2} \text{trace}((\mathbf{S}_i^{-1} + \mathbf{S}_j^{-1})(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T) + \frac{1}{2} \text{trace}(\mathbf{S}_i^{-1}\mathbf{S}_j + \mathbf{S}_j^{-1}\mathbf{S}_i - 2\mathbf{I})
 \end{aligned} \tag{2.20}$$

Accordingly, when the original feature space is mapped into an m -dimensional subspace by a linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$), the average interclass symmetric divergence defined in the transformed subspace, i.e. $\hat{Div}$, can be expressed by:

$$\begin{aligned}
 \hat{Div} &= \frac{2}{C(C-1)} \sum_{1 \leq i < j \leq C} \hat{Div}(i, j) \\
 &= \frac{1}{2} \text{trace}((\hat{\mathbf{S}}_i^{-1} + \hat{\mathbf{S}}_j^{-1})(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T) + \frac{1}{2} \text{trace}(\hat{\mathbf{S}}_i^{-1}\hat{\mathbf{S}}_j + \hat{\mathbf{S}}_j^{-1}\hat{\mathbf{S}}_i - 2\mathbf{I})
 \end{aligned} \tag{2.21}$$

$$\text{with } \hat{\mathbf{m}}_i = \mathbf{P}^T \mathbf{m}_i, \hat{\mathbf{m}}_j = \mathbf{P}^T \mathbf{m}_j, \hat{\mathbf{S}}_i = \mathbf{P}^T \mathbf{S}_i \mathbf{P} \text{ and } \hat{\mathbf{S}}_j = \mathbf{P}^T \mathbf{S}_j \mathbf{P}$$

By numerically maximizing $\hat{Div}$ with respect to $\mathbf{P}$, an optimal linear transformation can be obtained. However, unless $\hat{Div} = Div$ is achieved, the relation between the obtained optimal linear transformation and the Bayes optimality is obscure.

Since posterior class density distributions captures all the discriminatory information, if it remains same after the transformation, the Bayes error in the transformed subspace will equal to that in the original feature space. Demuynck *et al.* [90] proposed a DLDR method, which uses the Kullback-Leibler divergence between the posterior class density distributions in the original feature space and that in the transformed subspace as the criterion function. It turns out that minimizing such Kullback-Leibler divergence is equivalent to minimization of the expectation of the posterior class density distributions over the whole transformed space, referred to as transformed equivocation [85], or minimization of the conditional mutual information between the original feature vectors and their class labels given the corresponding transformed feature vectors. The implementation of the transformed equivocation under the heteroscedastic Gaussian mixture model assumption has been presented in [85]. It is worth noting that the equivocation is related to the Bayes error via certain upper bound in two-class cases [85].

Mutual Information (MI) between features vectors and their class labels can be directly related to the Bayes error via the lower and upper bounds [84]. Use of the MI as the DLDR criterion function may avoid the parametric assumption of class distributions. However, the high computational complexity may prevent its practical applicability. Petridis and Perantonis [98] demonstrated that the Bayes optimality could be achieved with respect to the optimal linear transformation matrix obtained by maximizing the MI criterion under the Zero Information Loss (ZIL) model, which assumes the existence of a totally noise subspace of the original feature space. They also established a hierarchy relationship between the three models: HOG, HDA_KA and ZIL with the former one being a special case of the later one. Model independent reformulations of the FLDA, HDA_KA and MI criterion functions were developed to give a unified view of their similarities and discrepancies. The mutual information $I(\mathbf{X}, \mathbf{Y})$ between two random variables $\mathbf{X}$ and $\mathbf{Y}$, which can be either continuous-valued or discrete-valued, is symmetric and defined by the difference between the entropy $H(\mathbf{X})$ of one variable $\mathbf{X}$ and its conditional entropy $H(\mathbf{X} | \mathbf{Y})$ given another variable $\mathbf{Y}$ as follows.

$$I(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X} | \mathbf{Y}) = \int \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x}d\mathbf{y} \quad (2.22)$$

$$\text{with } H(\mathbf{X}) = -\int p(\mathbf{x}) \log(p(\mathbf{x})) d\mathbf{x} \text{ and}$$

$$H(\mathbf{X} | \mathbf{Y}) = -\int p(\mathbf{y}) \left(\int p(\mathbf{x} | \mathbf{y}) \log(p(\mathbf{x} | \mathbf{y})) d\mathbf{x} \right) d\mathbf{y}$$

In the context of the DLDR, the variables $\mathbf{X}$ and $\mathbf{Y}$ can denote the transformed feature vectors and their class labels, respectively. The practical calculation of the MI is difficult since it involves numerical integration of the functional of several probability density functions. A typical solution is to resort to some simple parametric assumptions, such as the single Gaussian or mixture of Gaussians.

2.2.1.2 Non-parametric BODLDR techniques

In many practical cases, suitable parametric assumptions of class distributions are hard to be pre-specified. Alternatively, we can resort to non-parametric methods, such as Parzen window density estimator and k -nearest neighbor density estimator [1], [2], to estimate class distributions in a non-parametric fashion. Although avoid subjectively specifying the parametric assumption of class distributions, the non-parametric BODLDR techniques suffer from a notorious disadvantage, that is, their computational complexity may grow dramatically with the increment of training sample size, which prevents their wider application on large-scale datasets. Some representative non-parametric BODLDR techniques are described as follows.

Buturović [103] developed a two-class DLDR technique, which use as the criterion function the estimate of the Bayes error in terms of the leave-one-out and re-substitution errors calculated by the k -nearest neighbor density estimator and an optimal threshold scheme. The simplex algorithm was employed to derive the optimal linear transformation matrix. Aladjem [104] described a non-parametric formulation of the PF distance via the Parzen window density estimator in the transformed subspace as the DLDR criterion function for two-class problems. A recursive

optimization procedure developed from the “structure removal” technique [20] was used to seek for the directions corresponding to large local optima of the criterion function. Lotlikar and Kothari [105] proposed an adaptive linear projection algorithm, which use the posterior class probabilities estimated by the Parzen window method to formulate the criterion function. Torkkola *et al.* [68] formulated a criterion function by the quadratic mutual information defined as the quadratic divergence [106] between the joint density of the transformed feature vectors and their class labels and the products of the corresponding marginal densities. Since the convolution of two Gaussian distributions is still Gaussian, the quadratic mutual information is formulated in a non-parametric fashion by using the Parzen window density estimator. Numerical optimization of the quadratic mutual information with respect to the linear transformation yields the optimal solution. The major limitation of this method is its high computational complexity on the large-scale datasets, and accordingly some schemes such as stochastic gradient approximation and Gaussian Mixture Model (GMM) density estimation have been devised to reduce its computational cost. Noting that the same quadratic divergence measure has been derived from the Renyi entropy formulation by Principle *et al.* [107], but with the different usage. Also, Hero *et al.* [108] mentioned a measure of the Renyi entropy difference in the context of image registration, which can converge to the Shannon MI in the extreme case and allow the same non-parametric treatment like the Torkkola’s work. The generalized log-likelihood ratio in the transformed subspace was proposed by Zhu *et al.* [109] as the DLDR criterion function in the two-class case, which includes the Fisher criterion as a special case under the Gaussian homoscedastic assumption. The optimization procedure is carried out by the local likelihood density estimation. Multiple features can be extracted via the orthogonalization or structure removal technique [20]. The relation between the generalized log-likelihood ratio and the Bayes error is obscure expect under the Gaussian homoscedastic assumption. An attractive merit of such criterion function is that it can capture very amount of the discriminatory information balancing the high-order and low-order difference among classes without over-emphasizing the high-order information when the low-order information is dominant. Moreover, by using the projection pursuit density estimation [110], the generalized log-likelihood ratio criterion can lead to the non-parametric discriminant analysis.

2.2.2 The SMDLDR techniques

Although direct optimization of the Bayes error related criterion functions in the BODLDR techniques might result in a subspace approximating the Bayes optimality, the computational costs of such class of techniques can be very expensive due to the density estimation, iterative numerical optimization and trial and error process. Intuitively, large distances between the transformed feature vectors belonging different classes can indicate the good class separability and accordingly small misclassification probability. Therefore, it is reasonable to construct some class separability measures in terms of the geometric structures of the transformed feature vectors in each class, e.g. the average distances between the transformed feature vectors belonging to the same and different classes. Such kind of measures may greatly reduce the computational complexity by exempting from the density estimation and numerical optimization at the expense of loss of the explicit link with the Bayes error. The most attractive merit of the SMDLDR techniques is their closed-form solutions obtained in an efficient manner. Before close study of the SMDLDR techniques, different types of scatter matrices are briefly reviewed, which constitute basic components of the SMDLDR.

The scatter matrix of a set of feature vectors is defined with respect to a chosen reference vector to reflect the spread of these feature vectors around such reference vector. It is common to use the determinant or trace of the scatter matrix to quantitatively evaluate the degree of the spread. In general, scatter matrices can be defined either on the global or on the local basis. The global scatter matrix uses the arithmetic mean of all available feature vectors as the reference vector and can characterize the global geometric structure of the available feature vectors if only they are sampled from the Gaussian-like distribution. However, if the available feature vectors contain many local geometric structures, e.g. multiple clusters, or have the irregular global geometric structure, e.g. sampled from some complex non-Gaussian distribution, the global scatter matrix may fail to capture some useful structure information. In contrast, the local scatter matrix is defined over local regions in the

feature space in order to capture the local structure information. In many existing literatures [2], the SMDLDR techniques based on global and local scatter matrices are referred to as the parametric and non-parametric discriminant analysis, respectively, since global scatter matrices based SMDLDR methods can theoretically perform well under certain parametric assumptions of class distributions, such as the Gaussian distribution, while the local scatter matrix based SMDLDR methods may work properly without any parametric assumptions. In this thesis, the SMDLDR techniques based on the global or local scatter matrices are used interchangeably with the parametric or non-parametric discriminant analysis.

Given a set of N n -dimensional feature vectors $\mathbf{x}_i \in \mathbf{R}^n, i = 1, \dots, N$ belonging to C classes $\omega_i, i = 1, \dots, C$, with N_i feature vectors $\mathbf{x}_j^{(i)}, i = 1, \dots, C, j = 1, \dots, N_i$ in each class ω_i and $\sum_{i=1}^C N_i = N$, a set of commonly used global and local scatter matrices⁶ are described as follows, respectively.

Global scatter matrices:

Global average within-class scatter matrix $\mathbf{S}_W^G$ characterizes the average of the spread of feature vectors belonging to each class around their respective class means over all classes and is estimated by:

$$\mathbf{S}_W^G = \sum_{i=1}^C P_i \mathbf{S}_i, \mathbf{S}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \mathbf{m}_i)(\mathbf{x}_j^{(i)} - \mathbf{m}_i)^T \text{ and } \mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)} \quad (2.23)$$

where $\mathbf{S}_i$, $\mathbf{m}_i$ and P_i represent the covariance matrix, the mean vector and the prior probability of class ω_i , respectively.

Global between-class scatter matrix $\mathbf{S}_B^G$ characterizes the spread of all class mean vectors around the overall mean vector and is estimated by:

⁶ The strict definition of the scatter matrix differs from the formulation used here in a factor N , i.e. the total number of feature vectors. In many literatures [2], they are used interchangeably without influencing any derived conclusions.

$$\mathbf{S}_B^G = \sum_{i=1}^C P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \text{ and } \mathbf{m} = \sum_{i=1}^C P_i \mathbf{m}_i \quad (2.24)$$

where $\mathbf{m}$ represents the overall mean vector.

Global total scatter matrix $\mathbf{S}_T^G$, i.e. covariance matrix of the available feature vectors, characterizes the spread of the available feature vectors around the overall mean vector and is estimated by:

$$\begin{aligned} \mathbf{S}_T^G &= \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \\ &= \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})(\mathbf{x}_j^{(i)} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})^T \\ &= \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \mathbf{m}_i)(\mathbf{x}_j^{(i)} - \mathbf{m}_i)^T + \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \quad (2.25) \\ &= \sum_{i=1}^C P_i \frac{1}{N_i} \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \mathbf{m}_i)(\mathbf{x}_j^{(i)} - \mathbf{m}_i)^T + \sum_{i=1}^C P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ &= \mathbf{S}_W^G + \mathbf{S}_B^G \end{aligned}$$

Note that if the underlying class distributions satisfy the Gaussian mixture model, clustering algorithms can be preceded to seek for the Gaussian components in each class. By assuming all the discovered Gaussian components in each class as the unique “pseudo” classes, the global scatter matrices based SMDLDR techniques may still work properly. However, under other complex distributions highly deviating from the Gaussian, only local scatter matrices may extract more useful information for class discrimination.

Local scatter matrices:

Local average within-class scatter matrix $\mathbf{S}_W^L$ denotes the average of local within-class scatter matrices over all classes, which is defined over each feature vector in every class and its associated intra-class k -nearest neighbor mean as follows.

$$\mathbf{S}_W^L = \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} \left[\left(\mathbf{x}_j^{(i)} - \mathbf{m}_k^I(\mathbf{x}_j^{(i)}) \right) \cdot \left(\mathbf{x}_j^{(i)} - \mathbf{m}_k^I(\mathbf{x}_j^{(i)}) \right)^T \right] \quad (2.26)$$

where $\mathbf{m}_k^I(\mathbf{x}_j^{(i)})$ represents the local intra-class k -nearest neighbor mean with respect to the j^{th} feature vector in the class ω_i , i.e. $\mathbf{x}_j^{(i)}$, and is defined as the mean vector of k nearest neighbors of $\mathbf{x}_j^{(i)}$ from the same class ω_i . Obviously, such local scatter matrix captures the spread of each feature vector around its nearest intra-class neighbors within the same class and thus characterizes the intra-class compactness on the local basis.

Local between-class scatter matrix $\mathbf{S}_B^L$ denotes the average of local between-class scatter matrices over all classes, which is defined over each feature vector in every class and its associated extra-class k -nearest neighbor mean as follows:

$$\mathbf{S}_B^L = \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} \left[\left(\mathbf{x}_j^{(i)} - \mathbf{m}_k^E(\mathbf{x}_j^{(i)}) \right) \cdot \left(\mathbf{x}_j^{(i)} - \mathbf{m}_k^E(\mathbf{x}_j^{(i)}) \right)^T \right] \quad (2.27)$$

where $\mathbf{m}_k^E(\mathbf{x}_j^{(i)})$ represents the local extra-class k -nearest neighbor mean with respect to the j^{th} feature vector in class ω_i , i.e. $\mathbf{x}_j^{(i)}$, which is defined as the mean vector of k -nearest neighbors of $\mathbf{x}_j^{(i)}$ from classes other than ω_i . Obviously, such local scatter matrix captures the spread of each feature vector around its nearest extra-class neighbors and thus characterizes the class separation on the local basis. The above definition of $\mathbf{S}_B^L$ has an explicit deficiency that the feature vectors faraway from the decision boundary may significantly dominate the scatter matrix $\mathbf{S}_B^L$ since the distances of these feature vectors from their extra-class k -nearest neighbor mean are much larger than those of the feature vectors near the decision boundary. Consequently, the scatter information around the decision boundary will be masked, which, however plays an essential role in class discrimination. Therefore, some weighting schemes have been devised in [2] to de-emphasize the contributions of feature vectors located remotely from the decision boundary to the computation of $\mathbf{S}_B^L$.

Local total scatter matrix $\mathbf{S}_T^L$ is class-label independent, which can be defined in terms of all available feature vectors and their associated k -nearest neighbor mean as follows:

$$\mathbf{S}_T^L = \frac{1}{N} \sum_{i=1}^N [(\mathbf{x}_i - \mathbf{m}_k(\mathbf{x}_i)) \cdot (\mathbf{x}_i - \mathbf{m}_k(\mathbf{x}_i))^T] \quad (2.28)$$

where $\mathbf{m}_k(\mathbf{x}_i)$ represents the local k -nearest neighbor mean with respect to the i^{th} feature vector $\mathbf{x}_i$. This local scatter matrix characterizes the local compactness of the available feature vectors regardless of their classes.

Note that the above three local scatter matrices can generally have the full rank and the equality $\mathbf{S}_T^L = \mathbf{S}_W^L + \mathbf{S}_B^L$ no longer holds true.

Decision boundary feature matrix $\mathbf{S}_{DBFM}$ is defined by:

$$\mathbf{S}_{DBFM} = \frac{1}{\int_S p(\mathbf{x}) d\mathbf{x}} \int_S \mathbf{N}(\mathbf{x}) \mathbf{N}(\mathbf{x})^T p(\mathbf{x}) d\mathbf{x} \quad (2.29)$$

where S , $\mathbf{N}(\mathbf{x})$ and $p(\mathbf{x})$ denote the effective decision boundary, the unit normal vector to the decision boundary at point $\mathbf{x}$, and the density function of $\mathbf{x}$, respectively. The $\mathbf{S}_{DBFM}$ captures the between-class scatter in the neighborhood of the decision boundary and its rank equals the intrinsic discriminant dimension of the given classification problem [111]. Although the use of $\mathbf{S}_{DBFM}$ for feature extraction can provide the theoretically optimal solution as well as the intrinsic discriminant dimension, its success strongly depends on the accurate estimation of the effective decision boundary, which is a particularly difficult task. Theoretically speaking, if the Bayes decision boundary is obtained, the resulting linear transformation can lead to the Bayes optimality in the transformed subspace provided that the number of extracted features is more than the intrinsic discriminant dimension. In practice, the

suboptimal decision boundary can be estimated in terms of certain classifiers, e.g. SVM [112].

Some applications of the local scatter matrices based SMDLDR techniques can be found in [113]-[115]. In this thesis, we will mainly focus on the global scatter matrices based SMDLDR techniques since the SMDLDR based on local scatter matrices can be readily obtained by substituting the global scatter matrices with the corresponding local versions in the global scatter matrices based SMDLDR. More importantly, the Bayes optimality analysis of the local scatter matrices based SMDLDR techniques can be fairly hardly. On the contrary, the Bayes optimality analysis of the global scatter matrices based SMDLDR techniques can be well conducted under the Gaussian assumption of class distributions. For convenient notation, we hereafter omit the superscript “ G ” when expressing the global scatter matrices unless the confusion occurs.

Intuitively, the good class separability means that classes cover compact regions in the feature space with little overlapping with each other, and accordingly can be reflected by the large between-class scatter and the small average within-class scatter. It is reasonable to maximize the class separability via maximizing the volume covered by individual class mean vectors while simultaneously minimizing the average volume covered by each class. In general, determinant or trace of a scatter matrix denoting the product or summation of variances along all principle directions of such matrix can characterize the volume covered by its component feature vectors under the Gaussian assumption. Assume that the original n -dimensional feature space is mapped into an m -dimensional subspace by the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$). Some class separability measures in the transformed subspace can be formulated in terms of the determinant or trace of the scatter matrices as follows, with (a)-(d) being suggested in [2].

$$(a) \quad J_1(\mathbf{P}) = \text{trace}\left((\mathbf{P}^T \mathbf{S}_2 \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_1 \mathbf{P}\right) \text{ where } \text{trace}(\cdot) \text{ denotes the trace of a matrix.}$$

$$(b) \quad J_2(\mathbf{P}) = \ln\left|(\mathbf{P}^T \mathbf{S}_2 \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_1 \mathbf{P}\right| = \ln|\mathbf{P}^T \mathbf{S}_1 \mathbf{P}| - \ln|\mathbf{P}^T \mathbf{S}_2 \mathbf{P}|$$

- (c) $J_3(\mathbf{P}) = \text{trace}(\mathbf{P}^T \mathbf{S}_1 \mathbf{P}) - \alpha \cdot (\text{trace}(\mathbf{P}^T \mathbf{S}_2 \mathbf{P}) - c)$ where α and c are the langrage multiplier and positive constant, respectively.
- (d) $J_4(\mathbf{P}) = \frac{\text{trace}(\mathbf{P}^T \mathbf{S}_1 \mathbf{P})}{\text{trace}(\mathbf{P}^T \mathbf{S}_2 \mathbf{P})}$
- (e) $J_5(\mathbf{P}) = \text{trace}(\mathbf{P}^T \mathbf{S}_1 \mathbf{P}) - \mu \cdot \text{trace}(\mathbf{P}^T \mathbf{S}_2 \mathbf{P})$ where μ is a balancing weight

Here, $\mathbf{S}_1$ and $\mathbf{S}_2$ can be chosen among the three scatter matrices $\mathbf{S}_W$, $\mathbf{S}_B$ and $\mathbf{S}_T$ under the assumption that the matrix $\mathbf{S}_2$ is non-singular. These class separability measures are commonly used as the criterion functions of the SMDLDR techniques.

Some remarkable properties of these criterion functions are described as follows.

1. Typically, the combinations $\{\mathbf{S}_B, \mathbf{S}_W\}$, $\{\mathbf{S}_T, \mathbf{S}_W\}$ and $\{\mathbf{S}_B, \mathbf{S}_T\}$ can be used for $\{\mathbf{S}_1, \mathbf{S}_2\}$ when the criterion function is to be maximized, or $\{\mathbf{S}_W, \mathbf{S}_T\}$ can be used for $\{\mathbf{S}_1, \mathbf{S}_2\}$ when the criterion function is to be minimized. All these optimization problems can lead to the same optimal linear transformation.
2. The criterion functions J_1 and J_2 are invariant to any invertible transformation. However, this fact does not hold true for the criterion functions J_3 , J_4 and J_5 , which are only invariant to the orthonormal transformation.
3. The criterion functions J_1 , J_2 and J_4 can be optimized without constraints. The resulting optimal linear transformation matrix is composed of the m eigenvectors of $\mathbf{S}_2^{-1} \mathbf{S}_1$ corresponding to the m largest or smallest eigenvalues, by which the optimal value of the criterion function equals the summation of the largest or smallest m eigenvalues for J_1 and J_4 or the summation of the logarithm of those eigenvalues for J_2 . The criterion J_2 is not suitable in the case of $\text{rank}(\mathbf{S}_1) < m$ where some eigenvalues of $\mathbf{S}_2^{-1} \mathbf{S}_1$ are zeros and thus cannot be handled by the logarithm operation. Note that some literatures also used the expression

$|\mathbf{P}^T \mathbf{S}_1 \mathbf{P}| / |\mathbf{P}^T \mathbf{S}_2 \mathbf{P}|$ as the criterion J_2 instead of the above definition. It is easy to verify that the nature of these two definitions is actually same. Optimization of the criterion function J_4 can equivalently lead to optimize $\text{trace}\left(\left(\mathbf{P}^T \left((\mathbf{S}_2 + \mathbf{S}_2^T)/2\right) \mathbf{P}\right)^{-1} \mathbf{P}^T \left((\mathbf{S}_1 + \mathbf{S}_1^T)/2\right) \mathbf{P}\right)$, which comes down to the criterion function J_1 due to the typical symmetric property of the three scatter matrices considered in our context.

4. The criterion function J_3 is a Lagrange multiplier expression as described in [2], which denotes optimization of $\text{trace}(\mathbf{P}^T \mathbf{S}_1 \mathbf{P})$ under the constraint $\text{trace}(\mathbf{P}^T \mathbf{S}_2 \mathbf{P}) = c$. Considering optimization of $J_3 = \text{trace}(\mathbf{P}^T \mathbf{S}_1 \mathbf{P}) - \alpha \cdot (\text{trace}(\mathbf{P}^T \mathbf{S}_2 \mathbf{P}) - c)$, where $\mathbf{p} \in \mathbf{R}^{n \times 1}$ is a linear transformation vector, the obtained optimal transformation vector is the eigenvector of $\mathbf{S}_2^{-1} \mathbf{S}_1$ that corresponds to the largest or smallest eigenvalue. However, if the original n -dimensional feature space is mapped into an m -dimensional subspace by the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$), the eigenvalue decomposition problem $\mathbf{S}_2^{-1} \mathbf{S}_1 \mathbf{P} = \alpha \mathbf{P}$ needs to be solved. Accordingly, the resulting optimal linear transformation matrix should be composed of the eigenvectors of $\mathbf{S}_2^{-1} \mathbf{S}_1$ corresponding to the same eigenvalue α . Since the common eigenvalue of $\mathbf{S}_2^{-1} \mathbf{S}_1$ usually equals zero, the obtained m -dimensional subspace will be the null space of $\mathbf{S}_2^{-1} \mathbf{S}_1$, which generally contains little discriminatory information. Therefore, we modify the criterion function J_3 to $J'_3 = \text{trace}(\mathbf{P}^T \mathbf{S}_1 \mathbf{P}) - \sum_{i=1}^m \alpha_i (\mathbf{p}_i^T \mathbf{S}_2 \mathbf{p}_i - c)$ with $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_m]$, $\alpha_i, i = 1, \dots, m$ and c being the linear transformation matrix, the Lagrange multipliers on the m projection directions $\mathbf{p}_i, i = 1, \dots, m$ and a constant, respectively. Consequently, optimization of the criterion function J'_3 with respect to each transformation vector $\mathbf{p}_i, i = 1, \dots, m$ will lead to the eigenvalue decomposition problems $\mathbf{S}_2^{-1} \mathbf{S}_1 \mathbf{p}_i = \alpha_i \mathbf{p}_i, i = 1, \dots, m$, which means that the optimal linear transformation matrix should be composed of the eigenvectors of $\mathbf{S}_2^{-1} \mathbf{S}_1$. By

substituting $\alpha_i = \mathbf{p}_i^T \mathbf{S}_1 \mathbf{p}_i / \mathbf{p}_i^T \mathbf{S}_2 \mathbf{p}_i$ calculated from $\mathbf{S}_2^{-1} \mathbf{S}_1 \mathbf{p}_i = \alpha_i \mathbf{p}_i$ into J_3' , the value of J_3' evaluates to $c \sum_{i=1}^m \alpha_i$. This indicates that the optimal linear transformation matrix should consist of the m eigenvectors of $\mathbf{S}_2^{-1} \mathbf{S}_1$ corresponding to the m largest or smallest eigenvalues.

5. The criterion function J_5 should be optimized under the orthonormal constraint $\mathbf{P}^T \mathbf{P} = \mathbf{I}$. By setting the derivative of the expression $J_5 - \alpha \cdot (\mathbf{P}^T \mathbf{P} - \mathbf{I})$ to zero via the Lagrange multiplier scheme with α being the multiplier coefficient, the optimal linear transformation matrix can be obtained, which is composed of the m eigenvectors of $\mathbf{S}_1 - \mu \cdot \mathbf{S}_2$ corresponding to the m largest or smallest eigenvalues. Here, the non-negative weighting factor μ is pre-specified by users, which can balance the contributions of the two scatter matrices $\mathbf{P}^T \mathbf{S}_1 \mathbf{P}$ and $\mathbf{P}^T \mathbf{S}_2 \mathbf{P}$ to J_5 .

In comparison, all aforementioned criterion functions attempt to maximize the between-class scatter while simultaneously minimizing the within-class scatter under the proper choice of $\mathbf{S}_1$ and $\mathbf{S}_2$. Among them, criterion functions J_1 , J_2 , J_3' and J_4 equivalently lead to determine the eigenvalue decomposition of $\mathbf{S}_2^{-1} \mathbf{S}_1$, but the criterion function J_5 results in solving the eigenvalue decomposition of the subtraction of weighted $\mathbf{S}_2$ from $\mathbf{S}_1$. Without loss of generality, we take J_1 as the representative among the criterion functions J_1 , J_2 , J_3' and J_4 and choose the combination $\{\mathbf{S}_B, \mathbf{S}_W\}$ for $\{\mathbf{S}_1, \mathbf{S}_2\}$. In contrast to the criterion function J_5 in the subtraction form, J_1 can be regarded as having the quotient form due to its intrinsic equivalence to J_4 .

2.3 Fisher and Chernoff linear discriminant analysis

2.3.1 Fisher linear discriminant analysis (FLDA)

The FLDA, commonly recognized as the ancestor of the discriminant analysis in its modern disguise, is one of the most popular DLDR techniques, which was developed by Fisher [116] in a model-free fashion for two-class problems and extended by Rao [117] to deal with multi-class problems. Since the FLDA can yield the globally optimal solution in a closed form via simple matrix algebras, and meanwhile the obtained optimal solution can result in a transformed subspace achieving the Bayes optimality under certain conditions, it can be regarded as an intersection of the BODLDR and SMDLDR techniques. Moreover, the FLDA is fairly easy to use because its execution only relies on the training samples. The eminent properties of the FLDA have attracted a large amount of research attention from different communities to further improve its performance.

2.3.1.1 Recapitulation of the FLDA

Assume a set of N n -dimensional feature vectors $\mathbf{x}_i \in \mathbf{R}^n, i = 1, \dots, N$ belongs to C distinct classes $\omega_i, i = 1, \dots, C$ with N_i patterns $\mathbf{x}_j^{(i)}, i = 1, \dots, C, j = 1, \dots, N_i$ in each class $\omega_i, i = 1, \dots, C$ and $\sum_{i=1}^C N_i = N$. The FLDA, also referred to as multiple discriminant analysis [1] or canonical variate analysis [82], seeks for an optimal linear transformation matrix $\mathbf{P}^* \in \mathbf{R}^{n \times m}$ ($m \leq n$) composed of m discriminant vectors $\mathbf{p}_i^* \in \mathbf{R}^{n \times 1}, i = 1, \dots, m$ as columns, i.e. $\mathbf{P}^* = [\mathbf{p}_1^*, \dots, \mathbf{p}_m^*]$, that maps the original n -dimensional feature space into an m -dimensional subspace, i.e. $\mathbf{x}_j^{(i)} \in \mathbf{R}^n \rightarrow \mathbf{P}^{*T} \mathbf{x}_j^{(i)} \in \mathbf{R}^m$, to maximize the Fisher criterion⁷ defined by:

$$J_F(\mathbf{P}) = \text{trace} \left((\mathbf{P}^T \mathbf{S}_W \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_B \mathbf{P} \right) \quad (2.30)$$

That is, $J_F(\mathbf{P}^*) = \max_{\mathbf{P} \in \mathbf{R}^{n \times m}} J_F(\mathbf{P})$. Here $\mathbf{S}_W$ and $\mathbf{S}_B$ denote the average within-class scatter matrix and the between-class scatter matrix, respectively, and are estimated by:

⁷ The Fisher criterion may have various formulations that can lead to the same optimal solution, see subsection 2.2.2.

$$\mathbf{S}_W = \sum_{i=1}^C P_i \mathbf{S}_i \quad (2.31)$$

$$\text{with } P_i = N_i/N \text{ and } \mathbf{S}_i = \frac{1}{N_i} \sum_j^{N_i} (\mathbf{x}_j^{(i)} - \mathbf{m}_i)(\mathbf{x}_j^{(i)} - \mathbf{m}_i)^T$$

$$\begin{aligned} \mathbf{S}_B &= \sum_{i=1}^C P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ &= \sum_{i=1}^C P_i \left(\sum_{j=1}^C P_j \right) (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ &= \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T + \sum_{i=1}^C P_i (\mathbf{m}_i - \mathbf{m}) \sum_{j=1}^C P_j (\mathbf{m}_j - \mathbf{m})^T \\ &= \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T + \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\ &= \frac{1}{2} \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T + \frac{1}{2} \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\ &\quad + \frac{1}{2} \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T + \frac{1}{2} \sum_{i=1}^C \sum_{j=1}^C P_i P_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ &= \frac{1}{2} \sum_{i=1}^C \sum_{j=1}^C P_i P_j ((\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T + (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\ &\quad + (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T + (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T) \\ &= \frac{1}{2} \sum_{i=1}^C \sum_{j=1}^C P_i P_j ((\mathbf{m}_i - \mathbf{m}) + (\mathbf{m} - \mathbf{m}_j))((\mathbf{m}_i - \mathbf{m}) + (\mathbf{m} - \mathbf{m}_j))^T \\ &= \frac{1}{2} \left(2 \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T + \sum_{i=1}^C P_i P_i (\mathbf{m}_i - \mathbf{m}_i)(\mathbf{m}_i - \mathbf{m}_i)^T \right) \quad (2.32) \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \end{aligned}$$

$$\text{with } P_i = N_i/N, \mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)} \text{ and } \mathbf{m} = \sum_{i=1}^C P_i \mathbf{m}_i = \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}$$

where P_i , $\mathbf{m}_i$, $\mathbf{m}$ and $\mathbf{S}_i$ represent the prior probability of class ω_i , the mean vector of class ω_i , the overall mean vector of the available feature vectors and the scatter matrix of class ω_i , respectively. The between-class scatter matrix $\mathbf{S}_B$ can be expressed by both its conventional formulation and the equivalent pairwise decomposition form. The total scatter matrix $\mathbf{S}_T$ is estimated by:

$$\mathbf{S}_T = \frac{1}{N_i} \sum_{i=1}^C \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \mathbf{m})(\mathbf{x}_j^{(i)} - \mathbf{m})^T \text{ with } \mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B \quad (2.33)$$

It is easy to verify that all three scatter matrices $\mathbf{S}_W$, $\mathbf{S}_B$ and $\mathbf{S}_T$ are positive semi-definite with their ranks as follows:

$$\begin{aligned} \text{rank}(\mathbf{S}_W) &\leq \min(n, N - C) \\ \text{rank}(\mathbf{S}_B) &\leq \min(n, C - 1) \\ \text{rank}(\mathbf{S}_T) &\leq \min(n, N - 1) \end{aligned} \quad (2.34)$$

Maximization of the Fisher criterion defined in Eq. (2.30) with respect to the linear transformation matrix will lead to solve a generalized eigenvalue problem $\mathbf{S}_B \mathbf{p} = \lambda \mathbf{S}_W \mathbf{p}$. Since both scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ are positive semi-definite with $\text{rank}(\mathbf{S}_B) \leq \min(n, C - 1)$, there exist at most $\min(n, C - 1)$ generalized eigenvectors corresponding to positive generalized eigenvalues. We can therefore take those $\min(m, n, C - 1)$ generalized eigenvectors as columns of $\mathbf{P}^*$, which results in the maximal Fisher criterion value equal to the summation of the largest $\min(m, n, C - 1)$ generalized eigenvalues. If $\mathbf{S}_W$ is non-singular, the generalized eigenvalue problem comes down to the conventional eigenvalue problem $\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{p} = \lambda \mathbf{p}$, whose $\min(m, n, C - 1)$ leading eigenvectors corresponding to the $\min(m, n, C - 1)$ largest positive eigenvalues constitute $\mathbf{P}^*$. Note that the same optimal solution $\mathbf{P}^*$ can be obtained by maximizing $\text{trace}((\mathbf{P}^T \mathbf{S}_T \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_B \mathbf{P})$ or minimizing $\text{trace}((\mathbf{P}^T \mathbf{S}_B \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_W \mathbf{P})$. It is also worth noting that the optimal transformation matrix $\mathbf{P}^*$ may be not unique since any invertible transformations further imposed on $\mathbf{P}^*$ do not change the Fisher criterion value. However, they can change the coordinate system to represent the optimal subspace spanned by the optimal linear transformation matrix. Some coordinate systems may have unique properties desirable for certain subsequent classification tasks.

To save the computational cost and improve the numerical stability when solving the eigenvalue decomposition, each of the three scatter matrices can be reformulated by the outer product of a matrix with the smaller scale as follows.

$$\mathbf{S}_W = \mathbf{H}_W \mathbf{H}_W^T, \mathbf{S}_B = \mathbf{H}_B \mathbf{H}_B^T \text{ and } \mathbf{S}_T = \mathbf{H}_T \mathbf{H}_T^T \quad (2.35)$$

$$\begin{aligned} \text{with } \mathbf{H}_W &= \left[\sqrt{\frac{1}{N}}(\mathbf{x}_1^{(1)} - \mathbf{m}_1), \dots, \sqrt{\frac{1}{N}}(\mathbf{x}_{N_C}^{(C)} - \mathbf{m}_C) \right] \in \mathbf{R}^{n \times N} \\ \mathbf{H}_B &= \left[\sqrt{\frac{N_i}{N}}(\mathbf{m}_1 - \mathbf{m}), \dots, \sqrt{\frac{N_C}{N}}(\mathbf{m}_C - \mathbf{m}) \right] \in \mathbf{R}^{n \times C} \\ \mathbf{H}_T &= \left[\sqrt{\frac{1}{N}}(\mathbf{x}_1^{(1)} - \mathbf{m}), \dots, \sqrt{\frac{1}{N}}(\mathbf{x}_{N_C}^{(C)} - \mathbf{m}) \right] \in \mathbf{R}^{n \times N} \end{aligned}$$

where $\mathbf{x}_i^{(j)}$ denotes the i^{th} feature vector belonging to class ω_j , $i=1, \dots, N_j$, $j=1, \dots, C$. Accordingly, the eigenvalue decomposition of $\mathbf{S}_W$, $\mathbf{S}_B$ and $\mathbf{S}_T$ can be obtained by performing the Singular Value Decomposition (SVD) [92] of $\mathbf{H}_W$, $\mathbf{H}_B$ and $\mathbf{H}_T$, respectively, with computation costs reduced. Recently, the QR decomposition [118] and Schur decomposition were also introduced [119] to address this issue.

2.3.1.2 Bayes optimality analysis of the FLDA solutions

To simplify analysis, we assume that the average-within class scatter matrix $\mathbf{S}_W$ is positive definite. Accordingly, maximization of the Fisher criterion in Eq. (2.30) leads to determine the conventional eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$, which can be addressed by the simultaneous diagonalization scheme [2]. Specifically, the eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$ can be performed by firstly whitening $\mathbf{S}_W$ to an identity matrix, i.e. $\mathbf{P}_W^T \mathbf{S}_W \mathbf{P}_W = \mathbf{I}_n$ with $\mathbf{P}_W = \mathbf{S}_W^{-1/2}$ where $\mathbf{S}_W^{-1/2}$ is the inverted square root of $\mathbf{S}_W$, and then determining the eigenvalue decomposition of

Chapter 2: An Overview of Discriminative Linear Dimensionality Reduction Techniques

$\tilde{\mathbf{S}}_B = (\mathbf{S}_W^{-1/2})^T \mathbf{S}_B \mathbf{S}_W^{-1/2}$, i.e. $\tilde{\mathbf{S}}_B = \tilde{\mathbf{P}}_B \mathbf{\Lambda}_B \tilde{\mathbf{P}}_B^T$ with $\text{rank}(\tilde{\mathbf{S}}_B) = C - 1$ in general and the diagonal matrix $\mathbf{\Lambda}_B = \text{diag}(\lambda_{b1}, \dots, \lambda_{bn})$, $\lambda_{b1} \geq \dots \geq \lambda_{bC-1} > 0 = \lambda_{bC} = \dots = \lambda_{bn}$. Eventually, the n eigenvectors of $\mathbf{S}_W^{-1} \mathbf{S}_B$ and their corresponding eigenvalues in the non-increasing order are obtained as $\mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bi}, i = 1, \dots, n$ with $\tilde{\mathbf{P}}_B = [\tilde{\mathbf{p}}_{b1}, \dots, \tilde{\mathbf{p}}_{bn}]$ and $\lambda_{b1} > \dots > \lambda_{bC-1} > 0 = \lambda_{bC} = \dots = \lambda_{bn}$. Consequently, we can obtain an optimal linear transformation matrix $\mathbf{P}^* = \mathbf{S}_B^{-1/2} \tilde{\mathbf{P}}_B^* = [\mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{b1}, \dots, \mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bm}]$ with $\tilde{\mathbf{P}}_B^* = [\tilde{\mathbf{p}}_{b1}, \dots, \tilde{\mathbf{p}}_{bm}]$ consisting of the first m columns of $\tilde{\mathbf{P}}_B$, by which the Fisher criterion is maximized at $J_F(\mathbf{P}^*) = \sum_{k=1}^m \lambda_{bk}$. Note that $\tilde{\mathbf{P}}_B$ is orthonormal due to the symmetric property of $\tilde{\mathbf{S}}_B$, and thus we have $\tilde{\mathbf{P}}_B^T \tilde{\mathbf{P}}_B = \tilde{\mathbf{P}}_B \tilde{\mathbf{P}}_B^T = \mathbf{I}_n$, i.e. $\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{p}}_{bk} = 1, k = 1, \dots, n$, $\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{p}}_{bl} = 0, k, l = 1, \dots, n, k \neq l$, and $\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{P}}_B^* = \mathbf{I}_m$.

Due to the non-singularity of $\mathbf{S}_W^{1/2}$, we can obtain the following expression:

$$\begin{aligned} \max_{\mathbf{P} \in \mathbf{R}^{n \times m}} J_F(\mathbf{P}) &= \max_{\mathbf{P} \in \mathbf{R}^{n \times m}} \text{trace}((\mathbf{P}^T \mathbf{S}_W \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_B \mathbf{P}) \\ &= \max_{\mathbf{P}' \in \mathbf{R}^{n \times m}} \text{trace}((\mathbf{P}'^T \mathbf{P}')^{-1} \mathbf{P}'^T \tilde{\mathbf{S}}_B \mathbf{P}') \\ &= \max_{\mathbf{P}' \in \mathbf{R}^{n \times m}} J'_F(\mathbf{P}') \end{aligned} \quad (2.36)$$

$$\text{with } \mathbf{P}' = \mathbf{P} \mathbf{S}_W^{1/2} \text{ and } J'_F(\mathbf{P}') = \text{trace}((\mathbf{P}'^T \mathbf{P}')^{-1} \mathbf{P}'^T \tilde{\mathbf{S}}_B \mathbf{P}')$$

It is easy to verify that $\tilde{\mathbf{P}}_B^*$ is an optimal solution of $\mathbf{P}'$ that satisfies Eq. (2.36) with the maximum Fisher criterion value equal to $J'_F(\tilde{\mathbf{P}}_B^*)$. Due to $\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{P}}_B^* = \mathbf{I}_m$, we can further obtain the following expression:

$$\max_{\mathbf{P}' \in \mathbf{R}^{n \times m}} J'_F(\mathbf{P}') = \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} J'_F(\mathbf{P}') = \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \text{trace}(\mathbf{P}'^T \tilde{\mathbf{S}}_B \mathbf{P}') \quad (2.37)$$

Some attractive properties related to the FLDA solutions are demonstrated as follows.

Lemma 2.1 Let the available feature vectors in the original feature space be separately projected along the $n+1$ - C vectors $\mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bk}, k = C, \dots, n$ defined earlier. Then, in each of the transformed one-dimensional subspaces, the Mahalanobis distances of any class pairs ω_i and ω_j , $i, j = 1, \dots, C, i \neq j$ come down to the Euclidean distances between the corresponding class means and equal zero, i.e.

$$(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\mathbf{S}}_w^{-1} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) = (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) = 0, k = C, \dots, n, i, j = 1, \dots, C, i \neq j$$

with $\hat{\mathbf{m}}_i = \tilde{\mathbf{p}}_{bk}^T (\mathbf{S}_W^{-1/2})^T \mathbf{m}_i$ and $\hat{\mathbf{S}}_w = \tilde{\mathbf{p}}_{bk}^T (\mathbf{S}_W^{-1/2})^T \mathbf{S}_W \mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bk}$ being the mean vector of class ω_i and the average within-class scatter matrix in the transformed one-dimensional subspace, respectively.

Proof: We have $\tilde{\mathbf{S}}_B = (\mathbf{S}_W^{-1/2})^T \mathbf{S}_B \mathbf{S}_W^{-1/2} = \tilde{\mathbf{P}}_B \mathbf{\Lambda}_B \tilde{\mathbf{P}}_B^T$ with $\tilde{\mathbf{P}}_B = [\tilde{\mathbf{p}}_{b1}, \dots, \tilde{\mathbf{p}}_{bn}]$ and $\mathbf{\Lambda}_B = \text{diag}(\lambda_{b1}, \dots, \lambda_{bC-1}, 0, \dots, 0)$. Obviously, the $n+1$ - C eigenvectors $\tilde{\mathbf{p}}_{bk}, k = C, \dots, n$ of $\tilde{\mathbf{S}}_B$ correspond to eigenvalues equal to zero. That is,

$$\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{S}}_B \tilde{\mathbf{p}}_{bk} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \tilde{\mathbf{p}}_{bk}^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{p}}_{bk} = 0, k = C, \dots, n, \tilde{\mathbf{m}}_i = (\mathbf{S}_W^{-1/2})^T \mathbf{m}_i$$

Since $\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{S}}_B \tilde{\mathbf{p}}_{bk}, k = C, \dots, n$ are scalar, we can obtain:

$$\begin{aligned} \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{S}}_B \tilde{\mathbf{p}}_{bk} &= \text{trace}(\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{S}}_B \tilde{\mathbf{p}}_{bk}) \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \text{trace}((\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j) (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j)^T) \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j) \\ &= 0 \quad k = C, \dots, n \end{aligned}$$

Furthermore, due to the fact that

$$(\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j) \geq 0, k = C, \dots, n, i, j = 1, \dots, C, i \neq j$$

and $P_i > 0, i = 1, \dots, C$, we can obtain:

$$(\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j) = 0, k = C, \dots, n, i, j = 1, \dots, C, i \neq j. \text{ Meanwhile, it is}$$

easy to verify that $\tilde{\mathbf{p}}_{bk}^T (\mathbf{S}_W^{-1/2})^T \mathbf{S}_W \mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bk} = 1, k = C, \dots, n$ due to $\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{p}}_{bk} = 1, k = C, \dots, n$.

Therefore, we can obtain:

$$\begin{aligned}
 (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\mathbf{S}}_W^{-1} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) &= (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \left(\tilde{\mathbf{p}}_{bk}^T (\mathbf{S}_W^{-1/2})^T \mathbf{S}_W \mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bk} \right)^{-1} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) \\
 &= (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) \\
 &= (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j) \\
 &= (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j) \\
 &= 0 \quad k = C, \dots, n, i, j = 1, \dots, C, i \neq j
 \end{aligned}$$

This completes the proof of the Lemma 2.1.

Theorem 2.1 Assume that the original n -dimensional feature space is mapped into an m -dimensional subspace by the linear transformation matrix $\mathbf{P}^* = \mathbf{S}_W^{-1/2} \tilde{\mathbf{P}}_B^* = [\mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{b1}, \dots, \mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bm}]$ defined earlier. Then, if $m \geq C - 1$, (1) Mahalanobis distances between any class pairs ω_i and ω_j , $i, j = 1, \dots, C, i \neq j$ in original feature space are preserved and come down to Euclidean distances between the corresponding class means in the transformed subspace, i.e.

$$(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j) = (\mathbf{P}^{*T} \mathbf{m}_i - \mathbf{P}^{*T} \mathbf{m}_j)^T (\mathbf{P}^{*T} \mathbf{m}_i - \mathbf{P}^{*T} \mathbf{m}_j).$$

(2) The optimal Fisher criterion value in the transformed subspace retains same as that in the original feature space; (3) Any invertible transformation matrix $\mathbf{Q} \in \mathbf{R}^{m \times m}$ further imposed on $\mathbf{P}^*$ does not change the pairwise Mahalanobis distances and Fisher criterion value in the transformed subspace; (4) Bayes errors between any class pairs ω_i and ω_j , $i, j = 1, \dots, C, i \neq j$ in the transformed subspace retains same as that in the original feature space.

Proof: According to lemma 2.1, when $m \geq C - 1$, we have

$$(\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bk}^T \tilde{\mathbf{m}}_j) = 0, k = m + 1, \dots, n, i, j = 1, \dots, C, i \neq j.$$

According to $\tilde{\mathbf{P}}_B^T \tilde{\mathbf{P}}_B = \tilde{\mathbf{P}}_B \tilde{\mathbf{P}}_B^T = \mathbf{I}_n$ and $\mathbf{P}^{*T} \mathbf{S}_W \mathbf{P}^* = \tilde{\mathbf{P}}_B^{*T} (\mathbf{S}_W^{-1/2})^T \mathbf{S}_W \mathbf{S}_W^{-1/2} \tilde{\mathbf{P}}_B^* = \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{P}}_B^* = \mathbf{I}_m$,

we can obtain:

$$\begin{aligned}
 (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j) &= \\
 &= (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) \\
 &= (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{P}}_B \tilde{\mathbf{P}}_B^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) \\
 &= \text{trace}(\tilde{\mathbf{P}}_B^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{P}}_B) \\
 &= \sum_{k=1}^m \tilde{\mathbf{p}}_{bk}^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{p}}_{bk} + \sum_{t=m+1}^n \tilde{\mathbf{p}}_{bt}^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{p}}_{bt} \\
 &= \sum_{k=1}^m \tilde{\mathbf{p}}_{bk}^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{p}}_{bk} + \sum_{t=m+1}^n (\tilde{\mathbf{p}}_{bt}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bt}^T \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{p}}_{bt}^T \tilde{\mathbf{m}}_i - \tilde{\mathbf{p}}_{bt}^T \tilde{\mathbf{m}}_j) \\
 &= \sum_{k=1}^m \tilde{\mathbf{p}}_{bk}^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{p}}_{bk} \\
 &= \text{trace}(\tilde{\mathbf{P}}_B^{*T} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{P}}_B^*) \\
 &= (\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_i - \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_j)^T (\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_i - \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_j) \\
 &= (\mathbf{P}^* \mathbf{m}_i - \mathbf{P}^{*T} \mathbf{m}_j)^T (\mathbf{P}^{*T} \mathbf{m}_i - \mathbf{P}^{*T} \mathbf{m}_j) \\
 &= (\mathbf{P}^{*T} \mathbf{m}_i - \mathbf{P}^{*T} \mathbf{m}_j)^T (\mathbf{P}^{*T} \mathbf{S}_W \mathbf{P}^*)^{-1} (\mathbf{P}^{*T} \mathbf{m}_i - \mathbf{P}^{*T} \mathbf{m}_j)
 \end{aligned}$$

Moreover, we can obtain $J_F(\mathbf{P}^*) = \sum_{k=1}^m \lambda_{bk} = \sum_{k=1}^{C-1} \lambda_{bk} = \sum_{k=1}^n \lambda_{bk} = J_F(\mathbf{S}_W^{-1/2} \tilde{\mathbf{P}}_B)$ due to

$\lambda_{bk} = 0, k \geq C$. For any invertible transformation matrix $\mathbf{Q} \in \mathbf{R}^{m \times m}$ further imposed on $\mathbf{P}^*$, we can obtain:

$$\begin{aligned}
 (\mathbf{Q}^T \mathbf{P}^{*T} \mathbf{m}_i - \mathbf{Q}^T \mathbf{P}^{*T} \mathbf{m}_j)^T (\mathbf{Q}^T \mathbf{P}^{*T} \mathbf{S}_W \mathbf{P}^* \mathbf{Q})^{-1} (\mathbf{Q}^T \mathbf{P}^{*T} \mathbf{m}_i - \mathbf{Q}^T \mathbf{P}^{*T} \mathbf{m}_j) &= (\tilde{\mathbf{P}}_B^{*T} \mathbf{m}_i - \tilde{\mathbf{P}}_B^{*T} \mathbf{m}_j)^T (\mathbf{P}^{*T} \mathbf{S}_W \mathbf{P}^*)^{-1} (\tilde{\mathbf{P}}_B^{*T} \mathbf{m}_i - \tilde{\mathbf{P}}_B^{*T} \mathbf{m}_j) \\
 J_F(\mathbf{P}^* \mathbf{Q}) &= \text{trace}((\mathbf{Q}^T \mathbf{P}^{*T} \mathbf{S}_W \mathbf{P}^* \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{P}^{*T} \mathbf{S}_B \mathbf{P}^* \mathbf{Q}) = \text{trace}((\mathbf{P}^{*T} \mathbf{S}_W \mathbf{P}^*)^{-1} \mathbf{P}^{*T} \mathbf{S}_B \mathbf{P}^*) = J_F(\mathbf{P}^*)
 \end{aligned}$$

According to Eq. (2.10), under the Gaussian homoscedastic assumption, the Bayes error between any two classes is completely determined by the Mahalanobis distance and prior probabilities of these two classes. Since all class prior probabilities and pairwise Mahalanobis distances after the linear transformation $\mathbf{P}^*$ remains same as those in the original feature space, the Bayes errors between any two classes in the transformed subspace retains same as that in the original feature space.

This completes the proof of the Theorem 2.1.

In general, the equivalence between all pairwise Bayes errors in a transformed subspace and those in the original feature space may not indicate that the overall Bayes error in that transformed subspace retains same as that in the original feature space. However, by using the linear transformation matrix $\mathbf{P}^*$ with $m \geq C - 1$, this statement is supported by the following theorem.

Theorem 2.2 [82] Under the Gaussian homoscedastic assumption, suppose that the original n -dimensional feature space is mapped into an m -dimensional subspace by the linear transformation matrix $\mathbf{P}^* = \mathbf{S}_W^{-1/2} \tilde{\mathbf{P}}_B^* = [\mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{b1}, \dots, \mathbf{S}_W^{-1/2} \tilde{\mathbf{p}}_{bm}]$ defined earlier. Then, if $m \geq C - 1$, the overall Bayes error in the transformed subspace retains same as that in the original feature space. In other words, the Bayes optimality is achieved in the transformed subspace.

Proof: Based on $\tilde{\mathbf{S}}_B = \sum_{i=1}^C (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^T$ with $\tilde{\mathbf{m}} = \sum_{i=1}^C P_i \tilde{\mathbf{m}}_i$ being the overall mean vector in the $\mathbf{S}_W$ whitened space, we have $\tilde{\mathbf{P}}_B^T \tilde{\mathbf{S}}_B \tilde{\mathbf{P}}_B = \mathbf{\Lambda}_B = \text{diag}(\lambda_{b1}, \dots, \lambda_{bC-1}, 0, \dots, 0)$.

Suppose that $\tilde{\mathbf{P}}_B = [\tilde{\mathbf{P}}_B^*, \tilde{\mathbf{P}}_B^*]$ with $\tilde{\mathbf{P}}_B^*$ defined earlier and $\tilde{\mathbf{P}}_B^* = [\tilde{\mathbf{p}}_{bm+1}, \dots, \tilde{\mathbf{p}}_{bm}]$. We can obtain:

$$\begin{bmatrix} \tilde{\mathbf{P}}_B^{*T} \\ \tilde{\mathbf{P}}_B^{*T} \end{bmatrix} \tilde{\mathbf{S}}_B \begin{bmatrix} \tilde{\mathbf{P}}_B^* \\ \tilde{\mathbf{P}}_B^* \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{S}}_B \tilde{\mathbf{P}}_B^*, \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{S}}_B \tilde{\mathbf{P}}_B^* \\ \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{S}}_B \tilde{\mathbf{P}}_B^*, \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{S}}_B \tilde{\mathbf{P}}_B^* \end{bmatrix} = \mathbf{\Lambda}_B.$$

Accordingly, if $m \geq C - 1$, we can obtain:

$$\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{S}}_B \tilde{\mathbf{P}}_B^* = \sum_{i=1}^C \tilde{\mathbf{P}}_B^{*T} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^T \tilde{\mathbf{P}}_B^* = \mathbf{0}_{(n-m) \times (n-m)},$$

Consequently, we can obtain:

$$\text{trace} \left(\sum_{i=1}^C \tilde{\mathbf{P}}_B^{*T} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^T \tilde{\mathbf{P}}_B^* \right) = \sum_{i=1}^C \left(\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_i - \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}} \right)^T \left(\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_i - \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}} \right) = 0 \text{ and thus}$$

$$\tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}_i = \tilde{\mathbf{P}}_B^{*T} \tilde{\mathbf{m}}, i = 1, \dots, C, \text{ i.e. } \tilde{\mathbf{P}}_B^{*T} (\mathbf{S}_W^{-1/2})^T \mathbf{m}_i = \tilde{\mathbf{P}}_B^{*T} (\mathbf{S}_W^{-1/2})^T \mathbf{m}, i = 1, \dots, C.$$

Under the Gaussian homoscedastic assumption, the Bayes classifier in the original feature space can be formulated by $g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) + \ln P_i, i = 1, \dots, C$

with $\mathbf{S} = \mathbf{S}_w$. Assume that $\mathbf{P} = [\mathbf{P}^*, \bar{\mathbf{P}}^*] = [\mathbf{S}_w^{-1/2} \tilde{\mathbf{P}}_B^*, \mathbf{S}_w^{-1/2} \tilde{\mathbf{P}}_B^*]$. We can obtain:

$$\bar{\mathbf{P}}^{*T} \mathbf{m}_i = \bar{\mathbf{P}}^{*T} \mathbf{m}, i = 1, \dots, C \text{ and}$$

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}_w^{-1}(\mathbf{x} - \mathbf{m}_i) - \ln P_i \\ &= -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{P}(\mathbf{P}^T \mathbf{S}_w \mathbf{P})^{-1} \mathbf{P}^T(\mathbf{x} - \mathbf{m}_i) - \ln P_i \\ &= -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{P} \mathbf{P}^T(\mathbf{x} - \mathbf{m}_i) - \ln P_i \\ &= -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T (\mathbf{P}^* \mathbf{P}^{*T} + \bar{\mathbf{P}}^* \bar{\mathbf{P}}^{*T})(\mathbf{x} - \mathbf{m}_i) - \ln P_i \\ &= -\frac{1}{2}(\mathbf{P}^{*T} \mathbf{x} - \mathbf{P}^{*T} \mathbf{m}_i)^T (\mathbf{P}^{*T} \mathbf{x} - \mathbf{P}^{*T} \mathbf{m}_i) - \frac{1}{2}(\bar{\mathbf{P}}^{*T} \mathbf{x} - \bar{\mathbf{P}}^{*T} \mathbf{m}_i)^T (\bar{\mathbf{P}}^{*T} \mathbf{x} - \bar{\mathbf{P}}^{*T} \mathbf{m}_i) - \ln P_i \\ &= -\frac{1}{2}(\mathbf{P}^{*T} \mathbf{x} - \mathbf{P}^{*T} \mathbf{m}_i)^T (\mathbf{P}^{*T} \mathbf{x} - \mathbf{P}^{*T} \mathbf{m}_i) - \frac{1}{2}(\bar{\mathbf{P}}^{*T} \mathbf{x} - \bar{\mathbf{P}}^{*T} \mathbf{m})^T (\bar{\mathbf{P}}^{*T} \mathbf{x} - \bar{\mathbf{P}}^{*T} \mathbf{m}) - \ln P_i \\ &= g_i(\mathbf{P}^{*T} \mathbf{x}) - \frac{1}{2}(\bar{\mathbf{P}}^{*T} \mathbf{x} - \bar{\mathbf{P}}^{*T} \mathbf{m})^T (\bar{\mathbf{P}}^{*T} \mathbf{x} - \bar{\mathbf{P}}^{*T} \mathbf{m}) \end{aligned}$$

It is obvious that the second term in the last line of the above expression is independent of class index i . Therefore, the Bayes decision made for the feature vector $\mathbf{x}$ according to $g_i(\mathbf{P}^{*T} \mathbf{x})$ retains same as that made in terms of $g_i(\mathbf{x})$, which indicates that the overall Bayes error in the transformed subspace spanned by $\mathbf{P}^*$ remains same as that in the original feature space. In other words, the Bayes optimality is achieved in the transformed subspace.

This completes the proof of the Theorem 2.2.

2.3.1.3 Conjugated orthogonality constraints on FLDA

According to Theorem 2.1, the Fisher criterion is invariant to any further imposed invertible transformations. Therefore, the FLDA technique can have many optimal solutions achieving the maximum Fisher criterion value, which actually span the same subspace under different coordinate systems with the coordinate axes being the

optimal discriminant vectors. In general, discriminant vectors with different properties may influence the performance of the subsequent supervised pattern analysis task. For example, discriminant vectors orthogonal to each other span a subspace where noise in the original feature space may be suppressed [124]; discriminant vectors leading to uncorrelated features span a subspace where correlation among features that is undesirable for some classifiers, e.g. k NN classifier, can be removed; discriminant vectors maximizing Eq. (2.31) span a subspace where the Bayes classifier under the Gaussian homoscedastic assumption can be simplified such that the computational cost due to matrix inversion operations is saved. Fisher discriminant vectors with these characteristics can be derived by maximizing the Fisher criterion under the constraints of conjugated orthogonality to matrices $\mathbf{I}$, $\mathbf{S}_T$ and $\mathbf{S}_W$, respectively. To simplify discussions, we assume that the matrix $\mathbf{S}_W$ is positive definite hereafter.

Fisher discriminant vectors subject to the conjugated orthogonality to identity matrix $\mathbf{I}$ correspond to a family of the so-called orthogonal FLDA (OFLDA) techniques, in which the extracted discriminant vectors are orthogonal to each other, i.e. $\mathbf{p}_i^T \mathbf{p}_j = 0, i, j = 1, \dots, m, i \neq j$. The well-known Foley-Sammon FLDA technique was invented by Foley and Sammon [120] for two-class problems and extended by Duchene and Leclercy [121] to handle multi-class problems, in which new discriminant vectors are sequentially extracted in the subspace complementary orthogonal to the subspace spanned by the already extracted discriminant vectors to maximize the Fisher criterion. Theoretical analysis presented in [122] yields an efficient way to obtain the orthogonal discriminant vectors. Specifically, suppose that a set of i discriminant vectors $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, i \geq 1$ have been obtained. The $(i+1)^{\text{th}}$ discriminant vector $\mathbf{p}_{i+1}$ can be extracted as the eigenvector of $\mathbf{S}_W^{-1} \mathbf{U} \mathbf{S}_B$ corresponding to the largest eigenvalue, where $\mathbf{U} = \mathbf{I} - \mathbf{P}^T (\mathbf{P} \mathbf{S}_W^{-1} \mathbf{P}^T)^{-1} \mathbf{P} \mathbf{S}_W^{-1}$, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i]^T$. To facilitate future experimental comparisons, we name this technique FLDA_{q1}. Besides the sequential implementation, the OFLDA technique can also be performed in the batch manner [123], [124], in which the Gram-Schmidt orthogonalization or the QR decomposition is applied to derive a set of orthogonal discriminant vectors based on the already obtained discriminant vectors not satisfying the orthogonality constraint.

Fisher discriminant vectors subject to the conjugated orthogonality to matrix $\mathbf{S}_T$ correspond to a family of the so-called uncorrelated FLDA (UFLDA) techniques. The earliest study of the UFLDA was conducted by Jin *et al.* [122], in which discriminant vectors were sequentially extracted to maximize the Fisher criterion under the constraint of conjugated orthogonality to $\mathbf{S}_T$, i.e. $\mathbf{p}_i^T \mathbf{S}_T \mathbf{p}_j = 0, i, j = 1, \dots, m, i \neq j$. Theoretical analysis in [122] has demonstrated that suppose that a set of i discriminant vectors $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, i \geq 1$, have been extracted. The $(i+1)^{\text{th}}$ discriminant vector $\mathbf{p}_{i+1}$ can be extracted as the eigenvector of $\mathbf{S}_W^{-1} \mathbf{U} \mathbf{S}_B$ corresponding to the largest eigenvalue, where $\mathbf{U} = \mathbf{I} - \mathbf{S}_T \mathbf{P}^T (\mathbf{P} \mathbf{S}_T \mathbf{S}_W^{-1} \mathbf{S}_T \mathbf{P}^T)^{-1} \mathbf{P} \mathbf{S}_T \mathbf{S}_W^{-1}$, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i]^T$ and $\mathbf{I}$ is the identity matrix. To facilitate future experimental comparisons, we name this technique FLDA_{q2}. Note that based on the discriminant vectors extracted by the FLDA_{q2} technique, the Gram-Schmidt orthogonalization or QR decomposition can be applied to generate a set of orthogonal discriminant vectors [124]. Such uncorrelated-to-orthogonal FLDA technique is hereby named FLDA_{q3}. Intrinsic equivalence between the UFLDA and the conventional FLDA, as demonstrated in [125], motivates an efficient UFLDA implementation [124] that can compute the optimal discriminant vectors in the batch manner via the simultaneous diagonalization scheme. A subtle distinction between the sequential and batch ULFDA implemented in the existing literatures is that the former one generally requires the unitary length of the extracted discriminant vectors and thus leads to $\mathbf{P}^T \mathbf{P} = \mathbf{I}$ while the later one satisfies $\mathbf{P}^T \mathbf{S}_T \mathbf{P} = \mathbf{I}$. The condition $\mathbf{P}^T \mathbf{S}_T \mathbf{P} = \mathbf{I}$ might be preferred in practice since it can normalize each transformed feature to the unitary variance. In fact, a post-processing step can be incorporated into the sequential technique to implement $\mathbf{P}^T \mathbf{S}_T \mathbf{P} = \mathbf{I}$.

Fisher discriminant vectors, subject to the conjugated orthogonality to $\mathbf{S}_W$, correspond to the conventional FLDA technique. In the conventional FLDA, an optimal solution can be obtained by determining the eigenvalue decomposition of $\mathbf{S}_W^{-1} \mathbf{S}_B$, which is equivalent to firstly whitening $\mathbf{S}_W$ and then performing the eigenvalue decomposition of $\tilde{\mathbf{S}}_B = (\mathbf{S}_W^{-1/2})^T \mathbf{S}_B \mathbf{S}_W^{-1/2}$. The symmetric property of $\tilde{\mathbf{S}}_B$ ensures its eigenvectors to be

orthonormal to each other, which consequently retains $\mathbf{S}_w$ whitened in the ultimately transformed subspace. Therefore, the Mahalanobis distances involved in the LDC formula come down to the Euclidean distances in the transformed subspace and accordingly save the computational cost.

Some remarkable properties of the Fisher discriminant vectors subject to the above conjugated orthogonality constraints are outlined as follows.

- For C -class problems, assume that $\text{rank}(\mathbf{S}_B) = C - 1$ and $\mathbf{S}_w$ is positive definite. Let $C-1$ nonzero eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_B$ be represented in the non-increasing order, i.e. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{C-1} > 0$ and suppose $\lambda_i \neq \lambda_j$, $i, j = 1, \dots, C, i \neq j$. Then, for $m \leq C - 1$, the m^{th} discriminant vector extracted by the UFLDA is the m^{th} eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_B$ corresponding to the m^{th} largest eigenvalue; for $m > C - 1$, the m^{th} discriminant vector of UFLDA corresponds to the Fisher criterion value equal to zero. See proof in [125].
- For C -class problems, assume that $\text{rank}(\mathbf{S}_B) = C - 1$ and $\mathbf{S}_w$ is positive definite. Let the eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_B$ be represented in the non-increasing order, i.e. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{C-1} > 0 = \lambda_C = \dots = \lambda_n$. Then, the m^{th} ($m = 1, \dots, n$) discriminant vector extracted by the OFLDA corresponds to the Fisher criterion value greater than or equal to the m^{th} eigenvalue λ_m . Consequently, the number of orthogonal discriminant vectors corresponding to non-zero Fisher criterion values may exceed $C-1$. See proof in [126].
- For C -class problems, assume that $\text{rank}(\mathbf{S}_B) = C - 1$ and $\mathbf{S}_w$ is positive definite. Let the eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_B$ be represented in the non-increasing order, i.e. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{C-1} > 0 = \lambda_C = \dots = \lambda_n$. Then, the Fisher criterion value corresponding to the m^{th} ($m = 1, \dots, n$) discriminant vector of the OFLDA is not less

than that corresponding to the m^{th} ($m = 1, \dots, n$) discriminant vector of the UFLDA. See proof in [126].

2.3.1.4 Pros and cons of the Fisher criterion

The FLDA has demonstrated promising performance in diverse applications. It is commonly used as a non-parametric tool since its execution merely relies on the first and second statistics estimated from the training samples with no need of the parametric assumptions of class distributions. Another particularly nice property of the FLDA is its globally optimal solution in the closed-form can significantly reduce the computational cost and overcome the dilemma of local optima. However, some potential deficiencies associated with the conventional Fisher criterion destroy its relation to the Bayes error in general sense and consequently prevent it to achieve better performance. As to be discussed in the following, the first three deficiencies are associated with the multi-class Fisher criterion while the last three deficiencies exist in both two-class and multi-class scenarios. Recall that C , n and m denote the class number, the original feature space dimensionality and the dimension to be reduced to, respectively.

Firstly, the Fisher criterion is not related to the Bayes error in a general sense. Although Theorem 2.2 has demonstrated that the solution of the FLDA can result in a transformed subspace achieving the Bayes optimality under the Gaussian homoscedastic assumption with $m \geq C - 1$, this does not hold true if $m < C - 1$. In fact, the obtained Fisher discriminant vectors will always bias towards the already well-separated class pairs in the original feature space while neglecting those neighboring class pairs that should indeed be much more separated in order to achieve better class discrimination. To elucidate this fact, let us revisit Eq. (2.37) as follows.

$$\begin{aligned}
 \max_{\mathbf{P}' \in \mathbf{R}^{n \times m}} J'_F(\mathbf{P}') &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} J'_F(\mathbf{P}') \\
 &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \text{trace}(\mathbf{P}'^T \tilde{\mathbf{S}}_B \mathbf{P}') \\
 &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \text{trace}((\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T) \quad (2.38) \\
 &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\hat{\sigma}_E^{ij})^2 \\
 &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\hat{\sigma}_M^{ij})^2
 \end{aligned}$$

where $\hat{\sigma}_E^{ij} = \sqrt{(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)}$ is the Euclidean distance between the mean vectors of two classes ω_i and ω_j in the transformed subspace, which is equivalent to the Mahalanobis distance $\hat{\sigma}_M^{ij} = \sqrt{(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\mathbf{S}}_W^{-1} (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)}$ between these two classes due to $\hat{\mathbf{S}}_W = \mathbf{P}'^T (\mathbf{S}_W^{-1/2})^T \mathbf{S}_W \mathbf{S}_W^{-1/2} \mathbf{P}' = \mathbf{I}_m$. It can be observed from Eq. (2.38) that maximization of the Fisher criterion is essentially equivalent to maximizing the average squared Mahalanobis distance over all available class pairs in the orthonormally transformed subspace, which is different from the goal of minimizing the Bayes error in the transformed subspace. Let us further examine Eq. (2.37) from another perspective.

$$\begin{aligned}
 \max_{\mathbf{P}' \in \mathbf{R}^{n \times m}} J'_F(\mathbf{P}') &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} J'_F(\mathbf{P}') \\
 &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \text{trace}(\mathbf{P}'^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \mathbf{P}') \quad (2.39) \\
 &= \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \sum_{k=1}^m \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\mathbf{p}'_k^T (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \mathbf{p}'_k)
 \end{aligned}$$

$$\text{with } \mathbf{P}' = [\mathbf{p}'_1, \dots, \mathbf{p}'_m]$$

It can be observed that to achieve the maximal Fisher criterion value, the direction of each Fisher discriminant vector $\mathbf{p}'_k, k=1, \dots, m$ extracted under the orthonormal

constraint will bias towards the vector difference of class means belonging to the well-separated class pairs with larger Euclidean distances in the $\mathbf{S}_w$ whitened space whilst diverging from those discriminative directions better separating neighboring classes. Specifically, the firstly extracted Fisher discriminant vector $\mathbf{p}'_1$ corresponding to the direction with the maximal Fisher criterion value will bias, to the largest extent, towards pairs of classes located remotely with each other. Each of the remaining Fisher discriminant vectors $\mathbf{p}'_k, k = 2, \dots, m$ is extracted subsequently from the subspace complementary orthogonal to the subspace spanned by the already extracted vectors, which still biases towards already well-separated class pairs. Since the Euclidean distance between mean vectors of a class pair in the $\mathbf{S}_w$ whitened space corresponds to the Mahalanobis distance of the corresponding class pair in the original feature space, class pairs with large Mahalanobis distances in the original feature space will dominate the eigenvalue decomposition. According to theorems 2.1 and 2.2, if $m \geq C - 1$, Mahalanobis distances between any two classes as well as the Bayes error in the transformed subspace remain same as those in the original feature space. Therefore, the last few Fisher discriminant vectors may possibly bias towards directions that can better discriminate neighboring classes, and thus minimize the classification error. However, if the dimension is reduced below $C-1$, these important directions will be thrown away due to their small contributions to the Fisher criterion value. As a result, the low classification accuracy may be obtained in the transformed subspace. In fact, according to theorem 2.1, under the Gaussian homoscedastic assumption with $m \geq C - 1$, maximization of the Fisher criterion can result in a transformed subspace, in which the Bayes errors of all available class pairs equal those in the original feature space. This constitutes a necessary condition to achieve the Bayes optimality in the transformed subspace. Theorem 2.2 actually demonstrates the sufficient condition is also satisfied under the same condition. However, when $m < C - 1$, even the necessary condition will be violated.

Such deficiency of the Fisher criterion may cause problems in some practical applications, where dimensionality of the pattern representation is expected to be as low as possible and often below $C-1$ in order to resolve the visualization,

computational complexity or estimation problems. For example, visualizing class structures underlying the available labeled feature vectors may facilitate better human-computer interaction, which, however, requires the original feature space be linearly transformed into a low-dimensional (usually 2D or 3D) subspace. As discussed above, the low-dimensional pattern representation obtained by the FLDA may possibly result in large overlaps of neighboring classes and thus prevent an effective illustration of class structures. Take another example, in speech recognition tasks, the computational complexity significantly depends on the feature space dimensionality. Although high-dimensional pattern representations can enable the acoustic model to carry more discriminatory information, the prohibitive computational cost may reduce the applicability of the whole system. Therefore, to make the system work normally, the DLDR should be executed in the front-end of the system to reduce the dimensionality of the original feature space. Since there usually exist hundreds of classes corresponding to different context-dependent/independent states, which may even exceed the original feature dimensionality, the low-dimensional subspace derived by the FLDA will not be the most discriminatory one due to the justifications explained above. The same situation may happen in other applications involving a fairly large number of classes such as image recognition or protein family classification. These practical difficulties motivate researchers to improve the conventional Fisher criterion.

Secondly, for any Fisher discriminant vector $\mathbf{p}_i \in \mathbf{R}^{n \times 1}, i = 1, \dots, m$ to be extracted, the idea behind the Fisher criterion requires that $\mathbf{p}_i^T \mathbf{S}_B \mathbf{p}_i$ can be as large as possible while simultaneously minimizing $\mathbf{p}_i^T \mathbf{S}_W \mathbf{p}_i$ under certain constraints of the relationship between $\mathbf{p}_i$ and those already extracted discriminant vectors. In general, the large value of $\mathbf{p}_i^T \mathbf{S}_B \mathbf{p}_i$ can indicate good class separability as long as the value of $\mathbf{p}_i^T \mathbf{S}_W \mathbf{p}_i$ is relatively smaller (but may be absolutely large) to avoid overlaps among classes. On the other hand, the small value of $\mathbf{p}_i^T \mathbf{S}_B \mathbf{p}_i$ may imply the existence of overlaps among some classes even if $\mathbf{p}_i^T \mathbf{S}_W \mathbf{p}_i$ is also rather small. However, the fairly small value of $\mathbf{p}_i^T \mathbf{S}_W \mathbf{p}_i$ at the denominator position in the Fisher criterion may undesirably magnify the small $\mathbf{p}_i^T \mathbf{S}_B \mathbf{p}_i$ value at the nominator position and consequently masks some

indeed more discriminative directions. We hereby use a simple 2D three-class example to illustrate this fact.

Example 2.1 Assume that three classes are normally distributed in the 2D space with equal class prior probabilities $P_1 = P_2 = P_3 = \frac{1}{3}$, equal diagonal class covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma_3 = \text{diag}([0.2, 0.005])$ and respective class mean vectors $\mathbf{m}_1 = [3, 0]$, $\mathbf{m}_2 = [-3, 0]$ and $\mathbf{m}_3 = [0, 1]$. A certain number of samples are randomly drawn from each class for an illustration in Figure 2.1.

Figure 2.1: Illustration of the 2D three classes with Gaussian distributions

Figure 2.2: One-dimensional class distributions after mapping the original feature space on p_1 (left) and p_2 (right), respectively.

Analysis: According to $\mathbf{S}_w^{-1}\mathbf{S}_B = \begin{bmatrix} 5, & 0 \\ 0, & 200 \end{bmatrix} \cdot \begin{bmatrix} 6, & 0 \\ 0, & 0.222 \end{bmatrix} = \begin{bmatrix} 30, & 0 \\ 0, & 44.4 \end{bmatrix}$, it is obvious to obtain the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_B$: $\mathbf{p}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\mathbf{p}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ that correspond to the eigenvalues $\lambda_1 = 44.4$ and $\lambda_2 = 30.0$. Figure 2.2 shows one-dimensional class distributions after mapping the original feature space on $\mathbf{p}_1$ and $\mathbf{p}_2$ respectively. In terms of the magnitude of eigenvalues, eigenvector $\mathbf{p}_1$ will be chosen as the first Fisher discriminant vector corresponding to the most discriminative direction. However, it is obvious from Figure 2.2 that all three classes are well separated on the direction of $\mathbf{p}_2$ while two classes are completely overlapped on the direction of $\mathbf{p}_1$. In fact, on the direction of $\mathbf{p}_2$ with the optimal class separability, the value of $\mathbf{p}_2^T \mathbf{S}_B \mathbf{p}_2 = 6$ is considerably larger than that of $\mathbf{p}_1^T \mathbf{S}_B \mathbf{p}_1 = 0.222$ on the direction of $\mathbf{p}_1$. However, $\mathbf{p}_1$ is wrongly amplified by a rather smaller value of $\mathbf{p}_1^T \mathbf{S}_w \mathbf{p}_1 = 0.005$ in comparison with $\mathbf{p}_2^T \mathbf{S}_w \mathbf{p}_2 = 0.2$ at the denominator position in the Fisher criterion. Therefore, the FLDA solution may contain some directions that lead to overlap among some classes and thus small between-class scatter values but are undesirably highlighted by comparatively much smaller average within-class scatter values at the denominator position in the Fisher criterion.

Thirdly, when a given problem satisfies the Gaussian homoscedastic assumption, the average within-class scatter matrix $\mathbf{S}_w$ will denote the common covariance matrix for all classes. However, in practice, individual classes usually have different covariance matrices. As a consequence, the pre-whitening step with respect to $\mathbf{S}_w$ can hardly whiten all the individual class covariance matrices simultaneously, and accordingly may be dominated by some classes with largely deviating covariance matrices. Therefore, the mean vector and spread of feature vectors within some other classes may be undesirably expanded or shrunk along directions preferred by the dominating covariance matrices. If those dominating covariance matrices coincidentally belong to the outlier classes locating remotely from the remaining classes, the ultimately

obtained transformation matrix may tend to minimize the spread of the outlier classes but bring some already adjacent classes much closer in the transformed subspace.

Fourthly, the Fisher criterion implicitly oversimplifies a given problem by assuming that all discriminatory information is contained in the difference of class means in the S_w whitened space and thus can be condensed into two scatter matrices, i.e. the between-class scatter matrix and average within-class scatter matrix while neglecting other important discriminatory information present in the difference of individual class distributions. Consequently, unless strictly satisfying the Gaussian homoscedastic assumption, the extracted Fisher discriminant vectors will inevitably discard certain amount of discriminatory information present beside the difference of class means. For example, when the underlying class distributions are Gaussian heteroscedastic, discriminant information present in the difference of the class covariance matrices will be totally ignored by the extracted Fisher discriminant vectors. As a result, the maximum number of effective Fisher discriminant vectors that can be extracted to optimize the Fisher criterion is bounded from above by $C-1$. This limitation becomes very apparent in the two-class case, in which only the reduction to a single dimension is possible.

Fifthly, the FLDA is not directly suitable for the multimodal class distributions, in which some classes cover multiple local regions located separately with each other. A typical solution is to resort to clustering algorithms to seek for the possibly existing local regions in each class. By assuming all the discovered local regions in each class as the unique “pseudo” classes, the FLDA can be executed. Some examples of such kind of methods are presented in [127], [128]. Due to its easy tractability, this deficiency will not be specifically addressed in this thesis.

Sixthly, minimization of the Fisher criterion will lead to an eigenvalue decomposition problem typically involving the inverse of some scatter matrix, which, however, will become singular when facing high-dimensional and undersampled problems. Consequently, the FLDA will become inapplicable in such case.

2.3.1.5 The subtraction Fisher criterion

The aforementioned second deficiency of the Fisher criterion originates from the quotient expression. The criterion function J_s , as discussed in section 2.2.2, can provide a reasonable solution. Specifically, by substituting $\mathbf{S}_1$ and $\mathbf{S}_2$ in J_s with $\mathbf{S}_B$ and $\mathbf{S}_W$, respectively, we can obtain the following Fisher criterion in the subtraction form:

$$J_{Fs}(\mathbf{P}) = \text{trace}(\mathbf{P}^T \mathbf{S}_B \mathbf{P}) - \mu \cdot \text{trace}(\mathbf{P}^T \mathbf{S}_W \mathbf{P}) \quad (2.40)$$

Since the quotient expression is no longer used, the undesirable amplification effect coming from the denominator naturally disappears. Therefore, the second deficiency associated with the conventional Fisher criterion in the quotient form can be eliminated. The criterion function $J_{Fs}(\mathbf{P})$ can be maximized with the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$) subject to the orthonormal constraint. The resulting optimal transformation matrix is composed of the first m leading eigenvectors of $\mathbf{S}_B - \mathbf{S}_W$ corresponding to the m largest eigenvalues. We name the FLDA based on the subtraction Fisher criterion FLDA_{s1}. It is worth noting that the optimal discriminant vectors extracted by the FLDA_{s1} are orthonormal to each other and generally cannot simultaneously diagonalize two matrices $\mathbf{S}_B$ and $\mathbf{S}_W$ in the transformed subspace. For clear notation, we hereafter use $J_{Fq}(\mathbf{P})$ to denote the conventional Fisher criterion in the quotient form as defined in Eq. (2.30).

The subtraction Fisher criterion has been recently studied in [129]-[131]. The encouraging performance on face recognition applications demonstrates its effectiveness. In [129], the maximum margin criterion was proposed, which is essentially the subtraction Fisher criterion but entitled with a different name. In our research work, we will simply take $\mu = 1$ as suggested in [129]. Recall Example 2.1, maximization of the subtraction Fisher criterion under the orthonormal constraint of the linear transformation matrix leads to determine the eigenvalue decomposition of

$\mathbf{S}_B - \mathbf{S}_W$. The eigenvectors of $\mathbf{S}_B - \mathbf{S}_W$ and their corresponding eigenvalues are $\mathbf{p}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\lambda_1 = 0.2172$ and $\mathbf{p}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\lambda_2 = 5.8000$, respectively. In terms of the magnitude of eigenvalues, the eigenvector $\mathbf{p}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ will be chosen as the first discriminant vector, which is actually the most discriminative direction.

The orthonormality of the extracted discriminant vectors with respect to $J_{Fs}(\mathbf{P})$ provides an efficient way to yield uncorrelated features in the transformed subspace. Specifically, the original feature space is first mapped into the $\mathbf{S}_T$ whitened space, and then the orthonormal discriminant vectors are extracted with respect to $J_{Fs}(\mathbf{P})$ in the $\mathbf{S}_T$ whitened space, which can keep $\mathbf{S}_T$ whitened and accordingly result in uncorrelated features with the unitary variances in the transformed subspace. Moreover, a set of orthogonal discriminant vectors can be obtained by orthogonalizing the already obtained set of discriminant vectors via the Gram-Schmidt orthogonalization or QR decomposition scheme. The uncorrelated FLDA and the uncorrelated-to-orthogonal FLDA based on the subtraction criterion are hereby named FLDA_{s2} and FLDA_{s3}, respectively, to facilitate future experimental comparisons.

2.3.2 Chernoff linear discriminant analysis (CLDA)

2.3.2.1 Related work

The intrinsic incapability of handling heteroscedastic data restricts the effectiveness of the FLDA in practical applications. In the past few decades, many methods have been specifically developed to resolve this problem. The BODLDR techniques devised under the Gaussian heteroscedastic assumption, as described in section 2.2.1, employ the criterion functions related to the Bayes error, which implicitly take into account the discriminatory information present in the difference of both class means and class covariance matrices. Due to the need of an iterative numerical optimization, solutions obtained by these techniques are not in the closed form and may even not be globally

optimal after the time-consuming search. Some approaches attempt to analytically derive optimal solutions as the FLDA does whilst extending the applicability of the FLDA into the heteroscedastic scenario. Analysis of the Bhattacharyya distance in the two-class case [2] indicated that the involved two terms denoting the Mahalanobis distance between two classes and the discrepancy of the two class covariance matrices respectively cannot be simultaneously optimized to yield a coherent linear transformation. Nevertheless, two optimal linear transformation matrices with regard to each of the two terms, which extract the discriminatory information present merely in the class means and merely in the class covariances respectively, can be analytically derived and then directly cascaded to establish the ultimate optimal linear transformation. Obviously, such solution is only a suboptimal one since the simple combination of the solutions obtained in the two extreme cases may definitely neglect the effect of the intermediate cases. A recent technique presented by Hsieh *et al.* [132] improve the above idea in the multi-class scenario, in which the discriminatory information present in the difference of class means and class covariance matrices are separately extracted by using the approximated pairwise accuracy criterion based FLDA and common-mean feature extraction techniques, respectively. Moreover, a spanning-tree classification accuracy estimation method is proposed to recursively substitute some previously extracted discriminant vector by the newly extracted one. Although this improved technique considers the interaction of the solutions obtained in the two extreme cases, it is achieved in the post-processing stage and thus may still lose some discriminatory information. Some other approaches [133]-[135] explicitly introduce the difference of class covariance matrices into the criterion functions without considering the link with the Bayes error. In these approaches, analytical solutions can be obtained in the fast and easy way as the FLDA does.

2.3.2.2 Recapitulation of the CLDA

Loog *et al.* [76], [77] proposed a novel DLDR technique via a heteroscedastic extension of the Fisher criterion to a so-called Chernoff criterion based on the Chernoff distance. Chernoff distance is a well-known probabilistic distance measuring the affinity of two class distributions, which is intimately linked with the upper bound

of the two-class Bayes error under the Gaussian heteroscedastic assumption. The key idea behind this extension is the use of a specific directed distance matrix (DDM) to generalize the between-class scatter matrix involved in the Fisher criterion. This proposed technique is hereafter named Chernoff linear discriminant analysis (CLDA). To elucidate the CLDA, let us revisit the between-class scatter matrix $\mathbf{S}_B^{ij}$ in the two-class case, which is the building block of the pairwise decomposition formulation of the between-class scatter matrix $\mathbf{S}_B$ in the multi-class scenario, i.e.

$$\mathbf{S}_B = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \mathbf{S}_B^{ij} \text{ with } \mathbf{S}_B^{ij} = (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T, i, j = 1, \dots, C, i \neq j \quad (2.41)$$

The matrix $\mathbf{S}_B^{ij}$ has only one nonzero eigenvalue equal to $(\mathbf{m}_i - \mathbf{m}_j)^T(\mathbf{m}_i - \mathbf{m}_j)$, with the associated eigenvector $\mathbf{m}_i - \mathbf{m}_j$. Note that this eigenvalue denotes the squared Euclidean distance $(\partial_{ij}^E)^2$ between the mean vectors of classes ω_i and ω_j , which equals the trace of $\mathbf{S}_B^{ij}$. Under the Gaussian homoscedastic assumption with the identity covariance matrices, the difference between the distributions of classes ω_i and ω_j can be characterized by the Euclidean distance between their class means, and $\mathbf{S}_B^{ij}$ can capture this Euclidean distance ∂_{ij}^E as well as the direction $\mathbf{m}_i - \mathbf{m}_j$ to attain this distance. Subject to the same assumption, any other directions apart from $\mathbf{m}_i - \mathbf{m}_j$ can only result in a complete overlap of classes ω_i and ω_j , and thus contain no discriminatory information. The two-class DDM aims to generalize $\mathbf{S}_B^{ij}$ so as to capture the extra discriminatory information present because of the heteroscedasticity of the two class distributions. The nonzero distances between two class distributions due to heteroscedasticity are generally in different directions from $\mathbf{m}_1 - \mathbf{m}_2$. Therefore, the two-class DDM can have more than one nonzero eigenvalues.

The two-class DDM serving as the building block of the Chernoff criterion is devised in terms of the Chernoff distance $\partial_{ij}^C(\alpha)$ (Eq. (2.15)) between two classes ω_i and ω_j under the Gaussian heteroscedastic assumption. Specifically, it is defined as a positive

semi-definite matrix $\mathbf{D}_C^{ij}$ with its trace equal to $\partial_{ij}^C(\alpha)$ multiplied by a constant as follows⁸ [77].

$$\mathbf{D}_C^{ij} = (\mathbf{S}_{ij})^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T (\mathbf{S}_{ij})^{-1/2} + \frac{1}{\alpha(1-\alpha)} (\log(\mathbf{S}_{ij}) - \alpha \log(\mathbf{S}_i) - (1-\alpha) \log(\mathbf{S}_j))$$

$$\text{trace}(\mathbf{D}_C^{ij}) = \frac{\alpha(1-\alpha)}{2} \partial_{ij}^C(\alpha) = (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{ij}^{-1} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{\alpha(1-\alpha)} \ln \left(\frac{|\mathbf{S}_{ij}|}{|\mathbf{S}_i|^\alpha |\mathbf{S}_j|^{1-\alpha}} \right)$$

with $\mathbf{S}_{ij} = \alpha \mathbf{S}_i + (1-\alpha) \mathbf{S}_j$ (2.42)

where $\alpha, 0 \leq \alpha \leq 1$ is a variable inherited from $\partial_{ij}^C(\alpha)$. The matrix $\mathbf{S}_{ij}$ is the weighted combination of the covariance matrices of classes ω_i and ω_j , and defined by $\alpha \mathbf{S}_i + (1-\alpha) \mathbf{S}_j$. The matrices $\mathbf{S}_i$ and $\mathbf{S}_j$ are the covariance matrices of classes ω_i and ω_j , respectively. To guarantee that the CLDA comes down to the FLDA in the Gaussian homoscedastic scenario, the variable α is suggested in [77] to take the value of $P_i/(P_i + P_j)$, the relative priori probability of class ω_i with respect to class ω_j , from which it directly follows that $1-\alpha$ equals $P_j/(P_i + P_j)$. In [77], the behavior of $\mathbf{D}_C^{ij}$ has been exemplified in two extreme cases and shown that: when a given problem satisfies the Gaussian homoscedastic assumption with the common covariance matrix being the identity matrix, the matrix $\mathbf{D}_C^{ij}$ comes down to $\mathbf{S}_B^{ij}$ and captures the difference of the class means; when a given problem satisfies the Gaussian heteroscedastic assumption with the common class mean and different diagonal covariance matrices, the matrix $\mathbf{D}_C^{ij}$ captures the difference of the class covariance matrices. In a certain weighted combination of both extreme cases, the matrix $\mathbf{S}_C^{ij}$ is expected to capture the difference of both class means and class covariance matrices in a somewhat balanced manner.

⁸ The function $f(\cdot)$, e.g. power or logarithm, of a symmetric positive definite matrix $\mathbf{A}$ can be defined in terms of its eigenvalue decomposition $\mathbf{A} = \mathbf{U} \text{diag}(\lambda_1, \dots, \lambda_n) \mathbf{U}^T$ as $f(\mathbf{A}) = \mathbf{U} \text{diag}(f(\lambda_1), \dots, f(\lambda_n)) \mathbf{U}^T$.

By replacing $\mathbf{S}_B^{ij}$ with $\mathbf{D}_C^{ij}$ in Eq. (2.41), the multi-class DDM in terms of the Chernoff distance, i.e. $\mathbf{D}_C$, can be expressed by:

$$\mathbf{D}_C = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j D_C^{ij} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \left((\mathbf{S}_{ij})^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T (\mathbf{S}_{ij})^{-1/2} + \frac{1}{\alpha(1-\alpha)} \left(\log(\mathbf{S}_{ij}) - \alpha \log(\mathbf{S}_i) - (1-\alpha) \log(\mathbf{S}_j) \right) \right) \quad (2.43)$$

with $\alpha = P_i / (P_i + P_j)$ and $1 - \alpha = P_j / (P_i + P_j)$

An implicit deficiency inherent in such multi-class DDM is that $\hat{\mathbf{D}}_C$ in the ultimately transformed subspace with respect to a linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$) cannot be expressed by $\hat{\mathbf{D}}_C = \mathbf{P}^T \mathbf{D}_C \mathbf{P}$ unless $\mathbf{P}$ is a unitary matrix, i.e. $\mathbf{P}^T \mathbf{P} = \mathbf{P} \mathbf{P}^T = \mathbf{I}_n$. In contrast, it always holds true that $\hat{\mathbf{S}}_B$ in the Fisher criterion can be expressed by $\hat{\mathbf{S}}_B = \mathbf{P}^T \mathbf{S}_B \mathbf{P}$ in the transformed subspace. This deficiency will prevent the direct substitution of $\mathbf{S}_B$ in the Fisher criterion (Eq. (2.30)) with $\mathbf{D}_C$ to construct the Chernoff criterion. Specifically, although the optimal solution maximizing the Fisher criterion can be derived by determining the eigenvalue decomposition of $\mathbf{S}_W^{-1} \mathbf{S}_B$, this procedure cannot be applied here since the linear transformation matrix $\mathbf{P}$ composed of even all the eigenvectors of $\mathbf{S}_W^{-1} \mathbf{D}_C$ cannot satisfy $\hat{\mathbf{D}}_C = \mathbf{P}^T \mathbf{D}_C \mathbf{P}$. Loog *et al.* [77] constructed the Chernoff criterion in a crafted way inspired by a two-stage method. In the two-stage method, the original feature space is firstly mapped into the $\mathbf{S}_W$ whitened space, and the eigenvalue decomposition of $\tilde{\mathbf{D}}_C$ estimated in the $\mathbf{S}_W$ whitened space is determined. Since $\tilde{\mathbf{D}}_C$ is a symmetric real matrix, the linear transformation matrix $\mathbf{P}'$ composed of all the eigenvectors of $\tilde{\mathbf{D}}_C$ can satisfy $\hat{\mathbf{D}}_C = \mathbf{P}'^T \tilde{\mathbf{D}}_C \mathbf{P}'$ in the ultimately transformed subspace. Consequently, the optimal linear transformation matrix can be obtained as $\mathbf{S}_W^{-1/2} \mathbf{P}'$. Specifically, Loog *et al.* [77] modified $\mathbf{D}_C$ to $\mathbf{S}_C$ as follows.

$$\begin{aligned}
 \mathbf{S}_C &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \mathbf{S}_W^{1/2} \tilde{\mathbf{D}}_C^{ij} \mathbf{S}_W^{1/2} \\
 &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \mathbf{S}_W^{1/2} \left(\left(\mathbf{S}_W^{-1/2} \mathbf{S}_{ij} \mathbf{S}_W^{-1/2} \right)^{-1/2} \mathbf{S}_W^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1/2} \left(\mathbf{S}_W^{-1/2} \mathbf{S}_{ij} \mathbf{S}_W^{-1/2} \right)^{-1/2} \right. \\
 &\quad \left. + \frac{1}{\alpha(1-\alpha)} \left(\log(\mathbf{S}_W^{-1/2} \mathbf{S}_{ij} \mathbf{S}_W^{-1/2}) - \alpha \log(\mathbf{S}_W^{-1/2} \mathbf{S}_i \mathbf{S}_W^{-1/2}) - (1-\alpha) \log(\mathbf{S}_W^{-1/2} \mathbf{S}_j \mathbf{S}_W^{-1/2}) \right) \right) \mathbf{S}_W^{1/2} \\
 &\text{with } \alpha = P_i / (P_i + P_j) \text{ and } 1 - \alpha = P_j / (P_i + P_j)
 \end{aligned} \tag{2.44}$$

where $\tilde{\mathbf{D}}_C^{ij}$ is the two-class DDM estimated in the in the $\mathbf{S}_W$ whitened space. According to the simultaneous diagonalization scheme, it is easy to verify that determining the eigenvalue decomposition of $\mathbf{S}_W^{-1} \mathbf{S}_C$ can result in the same solution as that obtained by the two-stage method. Accordingly, the Chernoff criterion was defined by:

$$J_C(\mathbf{P}) = \text{trace} \left((\mathbf{P}^T \mathbf{S}_W \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_C \mathbf{P} \right) \tag{2.45}$$

Maximization of this Chernoff criterion will lead to determine the eigenvalue decomposition of $\mathbf{S}_W^{-1} \mathbf{S}_C$.

2.3.2.3 Conjugated orthogonality constraints on CLDA

Analogous to the discussion in subsection 2.3.1.3, Chernoff discriminant vectors can be extracted subject to the constraints of conjugated orthogonality to the matrices $\mathbf{I}$, $\mathbf{S}_T$ and $\mathbf{S}_W$, respectively, to maximize the Chernoff criterion defined in Eq. (2.45). The resulting orthogonal CLDA (OCLDA), uncorrelated CLDA (UCLDA) and conventional CLDA (CLDA) techniques can be implemented in the same way as their Fisher criterion based counterparts: OFLDA, UCLDA and FLDA by merely substituting the Fisher criterion by the Chernoff criterion. Especially, the FLDA_{q1}, FLDA_{q2} and FLDA_{q3} techniques can be extended to the CLDA_{q1}, CLDA_{q2} and CLDA_{q3} techniques, which denote the orthogonal, uncorrelated, uncorrelated-to-orthogonal CLDA based on the quotient Chernoff criterion, respectively.

2.3.2.4 Pros and cons of the Chernoff criterion

As a specific heteroscedastic extension of the Fisher criterion, the Chernoff criterion can capture the discriminatory information present in the difference of both class means and class covariance matrices in a somewhat balanced manner, which enables the maximal number of effective discriminant vectors to exceed $C-1$ when dealing with the C -class problems. The effectiveness of the Chernoff criterion has been demonstrated on many real-world datasets in [76], [77]. However, some potential deficiencies that are either inherited from the Fisher criterion or uniquely associated with the Chernoff criterion may prevent it achieving better performance. As to be discussed in the following, the first deficiency is associated with the multi-class Chernoff criterion while the last two deficiencies are present in both the two-class and multi-class scenarios. Recall that C , n and m denote the class number, the original feature space dimensionality and the dimension to be reduced to, respectively.

Firstly, the Chernoff criterion inherits several deficiencies of the Fisher criterion as described in subsection 2.3.1.4. Specifically, it suffers from the problem that the already well-separated class pairs with larger Chernoff distances in the $\mathbf{S}_w$ whitened space may dominate the eigenvalue decomposition of $\tilde{\mathbf{D}}_C$, and consequently discriminative directions that can separate some neighboring class pairs with smaller Chernoff distances will be discarded. Moreover, Chernoff discriminant vectors may contain some directions that lead to the overlapping among some classes but are undesirably highlighted by the comparatively smaller average within-class variances along those directions. Furthermore, if whitening $\mathbf{S}_w$ is dominated by some existing outlier classes locating far from the remaining classes, the obtained optimal linear transformation may bias minimizing the spread of the outlier classes but bringing some already adjacent classes in the original feature space much closer in the transformed subspace.

Secondly, calculation of $\mathbf{S}_C$ in the Chernoff criterion may encounter singularity or numerical instability problems. Specifically, if the training sample size within class ω_i

is smaller than the feature space dimensionality, the within-class scatter matrix $\mathbf{S}_i$ will become singular or nearly singular, which disables the logarithm operation on $\mathbf{S}_i$ and accordingly the calculation of $\mathbf{S}_C$. As a result, even although $\mathbf{S}_W$ can be non-singular, the Chernoff criterion may still be inapplicable because of some ill-posed $\mathbf{S}_i$. To resolve this problem, the regularization schemes [136]-[142] can be employed. In our implementation, we take a simple regularization scheme, that is, we regularize $\mathbf{S}_i$ by adding to it an identity matrix multiplied with a constant equal to the one millionth of the trace of $\mathbf{S}_i$. Testing and comparing other regularization schemes, such as singular value perturbation [136], shrink towards the identity matrix [137], [138], multiple parameter regularization [141] and maximum entropy covariance selection [142], etc, is among our future plans. Furthermore, we use the singular value decomposition technique, instead of the eigenvalue decomposition, to stabilize the inversion and logarithm operations.

Thirdly, maximization of the Chernoff criterion (Eq. (2.45)) with respect to the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$) will lead to determine the eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_C$. When $m = n$, the obtained optimal linear transformation matrix $\mathbf{P}^*$ composed of all the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_C$ satisfies that $\hat{\mathbf{S}}_C = \mathbf{P}^{*T}\mathbf{S}_C\mathbf{P}^*$ as discussed previously. However, if $m < n$, the obtained optimal linear transformation matrix $\mathbf{P}^*$ composed of the first m leading eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_C$ corresponding to the largest m eigenvalues no longer satisfies $\hat{\mathbf{S}}_C = \mathbf{P}^{*T}\mathbf{S}_C\mathbf{P}^*$. Consequently, the trace of $\hat{\mathbf{S}}_C$ estimated in the transformed subspace with respect to $\mathbf{P}^*$ may not correspond to the maximal Chernoff criterion value equal to the summation of the m largest eigenvalues. However, intuitively, the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_C$ corresponding to smaller eigenvalues may contribute less to preserve the average pairwise Chernoff distance in the transformed subspace, and thus can be supposed to contain less discriminatory information. Although the CLDA technique bears such theoretical deficiency, its effectiveness has been verified by experiments [76], [77]. Therefore, strictly speaking,

the CLDA should be regarded as a heuristic technique instead of an analytical technique.

2.3.2.5 The subtraction Chernoff criterion

According to the subtraction Fisher criterion $J_{Fs}(\mathbf{P})$ defined in Eq. (2.40), we propose the subtraction Chernoff criterion $J_{Cs}(\mathbf{P})$ by substituting $\mathbf{S}_B$ with $\mathbf{D}_C$ as follows.

$$J_{Cs}(\mathbf{P}) = \text{trace}(\mathbf{P}^T \mathbf{D}_C \mathbf{P}) - \mu \cdot \text{trace}(\mathbf{P}^T \mathbf{S}_W \mathbf{P}) \quad (2.46)$$

The criterion $J_{Cs}(\mathbf{P})$ can be maximized with the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$) subject to the orthogonality constraint. The resulting optimal linear transformation matrix is composed of the first m leading eigenvectors of $\mathbf{D}_C - \mathbf{S}_W$ corresponding to the m largest eigenvalues. We name this technique CLDA_{s1}. It is worth noting that discriminant vectors obtained with respect to $J_{Cs}(\mathbf{P})$ are orthonormal to each other and generally cannot simultaneously diagonalize two matrices $\mathbf{D}_C$ and $\mathbf{S}_W$ in the projected subspace. By directly extending the FLDA_{s2} and FLDA_{s3} techniques by substituting the subtraction Fisher criterion with the subtraction Chernoff criterion, we can obtain the CLDA_{s2} and CLDA_{s3} techniques, which denote the uncorrelated and uncorrelated-to-orthogonal CLDA based on the subtraction Chernoff criterion, respectively.

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

3.1 Introduction

As described in the previous chapter, criterion functions of the BODLDR techniques are generally designed from the perspectives of statistics and information theory and can be related to the Bayes error under certain assumptions. On the other hand, criterion functions of the SMDLDR techniques are usually derived by intuitively evaluating the separation between classes without considering the link with the Bayes error. However, the analytical solutions can be computed efficiently by the SMDLDR via simple matrix algebras while the iterative numerical optimization is generally required in the BODLDR to derive the numerical solutions. Accordingly, the computational cost of the SMDLDR techniques is in general much lower than that of the BODLDR techniques. Moreover, in practice, due to the finite training sample size with the possible noise contamination, the mismatch between the actual and assumed class distributions and heuristic optimization schemes, the performance of the BODLDR techniques may be unsatisfactory. Noticing the pros and cons of BODLDR and SMDLDR, it is desirable to design some DLDR techniques that can unify the strengths of both BODLDR and SMDLDR such that the optimal linear transformations can be derived in a simple and efficient manner with respect to certain criterion functions related to the Bayes error.

In this chapter, we present several enhanced parametric BODLDR techniques along with some of their archetypes. A common characteristic of the parametric BODLDR techniques is that their criterion functions are related to the Bayes error to some extent under certain parametric assumptions of class distributions. Since the Bayes error or its upper bounds can be analytically formulated under specific Gaussian assumptions, two categories of parametric BODLDR techniques are studied under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. In each category, if

higher priority is given to the performance, certain criterion functions closely related to the Bayes error are directly minimized with respect to the linear transformation matrix via the iterative numerical optimization methods such as the gradient descent at the expense of the high computational cost. On the other hand, if higher priority is given to the efficiency, certain criterion functions somewhat related to the Bayes error will be used in the SMDLDR techniques (specifically, the FLDA and CLDA techniques) so as to derive closed-form solutions in an efficient manner at the expense of loss of the link with the Bayes error. Accordingly, we propose, separately under the Gaussian homoscedastic and heteroscedastic assumptions, two classes of enhanced parametric BODLDR techniques that can result in numerical and analytical solutions, respectively. To simplify analysis, we assume that the average within-class scatter matrix $\mathbf{S}_w$ is non-singular in the context of this chapter. The singularity problem [2], [70] will be specifically addressed in the next chapter.

The remainder of this chapter is organized as follows. Sections 3.2 describes the direct BODLDR techniques developed under the Gaussian homoscedastic assumption. A group of class-wise weighted FLDA techniques motivated under the Gaussian homoscedastic Bayes optimality is presented in section 3.3. Sections 3.4 describes the direct BODLDR techniques devised under the Gaussian heteroscedastic assumption. A group of class-wise weighted CLDA techniques motivated under the Gaussian heteroscedastic Bayes optimality is presented in section 3.5. Experimental results are discussed in section 3.6. Finally, conclusions are drawn in section 3.7.

3.2 Gaussian homoscedastic direct BODLDR

Under the Gaussian homoscedastic assumption of class distributions, the Bayes error between two classes ω_i and ω_j can be analytically formulated in Eq. (2.10). Accordingly, assume that the feature vectors belonging to two classes ω_i and ω_j in the original n -dimensional feature space are mapped into an m -dimensional subspace via the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$). Then, the Bayes error $\hat{P}_e(i, j)$ between classes ω_i and ω_j in the transformed subspace can be expressed by:

$$\hat{P}_e(i, j) = \frac{P_j}{\sqrt{2\pi}} \int_{-\infty}^{\hat{k}_1} e^{-\frac{t^2}{2}} dt + \frac{P_i}{\sqrt{2\pi}} \int_{\hat{k}_2}^{+\infty} e^{-\frac{t^2}{2}} dt \quad (3.1)$$

$$\text{with } \hat{k}_1 = \left(\ln \frac{P_i}{P_j} - \frac{1}{2} (\hat{\sigma}_M^{ij})^2 \right) / \hat{\sigma}_M^{ij}, \hat{k}_2 = \left(\ln \frac{P_i}{P_j} + \frac{1}{2} (\hat{\sigma}_M^{ij})^2 \right) / \hat{\sigma}_M^{ij}$$

$$\hat{\sigma}_M^{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P} (\mathbf{P}^T \mathbf{S} \mathbf{P})^{-1} \mathbf{P}^T (\mathbf{m}_i - \mathbf{m}_j)}$$

where $\mathbf{S}$ and $\hat{\sigma}_M^{ij}$ are the common class covariance matrix in the original feature space and the Mahalanobis distance between classes ω_i and ω_j in the transformed subspace, respectively. As demonstrated in [83], the Bayes error in the multi-class case can be bounded from above via the union of the pairwise Bayes errors over all class pairs. Therefore, it is reasonable to directly use this upper bound as the criterion function. Starting from a properly chosen initial linear transformation, numerical optimization methods can be applied to iteratively minimize such criterion function with respect to the linear transformation matrix.

3.2.1 Existing technique

Lotlikar and Kothari [91] presented an adaptive linear dimensionality reduction (ALDR) algorithm, which is subject to three assumptions: (1) Gaussian class distributions; (2) covariance matrices of all classes commonly equal to $\sigma^2 \mathbf{I}_n$; (3) equal class prior probabilities. The Bayes error in the multi-class case is approximated by the sum total of all pairwise two-class Bayes errors. This multi-class Bayes error approximation in the ultimate transformed subspace is used as the criterion function in the ALDR algorithm and defined by:

$$\sum_{i=1}^{C-1} \sum_{j=i+1}^C \frac{1}{\sqrt{2\pi}\sigma} \int_{\hat{\sigma}_E^{ij}/2}^{+\infty} \exp(-t^2/2\sigma^2) dt \quad \text{with } \hat{\sigma}_E^{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P} \mathbf{P}^T (\mathbf{m}_i - \mathbf{m}_j)} \quad (3.2)$$

where $\hat{\sigma}_E^{ij}$ and $\mathbf{P}$ denote the Euclidean distance between the mean vectors of classes ω_i and ω_j in the transformed subspace and the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$

($m \leq n$) subject to the orthonormal constraint, respectively. Note that such orthonormal constraint can avoid the degenerated linear transformation. The gradient descent scheme is applied to minimize this criterion function with respect to $\mathbf{P}$ with the following updating rule:

$$\mathbf{P}_{new} = \mathbf{P}_{old} + \eta \left(\frac{1}{2\sqrt{2\pi}\sigma} \right) \sum_{i=1}^{C-1} \sum_{j=i+1}^C \frac{1}{\hat{\sigma}_E^{ij}} \exp \left(-(\hat{\sigma}_E^{ij})^2 / 8\sigma^2 \right) (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P}_{old} \quad (3.3)$$

$$\hat{\sigma}_E^{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P}_{old} \mathbf{P}_{old}^T (\mathbf{m}_i - \mathbf{m}_j)}$$

where η , $\mathbf{P}_{old}$ and $\mathbf{P}_{new}$ denote the learning step size, the transformation matrix before and after updating, respectively. To satisfy the orthonormal constraint on the linear transformation matrix, the Gram-Schmidt orthogonalization scheme [92] will be applied to the transformation matrix whenever it is updated. In the implementation of the ALDR, the average within-class scatter matrix $\mathbf{S}_w$ will be pre-whitened before performing the optimization process in order to satisfy the second assumption and accordingly it simplifies σ to equal “1”. Therefore, the discriminant vectors extracted by the ALDR algorithm is actually subject to the conjugated orthogonality constraint to $\mathbf{S}_w$. The optimization process performed in the $\mathbf{S}_w$ -whitened subspace is initialized with a randomly generated orthonormal matrix.

3.2.2 Proposed technique

According to Eq. (3.1), we propose a Gaussian homoscedastic direct BODLDR (DBODLDRho) technique, in which the criterion function is formulated as the union of the pairwise Bayes errors in the transformed subspace over all class pairs under the Gaussian homoscedastic assumption and expressed by:

$$\sum_{i=1}^{C-1} \sum_{j=i+1}^C \hat{P}_e(i, j) \text{ with } \hat{P}_e(i, j) \text{ defined by Eq. (3.1)} \quad (3.4)$$

The gradient descent scheme is applied to minimize this criterion function with respect to the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$) subject to the orthonormal constraint with the following updating rule:

$$\begin{aligned} \mathbf{P}_{new} &= \mathbf{P}_{old} + \frac{\eta}{\sqrt{2\pi}} \sum_{i=1}^{C-1} \sum_{j=i+1}^C \left(P_j \exp\left(-\frac{\hat{k}_i^2}{2}\right) \left(0.5 \ln\left(\frac{P_i}{P_j}\right) (\hat{\sigma}_M^{ij})^{-3} + 0.25 (\hat{\sigma}_M^{ij})^{-1} \right) \right. \\ &\quad \left. + P_i \exp\left(-\frac{\hat{k}_j^2}{2}\right) \left(-0.5 \ln\left(\frac{P_i}{P_j}\right) (\hat{\sigma}_M^{ij})^{-3} + 0.25 (\hat{\sigma}_M^{ij})^{-1} \right) \right) \mathbf{A}_{ij} \\ \mathbf{A}_{ij} &= 2 \left((\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P}_{old} \right. \\ &\quad \left. - \mathbf{S}_W \mathbf{P}_{old} (\mathbf{P}_{old}^T \mathbf{S}_W \mathbf{P}_{old})^{-1} \mathbf{P}_{old}^T (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P}_{old} \right) (\mathbf{P}_{old}^T \mathbf{S}_W \mathbf{P}_{old})^{-1} \\ \hat{\sigma}_M^{ij} &= \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P}_{old} (\mathbf{P}_{old}^T \mathbf{S}_W \mathbf{P}_{old})^{-1} \mathbf{P}_{old}^T (\mathbf{m}_i - \mathbf{m}_j)} \end{aligned} \quad (3.5)$$

where η , $\mathbf{P}_{old}$ and $\mathbf{P}_{new}$ denote the learning step size, the transformation matrix before and after updating, respectively. To satisfy the orthonormal constraint of the linear transformation matrix, the Gram-Schmidt orthogonalization scheme will be applied to the transformation matrix whenever it is updated. Moreover, an optimal solution obtained by the CWFLDA_{q0} technique to be described later, after experiencing the Gram-Schmidt orthogonalization, will be used to initialize our proposed DBODLDRho method.

In comparison with the ALDR, our proposed DBODLDRho method uses a more general formula to approximate the multi-class Bayes error via the union of all pairwise Bayes errors calculated under the Gaussian homoscedastic assumption. The DBODLDRho does not pre-whiten the average within-class scatter matrix before performing the optimization process, i.e. relaxing the second assumption in the ALDR. Therefore, the extracted discriminant vectors are subject to the conjugated orthogonality constraint to $\mathbf{I}$, that is, they are orthogonal to each other. Moreover, the DBODLDRho relaxes the third assumption of equal class prior probabilities in the ALDR algorithm. Furthermore, a carefully chosen initial solution is used to reduce the risk of getting struck into local minima during the optimization.

3.3 Class-wise weighted FLDA motivated by Gaussian homoscedastic Bayes optimality

Due to direct minimization of the Bayes error related criterion function, the optimal linear transformation obtained by the DBODLDRho method can result in a transformed subspace approximating the Bayes optimality when the given problem satisfies the Gaussian homoscedastic assumption. However, the numerical optimization methods tend to be very time-consuming. Moreover, the finally obtained optimal linear transformation is sensitive to initializations and learning parameter settings and thus may possibly get stuck into inferior local optima. In contrast, the SMDLDR techniques can efficiently derive globally optimal solutions in the closed form. However, criterion functions of the SMDLDR techniques generally lack the link with the Bayes error. Consequently, the obtained optimal linear transformation may result in a transformed subspace significantly deviating from the Bayes optimality. As described in section 2.3, the FLDA is an interaction of the BODLDR and SMDLDR techniques, which can yield the globally optimal solution in the closed form with respect to the Fisher criterion. As demonstrated in Theorem 2.2, the optimal linear transformation obtained by the FLDA can actually result in a transformed subspace achieving the Bayes optimality under certain conditions. However, the potential deficiencies of the conventional Fisher criterion, as analyzed in subsection 2.3.1.4, may destroy the link between the solutions of the FLDA and the Bayes error in general sense, e.g. when the dimension to be reduced to is below $C-1$. In this subsection, we propose two class-wise weighted Fisher criteria in the quotient and subtraction forms, respectively. The resulting class of DLDR techniques named class-wise weighted FLDA (CWFLDA) attempt to compensate the first three deficiencies associated with the conventional FLDA.

3.3.1 Existing techniques

Loog *et al.* [143], [144] reformulated the conventional Fisher criterion into a pairwise decomposition form and unveiled that the correspondingly obtained optimal linear transformation will prefer preserving the larger distances of the already well-separated class pairs (especially those involving the outlier classes) since they can contribute

more to the maximization of the Fisher criterion value, and neglect those neighboring class pairs that should indeed be much more separated in order to achieve better class discrimination. To address this problem, they introduced specific weighting factors into the pairwise between-class scatter matrix such that the contributions from individual class pairs to the maximization of the Fisher criterion rely on the two-class Bayes accuracies of those class pair instead of on the distances of those class pair. The resulting weighted pairwise Fisher criterion, also named Approximate Pairwise Accuracy Criterion (aPAC) [145], attempts to approximate the average pairwise two-class Bayes accuracy over all class pairs in the transformed subspace. The FLDA technique based on the aPAC is called aPAC-FLDA.

The weighted pairwise between-class scatter matrix $\mathbf{S}_{Bw}$ and the aPAC $J_{aPAC}(\mathbf{P})$ are expressed by:

$$\mathbf{S}_{Bw} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \mathbf{S}_{Bw}^{ij} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j w_{ij} (\partial_M^{ij}) (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \quad (3.6)$$

$$\text{with } w_{ij}(\partial_M^{ij}) = \frac{1}{(\partial_M^{ij})^2} \left(\frac{1}{\sqrt{2\pi}} \int_0^{\partial_M^{ij}/2} \exp\left(-\frac{t^2}{2}\right) dt \right) \text{ and } \partial_M^{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j)}$$

$$J_{aPAC}(\mathbf{P}) = \text{trace} \left((\mathbf{P}^T \mathbf{S}_W \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_{Bw} \mathbf{P} \right) \quad (3.7)$$

where $\mathbf{S}_{Bw}^{ij}$ and ∂_M^{ij} are the weighted between-class scatter matrix and the Mahalanobis distance for two classes ω_i and ω_j , respectively. The weighting factor $w_{ij}(\partial_M^{ij})$, as a function of the Mahalanobis distance ∂_M^{ij} , can be interpreted as the ratio of the two-class Bayes accuracy between classes ω_i and ω_j estimated under the Gaussian homoscedastic assumption with equal class prior probabilities in the original feature space, to the squared Mahalanobis distance between these two classes calculated in the original feature space, i.e.

$$w_{ij}(\partial_M^{ij}) = \frac{0.5 - P_e^{ij}}{(\partial_M^{ij})^2}, \quad P_e^{ij} = \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^{\partial_M^{ij}/2} \exp\left(-\frac{t^2}{2}\right) dt \quad (3.8)$$

Therefore, the adjacent class pairs with smaller Mahalanobis distances can receive larger weights than the distant class pairs with larger Mahalanobis distances. Maximizing the aPAC with respect to the linear transformation will lead to determine the eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_{Bw}$, which is equivalent to determining the eigenvalue decomposition of $\tilde{\mathbf{S}}_{Bw} = \mathbf{S}_W^{-1/2}\mathbf{S}_{Bw}\mathbf{S}_W^{-1/2}$. By closely examining $\tilde{\mathbf{S}}_{Bw}$, we can obtain:

$$\begin{aligned}\tilde{\mathbf{S}}_{Bw} &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j w_{ij} (\partial_M^{ij}) \mathbf{S}_W^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1/2} \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \frac{(0.5 - P_e^{ij})}{(\tilde{\partial}_E^{ij})^2} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \\ \text{with } \tilde{\partial}_E^{ij} &= \sqrt{(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j)} = \partial_M^{ij}\end{aligned}\quad (3.9)$$

It can be observed from Eq. (3.9) that in $\tilde{\mathbf{S}}_{Bw}$ the Euclidean distance metric associated with each pairwise scatter matrix $(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j) (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T$ is replaced by the two-class Bayes accuracy of the corresponding class pair.

Analogous to Eqs. (2.36) and (2.37), we can obtain:

$$\max_{\mathbf{P} \in \mathbf{R}^{n \times m}} J_{aPAC}(\mathbf{P}) = \max_{\mathbf{P} \in \mathbf{R}^{n \times m}} \text{trace}((\mathbf{P}^T \mathbf{S}_W \mathbf{P})^{-1} \mathbf{P}^T \mathbf{S}_{Bw} \mathbf{P}) = \max_{\substack{\mathbf{P}' \in \mathbf{R}^{n \times m} \\ \mathbf{P}'^T \mathbf{P}' = \mathbf{I}_m}} \text{trace}(\mathbf{P}'^T \tilde{\mathbf{S}}_{Bw} \mathbf{P}') \quad (3.10)$$

Note that to make Eq. (3.10) equivalent to maximization of the average pairwise two-class Bayes accuracy over all class pairs in the transformed subspace with respect to the transformation matrix subject to the orthonormal constraints, the weighting factors should be calculated in the ultimately transformed subspace instead of in the original feature space. However, such implementation will unfortunately prevent the acquirement of analytical solutions. Therefore, the aPAC implicitly assumes that pairwise two-class Bayes accuracies for all class pairs in the original feature space do not change after the linear transformation. Accordingly, the weighting factors

calculated in the original feature space can be used to approximate their values in the transformed subspace. Although this assumption can be seldom satisfied in practice, the encouraging empirical performance of the aPAC-FLDA has validated its practical effectiveness [143], [144]. Some alternative weighting factors, such as unit factor “1”, negative power of the pairwise Euclidean distance and the pairwise Kullback-Leibler distance, were suggested by Li *et al.* in [146], which however lack the link with the Bayes error.

Lotlikar and Kothari [147] developed the Fractional-step Fisher Linear Discriminant Analysis (FS-FLDA) technique, which incrementally reduces the dimensionality of the feature space in fractional steps. The FS-FLDA uses the weighted pairwise Fisher criterion, in which the weighting factor of each class pair is defined as the negative power function of the Mahalanobis distance between the corresponding two classes. Accordingly, neighboring class pairs will be highlighted with large weights. This is opposite to the conventional Fisher criterion where faraway class pairs will dominate. Therefore, maximization of such weighted pairwise Fisher criterion may result in an optimal linear transformation that can lead to better separation of the originally neighboring class pairs in the transformed subspace. In the FS-FLDA, the training samples are gradually squeezed by a small amount along the direction of the Fisher discriminant vector corresponding to the smallest Fisher criterion value, i.e. the least discriminative direction. In each incremental step, the weighting factors as well as the weighted pairwise between-class scatter matrix will be re-computed in the current feature space. After the pre-specified r steps, the least discriminative direction is removed, which results in the reduction of the dimensionality by “1”. The dimensionality of the feature space is repeatedly reduced in the same manner until a pre-defined dimension is achieved.

Compared with the aPAC-FLDA, the FS-FLDA is more time-consuming because of the iterative reduction. However, in each fractional step, the FS-FLDA can highlight the worst-separated class pairs in the current feature space, and then adjust the linear transformation to better separate those currently much confused class pairs. Consequently, it may reduce the mismatch between the class separability in the original

feature space with that in the transformed subspace. Such mismatch problem is implicitly involved in the non-iterative aPAC-FLDA technique, in which the weighting factors computed in the original feature space are used as the surrogates of those calculated in the transformed subspace. However, since some well-separated class pairs in the original feature space may possibly overlap or even occlude after the linear transformation, the surrogates may wrongly estimate the actual weighting factors in the transformed subspace. The detailed implementation of the FS-FLDA can be referred to [147].

3.3.2 Proposed techniques

The weighted pairwise between-class scatter matrix replaces the distance metric associated with each component pairwise scatter matrix by some special weighting factors such that the resulting weighted pairwise Fisher criterion can be somewhat related to the Bayes error. For example, the aPAC substitutes the distance metric associated with a class pair with the two-class Bayes accuracy of the corresponding two classes estimated under the Gaussian homoscedastic assumption with equal class prior probabilities. Different from the distance metric, the Bayes accuracy is bounded from above by a constant. Accordingly, if two well-separated class pairs both achieve the maximal Bayes accuracy but have the different between-class distances, they will be equivalently treated in the aPAC instead of biasing towards the class pair with larger between-class distance as in the conventional Fisher criterion. Consequently the class pairs with larger between-class distance will not dominate the eigenvalue decomposition. This can effectively suppress the influence from the outlier classes in the optimization process. However, the weighted pairwise Fisher criterion can merely compensate the first deficiency of the conventional Fisher criterion and leaves the remaining deficiencies unsolved. This fact motivates us to further improve the weighted pairwise Fisher criterion so as to compensate more deficiencies inherent in the conventional Fisher criterion. We therefore devise two class-wise weighted Fisher criteria in the quotient and subtraction forms respectively, which can tackle the first third deficiencies associated with the conventional Fisher criterion. Our newly proposed criteria are defined as follows.

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Quotient class-wise weighted Fisher criterion:

$$J_{cwFq}(\mathbf{P}) = \text{trace}\left(\left(\mathbf{P}^T \mathbf{S}_{Ww} \mathbf{P}\right)^{-1} \mathbf{P}^T \mathbf{S}_{Bw} \mathbf{P}\right) \quad (3.11)$$

Subtraction class-wise weighted Fisher criterion:

$$J_{cwFs}(\mathbf{P}) = \text{trace}(\mathbf{P}^T \mathbf{S}_{Bw} \mathbf{P}) - \text{trace}(\mathbf{P}^T \mathbf{S}_{Ww} \mathbf{P}) \quad (3.12)$$

Two key matrices $\mathbf{S}_{Ww}$ and $\mathbf{S}_{Bw}$ involved in the above criteria are calculated orderly as follows.

$$\begin{aligned} \mathbf{S}_W &= \sum_{i=1}^C P_i \mathbf{S}_i \\ \partial_M^{ij} &= \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j)} \\ k_i &= \left(\ln \frac{P_i}{P_j} - \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}, \quad k_j = \left(\ln \frac{P_i}{P_j} + \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}, \\ P_e^{ij} &= \frac{P_j}{(P_i + P_j)\sqrt{2\pi}} \int_{-\infty}^{k_1} e^{-\frac{t^2}{2}} dt + \frac{P_i}{(P_i + P_j)\sqrt{2\pi}} \int_{k_2}^{+\infty} e^{-\frac{t^2}{2}} dt \quad \text{and} \quad tw_{ij}^B = \frac{0.5 - P_e^{ij}}{(\partial_M^{ij})^2} \\ w_i^W &= \frac{\sum_{\substack{j=1 \\ j \neq i}}^C tw_{ij}^B}{\max_{1 \leq k \leq C} \left(\sum_{\substack{j=1 \\ j \neq k}}^C tw_{kj}^B \right)} \\ \boxed{\mathbf{S}_{Ww} = \sum_{i=1}^C P_i w_i^W \mathbf{S}_i} \end{aligned} \quad (3.13)$$

$$\begin{aligned} \partial_M^{ij} &= \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{Ww}^{-1} (\mathbf{m}_i - \mathbf{m}_j)} \\ \boxed{\text{For } J_{cwFq}(\mathbf{P}):} \quad k_i &= \left(\ln \frac{P_i}{P_j} - \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}, \quad k_j = \left(\ln \frac{P_i}{P_j} + \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}, \\ P_e^{ij} &= \frac{P_j}{(P_i + P_j)\sqrt{2\pi}} \int_{-\infty}^{k_1} e^{-\frac{t^2}{2}} dt + \frac{P_i}{(P_i + P_j)\sqrt{2\pi}} \int_{k_2}^{+\infty} e^{-\frac{t^2}{2}} dt \quad \text{and} \quad w_{ij}^B = \frac{0.5 - P_e^{ij}}{(\partial_M^{ij})^2} \end{aligned}$$

$$\begin{aligned}
 & \boxed{\text{For } J_{cwFs}(\mathbf{P}): k_i = \left(\ln \frac{P_i}{P_j} - \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij}, k_j = \left(\ln \frac{P_i}{P_j} + \frac{1}{2} (\partial_M^{ij})^2 \right) / \partial_M^{ij},} \\
 & P_e^{ij} = \frac{P_j}{(P_i + P_j)\sqrt{2\pi}} \int_{-\infty}^{k_1} e^{-\frac{t^2}{2}} dt + \frac{P_i}{(P_i + P_j)\sqrt{2\pi}} \int_{k_2}^{+\infty} e^{-\frac{t^2}{2}} dt \text{ and } w_{ij}^B = \frac{0.5 - P_e^{ij}}{(\partial_E^{ij})^2} \\
 & \boxed{\mathbf{S}_{Bw} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j w_{ij}^B (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T} \tag{3.14}
 \end{aligned}$$

Here, $\mathbf{S}_{Bw}$, w_{ij}^B , $\mathbf{S}_{Ww}$ and w_i^W denote weighted pairwise between-class scatter matrix, between-class weight, average weighted within-class scatter matrix and within-class weight, respectively. In our proposed criteria, a set of temporary between-class weights tw_{ij}^B is first computed in terms of the conventional average within-class scatter matrix $\mathbf{S}_w$. Then, the within-class weights w_i^W are calculated based on tw_{ij}^B . Subsequently, the average weighted within-class scatter matrix $\mathbf{S}_{Ww}$ is established. Finally, the between-class weights w_{ij}^B are computed in terms of $\mathbf{S}_{Ww}$, which lead to the weighted pairwise between-class scatter matrix $\mathbf{S}_{Bw}$. There exists a major distinction between $J_{cwFq}(\mathbf{P})$ and $J_{cwFs}(\mathbf{P})$. For $J_{cwFq}(\mathbf{P})$, according to the simultaneous diagonalization scheme, the minimization is performed in two stages, in which the original feature space is first mapped into the $\mathbf{S}_{Ww}$ whitened space and then the eigenvalue decomposition of $\tilde{\mathbf{S}}_{Bw} = \mathbf{S}_{Ww}^{-1/2} \mathbf{S}_{Bw} \mathbf{S}_{Ww}^{-1/2}$ is determined. Accordingly, the distance metric associated with each component pairwise between-class scatter matrix in $\tilde{\mathbf{S}}_{Bw}$ is the Euclidean distance in the $\mathbf{S}_{Ww}$ whitened space, which is equivalent to the Mahalanobis distance in the original feature space. This explains why in $J_{cwFq}(\mathbf{P})$ the denominator of w_{ij}^B employs the squared Mahalanobis distance $(\partial_M^{ij})^2$. However, for $J_{cwFs}(\mathbf{P})$, the minimization is performed in a single stage, in which the eigenvalue decomposition of $\mathbf{S}_{Bw} - \mathbf{S}_{Ww}$ is determined. Accordingly, the distance metric associated with each component pairwise between-class scatter matrix in $\mathbf{S}_{Bw}$ should

be the Euclidean distance in the original feature space. This accounts for why in $J_{cwFs}(\mathbf{P})$ the denominator of w_{ij}^B employs the squared Euclidean distance $(\mathcal{D}_E^{ij})^2$.

Attractive properties of the proposed two class-wise weighted Fisher criteria in Eqs. (3.11) and (3.12) are highlighted as follows.

1. The quotient class-wise weighted Fisher criterion $J_{cwFq}(\mathbf{P})$ generalizes $J_{aPAC}(\mathbf{P})$ in two aspects: Firstly, under the Gaussian homoscedastic assumption, $J_{cwFq}(\mathbf{P})$ relaxes the assumption of equal class prior probabilities associated with $J_{aPAC}(\mathbf{P})$. Secondly, $J_{cwFq}(\mathbf{P})$ introduces the within-class weight w_i^W to adjust the contribution from an individual class covariance matrix to the average within-class scatter matrix according to its confusability with the other classes. Therefore, the outlier classes with little confusability with the remaining classes will be assigned with the very small within-class weights, and accordingly their influences on the eigenvalue decomposition of the average within-class scatter matrix will be much suppressed. Consequently, the third deficiency associated with the conventional Fisher criterion can be compensated. Since the temporary between-class weight tw_{ij}^B can evaluate the confusability of classes ω_i and ω_j , it is reasonable to use the summation of all temporary between-class weights involving a certain class to represent the confusability of such class with the other classes, which results in the definition of w_i^W . Obviously, when the within-class weights for all classes take values “1” and all classes have the same prior probability, $J_{cwFq}(\mathbf{P})$ comes down to $J_{aPAC}(\mathbf{P})$.
2. Besides the conventional Fisher criterion formulated in the Raleigh quotient form, the Fisher criterion in the subtraction form was recently studied in [129]-[131] with the concept of maximum margin criterion. Since the quotient expression is no longer used, the undesirable amplification effect coming from the denominator naturally disappears. Therefore, the second deficiency associated with the conventional quotient Fisher criterion can be compensated. Our proposed

subtraction class-wise weighted Fisher criterion $J_{cwFs}(\mathbf{P})$ aims at enhancing the conventional subtraction Fisher criterion by incorporating the weighted pairwise between-class scatter matrix and the average weighted within-class scatter matrix. One attractive property of $J_{cwFs}(\mathbf{P})$ is that it can yield the optimal orthonormal linear transformation matrix due to the symmetric property of $\mathbf{S}_{Bw} - \mathbf{S}_{Ww}$. Consequently, the uncorrelated features can be extracted in the batch manner from the $\mathbf{S}_T$ whitened space. In contrast, since the equality $\mathbf{S}_T = \mathbf{S}_{Bw} + \mathbf{S}_{Ww}$ no longer holds true, simultaneous diagonalization of $\mathbf{S}_{Bw}$ and $\mathbf{S}_{Ww}$ cannot result in the diagonalized $\mathbf{S}_T$ any more. Therefore, the uncorrelated features cannot be extracted in the batch manner with respect to $J_{cwFq}(\mathbf{P})$.

As discussed in subsection 2.3.1.3, discriminant vectors subject to three different conjugated orthogonality constraints may facilitate the different subsequent tasks. Therefore, it is desirable to examine the effectiveness of the discriminant vectors extracted subject to the conjugated orthogonality to matrices $\mathbf{I}$, $\mathbf{S}_T$ and $\mathbf{S}_{Ww}$ with respect to the newly proposed class-wise weighted Fisher criteria, respectively. According to the way of extracting the discriminant vectors, there exist the batch and sequential manners, among which the sequential manner can be further dichotomized into the sequential forward extraction and sequential backward reduction. In the following, we present several CWFLDA techniques in terms of $J_{cwFq}(\mathbf{P})$ and $J_{cwFs}(\mathbf{P})$, respectively.

- $J_{cwFq}(\mathbf{P})$ based CWFLDA techniques

1. The CWFLDA_{q0} technique (batch)

By substituting $J_{aPAC}(\mathbf{P})$ used in the aPAC-FLDA with $J_{cwFq}(\mathbf{P})$, we obtain the CWFLDA_{q0} technique, in which a set of discriminant vectors subject to the

conjugated orthogonality to the matrix $\mathbf{S}_{Ww}$ can be extracted in the batch manner to maximize $J_{cwFq}(\mathbf{P})$ by determining the eigenvalue decomposition of $\mathbf{S}_{Ww}^{-1}\mathbf{S}_{Bw}$.

2. The F-CWFLDA_{q0} technique (sequential backward reduction)

The CWFLDA_{q0} inherits the class separability mismatch problem associated with the aPAC-FLDA since the between-class and within-class weights computed in the original feature space are directly used as the surrogates of those calculated in the transformed subspace in $J_{cwFq}(\mathbf{P})$. The FS-FLDA provides a solution to relieve such problem by fractionally reducing the dimensionality of the feature space and re-evaluating the weights in each fraction step. However, as analyzed previously, the FS-FLDA can be very time-consuming since many fractional steps are required when reducing a single dimension. Moreover, the weighted pairwise Fisher criterion used in the FS-FLDA does not link with the Bayes error. To address these problems, we extend the CWFLDA_{q0} into the fractional CWFLDA_{q0} technique, named F-CWFLDA_{q0}, in which the feature space dimensionality is reduced one by one and the between-class and within-class weights are re-calculated after each reduction step. In comparison with the FS-FLDA technique, the F-CWFLDA_{q0} only needs one step to reduce a single dimension and thus saves the computational cost. Furthermore, the quotient class-wise weighted Fisher criterion used in the F-CWFLDA_{q0} is somewhat related to the Bayes error. The F-CWFLDA_{q0} reduces the dimensionality of the feature space in the sequential backward manner. In each current feature space, a set of discriminant vectors subject to the conjugated orthogonality to the matrix $\mathbf{S}_{Ww}$ are extracted in the batch manner to maximize the updated $J_{cwFq}(\mathbf{P})$. It is worth noting that, in the sequential backward reduction, the set of discriminant vectors extracted to span a higher-dimensional transformed subspace may not contain the set of discriminant vectors extracted to span a lower-dimensional transformed subspace.

3. The CWFLDA_{q1} technique (sequential forward extraction)

By substituting $J_{Fq}(\mathbf{P})$ used in the FLDA_{q1} with $J_{cwFq}(\mathbf{P})$, we obtain the CWFLDA_{q1} technique, in which the discriminant vectors subject to the conjugated orthogonality to the matrix $\mathbf{I}$ can be extracted in the sequential forward manner to maximize the quotient class-wise weighted Fisher criterion. It is worth noting that, in the sequential forward extraction, the set of discriminant vectors extracted to span a higher-dimensional transformed subspace must contain the set of discriminant vectors extracted to span a lower-dimensional transformed subspace.

4. The CWFLDA_{q2} technique (sequential forward extraction)

By substituting $J_{Fq}(\mathbf{P})$ used in the FLDA_{q2} with $J_{cwFq}(\mathbf{P})$, we obtain the CWFLDA_{q2} technique, in which the discriminant vectors subject to the conjugated orthogonality to the matrix $\mathbf{S}_T$ can be extracted in the sequential forward manner to maximize the quotient class-wise weighted Fisher criterion, with the features in the transformed subspace uncorrelated to each other. The extracted discriminant vectors will be further normalized to whiten $\mathbf{S}_T$.

5. The CWFLDA_{q3} technique (sequential forward extraction)

By applying the QR decomposition to the optimal linear transformation matrix obtained by the CWFLDA_{q3}, we obtain the CWFLDA_{q3} technique, in which a set of orthonormal discriminant vectors can be extracted based on the already extracted discriminant vectors by applying the CWFLDA_{q2}. Such uncorrelated-to-orthogonal transformation may possibly suppress noise [124].

- $J_{cwFs}(\mathbf{P})$ based CWFLDA techniques

In general, maximization of $J_{cwFs}(\mathbf{P})$ cannot result in an optimal linear transformation that simultaneously diagonalizes matrices $\mathbf{S}_{Bw}$ and $\mathbf{S}_{Ww}$ in the transformed subspace. If $J_{cwFs}(\mathbf{P})$ is maximized with respect to the linear transformation matrix subject to the conjugated orthogonality to the matrix $\mathbf{S}_{Ww}$,

$J_{cwFs}(\mathbf{P})$ will be equivalent to $J_{cwFq}(\mathbf{P})$. Therefore, the constraint of the conjugated orthogonality to the matrix $\mathbf{S}_{Ww}$ will not be considered in the context of $J_{cwFs}(\mathbf{P})$. On the other hand, the symmetric property of $\mathbf{S}_{Bw} - \mathbf{S}_{Ww}$ can ensure the extracted discriminant vectors to be orthonormal to each other. This fact provides us an efficient way to extract the uncorrelated features in the batch manner. Specifically, the original feature space is mapped into the $\mathbf{S}_T$ whitened space, and thus the orthonormal discriminant vectors extracted with respect to the criterion $J_{cwFs}(\mathbf{P})$ in the $\mathbf{S}_T$ whitened space will keep $\mathbf{S}_T$ whitened. Therefore, uncorrelated features with the unitary variances can be obtained. To relieve the class separability mismatch problem, it is desirable to sequentially reduce the dimensionality of the feature space and re-calculate the between-class and within-class weights after reducing per dimension. Accordingly, we propose a group of fractional CWFLDA techniques based on $J_{cwFs}(\mathbf{P})$ as follows.

a. The F-CWFLDA_{s1} technique (sequential backward reduction)

In the F-CWFLDA_{s1}, the eigenvalue decomposition of $\mathbf{S}_{Bw} - \mathbf{S}_{Ww}$ is determined. By discarding the eigenvector with the smallest eigenvalue, the remaining eigenvectors constitute a linear transformation matrix. The feature space is mapped into a subspace by using this transformation matrix with the dimensionality reduced by “1”. In this new feature space, the between-class weights and within-class weights are re-calculated, and then the eigenvalue decomposition of $\mathbf{S}_{Bw}^{new} - \mathbf{S}_{Ww}^{new}$ based on the newly updated matrices $\mathbf{S}_{Bw}^{new}$ and $\mathbf{S}_{Ww}^{new}$ in terms of the re-calculated weights is determined, followed by the dimensionality of the feature space being further reduced by “1” in the same manner. In this sequential backward reduction, the set of discriminant vectors extracted at each dimension is generally not nested.

b. The F-CWFLDA_{s2} technique (sequential backward reduction)

In this technique, the original feature space is first mapped into the $\mathbf{S}_T$ whitened space by using the transformation matrix $\mathbf{S}_T^{-1/2}$. Then the F-CWFLDA_{s2} is applied in the $\mathbf{S}_T$ whitened space with the extracted discriminant vectors at each reduced dimension being multiplied by the matrix $\mathbf{S}_T^{-1/2}$ at the left-hand side.

c. The F-CWFLDA_{s3} technique (sequential backward reduction)

By applying the QR decomposition to the optimal linear transformation matrix obtained by applying the F-CWFLDA_{s2} at each reduced dimension, we obtain the F-CWFLDA_{s3} technique, in which a set of orthonormal discriminant vectors at each reduced dimension can be extracted in the batch manner based on those already extracted discriminant vectors by applying the F-CWFLDA_{s2} at the corresponding dimension.

3.4 Gaussian heteroscedastic direct BODLDR

The previously described DBODLDRho method was developed under the Gaussian homoscedastic assumption, which restricts its utility to handle heteroscedastic data. According to the Chernoff bound of the Bayes error between two classes ω_i and ω_j that is analytically formulated in Eq. (2.15) under the Gaussian heteroscedastic assumption, the DBODLDRho can be extended into the heteroscedastic version.

Under the Gaussian heteroscedastic assumption of class distributions, assume that the feature vectors belonging to two classes ω_i and ω_j in the original n -dimensional feature space are mapped into an m -dimensional subspace via the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$). Then, the Chernoff bound $\hat{U}_c(i, j)$ of the Bayes error between classes ω_i and ω_j in the transformed subspace can be expressed by:

$$\hat{U}_c(i, j) = P_i^\alpha P_j^{1-\alpha} e^{-\hat{\delta}_c^{ij}(\alpha)}, \quad 0 \leq \alpha \leq 1 \quad (3.15)$$

$$\text{with } \hat{\delta}_c^{ij}(\alpha) = \frac{\alpha(1-\alpha)}{2} (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{P} (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} \mathbf{P}^T (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2} \ln \left(\frac{|\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P}|}{|\mathbf{P}^T \mathbf{S}_i \mathbf{P}|^\alpha |\mathbf{P}^T \mathbf{S}_j \mathbf{P}|^{1-\alpha}} \right)$$

$$\mathbf{S}_{ij} = (1-\alpha) \mathbf{S}_i + \alpha \mathbf{S}_j$$

where $\hat{\delta}_c^{ij}(\alpha)$ is the Chernoff distance between two classes ω_i and ω_j in the transformed subspace, with α being a variable to be specified by users. Note that the special choice of α equal to 0.5 can result in the Bhattacharyya distance and the corresponding Bhattacharyya bound. In the multi-class case, the upper bound of the Bayes error can be estimated by the union of the pairwise Chernoff bounds over all class pairs, which can serve as the criterion function to be optimized. Starting from a properly chosen initial linear transformation, numerical optimization methods can be applied to iteratively minimize such criterion function with respect to the linear transformation matrix.

3.4.1 Existing technique

Saon *et al.* [93] presented a Bhattacharyya linear dimensionality reduction (BLDR) method, in which the union of the pairwise Bhattacharyya bounds over all class pairs is used as the criterion function to be minimized. Starting from an FLDA solution, the optimal linear transformation matrix is iteratively derived by using the conjugate gradient decent scheme. The updating rule of the linear transformation matrix is expressed by:

$$\mathbf{P}_{new} = \mathbf{P}_{old} + \eta \sum_{i=1}^{C-1} \sum_{j=i+1}^C \sqrt{P_i P_j} e^{-\hat{\delta}_B^{ij}} \mathbf{A}_{ij} \quad (3.16)$$

$$\begin{aligned} \mathbf{A}_{ij} = & \frac{1}{4} (\mathbf{B}_{ij} \mathbf{P} - \mathbf{S}_{ij} \mathbf{P} (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} \mathbf{P}^T \mathbf{B}_{ij} \mathbf{P}) (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} + \mathbf{S}_{ij} \mathbf{P} (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} \\ & - \frac{1}{2} \mathbf{S}_i \mathbf{P} (\mathbf{P}^T \mathbf{S}_i \mathbf{P})^{-1} - \frac{1}{2} \mathbf{S}_j \mathbf{P} (\mathbf{P}^T \mathbf{S}_j \mathbf{P})^{-1} \end{aligned}$$

$$\text{with } \mathbf{B}_{ij} = (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \text{ and } \mathbf{S}_{ij} = \frac{\mathbf{S}_i + \mathbf{S}_j}{2}$$

where η , $\mathbf{P}_{old}$ and $\mathbf{P}_{new}$ denote the learning step size, the transformation matrix before and after updating, respectively.

3.4.2 Proposed technique

According to Eq. (3.16), we propose a Gaussian heteroscedastic direct BODLDR (DBODLDRhe) method, in which the criterion function is defined as the union of the pairwise Chernoff bounds in the transformed subspace over all class pairs under the Gaussian heteroscedastic assumption and expressed by:

$$\sum_{i=1}^{C-1} \sum_{j=i+1}^C \hat{U}_c(i, j) \text{ with } \hat{U}_c(i, j) \text{ defined by Eq. (3.15)} \quad (3.17)$$

The gradient descent scheme is applied to minimize this criterion function with respect to the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m \leq n$) subject to the orthonormal constraint with the following updating rule:

$$\mathbf{P}_{new} = \mathbf{P}_{old} + \eta \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i^\alpha P_j^{1-\alpha} e^{-\hat{\phi}_c^{ij}(\alpha)} \mathbf{A}_{ij} \quad (3.18)$$

$$\begin{aligned} \mathbf{A}_{ij} = & \alpha(1-\alpha) \left(\mathbf{B}_{ij} \mathbf{P} - \mathbf{S}_{ij} \mathbf{P} (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} \mathbf{P}^T \mathbf{B}_{ij} \mathbf{P} \right) (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} \\ & + \mathbf{S}_{ij} \mathbf{P} (\mathbf{P}^T \mathbf{S}_{ij} \mathbf{P})^{-1} - (1-\alpha) \mathbf{S}_i \mathbf{P} (\mathbf{P}^T \mathbf{S}_i \mathbf{P})^{-1} - \alpha \mathbf{S}_j \mathbf{P} (\mathbf{P}^T \mathbf{S}_j \mathbf{P})^{-1} \\ \text{with } \mathbf{B}_{ij} = & (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \text{ and } \mathbf{S}_{ij} = (1-\alpha) \mathbf{S}_i + \alpha \mathbf{S}_j \end{aligned}$$

where η , α , $\mathbf{P}_{old}$ and $\mathbf{P}_{new}$ denote the learning step size, the variable inherited from the Chernoff distance, the transformation matrix before and after updating, respectively. We hereby adopt $\alpha = P_i / (P_i + P_j)$ and $(1-\alpha) = P_j / (P_i + P_j)$ as suggested in the CLDA technique in order to fairly compare with the CLDA related techniques. Moreover, in our proposed DBODLDRhe method, the orthogonality constraint is imposed on the linear transformation matrix so as to avoid the degeneration of the linear transformation. Accordingly, the Gram-Schmidt orthogonalization scheme will be applied to the transformation matrix whenever it is updated. Similar to the

DBODLDRho, the discriminant vectors extracted by the DBODLDRhe method is subject to the conjugated orthogonality constraint to $\mathbf{I}$. Furthermore, an optimal solution obtained by the CWCLDA_{q0} technique to be described later, after experiencing the Gram-Schmidt orthogonalization, will serve as the initial solution of our proposed DBODLDRhe method.

3.5 Class-wise weighted CLDA motivated by Gaussian heteroscedastic Bayes optimality

The CWFLDA techniques can efficiently derive globally optimal solutions in the closed form with respect to two newly proposed class-wise weighted Fisher criteria, which is motivated by Bayes optimality under the Gaussian homoscedastic assumption. However, it cannot well handle heteroscedastic data, which motivates us to consider their heteroscedastic extensions. The newly proposed CLDA technique provides a very archetype.

In the CLDA, the Chernoff distance based directed distance matrix $\mathbf{D}_C$ can capture the discriminatory information present in the difference of both class means and class covariance matrices in a somewhat balanced manner. However, several deficiencies associated with the conventional Chernoff criterion, as discussed in subsection 2.3.2.4, prevent it achieving better performance. Among them, some deficiencies inherited from the Fisher criterion can be compensated by extending the CWFLDA techniques to the heteroscedastic versions. The numerical singularity and instability problems can be resolved via some regularization scheme. Moreover, the deficiency coming from the inequality $\hat{\mathbf{D}}_C \neq \mathbf{P}^T \mathbf{D}_C \mathbf{P}$ with respect to the linear transformation matrix $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m < n$) is attempted to be relieved via the sequential backward reduction. In this section, we present a class of Class-wise Weighted CLDA (CWCLDA) techniques motivated by Gaussian heteroscedastic Bayes optimality, which can compensate several deficiencies inherent in the conventional Chernoff criterion. To our best knowledge, no existing literatures have addressed these issues.

Inspired by the FS-FLDA technique, we develop the Fraction-step CLDA (FS-CLDA) technique. Different from the FS-FLDA, the distance metric associated with each class pair in the FS-CLDA is proportional to Chernoff distance. Accordingly, we substitute the Euclidean distances used for constructing the weights in the weighted pairwise Fisher criterion of the FS-FLDA with the squared root of the Chernoff distance to establish the FS-CLDA technique. Other implementation steps remain same as those in the FS-FLDA. Besides the advantage of suppressing the class separability mismatch problem, the fractional reduction scheme may also somewhat relieve the problem coming from the inequality $\hat{\mathbf{D}}_C \neq \mathbf{P}^T \mathbf{D}_C \mathbf{P}$ with $\mathbf{P} \in \mathbf{R}^{n \times m}$ ($m < n$) being the linear transformation matrix. Specifically, in the FS-CLDA, the original feature space is first mapped into the $\mathbf{S}_w$ whitened space, and then discriminant vectors can be extracted in fractional steps based on the eigenvalue decomposition of $\tilde{\mathbf{D}}_C$ updated at each fractional step in the $\mathbf{S}_w$ whitened space. It has been verified in subsection 2.3.2.4 that the linear transformation matrix $\mathbf{P}'$ composed by all the eigenvectors of $\tilde{\mathbf{D}}_C$ can satisfy the equality $\hat{\mathbf{D}}_C = \mathbf{P}'^T \tilde{\mathbf{D}}_C \mathbf{P}'$. Therefore, intuitively, the transformation matrix composed of the first $n-1$ leading eigenvectors of $\tilde{\mathbf{D}}_C$ corresponding to the $n-1$ largest eigenvalues should approximate such equality to the large extent. The theoretical study of the precision of such approximation is among our further plans.

Under the Gaussian heteroscedastic assumption, we extend the class-wise weighted Fisher criteria in the quotient and subtraction forms to the quotient and subtraction class-wise weighted Chernoff criteria, respectively, which can compensate several CLDA deficiencies inherited from the FLDA technique. The newly proposed criteria are defined as follows.

Quotient class-wise weighted Chernoff criteria:

$$J_{cwCq}(\mathbf{P}) = \text{trace}\left(\left(\mathbf{P}^T \mathbf{S}_{ww} \mathbf{P}\right)^{-1} \mathbf{P}^T \mathbf{S}_{Cw} \mathbf{P}\right) \quad (3.19)$$

Subtraction class-wise weighted Chernoff criteria:

$$J_{cwCs}(\mathbf{P}) = \text{trace}(\mathbf{S}_{Cw}) - \text{trace}(\mathbf{S}_{ww}) \quad (3.20)$$

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Two key matrices $\mathbf{S}_{Ww}$ and $\mathbf{S}_{Cw}$ involved in the above criteria are calculated orderly as follows.

$$\begin{aligned}
 \mathbf{S}_i &= \text{Reg}(\mathbf{S}_i) \\
 \alpha &= P_i / (P_i + P_j), \quad (1 - \alpha) = P_j / (P_i + P_j) \\
 \mathbf{S}_{ij} &= (1 - \alpha)\mathbf{S}_i + \alpha\mathbf{S}_j, \quad \partial_C^{ij}(\alpha) = \frac{\alpha(1 - \alpha)}{2} (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{ij}^{-1} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2} \ln \left(\frac{|\mathbf{S}_{ij}|}{|\mathbf{S}_i|^\alpha |\mathbf{S}_j|^{1 - \alpha}} \right) \\
 U_C^{ij} &= (P_i / (P_i + P_j))^\alpha (P_j / (P_i + P_j))^{1 - \alpha} e^{-\partial_C^{ij}(\alpha)}, \quad tw_{ij}^B = \frac{1 - U_C^{ij}}{\partial_C^{ij}} \\
 w_i^W &= \sum_{\substack{j=1 \\ j \neq i}}^C tw_{ij}^B / \max_{1 \leq k \leq C} \left(\sum_{\substack{j=1 \\ j \neq k}}^C tw_{kj}^B \right) \\
 \mathbf{S}_{Ww} &= \sum_{i=1}^C P_i w_i^W \mathbf{S}_i
 \end{aligned} \tag{3.21}$$

$$\begin{aligned}
 \tilde{\partial}_C^{ij}(\alpha) &= \frac{\alpha(1 - \alpha)}{2} (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{Ww}^{-1/2} (\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_{ij} \mathbf{S}_{Ww}^{-1/2})^{-1} \mathbf{S}_{Ww}^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2} \ln \left(\frac{|\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_{ij} \mathbf{S}_{Ww}^{-1/2}|}{|\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_i \mathbf{S}_{Ww}^{-1/2}|^\alpha |\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_j \mathbf{S}_{Ww}^{-1/2}|^{1 - \alpha}} \right) \\
 &= \frac{\alpha(1 - \alpha)}{2} (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{ij}^{-1} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2} \ln \left(\frac{|\mathbf{S}_{ij}|}{|\mathbf{S}_i|^\alpha |\mathbf{S}_j|^{1 - \alpha}} \right) \\
 &= \partial_C^{ij}(\alpha) \\
 \text{For: } J_{cwCq}(\mathbf{P}): w_{ij}^B &= \frac{1 - \tilde{U}_C^{ij}}{(2\partial_C^{ij})/\alpha(1 - \alpha)}
 \end{aligned}$$

$$\tilde{U}_C^{ij} = (P_i / (P_i + P_j))^\alpha (P_j / (P_i + P_j))^{1 - \alpha} e^{-\tilde{\partial}_C^{ij}(\alpha)} = (P_i / (P_i + P_j))^\alpha (P_j / (P_i + P_j))^{1 - \alpha} e^{-\partial_C^{ij}(\alpha)} = U_C^{ij}$$

$$\begin{aligned}
 \mathbf{S}_{Cw} &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j w_{ij}^B \mathbf{S}_{Ww}^{1/2} \left((\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_{ij} \mathbf{S}_{Ww}^{-1/2})^{-1/2} \mathbf{S}_{Ww}^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{Ww}^{-1/2} (\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_{ij} \mathbf{S}_{Ww}^{-1/2})^{-1/2} \right. \\
 &\quad \left. + \frac{1}{\alpha(1 - \alpha)} (\log(\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_{ij} \mathbf{S}_{Ww}^{-1/2}) - \alpha \log(\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_i \mathbf{S}_{Ww}^{-1/2}) - (1 - \alpha) \log(\mathbf{S}_{Ww}^{-1/2} \mathbf{S}_j \mathbf{S}_{Ww}^{-1/2})) \right) \mathbf{S}_{Ww}^{1/2}
 \end{aligned} \tag{3.22}$$

$$\text{For } J_{cwCs}(\mathbf{P}): U_C^{ij} = (P_i / (P_i + P_j))^\alpha (P_j / (P_i + P_j))^{1 - \alpha} e^{-\partial_C^{ij}(\alpha)}, \quad w_{ij}^B = \frac{1 - U_C^{ij}}{(2\partial_C^{ij})/\alpha(1 - \alpha)}$$

$$\mathbf{S}_{Cw} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j w_{ij}^B \left((\mathbf{S}_{ij})^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T (\mathbf{S}_{ij})^{-1/2} + \frac{1}{\alpha(1-\alpha)} (\log(\mathbf{S}_{ij}) - \alpha \log(\mathbf{S}_i) - (1-\alpha) \log(\mathbf{S}_j)) \right) \quad (3.23)$$

Here $\mathbf{S}_{Cw}$, w_{ij}^B , $\mathbf{S}_{Ww}$ and w_i^W denote weighted pairwise directed distance matrix, between-class weight, average weighted within-class scatter matrix and within-class weight, respectively. Note that the matrix $\mathbf{S}_{Cw}$ is differently defined in the quotient and subtraction criteria. Our newly proposed criteria have the following attractive properties.

1. In the beginning, the singularity and numerical instability problems are checked for each individual covariance matrix $\mathbf{S}_i$. If some individual covariance matrices are singular, they will be regularized. In our work, if the matrix $\mathbf{S}_i$ has any eigenvalue smaller or equal to one millionth of $\text{trace}(\mathbf{S}_i)$, it will be regularized by addition with an identity matrix multiplied with a constant equal to the one millionth of $\text{trace}(\mathbf{S}_i)$. This step can eliminate the singularity problem while increasing the numerical stability.
2. In the class-wise weighed Chernoff criteria, due to the extension of the between-class scatter matrix to the directed distance matrix, the associated distance metric is correspondingly changed from the squared Mahalanobis distance to the scaled Chernoff distance. Under the Gaussian heteroscedastic assumption, the upper bound of the two-class Bayes error can be analytically formulated as the two-class Chernoff bound, which are used to calculate the between-class weights in the class-wise weighted Chernoff criteria. Consequently, the scaled Chernoff distance associated with each pairwise directed distance matrix is replaced by the two-class Chernoff bound of the corresponding two classes, which can significantly reduce the contributions from the excessively separated class pairs involving outlier classes to the estimate of the weighted pairwise directed distance matrix.

3. The introduction of the within-class weights can reduce the contributions from outlier classes to the estimate of the average within-class scatter matrix. The justifications have been given when describing the class-wise weighted Fisher criteria.
4. In $J_{cwCq}(\mathbf{P})$, the between-class weights should be calculated in the $\mathbf{S}_{Ww}$ whitened space due to the two-stage maximization procedure. However, in $J_{cwCs}(\mathbf{P})$, two matrices $\mathbf{S}_{Cw}$ and $\mathbf{S}_{Ww}$ generally cannot be simultaneously diagonalized. Therefore, the between-class weights should be calculated in the original feature space. Furthermore, the directed distance matrix can be formulated directly with no need of the whitening and anti-whitening procedure as described in subsection 2.3.2.2. Note that since the Chernoff distance is invariant to any invertible transformation, the between-class weights calculated in the $\mathbf{S}_{Ww}$ whitened space equal those computed in the original feature space.

Based on the newly proposed class-wise weighted Chernoff criteria in the quotient and subtraction forms, we can obtain a class of CWCLDA techniques, which extend the corresponding CWFLDA techniques with respect to the class-wise weighted Fisher criteria to their heteroscedastic versions. Specifically, the CWCLDA_{q0}, F-CWCLDA_{q0}, CWCLDA_{q1}, CWCLDA_{q2} and CWCLDA_{q3} techniques are derived with respect to $J_{cwCq}(\mathbf{P})$ as the heteroscedastic extension of the CWFLDA_{q0}, F-CWFLDA_{q0}, CWFLDA_{q1}, CWFLDA_{q2} and CWFLDA_{q3} techniques based on $J_{cwFq}(\mathbf{P})$. The F-CWCLDA_{s1}, F-CWCLDA_{s2} and F-CWCLDA_{s3} techniques are derived with respect to $J_{cwCs}(\mathbf{P})$ as the heteroscedastic extension of the F-CWFLDA_{s1}, F-CWFLDA_{s2} and F-CWFLDA_{s3} techniques based on $J_{cwFs}(\mathbf{P})$. Here, we do not give unnecessary details about the proposed CWCLDA techniques since they can be easily implemented by substituting the class-wise weighted Fisher criteria and therein involved parameters with the class-wise weighted Chernoff criteria and the corresponding parameters while remaining same all the other implementation steps.

3.6 Experimental results

In this section, we evaluate the performance of the proposed enhanced parametric BODLDR techniques in comparison with their archetypes⁹ for multi-class classification problems. Note that some archetype techniques based on the Chernoff criterion were actually proposed by us for the first time. For convenience, we summarize all the 36 experimental techniques in Table 3.1, with their indices and names interchangeable in the following context. The previously existing techniques are marked by the superscript “*”. Our proposed 18 enhanced parametric BODLDR techniques are typeset in bold.

Table 3.1: The 36 DLDR techniques involved in the experiments. Information is provided on technique index, technique name, dimensionality reduction manner, conjugated orthogonality constraint and technique provenance in this thesis.

Technique index	Technique Name	Reduction Manner	Conjugated Orthogonality	Provenance
1	FLDA [*]	batch	$\mathbf{S}_W$	pp. 50
2	aPAC-FLDA [*]	batch	$\mathbf{S}_W$	PP. 86
3	CWFLDA_{q0}	batch	$\mathbf{S}_{W_W}$	pp. 93
4	FS-FLDA [*]	sequential backward	$\mathbf{S}_W$	pp. 88
5	F-CWFLDA_{q0}	sequential backward	$\mathbf{S}_{W_W}$	pp. 94
6	ALDR [*]	gradient descent	$\mathbf{S}_W$	pp. 82
7	DBODLDRho	gradient descent	$\mathbf{I}$	pp. 83
8	FLDA _{q1} [*]	sequential forward	$\mathbf{I}$	pp. 60
9	CWFLDA_{q1}	sequential forward	$\mathbf{I}$	pp. 95
10	FLDA _{s1} [*]	sequential backward	$\mathbf{I}$	pp. 70
11	F-CWFLDA_{s1}	sequential backward	$\mathbf{I}$	pp. 96
12	FLDA _{q2} [*]	sequential forward	$\mathbf{S}_T$	pp. 61
13	CWFLDA_{q2}	sequential forward	$\mathbf{S}_T$	pp. 95

⁹ The BLDR technique is omitted in our current experiments due to its similarity to the DBODLDRhe technique

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

14	FLDA _{s2}	sequential backward	$\mathbf{S}_T$	pp. 71
15	F-CWFLDA_{s2}	sequential backward	$\mathbf{S}_T$	pp. 96
16	FLDA _{q3} *	sequential forward	$\mathbf{I}$	pp. 61
17	CWFLDA_{q3}	sequential forward	$\mathbf{I}$	pp. 95
18	FLDA _{s3}	sequential backward	$\mathbf{I}$	pp. 71
19	F-CWFLDA_{s3}	sequential backward	$\mathbf{I}$	pp. 97
20	CLDA*	batch	$\mathbf{S}_W$	pp. 72
21	CWCLDA_{q0}	batch	$\mathbf{S}_{Ww}$	pp. 104
22	FS-CLDA	sequential backward	$\mathbf{S}_W$	pp. 101
23	F-CWCLDA_{q0}	sequential backward	$\mathbf{S}_{Ww}$	pp. 104
24	DBODLDRhe	gradient descent	$\mathbf{I}$	pp. 99
25	CLDA _{q1}	sequential forward	$\mathbf{I}$	pp. 76
26	CWCLDA_{q1}	sequential forward	$\mathbf{I}$	pp. 104
27	CLDA _{s1}	sequential backward	$\mathbf{I}$	pp. 79
28	F-CWCLDA_{s1}	sequential backward	$\mathbf{I}$	pp. 104
29	CLDA _{q2}	sequential forward	$\mathbf{S}_T$	pp. 76
30	CWCLDA_{q2}	sequential forward	$\mathbf{S}_T$	pp. 104
31	CLDA _{s2}	sequential backward	$\mathbf{S}_T$	pp. 79
32	F-CWCLDA_{s2}	sequential backward	$\mathbf{S}_T$	pp. 104
33	CLDA _{q3}	sequential forward	$\mathbf{I}$	pp. 76
34	CWCLDA_{q3}	sequential forward	$\mathbf{I}$	pp. 104
35	CLDA _{s3}	sequential backward	$\mathbf{I}$	pp. 79
36	F-CWCLDA_{s3}	sequential backward	$\mathbf{I}$	pp. 104

3.6.1 Experimental setup

To demonstrate the effectiveness of between-class and within-class weights in the proposed class-wise weighted Fisher and Chernoff criteria, we compare the performance of several quotient class-wise weighted Fisher or Chernoff criterion based DLDR techniques with that of their conventional counterparts over some

artificial datasets. Each of these artificial datasets consists of 30 classes including 2 outlier classes subject to Gaussian homoscedastic and heteroscedastic distributions, respectively. For per artificial dataset, 3000 training samples and 1500 testing samples are generated 10 times according to the pre-specified class distributions to establish 10 pairs of training and testing sets. With respect to each training and testing pair and for every possible m to reduce the dimension to, the DLDR is carried out on the training set. The resulting optimal linear transformation is applied to both the training and testing sets, followed by classification on the transformed testing set via the classifier constructed on the transformed training set. The misclassification rates of the 10 pairs are averaged to obtain the mean classification error at each reduced dimension m . In this experiment, we choose between LDC¹⁰ and QDC according to the type of the already known class distributions since they stay close to the Gaussian homoscedastic and heteroscedastic assumptions, respectively.

To evaluate the performance of the proposed techniques to address real-world multi-class problems, we apply the 18 enhanced BODLDR techniques and their 18 archetypes on 10 multi-class datasets chosen from the UCI machine learning repository [78]. Note that two datasets named “Noiseiris” and “Noisewine” are created based on two UCI datasets “Iris” and “Wine”, respectively, by artificially adding an outlier class with 100 samples located remotely from the originally existing classes so as to contaminate the original datasets and increase the class number. The other 8 datasets are: Vowel context, Glass identification, Ecoli, Image segmentation, Landsat satellite, Vehicle, Pendigits and Optdigits. Table 3.2 summarizes the statistics of these 10 experimental datasets, which are normalized to $[-1, 1]$ before use.

Table 3.2: The 10 datasets used in our experiments. Information is provided on dataset index, dataset name, number of classes (C), feature space dimensionality (n), total number of samples (N) and dimensionality after applying the PCA (PC).

¹⁰ For some enhanced parametric BODLDR techniques with the extracted discriminant vectors subject to the conjugated orthogonality to the average (weighted) within-class scatter matrix, the LDC can be simplified to the nearest center classifier biased with class prior probabilities.

Index	Datasets	C	n	N	PC
a	Noiseiris	4	4	250	4
b	Noisewine	4	13	278	12
c	Vehicle	4	18	846	18
d	Glass	6	9	214	9
e	Landsat satellite	6	36	6435	36
f	Image segmentation	7	19	2310	14
g	Ecoli	8	7	336	6.9
h	Pendigits	10	16	10992	16
i	Optdigits	10	64	5620	62
j	Vowel	11	10	990	10

For these 10 UCI datasets, we use the 10-fold distribution-balanced stratified cross-validation method [148] to obtain 10 pairs of training and testing sets. For each pair of training and testing sets and for every possible m to reduce the dimension to, the DLDR is carried out on the training set. The resulting optimal linear transformation is applied to both training and testing sets, followed by classification on the transformed testing set via the classifier constructed on the transformed training set. The misclassification rates of the 10 pairs are averaged to obtain the mean classification error at each reduced dimension m . Note that to avoid the possible numerical singularity and instability, the PCA is performed on the training set before applying the DLDR, and the principal vectors with eigenvalues smaller than one millionth of the trace of the total scatter matrix are discarded. The linear transformation matrix composed of the remaining principal vectors is used to yield the transformed training and testing sets, on which the DLDR followed by classification is performed. In this experiment, three classifiers LDC, QDC and k NN ($k=1$) are used for classification since they are related to the empirical risk minimization to some extent.

3.6.2 Experiments on artificial data

Three artificial datasets with known class distributions subject to Gaussian homoscedastic or heteroscedastic assumption are generated as follows.

D_1 : 30 30D classes subject to Gaussian homoscedastic distributions. Prior probabilities are assumed to be equal for all classes. A set of 28 class mean vectors are generated according to the uniform distribution between -1 and 1 per dimension, and the remaining two class mean vectors are generated according to the uniform distribution between 100 and 150 per dimension. These two classes are regarded as outlier classes. All classes are assumed to have the same covariance matrices equal to $\mathbf{I}$. A training set containing 3000 training samples with 100 samples per class and a testing set containing 1500 testing samples with 50 samples per class can be generated from this class distribution. We generate 10 pairs of training and testing sets for our experiments.

D_2 : 30 30D classes subject to Gaussian heteroscedastic distributions. Prior probabilities are assumed to be equal for all classes. Class mean vectors are set same as those in D_1 . The 28 classes with mean vectors distributed between -1 and 1 per dimension are assumed to have the same covariance matrices equal to $\mathbf{I}$, while the remaining two outlier classes are assumed to have the randomly generated covariance matrices with the maximum variance along each principle axe less than or equal to 50. A training set containing 3000 training samples with 100 samples per class and a testing set containing 1500 testing samples with 50 samples per class can be generated from this class distribution. We generate 10 pairs of training and testing sets for our experiments.

D_3 : 30 30D classes subject to Gaussian heteroscedastic distributions. Prior probabilities are assumed to be equal for all classes. A set of 28 class mean vectors are generated according to the uniform distribution between -1 and 1 per dimension, and the remaining two class mean vectors are generated according to the uniform distribution between 100 and 150 per dimension. The 28 classes with mean vectors distributed between -1 and 1 per dimension are assumed to have the randomly generated covariance matrices with the maximum variance along each principle axe less than or equal to 4, while the remaining two outlier classes are assumed to have the randomly generated covariance matrices with the maximum variance along each principle axe less than or equal to 50. A training set containing 3000 training samples

with 100 samples per class and a testing set containing 1500 testing samples with 50 samples per class can be generated from this class distribution. We generate 10 pairs of training and testing sets for our experiments.

To demonstrate the utility of between-class and within-class weights, we choose several representative CWFLDA and CWCLDA techniques, namely CWFLDA_{q0}, F-CWFLDA_{q0}, CWCLDA_{q0} and F-CWCLDA_{q0}, and compare their performance with the corresponding archetypes based on the conventional Fisher or Chernoff criterion, i.e. FLDA, aPAC-FLDA, FS-FLDA, CLDA and FS-CLDA. The experiments are conducted within two groups of DLDR techniques based on Fisher and Chernoff criteria, respectively. The performance of the Fisher criteria based techniques including FLDA, aPAC-FLDA, FS-FLDA, CWFLDA_{q0} and F-CWFLDA_{q0} are compared together in terms of the LDC classifier. On the other hand, we compare the performance of the Chernoff criteria based techniques including CLDA, FS-CLDA, CWCLDA_{q0} and F-CWCLDA_{q0} in terms of the QDC classifier. The experiments are carried out on three datasets D_1 to D_3 with the results shown in Figures 3.1 to 3.3.

In the **first** experiment, we test the performance of the Fisher and Chernoff criteria based techniques on dataset D_1 subject to Gaussian homoscedastic assumption. It is easy to verify that all techniques in comparison when combined with the QDC classifier will achieve the same classification performance as when combined with the LDC classifier since the Gaussian homoscedasticity is preserved at each dimension to reduce to, which degrades the QDC to LDC. Therefore, we can show the performance of both two groups of techniques in terms of the LDC classifier. Figures 3.1a and 3.1b illustrate the mean classification error versus the reduced dimension by using Fisher and Chernoff criteria based techniques followed by the LDC classifier, respectively.

Some observations are outlined as follows.

1. For Fisher criteria based techniques shown in Figure 3.1a, the aPAC-FLDA performs better than FLDA, especially when the dimension is gradually reduced. This fact verifies the utility of the weighted pairwise between-class scatter matrix

used in the aPAC to suppress the influence from outlier classes. The proposed $CWFLDA_{q0}$ demonstrates the similar performance as the aPAC-FLDA, which indicates the fact that the $CWFLDA_{q0}$ can be regarded as the generalization of the aPAC-FLDA. The FS-FLDA demonstrates better or competitive performance at each reduced dimension in comparison with other techniques. However, the fractional reduction steps involved in the FS-FLDA make it expense the highest computational cost. In comparison with the $CWFLDA_{q0}$, the $F-CWFLDA_{q0}$ achieves the similar performance in most reduced dimensions while performing better at some lower dimensions, which is attributed to the stepwise reduction of the dimensionality.

2. For Chernoff criteria based techniques shown in Figure 3.1b, the $CWCLDA_{q0}$ performs better than CLDA when the dimension is gradually reduced, which verifies the utility of the class-wise weighted Chernoff criterion to suppress the influence from outlier classes. The FS-CLDA demonstrates better or competitive performance at each reduced dimension in comparison with the CLDA, which verify the utility of the fractional reduction steps to suppress the outlier influence. In comparison with the $CWCLDA_{q0}$, the $F-CWCLDA_{q0}$ achieves the similar performance in most reduced dimensions while performing better at some lower dimensions, which is attributed to the stepwise reduction of the dimensionality. It is worth noting that the FS-CLDA does not demonstrate better performance than the $F-CWCLDA_{q0}$ in many lower dimensions, which may be attributed to the fact that the FS-CLDA put much emphasis on some excessively overlapped classes regardless of the Bayes accuracy between them. Consequently, the resulting discriminant vectors biasing towards preserving some nearly inseparable classes may possibly lead to worse overall class separation in the transformed subspace.

In the **second** experiment, the performance test is carried out on dataset D_2 . In contrast to D_1 , D_2 retains the positions of all class mean vectors in D_1 and the identity covariance matrix of the 28 ordinary classes in D_1 , while making the two outlier classes to have randomly generated covariance matrices with the maximum

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.1a

Figure 3.1b

Figure 3.1: (a) The mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier on D_1 . (b) The mean classification error versus the reduced dimension by using Chernoff criteria based techniques followed by the LDC classifier on D_1 .

variance along each principle axe less than or equal to 50. With the existence of outlier classes having largely deviated covariance matrices, dataset D_2 is used to demonstrate the effectiveness of the average weighted within-class scatter matrix used in the class-wise weighted Fisher and Chernoff criteria. We show the performance of the group of Fisher criteria based techniques and the group of Chernoff criteria based techniques in terms of LDC and QDC classifiers, respectively. Figures 3.2a and 3.2b illustrate the mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier and by using Chernoff criteria based techniques followed by the QDC classifier, respectively.

Some observations are described as follows:

1. For Fisher criteria based techniques shown in Figure 3.2a, the aPAC-FLDA still outperforms the FLDA when the dimension is gradually reduced. However, it performs worse at many reduced dimensions than the proposed CWFLDA_{q0}. Since the only distinction between D_1 and D_2 is the covariance matrices of the two outlier classes, the performance difference between CWFLDA_{q0} and aPAC-FLDA can be attributed to the use of average weighted within-class scatter matrix. By incorporating such scatter matrix, the CWFLDA_{q0} can suppress undesirable contributions from outlier classes with largely deviating covariance matrices to the estimate of the assumed homoscedastic covariance matrix, i.e. the average within-class scatter matrix, and therefore result in better classification performance. The F-CWFLDA_{q0} achieves the similar performance to the CWFLDA_{q0} in most reduced dimensions while performing better at some lower dimensions, which is attributed to the stepwise reduction of the dimensionality. The FS-FLDA again outperforms the FLDA but performs worse than CWFLDA_{q0} and F-CWFLDA_{q0} at most reduced dimensions due to the influence from the outlier classes to the estimate of the average within-class scatter matrix.
2. For Chernoff criteria based techniques shown in Figure 3.2b, the CWCLDA_{q0} outperforms the CLDA when the dimension is gradually reduced. In contrast, the

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.2a

Figure 3.2b

Figure 3.2: (a) The mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier on D_2 . (b) The mean classification error versus the reduced dimension by using Chernoff criteria based techniques followed by the QDC classifier on D_2 .

F-CWCLDA_{q0} achieves the similar performance at most reduced dimensions while performing better at some lower dimensions, which is attributed to the stepwise reduction of the dimensionality. The FS-CLDA demonstrates better or competitive performance at each reduced dimension in comparison with the CLDA, but performs worse than both CWCLDA_{q0} and F-CWCLDA_{q0} at many reduced dimensions.

In the **third** experiment, the performance of the Fisher and Chernoff criteria based techniques is evaluated on dataset D_3 , in which all classes have different covariance matrices, with the two outlier classes having largely deviated covariance matrices. Relative to D_1 and D_2 , dataset D_3 may stay more closely to the real-world problem. We show the performance of the group of Fisher criteria based techniques and the group of Chernoff criteria based techniques in terms of the LDC and QDC classifiers, respectively. Figures 3.3a to 3.3b illustrate the mean classification error versus the reduced dimension by using Fisher criteria based techniques coupled with the LDC classifier and by using Chernoff criteria based techniques coupled with the QDC classifier, respectively.

Some observations are described as follows:

1. For Fisher criteria based techniques shown in Figure 3.3a, the aPAC-FLDA outperforms the FLDA when the dimension is gradually reduced. The FS-FLDA demonstrates better performance than the aPAC-FLDA at many reduced dimensions. Both CWFLDA_{q0} and F-CWFLDA_{q0} outperforms FLDA, aPAC-FLDA and FS-FLDA. Among them, the F-CWFLDA_{q0} achieves the similar performance to the CWFLDA_{q0} in most reduced dimensions while performing better at some lower dimensions.
2. For Chernoff criteria based techniques shown in Figure 3.3b, the CWCLDA_{q0} outperforms the CLDA while performing similarly in comparison with the FS-CLDA. The F-CWCLDA_{q0} achieves the better performance in comparison with both CWCLDA_{q0} and FS-CLDA at most reduced dimensions.

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.3a

Figure 3.3b

Figure 3.3: (a) The mean classification error versus the reduced dimension by using Fisher criteria based techniques followed by the LDC classifier on D_3 . (b) The mean classification error versus the reduced dimension by using Chernoff criteria based techniques followed by the QDC classifier on D_3 .

In summary, we demonstrate the effectiveness of the weighted pairwise between-class scatter matrix for suppressing the undesirable influences from the class pairs involving outlier classes to the eigenvalue decomposition of the estimated between-class scatter matrix on D_2 . The effectiveness of the average weighted within-class scatter matrix for suppressing the undesirable contributions from outlier classes having largely deviating covariance matrices to the estimate of the assumed homoscedastic covariance matrix (i.e. the average within-class scatter matrix) is verified on D_2 . The advantage of the quotient class-wise weighted Fisher or Chernoff criterion based DLDR techniques in comparison with their conventional counterparts are further demonstrated on a more general dataset D_3 .

3.6.3 Experiments on UCI data

We apply the proposed 18 enhanced BODLDR techniques as well as their 18 archetypes on 10 multi-class datasets described in Table 3.2 to evaluate the performance of our proposed techniques in comparison with their archetypes to address real-world multi-class problems.

To comprehensively evaluate the performance of a DLDR technique, we take into account both the minimum and average mean classification errors over all possible dimensions that can be reduced to. Intuitively, a good DLDR technique should show its superiority in both the overall and average performance. Moreover, the performance of a DLDR technique is generally data-dependent. Therefore, it is irrational to claim the superiority of one DLDR technique over the others only because it beats other competitors on a single dataset. In our experiments, we employ the Wilcoxon signed rank test [81] to calculate the relative performance indices of each experimental DLDR technique with respect to the optimal one on per dataset, and then average the relative performance indices over all the experimental datasets to give the final evaluation. Three relative performance indices devised for this purpose are described as follows.

1. Relative Best Performance Index (RBPI)

Assume a group of DLDR techniques followed by a certain classifier are applied on a specific dataset. For every group member, the mean classification error at each possible dimension that can be reduced to is calculated, and accordingly the minimum mean classification error and the corresponding optimal dimension¹¹ over all possible dimensions can be determined. Among all group members, the one with the smallest minimum mean classification error is chosen as the optimal DLDR technique. We compare the set of classification errors on the 10 cross-validation pairs at the optimal dimension obtained by each DLDR technique with that obtained by the optimal DLDR technique via the Wilcoxon signed rank test with the resulting probability value indicating the probability that the null hypothesis i.e. median difference of two samples under test is zero, is true. Therefore, each DLDR technique can be assigned with a probability value, named RBPI, to indicate its performance relative to that of the optimal DLDR technique on a specific dataset. The RBPI for the optimal DLDR technique will equal “1”. Since the RBPI implicitly normalizes the performance of all the experimental DLDR techniques on a specific dataset, for each DLDR technique, its average RBPI over all the experimental datasets can reflect its relative ranking from the perspective of the best performance.

2. Relative Average Performance Index (RAPI)

Assume a group of DLDR techniques followed by a certain classifier are applied on a specific dataset. For every group member, the mean classification error at each possible dimension that can be reduced to is calculated. At each reduced dimension, we calculate the mean classification errors obtained by all the experimental DLDR techniques, and determine the optimal DLDR technique with the minimum mean classification error at that dimension. We compare the set of classification errors on the 10 cross-validation pairs at such dimension obtained by each DLDR technique with that obtained by the optimal DLDR technique via the Wilcoxon signed rank test, with the resulting probability value denoting the performance of each DLDR technique relative to the optimal one at that dimension. For each DLDR technique,

¹¹ If the minimum mean classification error can be equally achieved at several dimensions, the lowest dimension is chosen as the optimal dimension.

the probability values calculated at all possible dimensions on a specific dataset are averaged to define the RAPI, which reflect the average dimensionality reduction performance of a DLDR technique relative to the optimal one. For each DLDR technique, its average RAPI over all the experimental datasets can reflect its relative ranking from the perspective of the average performance.

3. Relative Overall Performance Index (ROPI)

For a DLDR technique, the average of its RBPI and RAPI on a specific dataset can yield another relative performance measure, called ROPI, which indicate the overall relative performance of such DLDR technique on this dataset. For each DLDR technique, its average RAPI over all the experimental datasets can reflect its relative ranking from the perspective of the overall performance.

In our experiments, for the enhanced parametric BODLDR techniques under the Gaussian homoscedastic assumption, the possible dimensions that can be reduced to are specified from 1 to $C-1$. For those under the Gaussian heteroscedastic assumption, the possible dimensions that can be reduced to are specified from 1 to PC . It is worth noting that although some FLDA variants, such as those with uncorrelated transformed features, those with orthogonal discriminant vectors and those with the subtraction criterion function, may extract more than $C-1$ features. However, $C-1$ is the smallest number requirement for the extracted discriminant vectors to result in a transformed subspace achieving the Bayes optimality under the Gaussian homoscedastic assumption. Therefore, our current analysis is restricted to such specification. Since the possibly reduced dimensions are distinctly specified for the two categories of DLDR techniques under the Gaussian homoscedastic and heteroscedastic assumptions, respectively, it would be reasonable to separately calculate the RAPI in each category. Consequently, among the three relative performance measures defined above, the RBPI is used to evaluate the relative best performance over all the DLDR techniques coming from both categories, while the RAPI and ROPI are used to evaluate the relative average and overall performance for the DLDR techniques in each category.

Some experimental DLDR techniques involve parameters to be specified. According to the preliminary experimental trials, the learning step size and maximum number of generations in the ALDR, DBODLDRho and DBODLDRhe techniques are set as 0.01 and 500, respectively. For the FS-FLDA and FS-CLDA techniques, the power of the weighting function is set as “-8”, as suggested in [147].

Tables 3.3 to 3.5 show that, by respectively using the LDC, QDC and k NN classifiers, the minimum mean classification error and the corresponding reduced dimension attained by applying each of the 36 DLDR techniques on per dataset. Note that, in our current experiments, although no direct ways are given to automatically determine an optimal dimension to reduce to, the obtained minimum mean classification error can give an indication of the optimally attainable performance and thus can be used to compare different techniques. For each dataset, the optimal DLDR technique with the overall smallest minimum mean classification error¹² among all the experimental techniques is typeset in bold. The mean classification errors in the original feature space without applying any DLDR techniques are also reported in Tables 3.3 to 3.5.

Table 3.3: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 36 experimental techniques as well as the mean classification error obtained in the original feature space (*org*) with respect to the LDC classifier. Note that each row denotes a specific experimental technique with its technique index indicated in Table 3.1, and each column represents a specific dataset with its dataset index indicated in Table 3.2.

	a	b	c	d	e	f	g	h	i	j
1	0.1200 (2)	0.0431 (3)	0.2103 (3)	0.3686 (5)	0.1593 (5)	0.0823 (6)	0.1245 (6)	0.1237 (9)	0.0468 (9)	0.3788 (6)
2	0.1240 (3)	0.0431 (3)	0.2103 (3)	0.3543 (2)	0.1593 (5)	0.0823 (6)	0.1245 (6)	0.1237 (9)	0.0468 (9)	0.3828 (5)
3	0.0120 (2)	0.0108 (3)	0.2115 (3)	0.3455 (5)	0.1490 (5)	0.0775 (4)	0.1307 (4)	0.1222 (9)	0.0443 (9)	0.3798 (5)
4	0.1200 (1)	0.0431 (2)	0.2103 (3)	0.3595 (3)	0.1593 (5)	0.0823 (5)	0.1276 (5)	0.1237 (9)	0.0459 (7)	0.3818 (8)
5	0.0120 (2)	0.0071 (3)	0.2126 (3)	0.3455 (5)	0.1490 (5)	0.0779 (4)	0.1307 (4)	0.1222 (9)	0.0443 (9)	0.3798 (5)
6	0.1160 (3)	0.0503 (2)	0.2233 (3)	0.3777 (4)	0.1649 (5)	0.0879 (6)	0.1306 (6)	0.1295 (9)	0.0562 (8)	0.3798 (6)
7	0.1200	0.0505	0.2174	0.3548	0.1591	0.0848	0.1275	0.1227	0.0452	0.3838

¹² When several techniques reach the overall smallest minimum mean classification error, the corresponding reduced dimensions are compared. The optimal technique should achieve the overall smallest minimum mean classification error at the smallest dimension. When the dimensions are same again, standard deviations of the classification error are compared. The optimal technique should have the smallest standard deviation.

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

	(2)	(3)	(3)	(3)	(5)	(6)	(4)	(9)	(9)	(5)
8	0.1200 (2)	0.2488 (2)	0.2482 (3)	0.4058 (4)	0.2353 (5)	0.1199 (6)	0.1458 (6)	0.1367 (9)	0.1189 (9)	0.3727 (7)
9	0.0520 (2)	0.1079 (3)	0.2435 (2)	0.3870 (5)	0.1843 (5)	0.1091 (6)	0.1365 (6)	0.1348 (9)	0.0975 (9)	0.3768 (6)
10	0.1160 (3)	0.0505 (3)	0.2565 (3)	0.4108 (5)	0.1796 (3)	0.1147 (4)	0.1486 (6)	0.1413 (9)	0.0633 (9)	0.3657 (8)
11	0.0920 (2)	0.0648 (3)	0.2741 (3)	0.3732 (3)	0.2744 (2)	0.1130 (4)	0.1427 (6)	0.1500 (9)	0.1635 (9)	0.3687 (9)
12	0.1200 (2)	0.0431 (3)	0.2103 (3)	0.3686 (5)	0.1593 (5)	0.0823 (6)	0.1245 (6)	0.1237 (9)	0.0468 (9)	0.3788 (6)
13	0.0480 (2)	0.0754 (3)	0.2162 (3)	0.3498 (3)	0.1573 (5)	0.0840 (6)	0.1245 (6)	0.1218 (9)	0.0452 (9)	0.3808 (5)
14	0.1200 (2)	0.0431 (3)	0.2103 (3)	0.3686 (5)	0.1593 (5)	0.0823 (6)	0.1245 (6)	0.1237 (9)	0.0468 (9)	0.3788 (6)
15	0.1240 (1)	0.0358 (2)	0.2576 (2)	0.3831 (4)	0.1664 (3)	0.0874 (5)	0.1335 (6)	0.1235 (8)	0.0459 (9)	0.3727 (4)
16	0.1200 (2)	0.0431 (3)	0.2103 (3)	0.3686 (5)	0.1593 (5)	0.0823 (6)	0.1245 (6)	0.1237 (9)	0.0468 (9)	0.3788 (6)
17	0.0480 (2)	0.0754 (3)	0.2162 (3)	0.3498 (3)	0.1573 (5)	0.0840 (6)	0.1245 (6)	0.1218 (9)	0.0452 (9)	0.3808 (5)
18	0.1200 (2)	0.0431 (3)	0.2103 (3)	0.3686 (5)	0.1593 (5)	0.0823 (6)	0.1245 (6)	0.1237 (9)	0.0468 (9)	0.3788 (6)
19	0.1240 (1)	0.0358 (2)	0.2576 (2)	0.3831 (4)	0.1664 (3)	0.0874 (5)	0.1335 (6)	0.1235 (8)	0.0459 (9)	0.3727 (4)
20	0.1200 (2)	0.0396 (10)	0.2032 (10)	0.3550 (6)	0.1568 (23)	0.0823 (14)	0.1575 (6)	0.1237 (16)	0.0429 (40)	0.3788 (6)
21	0.0080 (3)	0.0108 (3)	0.2056 (12)	0.3498 (8)	0.1520 (35)	0.0818 (14)	0.1277 (6)	0.1204 (16)	0.0411 (30)	0.3727 (5)
22	0.1200 (3)	0.0253 (2)	0.2103 (18)	0.3680 (7)	0.1591 (35)	0.0823 (14)	0.1306 (6)	0.1237 (16)	0.0463 (59)	0.3788 (8)
23	0.0080 (4)	0.0108 (4)	0.2055 (13)	0.3498 (8)	0.1512 (19)	0.0805 (13)	0.1248 (6)	0.1204 (16)	0.0415 (32)	0.3758 (5)
24	0.0520 (1)	0.0431 (12)	0.2103 (18)	0.3634 (5)	0.1573 (16)	0.0823 (13)	0.1276 (6)	0.1235 (15)	0.0415 (28)	0.3646 (6)
25	0.1240 (2)	0.0431 (12)	0.2103 (17)	0.3686 (9)	0.1591 (25)	0.0823 (14)	0.1603 (6)	0.1237 (16)	0.0429 (36)	0.3747 (4)
26	0.1240 (4)	0.0360 (11)	0.2103 (18)	0.3684 (6)	0.1584 (32)	0.0823 (14)	0.1665 (6)	0.1237 (16)	0.0422 (38)	0.3646 (4)
27	0.1160 (3)	0.0397 (7)	0.2103 (18)	0.3686 (9)	0.1588 (34)	0.0823 (14)	0.1545 (6)	0.1237 (16)	0.0468 (62)	0.3778 (5)
28	0.1200 (3)	0.0360 (7)	0.2103 (18)	0.3498 (5)	0.1593 (36)	0.0823 (14)	0.1845 (6)	0.1101 (14)	0.0463 (61)	0.3646 (6)
29	0.1120 (2)	0.0431 (6)	0.2080 (10)	0.3593 (8)	0.1571 (21)	0.0823 (14)	0.1577 (6)	0.1237 (16)	0.0425 (28)	0.3818 (3)
30	0.1160 (2)	0.0393 (9)	0.2091 (16)	0.3403 (5)	0.1577 (27)	0.0823 (13)	0.1574 (6)	0.1237 (16)	0.0420 (41)	0.3707 (6)
31	0.1120 (2)	0.0288 (4)	0.2080 (9)	0.3550 (6)	0.1574 (17)	0.0823 (14)	0.1364 (6)	0.1237 (16)	0.0420 (32)	0.3828 (7)
32	0.0280 (1)	0.0396 (9)	0.2103 (18)	0.3500 (6)	0.1582 (26)	0.0823 (14)	0.1365 (6)	0.1224 (9)	0.0457 (51)	0.3707 (4)
33	0.1120 (2)	0.0431 (6)	0.2080 (10)	0.3593 (8)	0.1571 (21)	0.0823 (14)	0.1577 (6)	0.1237 (16)	0.0425 (28)	0.3818 (3)
34	0.1160 (2)	0.0396 (10)	0.2091 (16)	0.3403 (5)	0.1577 (27)	0.0823 (13)	0.1603 (6)	0.1237 (16)	0.0420 (41)	0.3707 (6)
35	0.1120 (2)	0.0288 (4)	0.2080 (9)	0.3550 (6)	0.1574 (17)	0.0823 (14)	0.1364 (6)	0.1237 (16)	0.0420 (32)	0.3828 (7)
36	0.0280 (1)	0.0360 (9)	0.2103 (18)	0.3500 (6)	0.1582 (26)	0.0823 (14)	0.1365 (6)	0.1224 (9)	0.0457 (51)	0.3707 (4)
org	0.1240 (4)	0.0396 (12)	0.2103 (18)	0.3686 (9)	0.1593 (36)	0.0823 (14)	0.1306 (6.9)	0.1237 (16)	0.0468 (62)	0.3919 (10)

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Table 3.4: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 36 experimental techniques as well as the mean classification error obtained in the original feature space (*org*) with respect to the QDC classifier. Note that each row denotes a specific experimental technique with its technique index indicated in Table 3.1, and each column represents a specific dataset with its dataset index indicated in Table 3.2.

	a	b	c	d	e	f	g	h	i	j
1	0.1160 (3)	0.0396 (3)	0.2020 (3)	0.4338 (2)	0.1388 (5)	0.0788 (4)	0.1514 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
2	0.1160 (2)	0.0396 (3)	0.2020 (3)	0.3918 (2)	0.1380 (4)	0.0792 (4)	0.1365 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
3	0.0120 (2)	0.0108 (3)	0.2020 (3)	0.3870 (2)	0.1355 (5)	0.0684 (6)	0.1395 (5)	0.0461 (9)	0.0290 (9)	0.0970 (10)
4	0.1160 (3)	0.0396 (3)	0.2020 (3)	0.3879 (2)	0.1388 (5)	0.0797 (4)	0.1366 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
5	0.0160 (2)	0.0071 (3)	0.2020 (3)	0.3870 (2)	0.1357 (5)	0.0675 (6)	0.1395 (5)	0.0461 (9)	0.0290 (9)	0.0970 (10)
6	0.0440 (3)	0.0398 (3)	0.2233 (3)	0.4104 (2)	0.1448 (4)	0.0658 (5)	0.1456 (5)	0.0480 (9)	0.0440 (7)	0.0970 (10)
7	0.0240 (3)	0.0183 (3)	0.2056 (3)	0.3868 (2)	0.1360 (4)	0.0697 (6)	0.1334 (4)	0.0464 (9)	0.0290 (9)	0.0970 (10)
8	0.1200 (3)	0.2020 (2)	0.2423 (3)	0.4203 (2)	0.2054 (4)	0.1147 (5)	0.1602 (6)	0.0586 (9)	0.1098 (9)	0.0970 (10)
9	0.0160 (2)	0.0251 (3)	0.2352 (3)	0.3976 (5)	0.1607 (5)	0.0671 (6)	0.1457 (5)	0.0587 (9)	0.0849 (9)	0.0970 (10)
10	0.0320 (3)	0.0327 (3)	0.2588 (2)	0.4069 (4)	0.1520 (4)	0.0745 (6)	0.1425 (5)	0.0519 (9)	0.0459 (7)	0.0970 (10)
11	0.0720 (2)	0.0544 (3)	0.2647 (3)	0.4024 (5)	0.2404 (5)	0.0697 (6)	0.1456 (5)	0.0700 (9)	0.1262 (9)	0.0970 (10)
12	0.1160 (3)	0.0396 (3)	0.2020 (3)	0.4338 (2)	0.1388 (5)	0.0788 (4)	0.1514 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
13	0.0120 (2)	0.0399 (2)	0.2020 (3)	0.3693 (3)	0.1352 (5)	0.0714 (6)	0.1365 (5)	0.0460 (9)	0.0290 (9)	0.0970 (10)
14	0.1160 (3)	0.0396 (3)	0.2020 (3)	0.4338 (2)	0.1388 (5)	0.0788 (4)	0.1514 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
15	0.0240 (3)	0.0180 (3)	0.2516 (3)	0.4208 (1)	0.1414 (3)	0.0714 (4)	0.1483 (6)	0.0439 (9)	0.0278 (9)	0.0970 (10)
16	0.1160 (3)	0.0396 (3)	0.2020 (3)	0.4338 (2)	0.1388 (5)	0.0788 (4)	0.1485 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
17	0.0120 (2)	0.0399 (2)	0.2020 (3)	0.3693 (3)	0.1352 (5)	0.0714 (6)	0.1365 (6)	0.0460 (9)	0.0290 (9)	0.0970 (10)
18	0.1160 (3)	0.0396 (3)	0.2020 (3)	0.4338 (2)	0.1388 (5)	0.0788 (4)	0.1485 (5)	0.0472 (9)	0.0304 (9)	0.0970 (10)
19	0.0240 (3)	0.0180 (3)	0.2516 (3)	0.4208 (1)	0.1414 (3)	0.0714 (4)	0.1424 (6)	0.0439 (9)	0.0278 (9)	0.0970 (10)
20	0.0120 (4)	0.0071 (10)	0.1288 (15)	0.4024 (9)	0.1395 (25)	0.0805 (7)	0.1991 (6)	0.0170 (16)	0.0358 (56)	0.0970 (10)
21	0.0080 (3)	0.0036 (5)	0.1276 (12)	0.4017 (8)	0.1391 (25)	0.0688 (5)	0.1693 (5)	0.0170 (16)	0.0361 (41)	0.0970 (10)
22	0.0120 (4)	0.0071 (12)	0.1371 (15)	0.4024 (9)	0.1394 (34)	0.1056 (13)	0.1457 (5)	0.0170 (16)	0.0297 (34)	0.0970 (10)
23	0.0080 (3)	0.0036 (7)	0.1276 (14)	0.4017 (8)	0.1383 (27)	0.0857 (4)	0.1693 (5)	0.0170 (16)	0.0358 (52)	0.0970 (10)
24	0.0120 (2)	0.0036 (9)	0.1311 (13)	0.3976 (9)	0.1385 (3)	0.0658 (3)	0.1809 (5)	0.0170 (16)	0.0324 (7)	0.0970 (10)
25	0.0120 (4)	0.0071 (12)	0.1418 (18)	0.3976 (9)	0.1368 (10)	0.0961 (8)	0.2020 (6)	0.0170 (16)	0.0324 (60)	0.0970 (10)
26	0.0120 (4)	0 (8)	0.1418 (18)	0.3929 (8)	0.1354 (15)	0.0792 (6)	0.1723 (6)	0.0170 (16)	0.0326 (62)	0.0970 (10)

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

27	0.0120 (4)	0.0071 (12)	0.1418 (18)	0.3976 (9)	0.1397 (36)	0.0740 (6)	0.1931 (6)	0.0170 (16)	0.0315 (56)	0.0970 (10)
28	0.0120 (4)	0.0071 (8)	0.1418 (18)	0.3790 (8)	0.1397 (36)	0.0623 (4)	0.1722 (6)	0.0170 (16)	0.0317 (55)	0.0970 (10)
29	0.0120 (4)	0.0036 (9)	0.1311 (12)	0.4024 (9)	0.1392 (26)	0.0701 (7)	0.1990 (6)	0.0169 (16)	0.0352 (52)	0.0970 (10)
30	0.0120 (3)	0 (11)	0.1253 (14)	0.3972 (8)	0.1382 (3)	0.0658 (7)	0.1693 (5)	0.0169 (16)	0.0349 (53)	0.0970 (10)
31	0.0120 (3)	0 (7)	0.1311 (15)	0.4017 (8)	0.1392 (26)	0.0814 (5)	0.1990 (6)	0.0169 (16)	0.0363 (53)	0.0970 (10)
32	0.0120 (4)	0.0036 (10)	0.1418 (18)	0.3981 (8)	0.1397 (36)	0.0736 (10)	0.1662 (6)	0.0169 (16)	0.0372 (62)	0.0970 (10)
33	0.0120 (4)	0.0036 (9)	0.1311 (12)	0.3976 (9)	0.1392 (26)	0.0701 (7)	0.1990 (6)	0.0170 (16)	0.0324 (61)	0.0970 (10)
34	0.0120 (3)	0 (6)	0.1253 (14)	0.3972 (8)	0.1382 (3)	0.0658 (7)	0.1722 (5)	0.0170 (16)	0.0326 (61)	0.0970 (10)
35	0.0120 (3)	0.0036 (7)	0.1311 (15)	0.3976 (9)	0.1392 (26)	0.0814 (5)	0.2170 (6)	0.0170 (16)	0.0326 (61)	0.0970 (10)
36	0.0120 (4)	0.0036 (11)	0.1418 (18)	0.3935 (8)	0.1397 (36)	0.0688 (12)	0.1484 (6)	0.0170 (16)	0.0326 (62)	0.0970 (10)
<i>org</i>	0.0120 (4)	0 (12)	0.1418 (18)	0.3976 (9)	0.1397 (36)	0.1035 (14)	0.1991 (6.9)	0.0170 (16)	0.0326 (62)	0.0970 (10)

Table 3.5: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 36 experimental techniques as well as the mean classification error obtained in the original feature space (*org*) with respect to the *k*NN classifier. Note that each row denotes a specific experimental technique with its technique index indicated in Table 3.1, and each column represents a specific dataset with its dataset index indicated in Table 3.2.

	a	b	c	d	e	f	g	h	i	j
1	0.1360 (3)	0.0324 (3)	0.2600 (3)	0.3649 (5)	0.1366 (5)	0.0238 (6)	0.1841 (6)	0.0143 (9)	0.0281 (9)	0.0061 (9)
2	0.1360 (3)	0.0324 (3)	0.2600 (3)	0.3649 (5)	0.1366 (5)	0.0238 (6)	0.1811 (5)	0.0143 (9)	0.0281 (9)	0.0061 (9)
3	0.0200 (2)	0.0071 (3)	0.2623 (3)	0.3649 (5)	0.1354 (5)	0.0216 (6)	0.1782 (6)	0.0142 (9)	0.0274 (9)	0.0071 (10)
4	0.1320 (1)	0.0324 (3)	0.2600 (3)	0.3649 (5)	0.1366 (5)	0.0229 (5)	0.1783 (6)	0.0143 (9)	0.0281 (9)	0.0061 (9)
5	0.0200 (2)	0.0071 (3)	0.2623 (3)	0.3604 (5)	0.1354 (5)	0.0221 (6)	0.1782 (6)	0.0142 (9)	0.0274 (9)	0.0071 (10)
6	0.0960 (3)	0.0471 (2)	0.2755 (3)	0.3693 (5)	0.1759 (4)	0.0251 (6)	0.1783 (6)	0.0133 (9)	0.0395 (9)	0.0071 (10)
7	0.0240 (2)	0.0325 (3)	0.2706 (3)	0.2803 (5)	0.1324 (5)	0.0277 (6)	0.1809 (6)	0.0174 (9)	0.0281 (8)	0.0071 (10)
8	0.1520 (2)	0.2524 (3)	0.2847 (3)	0.3277 (5)	0.2448 (5)	0.0424 (6)	0.2110 (6)	0.0182 (9)	0.1493 (9)	0.0071 (10)
9	0.0200 (2)	0.0181 (3)	0.2859 (2)	0.3119 (5)	0.1827 (5)	0.0290 (6)	0.1750 (6)	0.0167 (9)	0.1210 (9)	0.0071 (10)
10	0.0400 (3)	0.0720 (3)	0.3108 (2)	0.2994 (5)	0.1543 (5)	0.0290 (6)	0.1961 (6)	0.0188 (9)	0.0388 (8)	0.0071 (10)
11	0.0800 (1)	0.0828 (3)	0.3368 (3)	0.2983 (5)	0.2838 (3)	0.0268 (6)	0.1841 (6)	0.0312 (9)	0.2160 (9)	0.0071 (10)
12	0.1480 (2)	0.0398 (3)	0.2718 (3)	0.3597 (3)	0.1464 (4)	0.0264 (6)	0.1785 (6)	0.0193 (9)	0.0327 (9)	0.0101 (9)
13	0.0160 (3)	0.0577 (3)	0.2824 (3)	0.4121 (5)	0.1357 (4)	0.0286 (6)	0.1755 (6)	0.0185 (9)	0.0320 (9)	0.0101 (9)
14	0.1480 (2)	0.0398 (3)	0.2718 (3)	0.3597 (3)	0.1464 (4)	0.0264 (6)	0.1785 (6)	0.0193 (9)	0.0327 (9)	0.0101 (9)

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

15	0.0160 (3)	0.0290 (3)	0.2896 (3)	0.3271 (4)	0.1489 (5)	0.0238 (5)	0.1726 (6)	0.0166 (9)	0.0301 (9)	0.0101 (10)
16	0.1360 (2)	0.0434 (3)	0.2494 (3)	0.2946 (5)	0.1313 (5)	0.0277 (6)	0.1809 (6)	0.0176 (9)	0.0281 (9)	0.0071 (10)
17	0.0160 (3)	0.0507 (2)	0.2766 (3)	0.2991 (5)	0.1355 (5)	0.0273 (6)	0.1809 (6)	0.0167 (9)	0.0279 (9)	0.0071 (10)
18	0.1360 (2)	0.0434 (3)	0.2494 (3)	0.2946 (5)	0.1313 (5)	0.0277 (6)	0.1809 (6)	0.0176 (9)	0.0281 (9)	0.0071 (10)
19	0.0280 (3)	0.0253 (3)	0.2766 (3)	0.2981 (5)	0.1371 (5)	0.0268 (5)	0.1750 (6)	0.0144 (9)	0.0270 (9)	0.0071 (10)
20	0.0920 (4)	0.0253 (6)	0.1786 (15)	0.3026 (9)	0.1298 (18)	0.0190 (13)	0.1993 (6)	0.0044 (16)	0.0141 (53)	0.0061 (8)
21	0.0160 (4)	0 (9)	0.1750 (11)	0.3071 (9)	0.1293 (19)	0.0165 (13)	0.1814 (6)	0.0042 (13)	0.0148 (44)	0.0061 (9)
22	0.0920 (4)	0.0216 (9)	0.1786 (13)	0.3026 (9)	0.1422 (14)	0.0203 (14)	0.1754 (6)	0.0042 (13)	0.0141 (42)	0.0071 (10)
23	0.0160 (4)	0 (11)	0.1786 (10)	0.3039 (6)	0.1282 (13)	0.0165 (13)	0.1784 (6)	0.0045 (16)	0.0148 (54)	0.0061 (9)
24	0.0240 (2)	0.0146 (5)	0.1891 (8)	0.2846 (6)	0.0883 (35)	0.0212 (12)	0.1814 (5)	0.0052 (13)	0.0101 (52)	0.0051 (8)
25	0.0320 (4)	0.0471 (9)	0.2093 (10)	0.2426 (7)	0.0864 (33)	0.0212 (12)	0.2139 (6)	0.0048 (15)	0.0114 (57)	0.0071 (10)
26	0.0240 (3)	0.0074 (8)	0.2151 (13)	0.2608 (8)	0.0880 (34)	0.0221 (12)	0.2141 (6)	0.0054 (14)	0.0116 (55)	0.0071 (10)
27	0.0320 (3)	0.0433 (8)	0.2140 (9)	0.2937 (9)	0.0912 (36)	0.0238 (9)	0.2111 (5)	0.0041 (14)	0.0101 (45)	0.0030 (9)
28	0.0240 (3)	0.0147 (7)	0.2057 (8)	0.2377 (7)	0.0912 (36)	0.0229 (8)	0.2201 (6)	0.0058 (15)	0.0119 (60)	0.0071 (10)
29	0.0400 (4)	0.0324 (12)	0.1904 (15)	0.3355 (8)	0.1467 (6)	0.0260 (10)	0.1965 (6)	0.0055 (13)	0.0254 (32)	0.0101 (10)
30	0.0280 (3)	0.0250 (9)	0.1690 (9)	0.3139 (4)	0.1448 (6)	0.0255 (6)	0.2084 (6)	0.0052 (12)	0.0247 (14)	0.0101 (10)
31	0.0320 (3)	0.0181 (7)	0.1868 (10)	0.3450 (8)	0.1433 (6)	0.0242 (8)	0.1816 (6)	0.0055 (16)	0.0279 (31)	0.0101 (10)
32	0.0240 (2)	0.0107 (5)	0.2129 (15)	0.3310 (4)	0.1469 (4)	0.0247 (5)	0.1725 (6)	0.0055 (16)	0.0263 (16)	0.0101 (10)
33	0.0320 (4)	0.0503 (7)	0.2009 (8)	0.2937 (9)	0.0883 (33)	0.0208 (12)	0.1904 (6)	0.0050 (14)	0.0101 (45)	0.0051 (9)
34	0.0280 (2)	0.0181 (9)	0.1915 (6)	0.2937 (9)	0.0887 (35)	0.0199 (12)	0.2020 (6)	0.0050 (14)	0.0096 (47)	0.0051 (9)
35	0.0320 (4)	0.0361 (5)	0.2069 (6)	0.2937 (9)	0.0884 (35)	0.0242 (11)	0.1931 (6)	0.0053 (15)	0.0109 (54)	0.0040 (9)
36	0.0320 (3)	0.0180 (5)	0.2056 (8)	0.2422 (7)	0.0912 (36)	0.0264 (13)	0.1722 (6)	0.0062 (16)	0.0098 (53)	0.0071 (10)
org	0.0320 (4)	0.0503 (12)	0.2944 (18)	0.2937 (9)	0.0912 (36)	0.0264 (14)	0.1871 (6.9)	0.0062 (16)	0.0121 (62)	0.0071 (10)

It can be observed from Tables 3.3 to 3.5 that the performance of each DLDR technique varies from both datasets and classifiers. For different classifiers, the minimum mean classification errors achieved by the optimal DLDR techniques on per dataset differ much but are commonly lower than or equal to those obtained in the original feature space without applying the DLDR techniques. Since we only focus on evaluating the performance of the DLDR techniques, such performance difference caused by the distinction of classifiers is out of the scope of our current analysis. In the

context of the same classifier, one DLDR technique can seldom perform best on the majority of the experimental datasets. As a result, it is hard to evaluate the performance of different experimental DLDR techniques merely according to the raw data in these tables. Moreover, these three tables can only show the optimally attainable performance of each DLDR technique on per dataset in terms of different classifiers. In fact, the average classification performance over all possible dimensions that can be reduced to provides another important measure that can reflect the average dimension reduction performance of the DLDR techniques. Therefore, we resort to the previously introduced three relative performance measures, i.e. RBPI, RAPI and ROPI, to evaluate the effectiveness of each experimental DLDR technique. In our experiment, the major objective is to demonstrate the effectiveness of the proposed DLDR techniques in comparison with their archetypes. Therefore, we will not provide a detailed analysis on the performance distinction among different types of DLDR techniques involved in the experiments, e.g. those subject to different conjugated orthogonality constraints, those based on quotient and subtraction criterion, etc., which however deserve to be further investigated as one of our future plans.

The RBPI values of each experimental DLDR technique averaged over all the experimental datasets in the context of the three classifiers are shown as bar charts in Figures 3.4a to 3.4c, respectively. To compare the performance of the enhanced parametric BODLDR techniques with that of their archetypes (if exist) in an explicit manner, we divide all the experimental DLDR techniques into 18 groups with each group consisting of one enhanced technique and its possibly existing archetypes. The group indices and the corresponding indices of its involved enhanced and archetype techniques are given in Table 3.6. A general observation is that many proposed enhanced parametric BODLDR techniques can outperform their corresponding archetypes in terms of the RBPI values with respect to each of the three classifiers. Some remarkable observations are discussed as follows.

1. Among three DLDR techniques requiring the numerical optimization to derive optimal solutions, i.e. ALDR with $Gidx$ 3, DBODLDRho with $Gidx$ 3 and DBODLDRhe with $Gidx$ 12, the proposed DBODLDRho performs better than the

ALDR in terms of the RBPI values with respect to each of the three classifiers. We can also observe from Tables 3.4 to 3.6 that the DBODLDRho demonstrates competitive or better performance than the ALDR on most of experimental datasets with regard to every classifier. On the other hand, the DBODLDRho is outperformed by the DBODLDRhe in terms of the RBPI values with regard to every classifier. This fact can also be verified from Tables 3.4 to 3.6. As described in sections 3.2.2 and 3.4.2, the major distinctions between ALDR and DBODLDRho are that the extracted discriminant vectors are respectively subject to the conjugated orthogonality constraints to $\mathbf{S}_w$ and $\mathbf{I}$, as well as whether the equal class prior assumption is required to be satisfied or not. The major distinction between DBODLDRho and DBODLDRhe is the criterion functions formulated under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. Due to the use of different conjugated orthogonality constraints on the extracted discriminant vectors, we cannot draw the conclusion that the DBODLDRho is superior to the ALDR although it relaxes the assumption of the equal class prior probabilities. On the other hand, we can generally attribute the performance improvement of the DBODLDRhe over the DBODLDRho to the ability of the DBODLDRhe to take into account the data heteroscedasticity since the discriminant vectors extracted by both of them are subject to the conjugated orthogonality constraint to $\mathbf{I}$. Relatively speaking, the Gaussian heteroscedastic assumption stays more closely to real-world datasets than the Gaussian homoscedastic assumption, which may account for the superior performance of the DBODLDRhe among the three techniques.

2. Among Fisher criteria based techniques with *Gidx* 1, 2, 4 to 9, the enhanced DLDR techniques based on the quotient class-wise weighted Fisher criterion, i.e. *Gidx* 1, 2, 4, 6 and 8, demonstrate competitive or better performance than their archetypes in terms of the RBPI values with respect to each of the three classifiers. However, the enhanced DLDR techniques based on the subtraction class-wise weighted Fisher criterion, i.e. *Gidx* 5, 7 and 9, do not demonstrate consistently competitive or higher RBPI values than their archetypes with regard to different classifiers. For example, in terms of the LDC classifier, the F-CWFLDA_{s2} and F-

CWFLDA_{s3} perform worse than their corresponding archetypes FLDA_{s2} and FLDA_{s3}. One plausible explanation is that for the subtraction class-wise weighted Fisher criterion based DLDR techniques, the mismatch problem between the class separability estimated in the original feature space and that estimated in the lower-dimensional subspace to be transformed into is more serious than that encountered by the quotient class-wise weighted Fisher criterion based DLDR techniques. Specifically, the pairwise Bayes accuracies estimated in the S_{ww} -whitened subspace may deviate much from those estimated in the indeed transformed subspace. This is mainly because that the optimization of the subtraction Fisher criteria cannot simultaneously diagonalize the two involved scatter matrices. Since the class-wised weighted Fisher criterion is originally derived to solve the problems associated with the conventional Fisher criterion in the quotient form, direct extensions to the subtraction Fisher criterion as done in this thesis may still require further improvements according to the characteristics of the subtraction criterion. This is not addressed in the current thesis but deserves investigations in future.

3. Among Chernoff criteria based techniques with *Gidx* 10, 11, 13 to 18, the quotient class-wise weighted Chernoff criterion based enhanced DLDR techniques, i.e. *Gidx* 10, 11, 13, 15 and 17, demonstrate better performance than their archetypes in terms of the RBPI values with respect to each of the three classifiers. However, the enhanced DLDR techniques based on the subtraction class-wise weighted Chernoff criterion, i.e. *Gidx* 14, 16 and 18, do not demonstrate consistently higher RBPI values than their archetypes with regard to different classifiers. For example, the F-CWCLDA_{s1} performs worse than the CLDA_{s1} in terms of the *k*NN classifier. The explanation is analogous to that indicated in the subtraction class-wise weighted Fisher criterion based DLDR techniques.
4. Comparatively speaking, several CWFLDA and CWCLDA techniques can demonstrate competitive or better performance than the ALDR, DBODLDRho and DBODLDRhe techniques, respectively, in terms of different classifiers. This fact further validates the effectiveness of the proposed class-wise weighted Fisher

and Chernoff criteria to address the real-world multi-class problems. With respect to the LDC classifier, many class-wise weighted Chernoff criterion based DLDR techniques can demonstrate competitive or better performance than the class-wise weighted Fisher criterion based DLDR techniques. This is because that the number of discriminant vectors extracted by the Fisher criteria based DLDR techniques is bounded from above by $C-1$, which, however, may be insufficient to yield even better performance when the datasets under consideration deviate much from the Gaussian homoscedastic assumption. With respect to the QDC classifier, all the DLDR techniques based on the class-wise weighted Chernoff criterion as well as their archetypes demonstrate better performance than all the DLDR techniques based on the class-wise weighted Fisher criterion as well as their archetypes, respectively. This validates the fact that the discriminatory information due to data heteroscedasticity can be effectively extracted by the DLDR techniques based on the Chernoff criteria. With respect to the k NN classifier, the Chernoff criteria based DLDR techniques demonstrate generally better performance than the Fisher criteria based DLDR techniques, among which $CWCLDA_{q0}$ and $F-CWCLDA_{q0}$ demonstrate the superior performance over all other techniques in terms of the RBPI values.

Table 3.6: The group indices ($Gidx$) and the indices of the enhanced parametric BODLDR techniques ($Eidx$) and their possibly existing archetypes techniques ($Aidx$) within each group.

$Gidx$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$Aidx$	1,2	4	6	8	10	12	14	16	18	20	22	-	25	27	29	31	33	35
$Eidx$	3	5	7	9	11	13	15	17	19	21	23	24	26	28	30	32	34	36

Generally speaking, most of our proposed enhanced parametric BODLDR techniques can improve the performance of their archetypes in terms of the RBPI values with respect to each of the three classifiers. Some exceptions such as those DLDR techniques based on the subtraction Fisher or Chernoff criteria deserve further investigations as one of our future plans.

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.4a

Figure 3.4b

Figure 3.4c

Figure 3.4: The RBPI value averaged over all the experimental datasets for each experimental DLDR technique in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.

Although the RBPI values can evaluate the optimally attainable performance of each experimental DLDR technique averaged over all the experimental datasets, it cannot reflect the average dimension reduction performance of each technique over all possible dimensions that can be reduced to. Therefore, we resort to the RAPI and ROPI values to measure the relative average and overall performance of every technique within each of the two categories of DLDR techniques, namely, the enhanced parametric BODLDR techniques as well as their archetypes under the Gaussian homoscedastic and heteroscedastic assumptions, respectively.

For the 9 enhanced parametric BODLDR techniques under the Gaussian homoscedastic assumption and their 10 corresponding archetypes with technique indices from 1 to 19, the possible dimensions that can be reduced to are specified from 1 to $C-1$. For explicit comparison, we divide these 19 DLDR techniques into 9 groups with each group consisting of one enhanced technique and its possibly existing archetypes. Figures 3.5a to 3.5c show the RAPI value of each DLDR technique within every group averaged over all the experimental datasets in terms of the 3 classifiers,

respectively. Figures 3.6a to 3.6c show the ROPI value of each DLDR technique within every group by using the 3 classifiers, respectively. The group indices and the corresponding indices of its component enhanced and archetype techniques are given in Table 3.7. It can be clearly observed that most of our proposed enhanced parametric DLDR techniques can yield consistently higher RAPI and ROPI values than their existing archetypes with regard to each of the three classifiers, except for those DLDR techniques based on the subtraction class-wise weighted Fisher criterion with $Gidx$ 5, 7 and 9. The possible explanations have been given when we analyze the performance of experimental techniques in terms of the RBPI values. It is also worth noting that the step-wise dimension reduction step adopted by the FS-FLDA and F-CWFLDA_{q0} techniques improves the average dimensionality reduction performance of these techniques relative to their respective counterparts, i.e. FLDA and CWFLDA_{q0}, in terms of the Bayes classifiers, i.e. LDC and QDC. The overall best DLDR technique having the highest ROPI value varies with different classifiers. With respect to the LDC classifier, the CWFLDA_{q0}, F-CWFLDA_{q0} (1st position in parallel), CWFLDA_{q2} and CWFLDA_{q3} (3rd position in parallel) techniques win the first three places. With respect to the QDC classifier, the first three places are taken by the CWFLDA_{q3}, CWFLDA_{q0}, F-CWFLDA_{q0} techniques. The first three winners with respect to the k NN classifier are the CWFLDA_{q0}, F-CWFLDA_{q0} and F-CWFLDA_{s3} techniques.

Table 3.7: The group indices ($Gidx$) and the indices of the enhanced parametric BODLDR techniques ($Eidx$) under the Gaussian homoscedastic assumption and their possibly existing archetypes techniques ($Aidx$) within each group.

$Gidx$	1	2	3	4	5	6	7	8	9
$Aidx$	1,2	4	6	8	10	12	14	16	18
$Eidx$	3	5	7	9	11	13	15	17	19

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.5a

Figure 3.5b

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.5c

Figure 3.5: The RAPI value averaged over all the experimental datasets for each of the enhanced parametric BODLDR techniques under the Gaussian homoscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.

Figure 3.6a

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.6b

Figure 3.6c

Figure 3.6: The ROPI value for each of the enhanced parametric BODLDR techniques under the Gaussian homoscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN.

For the 9 enhanced parametric BODLDR techniques under the Gaussian heteroscedastic assumption and their 8 corresponding archetypes with technique indices 20 to 36, the possible dimensions that can be reduced to are specified from 1 to PC . For explicit comparison, we divide these 17 DLDR techniques into 9 groups with each group consisting of one enhanced technique and its possibly existing archetypes. Figures 3.7a to 3.7c show the RAPI value of each DLDR technique within each group averaged over all the experimental datasets in terms of the 3 classifiers, respectively. Figures 3.8a to 3.8c show the ROPI value of each DLDR technique within each group by using the 3 classifiers, respectively. The group indices and the corresponding indices of its component enhanced and archetype techniques are given in Table 3.8. It can be observed that most of our proposed enhanced parametric DLDR techniques under the Gaussian heteroscedastic assumption can yield consistently higher RAPI and ROPI values than their existing archetypes with regard to each of the three classifiers. The exception happens for those DLDR techniques based on the subtraction class-wise weighted Chernoff criterion with $Gidx$ 5, 7 and 9, where the explanations have been described when we analyze the performance of experimental techniques in terms of the RBPI values. It is also worth noting that the step-wise dimension reduction step adopted by the FS-CLDA and F-CWCLDA_{q0} techniques improves the average dimensionality reduction performance of these techniques relative to their respective counterparts, i.e. CLDA and CWCLDA_{q0}, in terms of the Bayes classifiers, i.e. LDC and QDC. The overall best DLDR technique having the highest ROPI value varies with different classifiers. With respect to the LDC classifier, the F-CWCLDA_{q0}, CWCLDA_{q0} and DBODLDRhe techniques win the first three places. With respect to the QDC classifier, the first three places are taken by the DBODLDRhe, F-CWCLDA_{q0} and F-CWCLDA_{s3} techniques. The first three winners with respect to the k NN classifier are the CWCLDA_{q0}, F-CWCLDA_{q0} and DBODLDRhe techniques.

Table 3.8: The group indices ($Gidx$) and the indices of the enhanced parametric BODLDR techniques ($Eidx$) under the Gaussian heteroscedastic assumption and their possibly existing archetype techniques ($Aidx$) within each group.

$Gnum$	1	2	3	4	5	6	7	8	9
$Aidx$	20	22	-	25	27	29	31	33	35
$Eidx$	21	23	24	26	28	30	32	34	36

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.7a

Figure 3.7b

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.7c

Figure 3.7: The RAPI value averaged over all the experimental datasets for each of the enhanced parametric BODLDR techniques under the Gaussian heteroscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN

Figure 3.8a

Chapter 3: Enhanced Parametric BODLDR Techniques for Multi-class Problems

Figure 3.8b

Figure 3.8c

Figure 3.8: The ROPI value for each of the enhanced parametric BODLDR techniques under the Gaussian heteroscedastic assumption and their archetypes in terms of the 3 classifiers: (a) LDC, (b) QDC and (c) k NN

3.7 Conclusions

In this chapter, we proposed two categories of enhanced parametric BODLDR techniques. Among them, two direct BODLDR techniques, i.e. DBODLDRho and DBODLDRhe, which employ the union of the pairwise Bayes errors or its Chernoff upper bounds over all class pairs as criterion functions, were devised under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. A suboptimal orthonormal transformation matrix can be obtained by numerically optimizing the criterion functions via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme. Two classes of DLDR techniques based on the newly derived quotient and subtraction class-wise weighted Fisher and Chernoff criteria were presented with the involved weighting terms motivated by Gaussian homoscedastic and heteroscedastic Bayes optimality, respectively. The discriminant vectors subject to different conjugated orthogonality constraints can be extracted by the proposed class-wise weighted DLDR techniques in the batch, sequential forward and sequential backward manner, respectively. The effectiveness of the between-class and within-class weighting terms involved in the class-wise weighted Fisher and Chernoff criteria was validated via three artificial datasets subject to the Gaussian homoscedastic or heteroscedastic assumption. The performance of the proposed techniques to address real-world problems was demonstrated on several UCI multi-class datasets and compared with their archetypes in terms of both optimal and average dimensionality reduction performance with regard to three classifiers, i.e. LDC, QDC, k NN. Three relative performance indices, namely RBPI, RAPI and ROPI, were designed to comprehensively evaluate each experimental technique. A general observation was that many proposed techniques could demonstrate better performance than their existing archetypes in terms of the three relative performance indices. Some techniques such as the subtraction criteria based DLDR techniques did not demonstrate consistently better performance and should be further improved as one of our future plans.

Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems

4.1 Introduction

With the rapid advances in sensing and storage technologies and apparatus, many nowadays active research fields, such as text categorization, multimedia retrieval and recognition, remotely sensing, microarray gene expression and so on, need to deal with high-dimensional data, which allow simultaneously observing and measuring patterns of interest from numerous aspects. The increased capacity of the pattern representation will generally bring along more descriptive and discriminatory information but may inevitably introduce some noise, redundant and irrelevant features. In fact, due to the expensive sensing cost and other practical limitations, many such high-dimensional applications can only provide severely fewer training samples relative to the feature space dimensionality. As a result, the *curse of dimensionality* [4], [5] problem may happen due to the limited training sample size, which will lead to the unreliable pattern analysis results or even disable some pattern analysis techniques. This phenomenon is commonly referred to as the undersampled problem, the small sample size problem or the singularity problem in existing literatures [2], [70]. Dimensionality reduction provides a solution to this problem, which can extract the meaningful information from the original high-dimensional feature space so as to facilitate the effective pattern analysis in a compact lower-dimensional space, e.g. visualizing data structures or performing classification.

In the past few decades, a huge amount of research attention from different communities has been attracted to develop novel pattern analysis techniques to handle high-dimensional and undersampled problems, among which supervised classification techniques received prominent emphasis due to the significance of discriminating genuine and imposter in the security control based on biometric data, of diagnosing a

symptom to be benign or malignant in terms of microarray gene expression profiles and of efficiently retrieving relevant multimedia information from large stand-alone or web-based databases. As a necessary and important preprocessing step for the supervised classification, many research efforts have been devoted to the development of the Discriminative Dimensionality Reduction (DDR) techniques, e.g. the Fisher Linear Discriminant Analysis (FLDA) related techniques. Moreover, the DDR techniques may also be used for visualizing class structures in an effective manner to assist the exploratory pattern analysis. However, implementation of some conventional DDR techniques may become infeasible when encountering high-dimensional and undersampled problems. For example, the SMDLDR techniques involving the inversion of scatter matrices may lose their applicability when dealing with image data represented in the vector space. Usually, the vectorization of image data by reading pixel values in the raster-scan manner will result in very high-dimensional image vectors, e.g. images of size 100×100 corresponds to a 10000-dimensional image space, the dimension of which may significantly exceeds the number of training samples. Hence, all scatter matrices in question will become singular in this high-dimensional image vector space, which consequently disables the SMDLDR techniques. Many improvements have been made to address this issue.

In this chapter, we generalize the utility of several representative CWFLDA techniques proposed in the previous chapter, namely $CWFLDA_{q2}$, $CWFLDA_{q3}$, $CWFLDA_{q0}$ and $F-CWFLDA_{q0}$ into the high-dimensional and undersampled scenario. The resulting Complementary Subspaces Based CWFLDA (CS-CWFLDA) techniques and Regularized Subspace Based CWFLDA (RS-CWFLDA) techniques can handle the high-dimensional and undersampled problems by taking advantage of the discriminatory information present in different subspaces, while retaining the superior performance in the well-sampled cases. Specifically, the discriminant vectors subject to different conjugated orthogonality constraints will be either extracted in the sequential forward manner from two complementary subspaces, i.e. the null and column spaces of the average within-class scatter matrix, or extracted in the batch or sequential backward fashion from a single subspace, i.e. the column space of the regularized average with-class scatter matrix, based on the quotient class-wise

weighted Fisher criterion. Note that the pairwise two-class Bayes accuracies of most class pairs may achieve the common maximum values in the high-dimensional and undersampled scenario. Consequently, the weighting terms involved in the quotient class-wise weighted Fisher criterion will mainly depend on the reciprocal of the squared Mahalanobis distances. However, such degenerated weighting terms can still take effect to suppress the undesirable influences due to those classes that are far away from the remaining classes. We have conducted extensive experiments to evaluate the proposed techniques on various types of data sources including handwritten character images, facial images, text documents and microarray gene expression profiles, and compare their performance with that of several state-of-the-art FLDA variants.

The rest of this chapter is organized as follows. Section 4.2 gives a thorough review on the existing DDR techniques capable of handling high-dimensional and undersampled problems. In section 4.3, a class of CS-CWFLDA techniques is presented, which can sequentially extract the discriminant vectors from the null and column spaces of the average within-class scatter matrix without need of regularization. Section 4.4 describes a class of RS-CWFLDA techniques, which allow the discriminant vectors to be extracted in the batch or sequential backward manner from the column space of the regularized average with-class scatter matrix. The experimental results are discussed in section 4.5. Section 4.6 concludes this chapter.

4.2 Review of DDR on high-dimensional and undersampled problems

Traditional pattern analysis techniques may face great challenges in terms of feasibility, time & space complexity and performance when dealing with high-dimensional and undersampled problems. To address these difficulties in the context of supervised pattern analysis, a vast number of DDR techniques have been developed during the past few decades to seek for a most discriminative lower-dimensional space that can best preserve class structures in the original high-dimensional feature space. There exist various ways to categorize the DDR techniques, e.g. linear & nonlinear

methods, single-stage & two-stage methods, projective & manifold methods, etc. In this section, we give a thorough review of the existing DDR techniques feasible in the high-dimensional and undersampled scenario along two avenues: discriminative subspace analysis techniques and discriminative manifold learning techniques.

4.2.1 Discriminative subspace analysis

Subspace analysis [149], [150], [151] covers a wide range of techniques to study on the subspace spanned by a subset of basis vectors of the original feature space, which can retain the useful information as much as possible. The resulting compact lower-dimensional pattern representation will facilitate the subsequent pattern analysis in an efficient and effective manner. The primary goal of discriminative subspace analysis is to seek for an optimal linear transformation to map the original high-dimensional feature space into a lower-dimensional subspace where the class discriminatory information is maximally retained. In general, discriminative subspace analysis techniques include two key components: a criterion function measuring the amount of discriminatory information loss in the transformed subspace and an optimization method to seek for the optimal linear transformation with respect to the chosen criterion. Three most representative classes of discriminative subspace analysis techniques: FLDA, Discriminative Locality Preserving Projection (DLPP) and Supervised Projection Pursuit (SPP) are reviewed in the following:

The **FLDA** techniques have a very long history yet still remain popular till now due to their easy-to-use nature, analytical solutions and encouraging empirical performance. To overcome a major shortcoming associated with the conventional FLDA, i.e. the infeasibility of dealing with undersampled problems, many modifications and extensions have been made to the conventional FLDA technique during the past few decades. In the following, we sweep some representative FLDA variants.

Regularized FLDA (RFLDA): The common idea behind the RFLDA techniques is to regularize some singular or nearly singular scatter matrices so as to augment their excessively small eigenvalues and accordingly enable and stabilize their inversions.

Many regularization methods [136]-[142] have been developed, such as singular value perturbation [136], shrink towards the identity matrix [137], [138], addition with a scaled identity matrix [139], [140], multiple parameter regularization [141] and maximum entropy covariance selection [142]. One limitation of the RFLDA techniques is that some parameters used to control the degree of regularization are hard to determine. In practice, the cross-validation can be used for estimating these parameters at the expense of the high computational cost.

Pseudo FLDA (PFLDA): Pseudo-inverse [92] is commonly used to deal with the problem of inversion of singular matrices. By substituting the common inverse involved in the eigenvalue decomposition problem of the FLDA with the pseudo-inverse, a class of PFLDA techniques comes out, which naturally overcomes the singularity problem. The PFLDA techniques have been studied for many years [152], [156]. Recent work [157] provided a theoretical insight into the PFLDA and unveiled its equivalence with some two-stage FLDA techniques under certain conditions.

Two-stage FLDA (TFLDA): The TFLDA techniques overcome the singularity problem via an intermediate dimension reduction stage, in which the original feature space is first mapped into a compact lower-dimensional space by some dimensionality reduction techniques not suffering from the singularity problem, such as PCA, kernel PCA, LSI, partial least square, QR decomposition and descriptive manifold learning¹³ etc. Then, the conventional FLDA is applied to further map this intermediate space to a lower-dimensional discriminative space. The earliest work related to the TFLDA can be traced back to Cheng *et al.* [158], in which only the two-class case is considered. Although some research work continued after that, e.g. [159], [160], the great popularity of such class of techniques was gained after the work by Pelhumeur *et al.* [161], in which a two-stage PCA+LDA technique, commonly referred to as “Fisherface”, was successfully applied to face recognition. Theoretical justifications for the PCA+LDA technique were presented by Yang *et al.* [162], [163]. Recent work [156], [157] demonstrated that some TFLDA techniques were theoretically equivalent to the PFLDA techniques when the scatter matrix to be inverted in the PFLDA was

¹³ Strictly speaking, the intermediate space obtained by the descriptive manifold learning techniques may not be a subspace of the original feature space.

properly defined. Note that all the kernelized FLDA techniques can be regarded as two-stage methods, in which the kernel PCA is first carried out to reduce the feature space dimensionality before applying the FLDA. The theoretical proof was given in [163].

Null space based FLDA (NFLDA): The common idea behind the NFLDA techniques is to extract the discriminatory information in the null space of the average within-class scatter matrix, and accordingly can implicitly avoid the singularity problem. The earliest work on the NFLDA can be traced back to Krzanowski [70], in which the NFLDA was developed under the name zero-variance discrimination. Pelhumeur *et al.* [161] also briefly mentioned the similar idea but without carrying out any further analysis. The great popularity of the NFLDA technique was gained after the work by Chen *et al.* [164], in which sufficient analysis and experiments was provided to justify this technique. Huang *et al.* [165] further improved its efficiency by removing the null space of the total scatter matrix before applying the NFLDA, based on the fact that the null space of the total scatter matrix is the intersection of the null space of the between-class scatter matrix and the null space of the average within-class scatter matrix. Therefore, such NFLDA is actually executed in the discriminative null space of the average within-class scatter matrix. Zheng *et al.* [166] developed the uncorrelated NFLDA (NUFLDA) technique to generate the uncorrelated features. The Discriminant Common Vectors (DCV) method, proposed by Cevikalp *et al.* [167], belongs to the family of the NFLDA techniques, which can provide an efficient way to obtain the discriminant vectors from the null space of the average within-class scatter matrix. Such technique in fact had been mentioned in [70] without the thorough analysis.

Direct FLDA (DLDA): The DLDA technique, proposed by Yu *et al.* [168], assumes that the null space of the between-class scatter matrix contains no discriminatory information. Accordingly, the optimal solution can be directly derived by determining the eigenvalue decomposition of the average within-class scatter matrix estimated in the column space of the between-class scatter matrix. Such technique has attracted considerable interests mainly because of its computational efficiency and encouraging

performance on some example datasets. Several extensions of the DLDA have been developed [169]–[171] in recent years. However, the theoretical deficiencies of the DLDA, as studied in a recent paper [172], challenge some merits of this technique advocated by its inventors. In fact, the DLDA technique only takes into account the discriminatory information present in the difference of class means estimated in the original feature space and totally ignores the within-class scatter. This is clearly different from the idea behind the conventional FLDA. As analyzed in [173], in undersampled problems, the insufficient number of training samples per class may significantly distort the estimate of the individual covariance matrices and consequently may possibly make the performance of FLDA worse than that of PCA. This fact may possibly provide a reasonable explanation for the promising empirical performance of the DLDA on some high-dimensional and undersampled datasets.

Generalized FLDA (GFLDA): The singularity problem encountered by the conventional FLDA essentially originates from the requirement of solving a generalized eigenvalue decomposition problem involving a singular scatter matrix. The optimization of a family of generalized Fisher criteria based on the pseudo-inverse [156] can avoid the matrix inversion operation by using the Generalized Singular Value Decomposition (GSVD) method [156], [174] or the simultaneous diagonalization scheme [124], and thus naturally solve the singularity problem. The resulting FLDA/GSVD technique [156] has been demonstrated to be equivalent to the PFLDA, where the pseudo-inverse is applied to the total scatter matrix. The simultaneous diagonalization scheme has resulted in the so-called uncorrelated linear discriminant analysis and orthogonal linear discriminant analysis techniques [124], [175], with the latter one demonstrated to be equivalent to the NFLDA technique under a mild condition [176].

Dual-space FLDA (DS-FLDA): Wang *et al.* [177] proposed a dual-space FLDA technique, which takes full advantage of the discriminatory information present in two complementary orthogonal subspaces, i.e. the null and column spaces of the average within-class scatter matrix. Based on a probabilistic visual model [178], the average eigenvalue in the null space of the average within-class scatter matrix is estimated.

The two sets of features, extracted by applying discriminant analysis separately in these two complementary orthogonal subspaces, are combined in a weighted manner for final classification, with the weighting factor being the estimated average eigenvalue. A similar technique was presented by Yang *et al.* [163], in which the original feature space is first mapped into a subspace by PCA, and then two sets of features are extracted separately from the null and column spaces of the average within-class scatter matrix estimated in this subspace. Finally, these two sets of features are fused in a weighted manner with the weighted factor determined by users. Although demonstrating competitive classification performance, the dual-space FLDA technique is not suitable for further dimensionality reduction since the full set of features extracted from each of the two subspaces will be used in the classification without considering the possibility of further condensation.

Random sampling based FLDA (RS-FLDA): Wang *et al.* [179] proposed a random sampling based FLDA technique, in which random subspace and bagging methods are used to improve the stability and performance of the FLDA and NFLDA techniques, respectively. Specifically, the random subspace method creates multiple subspaces with small dimensions, in which the FLDA can be applied without encountering the singularity problem. The bagging method creates multiple training sample sets with small sizes, on which the existence of the null space can be guaranteed such that the NFLDA can be carried out. The discriminative features extracted by the two kinds of complementary FLDA techniques are integrated via a fusion rule, and thus most of the discriminatory information can be preserved. This method provides a general framework in which other FLDA variants can replace the FLDA or NFLDA.

Unified subspace analysis (USA): The USA framework, presented by Wang *et al.* [180], is build upon a face difference model, which decomposes the face difference into three components, i.e. intrinsic difference, transformation difference and noise. The inherent connection among three popular subspace methods: PCA, Bayesian matching and FLDA, and their unique contributions to the extraction of the discriminatory information related to the intrinsic difference are analyzed. This eventually results in a 3D parameter space, in which the dimensions of the subspaces

spanned by the above three popular subspace methods are used as axis. Searching through this parameter space, improved subspace analysis techniques can be obtained.

Non-parametric discriminant analysis (NDA): All aforementioned techniques commonly operate on global scatter matrices that can be anyway estimated regardless of the underlying class distributions. However, the best performance in terms of the Bayes optimality can only be achieved only if the Gaussian homoscedastic assumption is satisfied. In fact, when the class distribution deviate much from Gaussian, the performance of these techniques may deteriorate since most useful discriminatory information can no longer be fully characterized by global scatter matrices. To address this problem, many NDA techniques have been developed, which reformulated the Fisher criterion by using local scatter matrices as described in subsection 2.2.2. The earliest work on NDA can be traced back to Fukunaga [2], which formulated local between-class and local average within-class scatter matrices for the two-class case. To enable the local between-class scatter matrix to more accurately specify a subspace in which the decision boundary regions are embedded, a heuristic weighting function was introduced to its formula such that feature vectors far from the decision boundary can exert little influence on the scatter matrix estimate. Based on the modified Fisher criterion as a function of local scatter matrices estimated in the transformed subspace, the optimal linear transformation matrix could be obtained. It is worth noting that, in the two-class case, when the local mean from the other classes coincides with the class mean of the actual class, the NDA will come down to the FLDA [2]. Although devised for two-class problems, the NDA can be easily extended into the multi-class scenario. Recently, an intimate relation between the NDA and the k -NN classifier in the transformed subspace has been disclosed [181]. Lee *et al.* [111] further developed the concept of NDA by directly constructing a decision boundary feature matrix in terms of norm vectors to the efficient decision boundary. Theoretically speaking, the subspace spanned by the leading eigenvectors of the decision boundary feature matrix can unveil the intrinsic discriminant dimensionality and retain all the discriminatory information in the original feature space. However, since the characteristics and complexity of the decision boundary usually depend on the employed classifier and the given problem, the optimal effective decision boundary may be highly complex or

even cannot be constructed, which consequently prevent the applicability of this method in practice. Recent attempt based on the SVM classifier [112] provides a promising way out for such method. In fact, from the perspective of practicability, the NDA techniques based on local weighted scatter matrices can be regarded as the simplification of the decision boundary feature matrix based NDA techniques. Specifically, the local average weighted between-class scatter matrix via the proper choice of the weighting function can approximate the decision boundary feature matrix. Motivated by this observation, a more effective weighting function was introduced in [182].

Two-dimensional FLDA (2DFLDA): Conventional FLDA related techniques work in the vector space, where patterns are presented as feature vectors. However, vectorization representation may not be a desirable choice for 2-D patterns such as images or maps since the resulting high-dimensional vector space will always lead to the singularity problem. To tackle the singularity problem, the subspace based FLDA variants described above typically involve the eigenvalue decomposition of matrices of huge size and thus significantly increase the time and memory costs. Furthermore, vectorization of 2-D patterns along either horizontal or vertical direction may destroy the correlation among either rows or columns of 2-D patterns, and thus lose certain amount of the discriminatory information. The recently proposed 2DFLDA techniques [183], [184], [186], [187] resolve these problems in which patterns are directly represented as matrices. The estimated scatter matrices based on the matrix pattern representation are of the size small enough to avoid the singularity. Developed from its preliminary version using unilateral projections, the latest 2DFLDA technique uses bilateral projections, which can take into account the correlation among both columns and rows of 2-D patterns and consequently much more discriminatory information can be retained. Note that 2DFLDA techniques essentially belong to the tensor subspace analysis [185] of order two. In order to further reduce the time and space costs, the 2DFLDA can be followed by the conventional FLDA to further reduce the dimensionality [183], [187]. Also, kernel tricks can be incorporated to handle the nonlinear pattern distributions [183]. In [186], the unilateral 2DFLDA is preceded by

the unilateral 2DPCA with a threshold on the Fisher criterion value to determine the intermediate and ultimate reduced dimensions.

The recently developed **DLPP** techniques based on the locality projection pursuit [188] attempt to find an optimal linear transformation matrix that can best preserve the local structure information while simultaneously increase the class separation in the lower-dimensional transformed subspace so as to make easier the subsequent supervised pattern analysis. The class label information can be incorporated either in the construction of the neighborhood graph or in the formulation of the criterion function. Chen *et al.* [189] proposed a local discriminant embedding technique and its kernel as well as 2D variants. This technique separately constructs the intra-class and inter-class affinity weight matrices and formulates a constraint optimization criterion aiming at bringing the neighboring feature vectors closer if they belong to the same class whereas preventing the feature vectors from different classes entering the same neighborhood. The kernel and 2D extension can be readily extended via the kernel trick [12] and the two-order tensor subspace analysis technique [185]. The locally discriminating projection, presented by Zhao *et al.* [190], modifies the affinity weight matrix defined in the LPP by considering the class label information. Specifically, the original affinity weight between two neighboring feature vectors will be artificially increased by multiplying with an intra-class discriminating weight if both feature vectors belong to the same class. Otherwise, the local weight will be decreased via the multiplication of an inter-class discriminating weight. The criterion function is defined same as that of the LPP but taking the distinct Laplacian matrix. Yu *et al.* [191] developed a discriminative locality preserving projection technique applied to face recognition. They formulate a new criterion function that can simultaneously preserve the intra-personal local structures while maximize the inter-personal variation defined by the mean face in each class. There exists intimate relation between the DLPP and FLDA techniques [192]. By certain definition of the affinity weight matrix and the criterion function, the DLPP technique can come down to the FLDA. More interestingly, some DLPP techniques can be interpreted by non-parametric weighted discriminant analysis, which may disclose the underlying theoretical foundation of such class of heuristic but effective non-parametric techniques. The DLPP also bears

strong link to the discriminative manifold learning techniques that will be described later since one of its ultimate goals is to preserve local structures of the manifold embedded in the ambient high-dimensional space in the transformed lower-dimensional subspace. The major distinction between DLPP and discriminative manifold learning is that DLPP finds a linear subspace that approximates the global geometry of the intrinsic manifold while discriminative manifold learning usually characterizes the underlying manifold via the locally linear but globally nonlinear mapping.

The **SPP** techniques [193] search for “interesting” low-dimensional linear projections by numerically optimizing certain criterion functions defined in the projected subspace, called the projection pursuit index, which can measure the interestingness of the projection. In the context of supervised pattern analysis, “interestingness” usually means good discrimination among classes and thus the projection pursuit index can be defined as a function of the linear projection matrix that characterizes class separability in the projected subspace. Although similar to the FLDA in terms of optimizing certain criterion functions defined in the projected subspace, the major distinction between SPP and FLDA lies in: the SPP starts from a pre-specified linear projection matrix and continues to seek for better solutions based on the currently obtained one by numerically optimizing certain criterion functions defined in the projected subspace, while the FLDA directly performs analytical optimization in terms of some statistic variables calculated in the original high-dimensional space, e.g. scatter matrices. Consequently, the applicability of the SPP remains well in the undersampled scenario since all the parameters involved in the computation are estimated in the projected subspace instead of in the original high-dimensional feature space. The projection pursuit index and the corresponding optimization method constitute two critical factors influencing a SPP technique. During the past few decades, many SPP techniques with different indices and optimization methods have been proposed.

According to their relationships to the subsequent classifier, projection pursuit indices can be dichotomized into classifier-independent and classifier-dependent ones. The

classifier-independent indices evaluate the class separability based on the characteristics of the training samples without considering the subsequent classifier. Accordingly, it may result in the inconsistency between dimensionality reduction and classification performance but may somewhat avoid overfitting. Santa Cruz and Dorronsoro [194] presented a nonlinear supervised feature extraction algorithm that combines the classifier-independent Fisher criterion with the perceptrons-like nonlinear projection. It first utilizes multilayer perceptrons to nonlinearly map the original feature space into a new space, and then the Fisher criterion is used as the projection pursuit index to derive an optimal projection from the current space to the final subspace. Lee *et al.* [195] proposed a group of new projection pursuit indices by extending the L_2 norm involved in the distance calculation of both between-class separation and within-class closeness to the L_r norm. The new indices L_r taking different values of r can result in projections with different characteristics. For example, with the L_1 index, one may obtain a projection robust to outliers. As r is increased, the L_r index tends to be more sensitive to outliers since the well-separated class pairs will be over-emphasized during the optimization. Different projected subspaces induced by the L_r index taking distinct values of r can be screened to better understand the class structures or to facilitate the exploratory supervised pattern analysis. Zhu *et al.* [109] developed a generalized log-likelihood ratio index to extend the Fisher criterion into arbitrary class distributions. A powerful non-parametric discriminant analysis technique was subsequently proposed by integrating such index with the projection pursuit density estimation. The geometric LDA technique, proposed by Ordowski and Meyer [199], uses the geometry of the contours of the constant class densities in the projected subspace to define the projection pursuit index. Kaskie and Peltonen [200] proposed an informative discriminate analysis technique, which employs as the projection pursuit index the log-likelihood of the generative probabilistic model of the class labels in the projected subspace. Such index is closely related to mutual information, Renyi mutual information, Fisher criterion and PCA reconstruction criterion defined in the learning or Fisher metrics. Some techniques reviewed in section 2.2 can also be categorized into the SPP methods with class-independent indices. The classifier-dependent indices directly take into consideration the performance of a specific classifier formulated as a function of the

projection matrix and classifier parameters, which result in a popular class of SPP methods using the classification error as indices, also referred to as the minimum classification error learning algorithm [196]-[198]. Such indices can be specifically formulated in the context of various classifiers [201]-[207], e.g. Regression based classifier, Mahalanobis distance based classifier, Gaussian Maximum likelihood classifier, k -nearest neighbor classifier, generalized learning vector quantization and neural networks based classifier, etc. The linear projection matrix, as an additional parameter in the index, will be jointly optimized along with the parameters of the classifier. In some cases, the classification error is calculated based on the projected training samples, which may result in a projected subspace with very low training error but unsatisfactory generalization performance due to overfitting. For example, Flick *et al.* [201] used projection pursuit regression to model a desired response (class) as a sum of ridge functions according to the minimum expected squared error as the projection pursuit index. Choi and Lee [203] proposed an algorithm for multi-class feature extraction and classification that searches for multiple projection directions in either sequential or batch manner to minimize the classification error on training samples by the Gaussian Maximum Likelihood classifier. To improve the generalization capability, additional factors related to the generalization performance can be combined with the training error to constitute the projection pursuit index, e.g. genetic algorithm based dimensionality reduction [208] and evolutionary pursuit [209]. In other cases, the classification error is defined over the whole projected subspace instead of over the limited number of projected training samples, and thus good generalization performance can be expected. For instance, many BODLDR methods reviewed in section 2.2 employ indices directly or indirectly related to the minimum probability of error achieved by the Bayes classifier, which cover the entire projected space. A drawback of such methods is that they will resort to the parametric or non-parametric class density estimation, which generally requires a large number of training samples and possibly leads to the high computational cost.

Many optimization techniques have been used to derive the optimal projection with respect to a chosen index, e.g. random search, gradient descent, stochastic approximation, structure removal, simulated annealing and evolutionary algorithms,

etc. A major problem of the SPP techniques is that the possibility of getting stuck into local optimal solutions as well as the computation & storage complexity will increase dramatically with the increment of the original feature space dimensionality. To mitigate such difficulty, some reasonable initialization schemes [207] of the projection matrix can be considered instead of a purely random choice, e.g. projections obtained by the FLDA can serve as good starting solutions for the SPP methods. Or one can resort to the two-stage dimensionality reduction methods, which first project the original high-dimensional feature space into a intermediate lower-dimensional subspace by some simple descriptive or discriminative dimensionality reduction techniques such as PCA or FLDA, then followed by the SPP techniques to further map the intermediate subspace to the eventually projected subspace. It is worthwhile to note that the two-stage approach also provides an effective way to generalize the conventional SPP techniques into their nonlinear counterparts. Specifically, in the first stage, the original feature vectors can be nonlinearly mapped into a new feature space, e.g. via kernel tricks or multiplayer perceptrons, where the complexity of class distributions may possibly be reduced. In the second stage, the SPP techniques can be carried out to project the new feature space obtained in the first stage to a lower-dimensional subspace.

4.2.2 Discriminative manifold learning

It is generally believed that high-dimensional data generated from many kinds of applications, such as facial and character images, lie on or nearly on the intrinsically lower-dimensional manifold embedded in the ambient high-dimensional space and governed by a few principal factors. Therefore, modeling manifold in the low-dimensional space can effectively reduce the feature space dimensionality without loss of essential information underlying the available feature vectors. Note that since many manifold learning algorithms perform only on a fixed set of training samples and return their corresponding lower-dimensional representations, there does not exist an explicit mapping from the original feature space into the obtained low-dimensional embedding space for newly presented data. To address this problem, some non-parametric and parametric methods have been employed [210] to construct the

mapping based on manifold learning results on the training samples. Recent years have seen rich progresses in manifold learning methods, which can be roughly dichotomized into: descriptive and discriminative ones. This thesis will only focus on the discriminative manifold learning methods, which differ from the descriptive manifold learning approaches in taking into account the class label information to guide the manifold learning. According to ways of incorporating the class label information into learning the manifold, the discriminative manifold learning methods can be divided into single-stage and two-stage approaches. Several representative examples in each category are described as follows.

Single-stage approaches use the class label information to artificially modify the Euclidean distances between two original feature vectors belonging to the same or different classes, which serve as building blocks to construct the neighborhood structure of the training samples in the original feature space. Some examples are briefly described as follows.

Supervised Locally Linear Embedding (SLLE) was introduced to handle labeled the training samples that possibly contain multiple manifolds corresponding to different classes. If the manifold associated with each class is fully disjoint with each other, the local neighborhood of a feature vector in one class should be totally composed of the feature vectors belonging to the same class. In practice, however, manifolds of different classes may overlap to some degree. Consequently, the local neighborhood of some feature vectors may contain the feature vectors from different classes. To yield good class separation in the low-dimensional embedding space so as to facilitate the supervised pattern analysis, the SLLE artificially increases the pre-calculated distance metric (e.g. Euclidean distance) between any two feature vectors belonging to different classes while remains unchanged those distances between the feature vectors within the same classes. Some SLLE algorithms [211] modify this distance metric before establishing the neighborhood relationship via the formula $\mathbf{D}' = \mathbf{D} + \alpha \cdot \max(\mathbf{D}) \cdot \mathbf{\Lambda}$, where $\mathbf{D}$ and $\max(\mathbf{D})$ represent a distance matrix composed of distances between any two original feature vectors and its maximum entry, respectively. The matrix $\mathbf{\Lambda}$ has binary entries with “1” and “0” denoting two feature

vectors with the corresponding matrix column and row indices belong to different and same classes, respectively. The parameter α controls the amount of class label information that is incorporated into the local neighborhood formation. When $\alpha = 0$, the conventional unsupervised LLE is obtained. When $\alpha = 1$, a fully supervised LLE is devised where nearest neighbors for a certain feature vector will definitely be selected from its own class and thus this method is actually α free. Varying α between 0 and 1 can generate a class of partially supervised LLE techniques, which attempt to increase the class separation while preserve the intrinsic manifold structure to some extent and thus may result in the better generalization performance. In contrast, the SLLE algorithm proposed by Bai *et al.* [212] performed the distance metric modification after establishing the neighborhood relationship. The distances from a specific feature vector to its neighbors belonging to the different class will be increased. Ridder *et al.* [213] presented a local Fisher embedding algorithm called Fisher LLE, which modified the key matrix formulated in the LLE embedding stage by combining both local geometric and global class information. This method has the flexibility to adjust the amount of supervision.

The weighted Isotropic Feature Maps (ISOMAP) algorithm, proposed by Vlachos *et al.* [214], rescales with a constant factor λ ($\lambda < 1$) the Euclidean distances between any two feature vectors belonging to the same class when constructing the neighborhood graph. Consequently, the feature vectors within the same class in the original feature space are forced to be closer in the obtained low-dimensional space where the classification becomes easier. However, Geng *et al.* [215] indicated that the weighted ISOMAP algorithm could be much sensitive to noise, sensitive to the proper choice of parameter λ and thus not suitable for the supervised visualization tasks. They accordingly proposed a supervised ISOMAP algorithm, in which the neighborhood graph is constructed according to a newly defined dissimilarity measure,

$$\text{i.e. } D(\mathbf{x}, \mathbf{y}) = \begin{cases} \sqrt{1 - e^{-d^2(\mathbf{x}, \mathbf{y})/\beta}} & \mathbf{x} \text{ and } \mathbf{y} \text{ belong to the same class} \\ \sqrt{e^{-d^2(\mathbf{x}, \mathbf{y})/\beta}} - \alpha & \mathbf{x} \text{ and } \mathbf{y} \text{ belong to different classes} \end{cases}, \quad \mathbf{x}, \mathbf{y} \in \mathbf{R}^{n \times 1}. \quad \text{The}$$

parameters β and α are used to control the noise influence and avoid the overfitting as well as the possibly occurred disconnected neighborhood graph, respectively. Such

dissimilarity measure, as two monotonically increasing functions of the Euclidean distance between the feature vectors belonging to the same and different classes respectively, can preserve intrinsic geometrical and topological structures underlying training samples and are capable of effectively identifying and controlling the noise influence.

Two-stage approaches firstly map in an unsupervised manner the original feature space into an intermediate feature space where the geometric characteristics of the manifold are maximally retained. Then discriminative subspace analysis techniques are carried out in such intermediate feature space to seek for an optimal linear transformation. We hereby briefly mention some representative examples. The MLA+FLDA technique, proposed by Zhang *et al.* [216], starts with a manifold learning algorithm (MLA) where the LLE is used to find the low-dimensional embedding of the training samples in the original high-dimensional feature space. The corresponding mapping is learnt via Gaussian RBF kernel functions according to the Weierstrass approximation theory. The FLDA is followed to further map the embedding space into its lower-dimensional discriminative subspace in order to improve the class separation. Zhang *et al.* [217] also proposed a unified algorithm ULLELDA based on LLE and FLDA techniques. Different from the MLA+FLDA, the newly presented test samples can be directly mapped into the ultimate discriminative space in an efficient way that avoids the time-consuming calculation of their locations in the embedding space. It is worth noting that the manifold learning algorithm, in the MLA+FLDA or ULLEFLDA framework, does not restrict to the LLE algorithm. Other popular manifold learning algorithms such as ISOMAP, diffusion maps, local tangent space alignment and Laplacian eigenmaps can be readily incorporated into this framework. Yan *et al.* [218] introduced a novel technique to perform discriminant analysis in terms of the embedded manifold structures. Firstly, an intra-cluster balanced k -means (ICBKM) algorithm is developed to obtain a number of clusters with balanced samples with respect to each class therein. Then, local discriminative features for each cluster are simultaneously calculated according to a global weighted Fisher criterion. The obtained nonlinear mapping function can easily map any newly presented test sample into the low-dimensional space. This method can be regarded as

a locally linear yet globally nonlinear discriminant analyzer with the geometry-adaptive kernel. Extended ISOMAP algorithm, developed by Yang [219], maps the original feature space into a new space where each training sample is represented as a vector of geodesic distances from it to all the training samples. The FLDA is executed subsequently to find an optimal transformation from the new space to its lower-dimensional subspace. For a test sample, geodesic distances from it to all the training samples are calculated to form the new feature vector. Then, the already obtained FLDA transformation is used to map it into the ultimate low-dimensional space. A similar work named discriminative manifold learning was proposed by Wu *et al.* [220]. Different from the extended ISOMAP, a novel CK-neighborhood graph construction scheme is devised to calculate the geodesic distance matrix, and the relevant component analysis [221] replaces the FLDA to find the optimal transformation from the new feature space to the ultimate lower-dimensional space.

4.3 Complementary subspaces based CWFLDA techniques

Due to easy-to-use nature, closed-form solutions and encouraging empirical performance to facilitate supervised pattern analysis, the FLDA related techniques remain a most popular and active research topic for dimensionality reduction. Review given in the previous section highlights their prominent roles in solving high-dimensional and undersampled problems: modifications and extensions of the conventional FLDA constitute an important class of discriminative subspace analysis techniques, provide good initializations for the SPP and play a crucial role in the two-stage discriminative manifold learning.

Many CWFLDA techniques proposed in the previous chapter have shown superior performance over their archetypes based on the conventional Fisher criteria when dealing with well-sampled datasets, in which the training sample size much exceeds the dimensionality of the feature space. Therefore, it is natural to evaluate their performance in the high-dimensional and undersampled scenario. However, singular scatter matrices encountered in high-dimensional and undersampled problems will

disable the utility of these techniques: the between-class weight and the within-class weight involved in the class-wise weighted Fisher criteria are functions of the Mahalanobis distance between two classes, which requires the average within-class scatter matrix to be non-singular. Moreover, optimization of the quotient Fisher criterion will lead to solve an eigenvalue decomposition problem involving the inverse of a scatter matrix, and accordingly requires the non-singularity of this scatter matrix. A simple way to solve the singularity problem is to resort to the two-stage dimensionality reduction, in which the original feature space is first mapped into a lower-dimensional space without suffering from the singularity problem, and then the CWFLDA is applied therein. However, certain amount of the useful discriminatory information can be undesirably discarded at the intermediate dimensionality reduction stage. On the other hand, many empirical results and some theoretical analysis [164]-[167] have highlighted the role played by the null space of the average within-class scatter matrix, i.e. $N(\mathbf{S}_w)$ in class discrimination. In fact, a large amount of useful discriminatory information resides in such subspace. Therefore, we can consider two complementary orthogonal subspaces with respect to the scatter matrix $\mathbf{S}_w$, i.e. its null space $N(\mathbf{S}_w)$ and column space $C(\mathbf{S}_w)$, and apply the CWFLDA in both $N(\mathbf{S}_w)$ and $C(\mathbf{S}_w)$. We name such class of techniques Complementary Subspaces Based CWFLDA (CS-CWFLDA) since the discriminatory information is extracted from two complementary subspaces. Some general characteristics on the null and column spaces with respect to a positive semi-definite square matrix are described in Theorem 4.1.

Theorem 4.1 Assume that $\mathbf{A} \in \mathbf{R}^{n \times n}$ is a positive semi-definite symmetric matrix with $\text{rank}(\mathbf{A}) = r$ and let $N(\mathbf{A})$ and $R(\mathbf{A})$ denote the null space and the column spaces of $\mathbf{A}$, respectively. Then, (1) $\mathbf{R}^n$ is a direct sum of the two subspaces $N(\mathbf{A})$ and $R(\mathbf{A})$, i.e. $\mathbf{R}^n = N(\mathbf{A}) \oplus R(\mathbf{A})$, with $\dim(R(\mathbf{A})) = r$ and $\dim(N(\mathbf{A})) = n - r$; (2) $N(\mathbf{A})$ and $R(\mathbf{A})$ are complementary orthogonal subspaces; (3) If $\mathbf{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n]$ consist of the eigenvectors of $\mathbf{A}$ corresponding to eigenvalues in the non-ascending order $\lambda_1 \geq \dots \geq \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_n$, then $R(\mathbf{A}) = \text{span}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_r\}$ and $N(\mathbf{A}) = \text{span}\{\boldsymbol{\beta}_{r+1}, \dots, \boldsymbol{\beta}_n\}$.

Proof. See [92].

It has been demonstrated [162], [176] that the null space of the total scatter matrix $N(\mathbf{S}_T)$ is the intersection of the null space of the between-class scatter matrix $N(\mathbf{S}_B)$ and the null space of the average within-class scatter matrix $N(\mathbf{S}_w)$, and thus contains no discriminatory information. Therefore, removal of $N(\mathbf{S}_T)$ before determining $N(\mathbf{S}_w)$ and $C(\mathbf{S}_w)$ can result in a much compact discriminative null space $N^d(\mathbf{S}_w)$ retaining all the discriminatory information in $N(\mathbf{S}_w)$. In the following discussion, we assume that $N(\mathbf{S}_T)$ has been removed, with $N^d(\mathbf{S}_w)$ and $C^d(\mathbf{S}_w)$ estimated in the column space of the total scatter matrix $C(\mathbf{S}_T)$. According to Theorem 4.1, it is easy to verify that $\dim(N^d(\mathbf{S}_w)) = \dim(C(\mathbf{S}_T)) - \dim(C^d(\mathbf{S}_w))$.

The CWFLDA techniques can be applied in $N^d(\mathbf{S}_w)$ and $C^d(\mathbf{S}_w)$ separately without suffering from the singularity problem. However, the discriminant vectors extracted from each of these two subspaces may not simultaneously satisfy some conjugated orthogonality constraint. For example, the CWFLDA may extract a set of uncorrelated features from each subspace but cannot ensure the two sets of extracted features to be uncorrelated to each other. As a result, some features extracted from one subspace may be highly correlated with some features extracted from the other subspace, which will lead to a less compact pattern representation. To address this issue, we choose the CWFLDA_{q2} technique proposed in the previous chapter and generalize it to the Complementary Subspaces Uncorrelated CWFLDA (CS-UCWLDA) technique, in which the CWFLDA_{q2} is first applied in $N^d(\mathbf{S}_w)$ to sequentially extract a set of discriminant vectors subject to the conjugated orthogonality constraint to $\mathbf{S}_T$ and then applied in $C^d(\mathbf{S}_w)$ to sequentially extract another set of discriminant vectors under the constraint that any newly extracted discriminant vector together with all the previously extracted discriminant vectors should satisfy the conjugated orthogonality to $\mathbf{S}_T$. Consequently, the eventually obtained discriminant vectors from the two subspaces can result in a set of uncorrelated features in the transformed subspace. The pseudo-code of the CS-UCWLDA technique is given in Table 4.1.

Table 4.1: Pseudo-code of the CS-UCWFLDA technique

-
- Remove $N(\mathbf{S}_T)$: perform Singular Value Decomposition (SVD) to obtain eigenvectors and the corresponding eigenvalues of $\mathbf{S}_T$. Let $\mathbf{P}_{Tc}$ be a linear transformation matrix composed of eigenvectors of $\mathbf{S}_T$ corresponding to the largest $r_t = \text{rank}(\mathbf{S}_T)$ eigenvalues. The average within-class scatter matrix defined in $C(\mathbf{S}_T)$ can be expressed by $\mathbf{P}_{Tc}^T \mathbf{S}_W \mathbf{P}_{Tc}$.
 - Determine $N^d(\mathbf{S}_W)$ and $C^d(\mathbf{S}_W)$: perform SVD to obtain eigenvectors and the corresponding eigenvalues of $\mathbf{P}_{Tc}^T \mathbf{S}_W \mathbf{P}_{Tc}$. Let $\mathbf{P}_{wc}$ and $\mathbf{P}_{wn}$ denote two linear transformation matrices composed of the eigenvectors of $\mathbf{P}_{Tc}^T \mathbf{S}_W \mathbf{P}_{Tc}$ corresponding to the largest $r_w = \text{rank}(\mathbf{P}_{Tc}^T \mathbf{S}_W \mathbf{P}_{Tc})$ eigenvalues and the smallest $\dim(C(\mathbf{S}_T)) - r_w$ eigenvalues, respectively. The pairwise weighted between-class scatter matrix defined in $N^d(\mathbf{S}_W)$ can be expressed by $\tilde{\mathbf{S}}_{Bw}^n$. The average weighted within-class scatter matrix and the pairwise weighted between-class scatter matrix defined in $C^d(\mathbf{S}_W)$ can be expressed by $\tilde{\mathbf{S}}_{Ww}^c$ and $\tilde{\mathbf{S}}_{Bw}^c$.
 - Sequentially extract m discriminant vectors from $N^d(\mathbf{S}_W)$ and $C^d(\mathbf{S}_W)$:
 - Case 1: $\dim(N^d(\mathbf{S}_W)) = 0$

The 1st discriminant vector ϕ_1 is obtained as $\phi_1 = \frac{\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_1\|}$, where ϕ_1 is the

eigenvector of $(\tilde{\mathbf{S}}_{Ww}^c)^{-1} \tilde{\mathbf{S}}_{Bw}^c$ corresponding to the largest eigenvalue. After i^{th} discriminant vector $\phi_i, i \geq 1$ is extracted, the $(i+1)^{\text{th}}$ discriminant vector ϕ_{i+1} is

obtained as $\phi_{i+1} = \frac{\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_{i+1}}{\|\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_{i+1}\|}$, where ϕ_{i+1} the eigenvector of $(\tilde{\mathbf{S}}_{Ww}^c)^{-1} \mathbf{M} \tilde{\mathbf{S}}_{Bw}^c$

corresponding to the largest eigenvalue, where

$$\mathbf{M} = \mathbf{I} - (\mathbf{P}_{Tc} \mathbf{P}_{wc})^T \mathbf{S}_T \mathbf{D}^T \left(\mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{wc} (\tilde{\mathbf{S}}_{Ww}^c)^{-1} (\mathbf{P}_{Tc} \mathbf{P}_{wc})^T \mathbf{S}_T \mathbf{D}^T \right)^{-1} \mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{wc} (\tilde{\mathbf{S}}_{Ww}^c)^{-1}$$

$$, \mathbf{D} = [\phi_1, \dots, \phi_i]^T = \left[\frac{\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_1\|}, \dots, \frac{\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_i}{\|\mathbf{P}_{Tc} \mathbf{P}_{wc} \phi_i\|} \right]^T \text{ and } \mathbf{I} \text{ is an identity matrix.}$$

- Case 2: $\dim(N^d(\mathbf{S}_W)) \neq 0 \wedge m \leq \dim(N^d(\mathbf{S}_W))$

The 1st discriminant vector φ_1 is obtained as $\varphi_1 = \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1\|}$, where ϕ_1 is the eigenvector of $\tilde{\mathbf{S}}_{Bw}^n$ corresponding to the largest eigenvalue. After i^{th} discriminant vector $\varphi_i, i \geq 1$ is extracted, the $(i+1)^{\text{th}}$ discriminant vector φ_{i+1} is obtained as $\varphi_{i+1} = \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_{i+1}}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_{i+1}\|}$, where ϕ_{i+1} the eigenvector of $\mathbf{M} \tilde{\mathbf{S}}_{Bw}^n$ corresponding to the largest eigenvalue, where

$$\mathbf{M} = \mathbf{I} - (\mathbf{P}_{Tc} \mathbf{P}_{Wn})^T \mathbf{S}_T \mathbf{D}^T (\mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{Wn} (\mathbf{P}_{Tc} \mathbf{P}_{Wn})^T \mathbf{S}_T \mathbf{D}^T)^{-1} \mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{Wn},$$

$$\mathbf{D} = [\varphi_1, \dots, \varphi_i]^T = \left[\frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1\|}, \dots, \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_i}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_i\|} \right]^T \text{ and } \mathbf{I} \text{ is an identity matrix.}$$

- Case 3: $\dim(N^d(\mathbf{S}_W)) \neq 0 \wedge m \geq \dim(N^d(\mathbf{S}_W))$

The 1st discriminant vector φ_1 is obtained as $\varphi_1 = \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1\|}$, where ϕ_1 is the eigenvector of $\tilde{\mathbf{S}}_{Bw}^n$ corresponding to the largest eigenvalue. After i^{th} discriminant vector $\varphi_i, i \geq 1$ is extracted:

- If $i < \dim(N^d(\mathbf{S}_W))$, the $(i+1)^{\text{th}}$ discriminant vector φ_{i+1} is obtained as

$$\varphi_{i+1} = \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_{i+1}}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_{i+1}\|}, \text{ where } \phi_{i+1} \text{ the eigenvector of } \mathbf{M} \tilde{\mathbf{S}}_{Bw}^n \text{ corresponding to}$$

the largest eigenvalue, where

$$\mathbf{M} = \mathbf{I} - (\mathbf{P}_{Tc} \mathbf{P}_{Wn})^T \mathbf{S}_T \mathbf{D}^T (\mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{Wn} (\mathbf{P}_{Tc} \mathbf{P}_{Wn})^T \mathbf{S}_T \mathbf{D}^T)^{-1} \mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{Wn},$$

$$\mathbf{D} = [\varphi_1, \dots, \varphi_i]^T = \left[\frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1\|}, \dots, \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_i}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_i\|} \right]^T \text{ and } \mathbf{I} \text{ is an identity matrix.}$$

- If $i \geq \dim(N^d(\mathbf{S}_W))$, the $(i+1)^{\text{th}}$ discriminant vector φ_{i+1} is obtained as

$$\varphi_{i+1} = \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wc} \phi_{i+1}}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wc} \phi_{i+1}\|}, \text{ where } \phi_{i+1} \text{ the eigenvector of } (\tilde{\mathbf{S}}_{Ww}^c)^{-1} \mathbf{M} \tilde{\mathbf{S}}_{Bw}^c$$

corresponding to the largest eigenvalue, where

$$\mathbf{M} = \mathbf{I} - (\mathbf{P}_{Tc} \mathbf{P}_{Wc})^T \mathbf{S}_T \mathbf{D}^T \left(\mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{Wc} (\tilde{\mathbf{S}}_{Ww}^c)^{-1} (\mathbf{P}_{Tc} \mathbf{P}_{Wc})^T \mathbf{S}_T \mathbf{D}^T \right)^{-1} \mathbf{D} \mathbf{S}_T \mathbf{P}_{Tc} \mathbf{P}_{Wc} (\tilde{\mathbf{S}}_{Ww}^c)^{-1}$$

$$, \mathbf{D} = [\phi_1, \dots, \phi_i]^T = \left[\frac{\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wn} \phi_1\|}, \dots, \frac{\mathbf{P}_{Tc} \mathbf{P}_{Wc} \phi_i}{\|\mathbf{P}_{Tc} \mathbf{P}_{Wc} \phi_i\|} \right]^T \text{ and } \mathbf{I} \text{ is an identity matrix.}$$

- Normalization: calculate the diagonal covariance matrix in the transformed subspace $\mathbf{D} \mathbf{S}_T \mathbf{D}^T = \mathbf{\Lambda}_t$. Normalize $\mathbf{D}$ to make the uncorrelated features in the transformed subspace to have unit variances. $\mathbf{D}' = \mathbf{D}^T \mathbf{\Lambda}_t^{-1/2}$.
- The m extracted discriminant vectors constitute the columns of $\mathbf{D}'$.

The justifications for the equations involved in the CS-UCWFLDA pseudo-code can be easily verified according to the proof in [222].

The CS-UCWFLDA technique is attractive from the following aspects:

1. When $N^d(\mathbf{S}_w)$ is empty, the CS-UCWFLDA comes down to the CWFLDA_{q2}, and accordingly inherits the eminent performance of the CWFLDA_{q2} when dealing with well-sampled datasets. From this perspective, the CS-UCWFLDA technique can be regarded as the generalization of CWFLDA_{q2}.
2. The dimensionality of $N^d(\mathbf{S}_w)$ is determined by $\dim(N^d(\mathbf{S}_w)) = \dim(C(\mathbf{S}_T)) - \dim(C^d(\mathbf{S}_w))$ which varies with the training sample size and the number of classes. When $\dim(N^d(\mathbf{S}_w))$ is fairly small, the conventional NFLDA techniques will deteriorate much since they can only extract very few discriminant vectors in $N^d(\mathbf{S}_w)$. In this sense, the CS-UCWFLDA can be regarded as a generalized NFLDA technique, which can extract additional discriminant vectors in $C^d(\mathbf{S}_w)$ so as to improve the performance of the subsequent pattern analysis task.
3. The class-wise weighted Fisher criterion used in CS-UCWFLDA compensates several deficiencies inherent in the conventional Fisher criterion. Consequently, a more compact lower-dimensional pattern representation may be obtained, which can retain most of the useful discriminatory information present in the original

high-dimensional feature space. It is worth noting that in the high-dimensional and undersampled case, most estimated pairwise two-class Bayes errors or their upper bounds under Gaussian assumptions may achieve the common minimum values since most class pairs in high-dimensional and undersampled problems are well separated in the sparse high-dimensional feature space. As a result, the weighting factors will be mainly determined by the reciprocal of the squared Mahalanobis distance between two classes. Obviously, outlier classes located remotely from the remaining classes will receive small weights and thus contribute less to the estimate of the average within-class and between-class scatter matrices.

4. Some of the existing techniques, such as DS-FLDA and GFLDA, can also explicitly or implicitly take into account the discriminatory information present in both $N^d(\mathbf{S}_w)$ and $C^d(\mathbf{S}_w)$. In comparison with our proposed CS-UCWFLDA technique, they have several limitations. The DS-FLDA techniques separately extract the maximum possible number of discriminative features from each subspace and then fuse the obtained two sets of features in a weighted manner for classification. Although demonstrating promising performance in classification, such kind of techniques cannot further reduce the dimensionality. The GFLDA techniques can implicitly extract additional discriminant vectors from $C^d(\mathbf{S}_w)$ to compensate the limited number of discriminant vectors extracted from the possibly low-dimensional $N^d(\mathbf{S}_w)$, and result in uncorrelated features in the transformed subspace. However, such kind of techniques is subject to the equality $\mathbf{S}_T = \mathbf{S}_w + \mathbf{S}_b$. Both DS-FLDA and GFLDA techniques employ the conventional Fisher criterion and accordingly suffer from the deficiencies described in the previous chapter.
5. Normalization applied in the transformed subspace can eliminate the problem caused by the magnitude discrepancy among different features in the transformed space.

Although uncorrelated features extracted by the CS-UCWFLDA technique can eliminate the undesirable correlatedness among features, they may overfit the training samples as indicated in [124]. As a solution, we perform the QR decomposition of the matrix composed of the discriminant vectors obtained by the CS-UCWFLDA to yield a set of orthogonal discriminant vectors. We name this technique Complementary Subspace Orthogonal CWFLDA (CS-OCWFLDA).

4.4 Regularized subspace based CWFLDA techniques

Besides the unified extraction of discriminatory information from the null and column spaces of the average within-class scatter matrix to avoid the singularity problem, direct regularization of the singular average within-class scatter matrix provides another effective way to generalize the utility of the CWFLDA techniques to handle high-dimensional and undersampled problems. Since $N(\mathbf{S}_T)$ does not contain any discriminatory information, it will be removed in the beginning and regularization is subsequently applied in $C(\mathbf{S}_T)$, a subspace of the original feature space. We therefore name this class of techniques as the Regularized Subspace Based CWFLDA (RS-CWFLDA).

In this section, we choose 2 representative CWFLDA techniques proposed in the previous chapter, namely CWFLDA_{q_0} and F-CWFLDA_{q_0} , and reactive their feasibility in the high-dimensional and undersampled scenario via regularization. Note that CWFLDA_{q_0} and F-CWFLDA_{q_0} techniques have demonstrated superior performance in the experiments of the previous chapter.

Even since the pioneering work by Hong and Yang [136], many regularization methods have been devised to compensate the singular or poorly estimated average within-class scatter matrix in the FLDA, such as singular value perturbation [136], shrink towards the identity matrix [137], [138], addition with a scaled identity matrix [139], [140], multiple parameter regularization [141] and maximum entropy

covariance selection [142] and so on. Among them, we choose two simple but effective schemes to generalize the 2 representative CWFLDA techniques.

The first regularization method is to add a multiple of the identity matrix to the average within-class scatter matrix as $\mathbf{S}_w + \sigma \mathbf{I}$ with $\sigma > 0$. It is easy to verify that the regularized average within-class scatter matrix $\mathbf{S}_w + \sigma \mathbf{I}$ is positive definite. To generalize the CWFLDA techniques, the average weighted within-class scatter matrix and the pairwise weighted between-class scatter matrix will be estimated in the regularized $C(\mathbf{S}_T)$. The resulting regularized CWFLDA_{q0} and F-CWFLDA_{q0} techniques are named R1CWFLDA_{q0} and R1F-CWFLDA_{q0} respectively. A limitation of such regularization method is that the optimal value of the multiplier σ is hard to determine. A very small value of σ may result in the poorly estimated scatter matrix and accordingly lead to numerical instability, while a very large value of σ may distort the useful information contained in the original scatter matrix. In practical, cross-validation is typically used to seek for the optimal value of σ at the expense of computational costs.

The second regularization method is based on the Maximum Entropy Covariance Selection (MECS), which was developed by Thomaz *et al.* [142] according to the maximum entropy (uncertainty) principle [223] to improve the performance of the quadratic discriminant classifier on undersampled problems. The core of MECS is to maximize the Gaussian entropy of the convex mixture of a singular or poorly estimated individual covariance matrix and a well-estimated pooled covariance matrix by selecting the larger variance between the two component covariance matrices along each of the orthonormal eigenvectors of the mixed matrix. The selected variances and the corresponding orthonormal eigenvectors of the mixed matrix constitute the re-estimated individual covariance matrix. The Maximum Uncertainty FLDA (MUFLDA) technique, proposed by Thomaz *et al.* [223], takes advantage of the MECS idea to regularize the singular average within-class scatter matrix based on the following equation suggested by Friedman [137].

$$\mathbf{S}_w = (1 - \gamma) \mathbf{S}_w + \gamma \frac{\text{trace}(\mathbf{S}_w)}{\text{dim}(\mathbf{S}_w)} \mathbf{I} \quad \text{with } \gamma \in [0, 1] \quad (4.1)$$

The MECS method used in the MUFLDA is described in Table 4.2.

Table 4.2: The MECS method used in the MUFLDA

<ul style="list-style-type: none"> • Perform eigenvalue decomposition of $\mathbf{S}_p = \mathbf{S}_w / (N - C) \in \mathbf{R}^{n \times n}$, i.e. $\mathbf{P}_p^T \mathbf{S}_p \mathbf{P}_p = \mathbf{\Lambda}_p = \text{diag}(\lambda_1, \dots, \lambda_n)$ <ul style="list-style-type: none"> • Calculate the average eigenvalue of $\mathbf{S}_p$, i.e. $\bar{\lambda} = \text{trace}(\mathbf{S}_p) / n$ • Establish a new diagonal matrix as: $\mathbf{\Lambda}^* = \text{diag}(\max(\bar{\lambda}, \lambda_1), \dots, \max(\bar{\lambda}, \lambda_n))$ • The re-estimated non-singular average within-class scatter matrix is defined by: $\mathbf{S}_w^* = (N - C) (\mathbf{P}_p \mathbf{\Lambda}^* \mathbf{P}_p^T)$

Compared with Friedman's idea in Eq. (4.1), the MECS method expands only the smaller and consequently less reliable eigenvalues of the average within-class scatter matrix while keeping unchanged most of its larger eigenvalues, instead of simultaneously decreasing the larger eigenvalues and increasing the smaller eigenvalues. Moreover, the MECS avoids the estimation of the parameter γ .

Furthermore, the MECS has a reasonable justification originating from the maximum entropy principle [223]: “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.” In fact, the MECS method can also be regarded as a special case of the singular value perturbation method [136] with the threshold value chosen as $\text{trace}(\mathbf{S}_w) / \text{dim}(\mathbf{S}_w)$, which is larger than that suggested in [136]. To generalize the CWFLDA techniques, the average weighted within-class scatter matrix and the pairwise weighted between-class scatter matrix will be estimated in the regularized $C(\mathbf{S}_T)$. The resulting regularized CWFLDA_{q0} and F-CWFLDA_{q0} techniques are named R2CWFLDA_{q0} and R2F-CWFLDA_{q0}, respectively. Obviously, in the well-sampled case, the RS-CWFLDA techniques come down to the CWFLDA.

4.5 Experiments

In this section, we conduct extensive experiments to evaluate the performance of the generalized CWFLDA techniques including 2 CS-CWFLDA techniques and 4 RS-CWFLDA techniques to handle high-dimensional and undersampled problems. We compare them with 10 state-of-the-art FLDA variants consisting of PCA+FLDA, regularized FLDA (RFLDA), MUFLDA, DLDA, Regularized DLDA (RDLDA), DCV, NUFLDA, ULDA, OLDA and FLDA_{s1}. Among them, PCA+FLDA is a famous two-stage FLDA technique, which is widely used as a baseline approach in many face recognition literatures. RFLDA and MUFLDA use the two simple but effective regularization methods as introduced in our proposed RS-CWFLDA techniques. DLDA and RDLDA have received remarkable attentions in recent computer vision literatures. DCV and NUFLDA stands for two popular null space based FLDA techniques. ULDA and OLDA are two recently proposed techniques with good theoretical foundations, which have demonstrated promising empirical performance. FLDA_{s1} in terms of the maximum margin criterion have demonstrated promising performance in face recognition applications. For convenience, we summarize all techniques involved in our experiments in Table 4.3 with their indices and names interchanged in the following context. Our proposed 10 generalized CWFLDA techniques are typeset in bold.

Table 4.3: The 20 DLDR techniques evaluated in our experiments. Information is provided on technique index, technique name and their provenance in this thesis or previous literatures.

Technique index	Technique name	Provenance
1	CS-UCWFLDA	pp. 161
2	CS-OCWFLDA	pp. 165
3	R1CWFLDA_{q0}	pp. 166
4	R2CWFLDA_{q0}	pp. 167
5	R1F-CWFLDA_{q0}	pp. 166
6	R2F-CWFLDA_{q0}	pp. 167
7	PCA_LDA	[161]

8	RFLDA	[139]
9	MUFLDA	[224]
10	DLDA	[168]
11	RDLDA	[140]
12	DCV	[167]
13	NUFLDA	[166]
14	ULDA	[124]
15	OLDA	[124]
16	FLDA _{s1}	[129]

4.5.1 Experimental setup

We choose 10 datasets from various data sources to perform the experimental study. Among them, the Multiple Feature (MultiFeat) dataset taken from the UCI machine learning repository [78] consists of 200 binary images for each of the ten handwritten digits “0” to “9” extracted from Dutch utility maps. Each image is characterized by 649 features in terms of Fourier coefficients, profile correlations, Karhunen-Love coefficients, pixel averages, Zernike moments and morphological features. Three facial image datasets, including AT&T¹⁴, YALE¹⁵, PIX¹⁶, have been used widely to evaluate face recognition algorithms. Another five bioinformatics datasets, including ALL [225], ALLAML4 [226], GCM [227], [228], NCI [229] and SRBCT [230], [231] belong to the microarray gene expression profiles, which record tens of thousands of gene expression levels simultaneously but usually contain a very few number of samples due to high sensing costs. We also include one text document dataset tr41, which is derived from the TREC-5, TREC-6 and TREC-7 collections [232]. Note that most of these datasets are provided by Dr. Jieping Ye in his academic webpage and we use the same preprocessing steps as suggested in [124], [156], [175]. The statistics of these datasets are summarized in Table 4.4. All the datasets are normalized to $[-1, 1]$ before use. In our experiments, the Nearest Center (NC) classifier and the k NN ($k=1$) classifier are used for classification due to their efficiency and effectiveness in the

¹⁴ <http://www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html>

¹⁵ <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>

¹⁶ <http://peipa.essex.ac.uk/ipa/pix/faces/manchester/test-hard/>

high-dimensional and undersampled scenario. The whole experiments are divided into two parts as follows.

Table 4.4: The 10 datasets used in our experiments. Information is provided on dataset index, dataset name, number of classes (C), feature space dimensionality (n) and total number of samples (N).

Index	Datasets	C	n	N
a	MultiFeat	10	649	2000
b	AT&T	40	10304	400
c	YALE	15	12573	165
d	PIX	30	10000	300
e	ALL	6	12558	248
f	ALLAML4	4	7129	72
g	NCI	9	6167	61
h	SRBCT	4	2308	83
i	GCM	14	11485	198
j	tr41	7	7454	210

The 1st experiment aims at verifying the promising performance of the CS-CWFLDA as the generalized null space based FLDA techniques. To achieve this goal, we choose the MultiFeat dataset and artificially change the number of samples per class so as to alter the dimensionality of the discriminative null space of the average within-class scatter matrix, i.e. $N^d(\mathbf{S}_w)$. With the dimensionality of $N^d(\mathbf{S}_w)$ gradually decreased from 9 to 0, we compare the performance of the proposed CS-UCWFLDA and CS-OCWFLDA techniques with several state-of-the-art null space related FLDA variants including NUFLDA, DCV, ULDA and OLDA. Among them, the NUFLDA and DCV can only extract discriminant vectors in $N^d(\mathbf{S}_w)$ and in the whole null space of the average within-class scatter matrix $N(\mathbf{S}_w)$, respectively, while the ULDA and OLDA can implicitly extract discriminant vectors in $N^d(\mathbf{S}_w)$ via the simultaneous diagonalization scheme. Moreover, when $N^d(\mathbf{S}_w)$ is empty, the NUFLDA will be switched to the conventional FLDA, while the DCV will obtain meaningless

discriminant vectors. In contrast, as long as the dimensionality of $N^d(\mathbf{S}_w)$ is below certain degree, the ULDA and OLDA can implicitly extract additional discriminant vectors in $C^d(\mathbf{S}_w)$ to augment the performance. Furthermore, both NUFLDA and ULDA can yield uncorrelated transformed features while both DCV and OLDA can generate orthogonal discriminant vectors. In this experiment, we also study the numerical instability problem in our proposed CS-CWFLDA techniques. In the subspace based FLDA techniques, determination of the null and column spaces of a scatter matrix plays a crucial role. However, the typical way to obtain the null and column spaces of a scatter matrix based on its rank (as described in Theorem 4.1) may possibly lead to the numerical instability problem in the resulting column space. Specifically, suppose the rank of a scatter matrix $\mathbf{A}$ is $r(\mathbf{A})$ and accordingly the eigenvectors of $\mathbf{A}$ corresponding the largest $r(\mathbf{A})$ eigenvalues span the column space of $\mathbf{A}$. Since $r(\mathbf{A})$ is generally defined as the number of eigenvalues above a specific threshold, if this threshold is too low, it may possibly happen that some eigenvectors of $\mathbf{A}$ corresponding to the relatively smaller eigenvalues may also serve as the basis vectors of the resulting column space, in which some estimated scatter matrix may become nearly singular. In fact, the column space determined by the matrix rank can only eliminate the singularity problem but may not avoid the numerical instability problem. To compensate this deficiency in our proposed CS-CWFLDA techniques, we introduce a matrix stability index $\theta_s(\mathbf{A})$, and accordingly the eigenvectors of $\mathbf{A}$ corresponding to eigenvalues above this index value will span the column space of $\mathbf{A}$ and the remaining eigenvectors will span the null space of $\mathbf{A}$. In our experiments, we set $\theta_s(\mathbf{A})$ equal to one millionth of $\text{trace}(\mathbf{A})$, and compare two groups of CS-CWFLDA techniques using the matrix rank and matrix stability index respectively to determine the null and column spaces of scatter matrices. We reserve names CS-UCWFOLDA and CS-OCWFLDA for the CS-CWFLDA techniques using the matrix stability index, while use oCS-UCWFLDA and oCS-OCWFLDA to denote the CS-CWFLDA techniques employing the matrix rank. In this experiment, we will randomly choose l samples from each class of the MultiFeat dataset to constitute the training set with the remaining samples serving as the testing test. This process will be repeated 20 times to obtain 20 pairs of training and testing sets. The mean

classification error at each possible dimension that can be reduced to is computed by averaging the classification errors at that dimension over the 20 pairs. The maximum possible dimension to reduce to is specified as $C-1$, hereby equal to 9. We choose l as 10, 30, 50, 55, 60, 65, 80 and 100, respectively, to reduce the dimensionality of $N^d(\mathbf{S}_w)$ from 9 to 0.

The 2nd experiment aims at extensively evaluating the effectiveness of our proposed generalized CWFLDA techniques to deal with various types of real-world problems involving high-dimensional and undersampled data, and compare their performance with that of the state-of-the-art FLDA variants. We employ the same performance evaluation scheme as that used in the previous chapter. The Wilcoxon signed rank test is carried out to compute the 3 relative performance indices, i.e. RBPI, RAPI and ROPI, for each experimental technique on per dataset. These relative performance indices are eventually averaged over all the experimental datasets to evaluate each experimental technique. In this experiment, the training and testing sets are generated in different ways with respect to different types of datasets. For facial image and text document datasets, due to its balanced sample size over all classes, we randomly choose around half of the total samples per class to constitute the training set with the remaining samples serving as the testing set. Specifically, we randomly choose 5 samples per class for ORL, YALE and PIX datasets, respectively and randomly choose 15 samples per class for dataset tr41. For microarray gene expression datasets, because of its unbalanced sample size over all classes, we randomly choose two third of the whole dataset as the training set with the remaining samples constituting the testing set. This process will be repeated 20 times to obtain 20 pairs of training and testing sets. The mean classification error at each possible dimension that can be reduced to is calculated by averaging the classification errors at that dimension over the 20 pairs. We specify the maximum possible dimension to reduce to as $C-1$.

4.5.2 Result analysis

For the 1st experiment, Tables 4.5 and 4.6 report, by using the k NN and NC classifiers respectively, the minimum mean classification errors and the corresponding reduced

dimensions attained by the null space related FLDA techniques including CS-UCWFLDA, CS-OBFLDA, oCS-UCWFLDA, oCS-OCWFLDA, NUFLDA, DCV, ULDA and OLDA, on various splits of the MultiFeat dataset with different numbers of training samples per class ($\#TSPC$). The average dimensionality of $C(\mathbf{S}_T)$ and $N^d(\mathbf{S}_W)$ over 20 pairs of training and testing sets with regard to different data splits, determined by the matrix rank and matrix stability index respectively, is shown in Table 4.7. For each data split, the optimal technique with the overall smallest minimum mean classification error¹⁷ among all the null space related FLDA techniques is typeset in bold. Note that we do not provide a direct way to determine the optimal dimensionality to reduce to, however, the observed minimum mean classification errors give an indication of the optimally attainable performance and thus can be used to compare different techniques.

Table 4.5: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by the null space related FLDA (NS-FLDA) techniques with respect to the k NN classifier.

$\#TSPC$ NS-FLDA	10	30	50	55	60	65	80	100
CS-UCWFLDA	0.0702 (7)	0.0504 (8)	0.0790 (9)	0.0622 (9)	0.0505 (9)	0.0416 (9)	0.0247 (8)	0.0173 (8)
CS-OCWFLDA	0.0596 (9)	0.0328 (9)	0.0511 (9)	0.0502 (9)	0.0399 (9)	0.0289 (9)	0.0204 (9)	0.0147 (9)
oCS-UCWFLDA	0.0702 (7)	0.0504 (8)	0.1016 (8)	0.1617 (8)	0.2803 (7)	0.6714 (9)	0.0402 (8)	0.0198 (8)
oCS-OCWFLDA	0.0596 (9)	0.0328 (9)	0.0574 (9)	0.0930 (9)	0.1784 (9)	0.5966 (9)	0.0736 (9)	0.0335 (9)
NUFLDA	0.0717 (8)	0.0524 (8)	0.1030 (9)	0.1620 (8)	0.2835 (7)	0.6966 (4)	0.0400 (9)	0.0197 (9)
DCV	0.0596 (9)	0.0328 (9)	0.0574 (9)	0.0930 (9)	0.1784 (9)	0.6826 (6)	0.5443 (9)	0.4824 (9)

¹⁷ When several techniques reach the overall smallest minimum mean classification error, the corresponding reduced dimensions are compared. The optimal technique should achieve the overall smallest minimum mean classification error at the smallest dimension. When the dimensions are same again, standard deviations of the classification error are compared. The optimal technique should have the smallest standard deviation.

Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems

ULDA	0.0742 (9)	0.0526 (9)	0.1030 (9)	0.1747 (9)	0.3049 (9)	0.7078 (9)	0.0400 (9)	0.0197 (9)
OLDA	0.0596 (9)	0.0328 (9)	0.0574 (9)	0.0930 (9)	0.1784 (9)	0.5810 (9)	0.0713 (9)	0.0327 (9)

Table 4.6: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by the null space related FLDA (NS-FLDA) techniques with respect to the NC classifier.

#TSPC \ NS-FLDA	10	30	50	55	60	65	80	100
CS-UCWFLDA	0.0702 (7)	0.0504 (8)	0.0810 (9)	0.0639 (9)	0.0519 (9)	0.0429 (9)	0.0263 (8)	0.0179 (8)
CS-OCWFLDA	0.0596 (9)	0.0328 (9)	0.0514 (9)	0.0509 (9)	0.0397 (9)	0.0295 (9)	0.0213 (9)	0.0164 (9)
oCS-UCWFLDA	0.0702 (7)	0.0504 (8)	0.1016 (8)	0.1617 (8)	0.2803 (7)	0.6717 (9)	0.0421 (8)	0.0207 (9)
oCS-OCWFLDA	0.0596 (9)	0.0328 (9)	0.0574 (9)	0.0930 (9)	0.1784 (9)	0.5981 (9)	0.0804 (9)	0.0388 (9)
NUFLDA	0.0717 (8)	0.0524 (8)	0.1030 (9)	0.1620 (8)	0.2835 (7)	0.6966 (4)	0.0424 (9)	0.0205 (9)
DCV	0.0596 (9)	0.0328 (9)	0.0574 (9)	0.0930 (9)	0.1784 (9)	0.6826 (6)	0.6061 (9)	0.5758 (9)
ULDA	0.0742 (9)	0.0526 (9)	0.1030 (9)	0.1747 (9)	0.3049 (9)	0.7086 (9)	0.0424 (9)	0.0205 (9)
OLDA	0.0596 (9)	0.0328 (9)	0.0574 (9)	0.0930 (9)	0.1784 (9)	0.5814 (9)	0.0788 (9)	0.0389 (9)

Table 4.7: The average dimensionality (standard deviation in bracket) of $C(S_T)$ and $N^d(S_w)$ over 20 pairs of training and testing sets with regard to different data splits, determined by the matrix rank and matrix stability index, respectively.

#TSPC	Dimension	Matrix rank	Matrix stability index
10	Ave $\dim(C(S_T))$	98.95 (0.22)	98.95 (0.22)

Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems

	Ave $\dim(N^d(\mathbf{S}_w))$	9 (0)	9 (0)
30	Ave $\dim(C(\mathbf{S}_T))$	298.75 (0.44)	298.75 (0.44)
	Ave $\dim(N^d(\mathbf{S}_w))$	9 (0)	9 (0)
50	Ave $\dim(C(\mathbf{S}_T))$	498.85 (0.49)	479.40 (0.88)
	Ave $\dim(N^d(\mathbf{S}_w))$	9 (0)	7 (0.65)
55	Ave $\dim(C(\mathbf{S}_T))$	548.70 (0.47)	501.45 (0.83)
	Ave $\dim(N^d(\mathbf{S}_w))$	9 (0)	3.95 (0.61)
60	Ave $\dim(C(\mathbf{S}_T))$	598.45 (0.61)	518.45 (0.89)
	Ave $\dim(N^d(\mathbf{S}_w))$	9 (0)	1.65 (0.67)
65	Ave $\dim(C(\mathbf{S}_T))$	646 (0)	531.95 (0.89)
	Ave $\dim(N^d(\mathbf{S}_w))$	6.20 (0.52)	0.55 (0.61)
80	Ave $\dim(C(\mathbf{S}_T))$	646 (0)	556.70 (0.66)
	Ave $\dim(N^d(\mathbf{S}_w))$	0 (0)	0 (0)
100	Ave $\dim(C(\mathbf{S}_T))$	646 (0)	572.8 (0.83)
	Ave $\dim(N^d(\mathbf{S}_w))$	0 (0)	0 (0)

Some remarkable observations are outlined as follows:

1. When the matrix rank is used to determine the null and column spaces of a scatter matrix, the dimensionality of $N^d(\mathbf{S}_w)$ will decrease dramatically from its maximum value to zero, starting from a certain number of training samples per class, hereby around 65. In contrast, when using the matrix stability index, the

dimensionality of $N^d(\mathbf{S}_w)$ will decrease gradually from its maximum value to zero, starting from a fewer number of training samples per class, hereby around 50. Moreover, the dimensionality of $C(\mathbf{S}_T)$ determined by the matrix stability index is always less than or equal to that determined by the matrix rank, which can remove instable factors.

2. The CS-UBOFLDA and CS-OCWFLDA techniques, which use the matrix instability index to determine the null and column spaces of scatter matrices, achieve remarkably better performance than other techniques employing the matrix rank. And the CS-OCWFLDA technique always wins the first place among all the experimental techniques in comparison over various data splits.

3. Among the techniques using the matrix rank to determine the column and null spaces of scatter matrices, when the dimensionality of $N^d(\mathbf{S}_w)$ reaches its maximum value of 9 in the cases of 10, 30, 50, 55 and 60 training samples per class, the oCS-OCWFLDA, DCV and OLDA techniques, which can generate the orthogonal discriminant vectors, demonstrate the same performance. The study of the intrinsic relationships between these techniques is among our future plans. When the dimensionality of $N^d(\mathbf{S}_w)$ starts to decrease in the case of 65 training sample per class, all the oCS-UCWFLDA, oCS-UCWFLDA, NUFLDA, DCV, ULDA and OLDA techniques demonstrate inferior performance. The reason might be that some estimated scatter matrices are numerically instable. Among oCS-OCWFLDA, DCV and OLDA, the DCV performs worst since it cannot extract additional discriminant vectors from $C^d(\mathbf{S}_w)$. When the dimensionality of $N^d(\mathbf{S}_w)$ decreases to zero in the cases of 70 and 100 training samples per class, the performance of DCV is fairly bad since it can no longer extract any discriminatory information. In contrast, the oCS-UCWFLDA, oCS-UCWFLDA, NUFLDA, ULDA and OLDA techniques can extract discriminant vectors in $C^d(\mathbf{S}_w)$, and thus demonstrate satisfactory performance. In comparison with the ULDA, the oCS-UCWFLDA technique can generally achieve the competitive performance at the smaller dimension.

For the 2nd experiment, Tables 4.8 and 4.9 report, by using the k NN and NC classifiers respectively, the minimum mean classification error and the corresponding reduced dimension attained by each of the 20 experimental techniques on per dataset. For each dataset, the optimal technique with the overall smallest minimum mean classification error¹⁸ among all the experimental techniques is typeset in bold. The mean classification errors in the original feature space without reducing the dimensionality are also reported in the tables. Analogous to the previous chapter, we resort to the 3 relative performance indices, i.e. RBPI, RAPI and ROPI, to evaluate the performance of each experimental technique.

Table 4.8: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 20 experimental techniques as well as the mean classification errors obtained in the original feature space (*org*) with respect to the k NN classifier.

	b	c	d	e	f	g	h	i	j
1	0.0722 (19)	0.0539 (14)	0.0323 (7)	0.0536 (5)	0.1042 (3)	0.4400 (7)	0.0405 (3)	0.2754 (13)	0.1010 (6)
2	0.0365 (39)	0.0883 (14)	0.0217 (14)	0.0265 (5)	0.0958 (3)	0.4550 (7)	0.0405 (3)	0.2722 (13)	0.1576 (6)
3	0.0365 (39)	0.0883 (14)	0.0213 (15)	0.0265 (5)	0.0958 (3)	0.4550 (7)	0.0405 (3)	0.2722 (13)	0.1576 (6)
4	0.0298 (33)	0.1061 (14)	0.0230 (27)	0.0410 (5)	0.1021 (3)	0.4675 (5)	0.0714 (3)	0.2937 (13)	0.1652 (6)
5	0.0628 (38)	0.1161 (14)	0.0280 (21)	0.0548 (5)	0.0979 (3)	0.4975 (8)	0.0476 (3)	0.2881 (13)	0.2643 (6)
6	0.0403 (30)	0.1100 (14)	0.0237 (28)	0.0566 (5)	0.1083 (3)	0.4825 (8)	0.0690 (3)	0.2873 (13)	0.2024 (6)
7	0.6060 (39)	0.6044 (14)	0.4453 (21)	0.2211 (5)	0.1875 (3)	0.6600 (6)	0.1571 (3)	0.5548 (13)	0.6157 (6)
8	0.0365 (39)	0.0889 (14)	0.0217 (28)	0.0265 (5)	0.0958 (3)	0.4700 (7)	0.0405 (3)	0.2722 (13)	0.1581 (6)
9	0.0298 (37)	0.1044 (14)	0.0233 (22)	0.0446 (5)	0.1063 (3)	0.4775 (8)	0.0667 (3)	0.3016 (13)	0.1686 (6)
10	0.0455 (38)	0.1417 (11)	0.0243 (19)	0.0819 (5)	0.1562 (3)	0.4850 (7)	0.1214 (3)	0.3270 (13)	0.2062 (6)
11	0.0682 (37)	0.1489 (11)	0.0423 (10)	0.0795 (5)	0.1542 (3)	0.4625 (7)	0.1214 (3)	0.3286 (12)	0.1662 (6)
12	0.0365 (39)	0.0883 (14)	0.0217 (28)	0.0265 (5)	0.0958 (3)	0.4650 (7)	0.0405 (3)	0.2722 (13)	0.1576 (6)
13	0.0827 (39)	0.0539 (14)	0.0340 (8)	0.0536 (5)	0.1021 (2)	0.4550 (7)	0.0405 (3)	0.2754 (13)	0.1010 (6)
14	0.0827 (39)	0.0539 (14)	0.0483 (29)	0.0536 (5)	0.1042 (3)	0.4600 (8)	0.0405 (3)	0.2754 (13)	0.1010 (6)
15	0.0365 (39)	0.0883 (14)	0.0227 (29)	0.0265 (5)	0.0958 (3)	0.4725 (8)	0.0405 (3)	0.2722 (13)	0.1576 (6)

¹⁸ When several techniques reach the overall smallest minimum mean classification error, the corresponding reduced dimensions are compared. The optimal technique should achieve the overall smallest minimum mean classification error at the smallest dimension. When the dimensions are same again, standard deviations of the classification error are compared. The optimal technique should have the smallest standard deviation.

Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems

16	0.0483 (39)	0.1628 (14)	0.0340 (29)	0.0578 (5)	0.1063 (3)	0.4825 (7)	0.0762 (3)	0.3262 (13)	0.1595 (6)
<i>Org</i>	0.0685	0.2722	0.0413	0.1337	0.1083	0.5400	0.1429	0.3921	0.7124

Table 4.9: The minimum mean classification errors and the corresponding reduced dimensions (in bracket) attained by 20 experimental techniques as well as the mean classification errors obtained in the original feature space (*org*) with respect to the NC classifier.

	b	c	d	e	f	g	h	i	j
1	0.0722 (19)	0.0539 (14)	0.0323 (7)	0.0536 (5)	0.1042 (3)	0.4400 (7)	0.0405 (3)	0.2754 (13)	0.1010 (6)
2	0.0365 (39)	0.0883 (14)	0.0217 (14)	0.0265 (5)	0.0958 (3)	0.4550 (7)	0.0405 (3)	0.2722 (13)	0.1576 (6)
3	0.0365 (39)	0.0883 (14)	0.0213 (15)	0.0265 (5)	0.0958 (3)	0.4550 (7)	0.0405 (3)	0.2722 (13)	0.1576 (6)
4	0.0355 (35)	0.1039 (14)	0.0257 (19)	0.0307 (5)	0.0833 (3)	0.4750 (6)	0.0619 (3)	0.2841 (13)	0.1657 (6)
5	0.0628 (38)	0.1161 (14)	0.0280 (21)	0.0548 (5)	0.0979 (3)	0.4975 (8)	0.0476 (3)	0.2881 (13)	0.2643 (6)
6	0.0445 (27)	0.1183 (14)	0.0243 (28)	0.0500 (5)	0.0896 (3)	0.4725 (8)	0.0571 (3)	0.2960 (13)	0.2157 (6)
7	0.0605 (34)	0.6033 (14)	0.4453 (21)	0.2229 (5)	0.1854 (3)	0.6550 (8)	0.1476 (3)	0.5524 (13)	0.6114 (6)
8	0.0365 (39)	0.0889 (14)	0.0217 (28)	0.0265 (5)	0.0958 (3)	0.4700 (7)	0.0405 (3)	0.2722 (13)	0.1581 (6)
9	0.0365 (38)	0.1056 (14)	0.0250 (21)	0.0289 (5)	0.0833 (3)	0.4800 (8)	0.0619 (3)	0.2849 (13)	0.1652 (6)
10	0.0508 (35)	0.1317 (12)	0.0263 (29)	0.0783 (5)	0.1417 (3)	0.4775 (7)	0.1238 (3)	0.3087 (13)	0.2324 (6)
11	0.0892 (36)	0.1300 (11)	0.0487 (10)	0.0765 (5)	0.1438 (3)	0.4400 (7)	0.1310 (3)	0.3008 (13)	0.1481 (6)
12	0.0365 (39)	0.0883 (14)	0.0217 (28)	0.0265 (5)	0.0958 (3)	0.4650 (7)	0.0405 (3)	0.2722 (13)	0.1576 (6)
13	0.0827 (39)	0.0539 (14)	0.0340 (8)	0.0536 (5)	0.1021 (2)	0.4550 (7)	0.0405 (3)	0.2754 (13)	0.1010 (6)
14	0.0827 (39)	0.0539 (14)	0.0483 (29)	0.0536 (5)	0.1042 (3)	0.4600 (8)	0.0405 (3)	0.2754 (13)	0.1010 (6)
15	0.0365 (39)	0.0883 (14)	0.0227 (29)	0.0265 (5)	0.0958 (3)	0.4725 (8)	0.0405 (3)	0.2722 (13)	0.1576 (6)
16	0.0745 (39)	0.1783 (14)	0.0547 (29)	0.0373 (5)	0.0833 (3)	0.4825 (8)	0.0786 (3)	0.3706 (13)	0.1600 (6)
<i>Org</i>	0.0975	0.2589	0.0737	0.0584	0.1333	0.6150	0.1262	0.4413	0.1829

The RBPI, RAPI and ROPI values of each experimental DLDR technique averaged over all the experimental datasets with respect to the k NN and NC classifiers respectively, are illustrated as box charts in Figures 4.1 and 4.2. Some remarkable observations are discussed as follows.

1. From the perspective of the relative best performance measured by the RBPI values, several techniques including R1CWFLDA_{q0} (technique index¹⁹ 3) CS-OCWFLDA (technique index 2), RFLDA (technique index 8) and DCV (technique index 12) demonstrate consistently high performance with respect to both k NN and NC classifiers. Among them, three regularized subspace based techniques by adding a multiple of identity matrix to the singular average within-class scatter matrix estimated in $C(\mathbf{S}_T)$, i.e. R1CWFLDA_{q0}, R1F-CWFLDA_{q0} (technique index 5) and RFLDA, demonstrate better performance than the other three regularized subspace based techniques by using the MECS method, i.e. R2CWFLDA_{q0} (technique index 4), R2F-CWFLDA_{q0} (technique index 6) and MUFLDA (technique index 9). The CS-OCWFLDA technique obtained from the CS-UCWFLDA (technique index 1) via the QR decomposition demonstrates improved performance. The similar fact can also be observed from the OLDA (technique index 15) and ULDA (technique index 14) techniques. The NUFLDA (technique index 13) technique demonstrates better performance than the ULDA technique. The DLDA (technique index 10) technique and its regularized version RDLDA (technique index 11) demonstrate inferior performance in comparison with many other techniques, among which the RDLDA outperforms the DLDA. The FLDA_{s1} (technique index 16) also perform not well in comparison with many other techniques. Among all the experimental techniques, the PCA+LDA (technique index 7) technique always performs worst since it discards most of the important discriminatory information present in $N^d(\mathbf{S}_w)$.
2. From the perspective of the relative average performance measured by the RAPI values, four of our proposed techniques, i.e. CS-UCWFLDA, CS-OCWFLDA, R1CWFLDA_{q0} and R2CWFLDA_{q0} demonstrate consistently better performance than the other techniques with respect to both k NN and NC classifiers.
3. From the perspective of the relative overall performance measured by the ROPI values, three of our proposed techniques, i.e. R1CWFLDA_{q0}, CS-OCWFLDA, and CS-UCWFLDA win the first three places with respect to both k NN and NC

¹⁹ See Table 4.2.

classifiers. The RFLDA and DCV techniques also demonstrate good performance. The QR decomposition makes the CS-OCWFLDA and OLDA always outperform the CS-UCWFLDA and ULDA, respectively. Three regularized subspace based techniques by adding a multiple of identity matrix to the singular average within-class scatter matrix estimated in $C(\mathbf{S}_T)$, i.e. R1CWFLDA_{q0}, R1F-CWFLDA_{q0} and RFLDA, demonstrate better performance than the other three regularized subspace based techniques by using the MECS method, i.e. R2CWFLDA_{q0}, R2F-CWFLDA_{q0} and MUFLDA. The DLDA technique and its regularized version RDLDA as well as the FLDA_{s1} techniques perform not well in comparison with many other experimental techniques. The PCA+LDA always perform worst among all the experimental techniques with regard to both classifiers.

4. Two of our proposed techniques, i.e. R1F-CWFLDA_{q0} and R2F-CWFLDA_{q0} do not show promising performance as they have behaved in the well-sampled datasets. The plausible explanation is that in the high-dimensional and undersampled scenario, the pairwise two-class Bayes accuracies of most class pairs may achieve the common maximum values. Consequently, due to the definition of the between-class weights, the direction associated with each component pairwise scatter matrix in the weighted pairwise between-class scatter matrix will be equivalently emphasized in the eigenvalue decomposition. Therefore, the step-wise dimensionality reduction step cannot effectively differentiate the relatively least important discriminative direction to remove. This problem might be addressed by increasing the power of the Mahalanobis distances in the denominator position of the between-class weights. We attempt to address this problem in the future research work.

Generally speaking, the two generalized CWFLDA techniques, i.e. R1CWFLDA_{q0} and CS-OCWFLDA, demonstrate consistently better performance relative to the other techniques. It is interesting to note that the regularization scheme by simply adding a multiple of identity matrix to the singular average within-class scatter matrix estimated in $C(\mathbf{S}_T)$ can make the resulting algorithms, e.g. R1CWFLDA_{q0} and RFLDA demonstrate promising performance.

Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems

Figure 4.1: The RBPI (top), RAPI (middle) and ROPI (bottom) values of each experimental technique averaged over all the experimental datasets in terms of the k NN classifier.

Chapter 4: Generalized CWFLDA on High-dimensional and Undersampled Problems

Figure 4.2: The RBPI (top), RAPI (middle) and ROPI (bottom) values of each experimental technique averaged over all the experimental datasets in terms of the NC classifier.

4.6 Conclusions

In this chapter, we provided an extensive overview of the existing discriminative dimensionality reduction techniques capable of handling high-dimensional and undersampled problems along two main avenues: discriminative subspace analysis and discriminative manifold learning. Two categories of generalized CWFLDA techniques were proposed to handle the high-dimensional and undersampled problems, which can take advantage of the discriminatory information present in different subspaces while retaining their superior performance in the well-sampled scenario. The CS-CWFLDA techniques extract discriminant vectors subject to different conjugated orthogonality constraints from the null and column spaces of the average within-class scatter matrix in the sequential forward manner. The RS-CWFLDA techniques extract discriminant vectors subject to different conjugated orthogonality constraints from the column space of the regularized average with-class scatter matrix in the batch or sequential backward fashion. Although most weighted terms in the quotient class-wise weighted Fisher criterion degenerate since the pairwise two-class Bayes accuracies of most class pairs may achieve the maximum values in the high-dimensional and undersampled scenario, the degenerated weighting terms can still suppress the undesirable influences from the possibly existing outlier classes. Experiments were carried out on various types of data sources to evaluate the performance of the proposed techniques in comparison with several state-of-the-art FLDA variants. Two of the proposed generalized CWFLDA techniques, i.e. $R1CWFLDA_{q0}$ and CS-OCWFLDA, demonstrated consistently better overall performance relative to the other techniques. Some proposed techniques such as $R1F-CWFLDA_{q0}$ and $R2F-CWFLDA_{q0}$ did not perform as well as they behaved in the well-sampled scenario due to the degeneration of the weighting terms.

Chapter 5: Robust Growing Neural Gas Algorithm with Application in Cluster Analysis

5.1 Introduction

Cluster analysis [44]-[46] is a crucial and powerful tool for exploring and discovering the underlying data substructures by crisply or softly partitioning a set of N available n -dimensional feature vectors (also called training samples or data points), i.e. $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \mid \forall \mathbf{x}_i \in \mathbf{R}^n, i = 1, 2, \dots, N\}$ into C ($1 \leq C \leq N$) natural groups (so-called clusters) such that each feature vector can be assigned to per cluster with a certain degree of membership. Cluster analysis has found applications in diverse fields [233]-[238] ranging from pattern recognition, data mining, computer vision and communication to information retrieval and bioinformatics.

There exist no definite categorization standards for grouping different clustering techniques. Traditionally, clustering methods can be dichotomized into hierarchical clustering and partitioning clustering. Hierarchical approaches sequentially build nested clusters with the graphical representation known as dendrogram. Partitioning approaches directly assign all the available feature vectors according to some criteria, such as similarity and density, into different clusters. Representative algorithms for the partitioning clustering include Prototype-based Clustering (PBC) [47], density-based clustering [48], [49] and graph theory based clustering [50], etc. Agglomerative [52] and divisive [53] clustering algorithms constitute two main types of hierarchical clustering. In recent years, many novel clustering techniques have been developed such as neural network-based clustering (e.g. self-organizing maps [18], Neural Gas (NG) networks [54] and adaptive resonance theory network [55], etc.), kernel clustering [56], [57], spectral clustering [59], information clustering [58], relationship-based clustering [60], subspace clustering [51], evolutionary clustering [61], and online clustering [62], [63], [64], etc. These clustering algorithms can either fit into

one of the two traditional categories or be regarded as a unique category of clustering techniques.

Research work in this thesis focuses on the PBC algorithms with point prototypes, which are simply characterized by reference vectors and the Euclidean distance metric. The PBC is probably the most popular class of partitioning clustering methods and can also be extended to the hierarchical approaches via superposition. Hard clustering and soft clustering can be regarded as two major branches of the PBC algorithms. Hard clustering assigns each feature vector to exactly one cluster with the nearest prototype, while soft clustering assigns each feature vector to all clusters with different degrees of membership. According to the updating strategies, the PBC algorithms can also be dichotomized into the batch and sequential methods.

Hard c -means (HCM), also called k -means [1], [2], [71], is a well-known hard clustering algorithm, which is suitable for well-separated clusters. However, in real applications, the underlying clusters usually overlap to some extent. Therefore, it is not reasonable to assign each feature vector to a single cluster with the full certainty. The famous Fuzzy c -means (FCM) algorithm [233], [239], devised by Bezdek [72], can well handle such case, in which prototypes are derived by using the Alternative Optimization (AO) method. The fuzzy competitive learning algorithm [240] is intrinsically same with the FCM but uses the sequential learning strategy and the gradient descent scheme to derive prototypes.

Although widely applied, the robustness of many existing PBC algorithms may not be satisfactory. For example, the performances of many conventional PBC algorithms are sensitive to initializations, input sequence orders, the presence of outliers and the correct assumption of the underlying cluster numbers for a given dataset. In the past few decades, many improved PBC versions have been developed to tackle these robustness issues. We describe some representative methods as follows.

- **Initialization and input sequence order problems**

Due to the use of winner-take-all strategy [41], the performance of the conventional HCM algorithm is very sensitive to initializations. If prototypes are initialized improperly, some of them may never win and get updated during the whole training process, which is known as the prototype-under-utilization or “dead nodes” problem. Comparatively, the FCM algorithm may mitigate the “dead node” problem by using the fuzzy membership concept. Specifically, any feature vector will, to some degree, contribute to the updating of all prototypes. However, due to the use of the AO scheme, different initializations for the FCM may possibly lead to different clustering results. In the sequential learning methods, the initialization problem is intertwined with the input sequential order problem. Specifically, with respect to a certain initial state, randomly shuffled input sequence orders may lead to different clustering results.

Many approaches have been devised to tackle the initialization and input sequence order problems. Frequency sensitive competitive learning algorithm [235] is a typical method that can eliminate the dead nodes by gradually increasing the winning chance for those infrequently winning prototypes. Observing that sub-sampling can provide information about the underlying joint probability density function that generate the available feature vectors, Bradley [242] refined the initial state of the HCM algorithm based on an effective approach estimating the modes of the distribution. The k -harmonic means algorithm [243], proposed by Zhang, modified the criterion function of the HCM algorithm by substituting the minimum distance from each feature vector to its nearest prototype with the harmonic average of the distances from such feature vector to all prototypes. Consequently, each available feature vector can influence the updating of all prototypes to effectively cope with the initialization problem. The hard robust clustering algorithm developed by Yang *et al.* [244] employs the sequential learning strategy. At the presence of a feature vector, its nearest prototype together with other prototypes are all updated but with distinct updating strengths. Although the so far mentioned robust variants claim to successfully solve the “dead nodes” problem, they may take a risk of getting stuck into some inferior local minima if the initial states are not carefully chosen or the input sequence are not properly ordered.

- **Outlier Problem**

In cluster analysis, outliers can be regarded as the feature vectors faraway from any clusters. Consequently, outliers should have low degrees of membership with respect to all clusters. On the other hand, inliers denote the feature vectors belonging to some cluster with high degrees of membership. The sum of squared quantization error (SSQE), defined by the summation of squared Euclidean distances from the available feature vectors to all the prototypes weighted by the corresponding degrees of membership, is a widely used criterion function in the PBC algorithms. However, without differentiating the inliers' and outliers' influences, minimization of SSQE may often lead to unsatisfactory clustering results. For both the HCM and FCM algorithms, outliers may significantly influence the prototype updating process. In order to yield satisfactory clustering results, it is desirable to identify the existing outliers and suppress their influences on the updating of the prototypes.

As an outlier resistant clustering algorithm, F should have the following characteristics [245]: P_1 : inliers and outliers should be differentiated during the minimization of criterion function such that outliers exert negligible influences on the updating of the prototypes. P_2 : final prototypes obtained by F on a contaminated dataset should not significantly deviate from those attained by F on the corresponding "clean" dataset. In recent years, many outlier resistant clustering algorithms have been developed.

The partition around medoides algorithm [45], as one of the famous k -medoides clustering techniques [246], aims at finding k medoides out of the available feature vectors and accordingly assigning each of the remaining feature vectors to its most similar medoid so as to minimize the SSQE. Here, the medoid means the feature vector most centrally located within a specific cluster. Since the criterion function of the partition around medoides cannot differentiate the influences coming from the inliers or the outliers during the training process, i.e. violating P_1 , the final clustering results may possibly deviate from the desirable one. The noise cluster approach [247], [248] defines an extra noise cluster, which aims at including all the existing outliers in the given dataset and assigning them with high degrees of membership. Therefore, the prescribed C clusters can be almost composed of the inliers. This algorithm can

perform well given a reasonable noise cluster assignment cost, which, however, is usually problem-dependent and should be determined by the trail and error scheme. Moreover, due to the use of the AO scheme, the final clustering result will inevitably depend on the initial state.

The Possibilistic c -means (PCM) developed by Krishnapuram and Keller [249] and the least biased fuzzy clustering algorithm proposed by Beni and Liu [250] belong to the possibilistic clustering methods, in which any available feature vector with the higher possibility to be an outlier will give little influence on updating prototypes. Although P_1 is carefully considered, these two algorithms are fairly sensitive to different initializations. Specifically, the PCM usually employs the FCM to yield its initial states. Since the FCM itself is sensitive to the presence of outliers, the PCM may be bad initialized. Moreover, PCM may generate coincident prototypes since all the prototypes are independently optimized during the training process. Timm *et al.* [251]-[253] proposed two possibility fuzzy clustering algorithms to avoid the coincident cluster problem. Pal *et al.* [254] proposed a mixed c -means algorithm, called the fuzzy possibilistic c -means algorithm, which simultaneously considers the membership (probability) and typicality (possibility) of each available feature vector with respect to a specific cluster. The typicality values of the outliers can be very small although their membership values may be large. Therefore, P_1 is satisfied. However, this algorithm may sometimes violate P_2 because the mixture of typicality and probability may still generate large membership values with respect to some existing outliers in the AO process. Moreover, when the dataset is of a very large size, the typical values of all the available feature vectors with respect to a specific cluster will become fairly small since the summation of the typical values over all the available feature vectors is required to equal “1”, and consequently the algorithm may perform similar to the FCM. To overcome this problem, Pal *et al.* [255] proposed a possibility fuzzy c -means algorithm recently, which removes such summation to “1” constraint. By introducing a credibility value for each available feature vector to represent the typicality of such feature vector with respect to all the existing clusters, Chintalapudi and Kam [245] presented a credibilistic fuzzy c -means clustering algorithm. Since calculation of the credibility value for each available feature vector

depends on the choice of the neighborhood of such feature vector, this method is sensitive to parameter settings. Density weighted fuzzy c -means algorithm, proposed by Chen and Wang [256], aims at eliminating the possible outliers before applying the clustering algorithms. As a byproduct, some inliers may also be wrongly discarded as outliers. Therefore, the performance of this method is problem-dependent.

By modifying the conventional criterion functions, many improved PBC algorithms were developed. For example, by replacing the Euclidean distance metric used in the SSQE criterion function with some robust distance metrics, alternative c -means algorithms [257], generalized fuzzy c -means clustering strategies using L_p norm [258], robust deterministic annealing algorithms by pruning outliers [259] were proposed. In general, the robust distance metrics can better suppress the influences from outliers than the Euclidean distance. However, these algorithms are actually sensitive to initializations since, in these batch learning methods, the initial states of the prototypes may determine which features vectors will be regarded as outliers in the subsequent training process. A bad initial state may mistake some inliers for outliers and wrongly suppress their influences on the updating of prototypes. Further, some parameters such as the scale factor are hard to determine in these robust distance metrics. Fuzzy robust clustering method, proposed by [260] can perform well on datasets with outliers only if the initial state is carefully configured.

To be a robust PBC algorithm, two conditions should be satisfied: (1) as training proceeds, prototypes should move towards the underlying clusters, and this procedure should be less sensitive to initializations and input sequence orders. (2) The influences from potential outliers whose distances from the current prototypes are large should be suppressed. Most of the existing robust PBC variants described above only satisfy the second condition while violating the first condition.

- **Choice of the number of prototypes**

Robustness of a PBC algorithm also depends on the proper choice of the number of prototypes. However, most of the conventional PBC algorithms and their robust

variants as described above can only subjectively presume this crucial number. In general, only if the number of prototypes is equal to the actual number clusters, each prototype may be able to detect and represent a meaningful cluster. Otherwise, the final clustering results may be unsatisfactory or even meaningless. Determining the number of underlying clusters remains a challenging problem. Some researchers suggested initializing the clustering algorithm with a large number of prototypes, and then employing the merging or pruning scheme during the training process to reduce the number of prototypes to an optimal value. Some representative examples are robust competitive clustering algorithm [261], [262], robust vector quantization algorithm [236], etc. However, due to lack of the prior knowledge about the given dataset, it might be hard to choose the starting prototype number. Excessively large number may significantly increase the computational cost. Another class of methods employs the growing strategy, which starts with a few prototypes, and then inserts new prototypes via certain rules. Comparatively, the growing methods often require less computational cost than the pruning methods. Usually the maximum number of prototypes to grow to and the extreme score of some specific validation measure are used to stop the growing process and determine the optimal number of clusters as well as their center positions. The X-means algorithm [263] and G-means algorithm [264] are two representative methods. However, the performance of most existing growing methods may significantly deteriorate when the given dataset is contaminated by outliers. Consequently, even if the number of prototype reaches the number of underlying clusters, the detected cluster center positions may deviate from the actual ones due to the presence of outliers. Although many robust PBC algorithms have been developed as described above, only a few of them using the pruning methods, such as the algorithms proposed by Bischof *et al.* [236] and by Frigui and Krishnapuram [262], claim to have the ability to detect the number of underlying clusters and their center positions at the presence of outliers.

In this chapter, a novel Robust Growing Neural Gas (RGNG) algorithm will be presented. By introducing into the Growing Neural Gas (GNG) framework [75] the outlier resistant strategy, the adaptive modulation of learning rates with respect to each prototype and the cluster repulsion scheme used to avoid the coincident prototypes, the

proposed RGNG algorithm can effectively suppress the sensitivity to initializations, input sequence orders and the presence of outliers at each growing stage. The RGNG starts with a small number of prototypes, usually 2, and gradually inserts new prototypes. The adaptive learning rates with respect to the winning prototype and its direct topological neighbors are introduced to differentiate the prototypes according to their insertion order such that newly inserted prototypes should be assigned with larger learning rates so as to better explore some unidentified clusters. At the completion of each growing stage, Minimum Description Length (MDL) measure [79] is calculated to evaluate the cluster validity with respect to the current number of prototypes and their positions. Consequently, the RGNG can output the optimal number clusters and their center positions with the minimal MDL value over the whole growing process. Furthermore, the topological relations among all the prototypes can be established [265]. Compared with the original GNG algorithm incorporating the MDL algorithm, named GNG-M, the proposed RGNG algorithm can effectively detect the number of underlying clusters when dealing with a dataset containing many outliers.

The remainder of this chapter is organized as follows. In section 5.2, we revisit the GNG algorithm. The proposed RGNG algorithm is described in section 5.3. Experimental results on 5 artificial and UCI benchmark datasets are discussed in section 5.4. Section 5.5 summarizes this chapter.

5.2 Review of growing neural gas algorithm

The Neural Gas (NG) network algorithm, proposed by Martinetz [54], has been successfully applied [266]-[270] to cluster analysis, vector quantization, pattern recognition and topology representation, etc. It is similar to Kohonen's self-organizing feature map [18] but exempted from the fixed topological arrangement of the network. As a single hidden-layer soft competitive learning neural network, it not only adjusts the nearest prototype with respect to an input feature vector but also updates the remaining prototypes according to their proximities to this input feature vector by using the soft-max updating rule [41]. Three attractive merits [54] of the NG algorithm

are: (1) faster convergence to the low distortion error; (2) distortion error lower than those resulting from HCM clustering, maximum-entropy clustering [271] and Kohonen's self-organizing feature map; (3) obeying the stochastic gradient descent [94] on an explicit energy surface.

In the NG algorithm, the updating strength for the C prototypes $\mathbf{w}_i, i = 1, \dots, C$ depends on their positions in the “neighborhood ranking” list. When a feature vector $\mathbf{x}$ is presented, we determine the neighborhood ranking $(\mathbf{w}_{i_0}, \mathbf{w}_{i_1}, \dots, \mathbf{w}_{i_{C-1}})$ of this prototypes with $\mathbf{w}_{i_0}$ being the closest to $\mathbf{x}$ in terms of the Euclidean distance, $\mathbf{w}_{i_1}$ being the second closest to $\mathbf{x}$, and $\mathbf{w}_{i_k}, k = 1, \dots, C-1$ being the prototype for which there are k prototypes $\mathbf{w}_j$ with $\|\mathbf{x} - \mathbf{w}_j\| < \|\mathbf{x} - \mathbf{w}_{i_k}\|$ where $\|\cdot\|$ denotes the Euclidean distance. The ranking index associated with each prototype $\mathbf{w}_i$ with respect to $\mathbf{x}$ is denoted by $k_i(\mathbf{x}, \mathbf{w})$. According to the Hebbian learning rule [41], the prototype $\mathbf{w}_i, i = 1, \dots, C$ can be updated by:

$$\Delta \mathbf{w}_i = \varepsilon(t) \cdot h_\lambda(k_i(\mathbf{x}, \mathbf{w})) \cdot (\mathbf{x} - \mathbf{w}_i), i = 1, \dots, C \quad (5.1)$$

The learning rate $\varepsilon(t) \in [0, 1]$ denotes the overall updating strength, which usually has an exponentially decreasing formula as $\varepsilon(t) = \varepsilon_i \cdot (\varepsilon_f / \varepsilon_i)^{t/(Maxiteration)}$ where t and $Maxiteration$ stand for the iteration step and the pre-specified maximum number of iterations, respectively. The term $h_\lambda(k_i(\mathbf{v}, \mathbf{w})) \in [0, 1]$ accounts for the topological arrangement of $\mathbf{w}_i$ within the feature space. The exponential function $\exp(-k/\lambda)$ was suggested in [54] for $h_\lambda(k)$ to obtain the overall best performance. The parameter λ controls the number of prototypes that can significantly change their positions in the updating step and usually monotonically decreases with the iteration step t by $\lambda(t) = \lambda_i \cdot (\lambda_f / \lambda_i)^{t/(Maxiteration)}$.

Acting as a PBC algorithm, the NG is closely related to the FCM. Instead of using the fuzzy membership u_{ij} , $1 \leq i \leq C, 1 \leq j \leq N$ as in the FCM, the NG employs $h_\lambda(k_i(\mathbf{x}, \mathbf{w})) / C(\lambda)$ to assign feature vector $\mathbf{x}$ to each prototype $\mathbf{w}_i, i = 1, \dots, C$. The criterion function of the NG algorithm is expressed by:

$$E_{ng} = \frac{1}{2C(\lambda)} \sum_{i=1}^C \int p(\mathbf{x}) h_\lambda(k_i(\mathbf{x}, \mathbf{w})) \|\mathbf{x} - \mathbf{w}_i\|^2 d\mathbf{x}$$

$$\text{with } C(\lambda) = \sum_{i=1}^C h_\lambda(k_i) = \sum_{k=0}^{C-1} h_\lambda(k) \quad (5.2)$$

The updating rule in Eq. (5.1) can be derived by applying the stochastic gradient descent scheme to minimize this criterion function with respect to each prototype $\mathbf{w}_i, i = 1, \dots, C$ [54]. To obtain a set of satisfactory prototypes, we start with a large λ value and gradually decrease it during the updating process and correspondingly the algorithm transits from a soft model to a hard one. By using this deterministic annealing [271], [273] like strategy, we expect good local minima of the criterion function to emerge slowly and thereby efficiently prevent the prototypes from getting stuck into the inferior suboptimal states. In addition, compared with batch learning algorithms such as HCM and FCM, the NG algorithm is less sensitive to initializations due to the use of the sequential learning scheme and the neighborhood cooperation rule. The detailed implementation of NG algorithm can be referred to [54].

Originating from the NG algorithm, Fritzke proposed an incremental self-organizing network with a variable topological structure, called Growing Neural Gas (GNG) algorithm [75]. It combines the growing mechanism inherited from the growing cell structures algorithm [275] with the topology formation via the competitive Hebbian learning scheme [241], [274]. For each prototype $\mathbf{w}_i, i = 1, \dots, C$, a set of edges emanating from it is defined to represent the connection between it and its direct topological neighbors. The GNG algorithm starts with a few prototypes, usually two. The new prototypes will be successively inserted around the prototype with the largest local accumulated error after a pre-specified number of training epochs. Here, one

training epoch denotes that each available feature vector has been used for training once time. The growing process stops when the pre-specified maximum number of prototypes is reached or some performance requirements are met. In the GNG, at the presence of a feature vector $\mathbf{x}$, the updating is only applied to the winning prototype $\mathbf{w}_{s_1}$ located closest to $\mathbf{x}$ and its direct topological neighbors $\mathbf{w}_i, \forall i \in N_{s_1}$, where N_{s_1} denotes the set of indices of the direct topological neighbors of $\mathbf{w}_{s_1}$ that are connected with $\mathbf{w}_{s_1}$ by edges. Furthermore, the updating strength is different for the winning prototype and its topological neighbors and remains constant during the whole training process. The updating rule of the GNG algorithm is defined by:

$$\Delta \mathbf{w}_{s_1} = \varepsilon_b \cdot (\mathbf{x} - \mathbf{w}_{s_1}), \quad \Delta \mathbf{w}_i = \varepsilon_n \cdot (\mathbf{x} - \mathbf{w}_i), \quad \forall i \in N_{s_1} \quad (5.3)$$

Here, ε_b and ε_n represent the constant learning rates for the winning prototype and its topological neighbors, respectively. The GNG algorithm can detect inactive prototypes that never win within a certain time interval by tracing the change of the age variable associated with each edge. Therefore, it is able to modify the network topology by removing the edges with their age variables not being refreshed within the time interval $\alpha_{\max}$ and the possibly resulting inactive prototypes. Unlike the NG algorithm, the neighborhood ranking is no longer used in the GNG since the growing process coupled with the local neighborhood updating rule can achieve the similar effect. The pseudo-code of the GNG algorithm is given in Table 5.1.

Table 5.1: Pseudo-code of the Growing Neural Gas (GNG) algorithm

Initialize a set of prototypes (usually 2) $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2\}$ randomly, the learning rates ε_b and ε_n and a connection set $\mathbf{C}$, $\mathbf{C} \subset \mathbf{W} \times \mathbf{W}$ to an empty set, i.e. $\mathbf{C} = \emptyset$. Set the maximum number of prototypes to grow to *pre_numnode* and the maximum training epoch *Max_iter* at each growing stage. Set the initial training epoch number: $m = 0$ and the iteration step in the training epoch m : $t=0$. Accordingly, the overall iteration

step $iter$ at one growing stage is: $iter = m \cdot N + t$. The set of available feature vectors is denoted by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

(a) While the current number of prototypes is smaller than or equal to $pre_numnode$ and some predefined performance requirement, if exists, is not met

(b) For $m = 0$ to $Max_iter - 1$

- Set $t=0$, and accordingly $iter = m \cdot N + t$
- Set $trainingset = \mathbf{X}$, i.e. include all available feature vectors into $trainingset$

(c) If $trainingset$ is not empty

- Draw randomly a feature vector $\mathbf{x}_t^m$, at the iteration step t in the training epoch m , from the $trainingset$
- Determine the nearest s_1 and the second nearest prototype s_2 ($s_1, s_2 \in \mathbf{W}$) by $s_1 = \arg \min_{i \in \mathbf{W}} \|\mathbf{x}_t^m - \mathbf{w}_i^{iter}\|$, $s_2 = \arg \min_{i \in \mathbf{W} \setminus \{s_1\}} \|\mathbf{x}_t^m - \mathbf{w}_i^{iter}\|$ with $iter = m \cdot N + t$
- If a connection between s_1 and s_2 does not exist already, create it: $\mathbf{C} = \mathbf{C} \cup \{(s_1, s_2)\}$. Set the *age* of the connection between s_1 and s_2 to 0, i.e. $age_{(s_1, s_2)} = 0$
- Add the squared Euclidean distance from feature vector $\mathbf{x}_t^m$ to the winning prototype to a local accumulated error variable: $\Delta \mathbf{E}_{s_1} = \|\mathbf{x}_t^m - \mathbf{w}_{s_1}\|^2$
- Update the winning prototype and its direct topological neighbors by $\Delta \mathbf{w}_{s_1} = \varepsilon_b \cdot (\mathbf{x} - \mathbf{w}_{s_1})$, $\Delta \mathbf{w}_i = \varepsilon_n \cdot (\mathbf{x} - \mathbf{w}_i)$, $\forall i \in N_{s_1}$ where N_{s_1} is the set of indices of the direct topological neighbors of s_1 .
- Increment the *age* of all edges emanating from s_1 by 1: $age_{(s_1, i)} = age_{(s_1, i)} + 1, \forall i \in N_{s_1}$
- Remove edges with *age* values larger than $\alpha_{\max}$. If this results in some prototypes with no emanating edges, remove those prototypes as well
- Increment the iteration step t by 1 and remove the used feature vector $\mathbf{x}_t^m$ from the set $trainingset$.

(c) End If

(b) End For

(d) If the current number of prototypes do not reach $pre_numnode$

- Determine the prototype q with the maximum accumulated error:

$$q = \arg \max_{i \in \mathbf{W}} \mathbf{E}_i .$$
- Determine, among the neighbors of q , the prototype f with the maximum accumulated error: $f = \arg \max_{i \in N_q} \mathbf{E}_i .$
- Add a new prototype r into the network and interpolate its prototype position from q and f :

$$\mathbf{W} = \mathbf{W} \cup \{\mathbf{w}_r\}, \quad \mathbf{w}_r = (2\mathbf{w}_q + \mathbf{w}_f)/3$$

- Insert edges connecting the new prototype r with the prototypes q and f , respectively, and remove the original edge between q and f :

$$\mathbf{C} = \mathbf{C} \cup \{(r, q), (r, f)\}, \quad \mathbf{C} = \mathbf{C} \setminus \{(q, f)\}$$

- Decrease the accumulated errors of q and f by a fraction α :

$$\Delta \mathbf{E}_q = -\alpha \cdot \mathbf{E}_q, \quad \Delta \mathbf{E}_f = -\alpha \cdot \mathbf{E}_f$$

- Interpolate the error value of r from q and f : $\mathbf{E}_r = (\mathbf{E}_q + \mathbf{E}_f)/2 .$
- Decrease the error values of all prototypes: $\Delta \mathbf{E}_i = -\beta \cdot \mathbf{E}_i, \quad \forall i \in \mathbf{W} .$

(d) End If

(a) End While

5.3 Robust growing neural gas algorithm

Robustness is an important characteristic of a good clustering algorithm. According to Huber [73], [74]: a robust algorithm should have the following properties: (1) it should achieve a reasonably good accuracy at the assumed model; (2) small deviations from the model assumptions should impair the performance only by a small amount; (3) larger deviations from the model assumption should not cause a catastrophe. Conventional clustering algorithms, such as the PBC algorithms, often satisfy

condition (1). The main robustness issues associated with them are the sensitivity to initializations, input sequence orders and the presence of outliers.

The GNG algorithm seldom suffers from the initialization problem due to its use of growing mechanism, sequential learning strategy and soft-max updating rule. However, if the input sequence order is not chosen carefully, unsatisfactory clustering results may be obtained. Specifically, if a given dataset contains outliers, the conventional prototype updating rule used in the GNG may fail to differentiate the influences from outliers and inliers and consequently the prototypes may be significantly attracted towards those sequentially presented outliers.

By retaining the merits of the conventional GNG algorithm while tackling its robustness problems, we propose a RGNG algorithm, which can be less sensitive to initializations, input sequence orders and the presence of outliers during the training process. Moreover, by using the MDL value calculated at the completion of each growing stage as the cluster validity index, we can automatically determine the optimal number clusters and their center positions, which correspond to the smallest MDL value over the whole growing process.

5.3.1 Outlier resistant strategy

Through retrospect of the updating rule (Eq. (5.3)) of the conventional GNG algorithm, we find that it is inherently sensitive to input sequence orders and the presence of outliers. To elucidate this problem, we reformulate Eq. (5.3) by:

$$\begin{aligned}\Delta \mathbf{w}_{s_1} &= \varepsilon_b \cdot \|\mathbf{x} - \mathbf{w}_{s_1}\| \cdot \frac{(\mathbf{x} - \mathbf{w}_{s_1})}{\|\mathbf{x} - \mathbf{w}_{s_1}\|} \\ \Delta \mathbf{w}_i &= \varepsilon_n \cdot \|\mathbf{x} - \mathbf{w}_i\| \cdot \frac{(\mathbf{x} - \mathbf{w}_i)}{\|\mathbf{x} - \mathbf{w}_i\|}, \quad \forall i \in N_{s_1}\end{aligned}\tag{5.4}$$

According to Eq. (5.4), it is clear that if an outlier $\mathbf{x}_o$ is fed to the GNG network, the Euclidean distances from $\mathbf{x}_o$ to each prototype to be updated, i.e. $\|\mathbf{x}_o - \mathbf{w}_k\|$,

$\forall k \in N_{s_1} \cup \{s_1\}$ become fairly large. Consequently, it can exert significant influence on the updating of prototypes. Moreover, when outliers occur at different positions in the input sequence with respect to a specific growing stage, the updating strength from these outliers to each prototype may differ and accordingly the final positions of the prototypes may become instable. For example, if most of the outliers are presented in the beginning, the updating strength from them to all prototypes is so strong that prototypes at the current growing stage may get stuck into the inferior local minima. To obtain satisfactory clustering results, the outliers should give little impact on the updating of prototypes. Therefore, any large distances from a feature vector $\mathbf{x}$ to the current prototypes to be updated $\|\mathbf{x} - \mathbf{w}_k\|$, $\forall k \in N_{s_1} \cup \{s_1\}$ at each growing stage should be modulated so as to eliminate the influence from the possible outliers. Accordingly, we modify the updating rule in Eq. (5.4) to a new formula as follows.

$$\begin{aligned}\Delta \mathbf{w}_{s_1} &= \varepsilon_b \cdot \sigma_{s_1}(iter) \cdot \frac{(\mathbf{x} - \mathbf{w}_{s_1})}{\|\mathbf{x} - \mathbf{w}_{s_1}\|} \\ \Delta \mathbf{w}_i &= \varepsilon_n \cdot \sigma_i(iter) \cdot \frac{(\mathbf{x} - \mathbf{w}_i)}{\|\mathbf{x} - \mathbf{w}_i\|}, \quad \forall i \in N_{s_1}\end{aligned}\quad (5.5)$$

In this new formula, we replace the Euclidean distance metric associated with the vector difference $\mathbf{x} - \mathbf{w}_k$, $\forall k \in N_{s_1} \cup \{s_1\}$ by an adaptive parameter $\sigma_k(iter)$, $\forall k \in N_{s_1} \cup \{s_1\}$ in order to control the influences from the possible outliers to the updating of prototypes. Here, the parameter $iter$ denotes the iteration step with respect to a certain growing stage, and once the number of prototypes is increased $iter$ will be reset. Assume that the maximum number of training epochs at each growing stage is specified as Max_iter . At the iteration $t, 1 \leq t \leq N$ in the training epoch $m, 0 \leq m \leq Max_iter - 1$, the overall iteration step $iter$ at the current growing stage is $iter = m \cdot N + t$ and the parameter $\sigma_k(iter)$ is defined by:

$$\sigma_k(iter) = \sigma_k^m(t) = \begin{cases} d_k^m(t) & \text{if } \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\| \geq d_k^m(t-1) \\ \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\| & \text{if } \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\| < d_k^m(t-1) \end{cases} \quad (5.6)$$

$$\text{with } d_k^m(t) = \begin{cases} 2 \left(d_k^m(t-1)^{-1} + \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\|^{-1} \right)^{-1} & \text{if } \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\| \geq d_k^m(t-1) \\ \frac{1}{2} \left(d_k^m(t-1) + \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\| \right) & \text{if } \|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\| < d_k^m(t-1) \end{cases} \quad (5.7)$$

$$d_k^m(0) = N \left(\sum_{j=1}^N \|\mathbf{x}_j - \mathbf{w}_k^{m \cdot N}\|^{-1} \right)^{-1} \quad (5.8)$$

where N , $\mathbf{x}_t^m$ and $\mathbf{w}_k^{iter}$ represent the total number of the available feature vectors, the feature vector presented at the iteration t in the training epoch m and the position of the prototype k at the overall iteration step $iter$, respectively. The term $d_k^m(t)$ denotes the restricting distance with respect to prototype $\mathbf{w}_k$, $k \in N_{s_1} \cup \{s_1\}$, which takes into account both historical and current moving distance information of $\mathbf{w}_k$ and is used to suppress the outlier influence. The initial restricting distance $d_k^m(0)$ is defined by Eq. (5.8) as the harmonic average of distances from all the available feature vectors $\mathbf{x}_j$, $j=1, \dots, N$ to the prototype $\mathbf{w}_k^{m \cdot N}$ before the first iteration in the training epoch m .

The meaning of $\sigma_k(iter)$ can be explained as follows. With respect to a specific growing stage, at $iter = m \cdot N$, i.e. before the first iteration step in the training epoch m , the harmonic average distance $d_k^m(0)$ for each prototype $\mathbf{w}_k$ is calculate by Eq. (5.8). In the training epoch m , when the iteration step t is increased by “1”, we average the current distance $\|\mathbf{x}_t^m - \mathbf{w}_k^{iter}\|$ and the historical restricting distance $d_k^m(t-1)$ to calculate the current restricting distance $d_k^m(t)$. Eq. (5.7) indicates that if the current Euclidean distance is greater than or equal to the historical restricting distance we take the harmonic average of them, otherwise we take the arithmetic average of them. In this way, the large distances from $\mathbf{x}_t^m$ to prototypes $\mathbf{w}_k^{iter}$, $\forall k \in N_{s_1} \cup \{s_1\}$ can be suppressed to some extent. Eq. (5.6) shows that if the current Euclidean distance is greater than or equal to the historical restricting distance, we take the current restricting distance as $\sigma_k(iter)$, otherwise the current Euclidean distance will be used as $\sigma_k(iter)$. Imagine that if one prototype is currently located at

a sparse region containing outliers, the influence from the inliers within actual cluster regions may be weakened. However, since the number of inliers in natural clusters is generally much larger than the number of training samples in the outlier regions, these inliers can gradually increase the restricting distance according to Eq. (5.7) for this prototype. Accordingly, this prototype can deviate from the outlier region step by step. Finally, if outliers are presented again later, the influence from these outliers on the updating of this prototype can be greatly suppressed. In this way, this prototype can gradually locate around an actual cluster center. Furthermore, when this prototype has entered a natural cluster, inliers therein can gradually decrease its restricting distance according to Eq. (5.7). Consequently, this prototype can be stabilized at the current cluster. Therefore, by taking into consideration both historical and current moving distance information to update each prototype, the RGNG algorithm can suppress the significant influence from the possible outliers at each growing stage and consequently mitigate the input sequence order problem.

In comparison with the conventional updating rule in Eq. (5.3), the new one in Eqs. (5.5)-(5.8) does not change the gradient descent direction while heuristically and adaptively adjusting the magnitude of the gradient descent along its original direction in order to resist the influence from the possible outliers.

5.3.2 Adaptive learning rates and cluster repulsion scheme

By using the above outlier resistant strategy, the training process at each growing stage with a certain number of prototypes become less sensitive to input sequence orders and the presence of outliers. Consequently, the prototypes at a certain growing stage can locate some meaningful clusters in the given dataset. When a new prototype is inserted, we expect that this new prototype should try to find some previously unidentified clusters while the already existing prototypes can refine the cluster center positions that they have found. However, in the conventional GNG algorithm, both the newly inserted prototype and the existing prototypes have the same learning rates ε_b and ε_n . Consequently, both of them may be significantly influenced by feature vectors belonging to the unidentified clusters. In fact, during the training process, most

existing prototypes have found some meaningful clusters and thus should gradually refine their positions towards the actual cluster center positions. On the other hand, the newly inserted prototypes should avoid finding the same clusters as those found by the existing prototypes and try to explore some previously unidentified clusters. To realize this idea, we introduce the adaptive learning rates ε_b and ε_n as follows, which rely on the insertion order the prototypes.

$$\begin{aligned}\varepsilon_b^l &= \varepsilon_{bi}^l \cdot \left(\varepsilon_{bf}^l / \varepsilon_{bi}^l \right)^{prenode^l / pre_numnode} \\ \varepsilon_n^l &= \varepsilon_{ni}^l \cdot \left(\varepsilon_{nf}^l / \varepsilon_{ni}^l \right)^{prenode^l / pre_numnode}, \quad l = 1, \dots, C\end{aligned}\quad (5.9)$$

where ε_{bi}^l , ε_{bf}^l , ε_{ni}^l , ε_{nf}^l , $pre_numnode$ and C denote the initial and final values of ε_b^l and ε_n^l with respect to prototype l , the pre-specified maximum number of prototypes to grow to and the number of prototypes at the current growing stage, respectively. The parameter $prenode^l$ denotes the ranking counter of prototype l . For a newly inserted prototype l , $prenode^l$ is initialized to “0” and increased by “1” once a new prototype is inserted subsequently. In this way, the “older” a prototype is, the closer its learning rates ε_b^l and ε_n^l approach the final values. Here, we choose the initial value to be ten times the final value for each of these two learning rates such that the newly inserted prototypes can have the larger flexibility to explore the previously unidentified clusters in the given dataset whilst those already existing prototypes can refine the cluster center positions that they have found.

However, it may possibly happen that two prototypes simultaneously exploit the same cluster since each inlier of this cluster can attract both the winning prototype and its direct topological neighbors. To address this prototype coincident problem [251]-[253], we introduce a mutual repulsion scheme for prototypes within the local neighborhood. Specifically, given a feature vector $\mathbf{x}$, the corresponding winning prototype and its direct topological neighbors can be determined. Then, the repulsive force is exerted along the direction from the winning prototype to its topological neighbors, the magnitude of which is defined by an integer multiple of the average distance from the

winning prototype to all its direct topological neighbors. Moreover, an exponential function is used to control the strength of the repulsive force according to the proximity between the topological neighbors and the corresponding winning prototype. We implement the above ideas and accordingly establish the updating rule used in the proposed RGNG algorithm as follows.

$$\Delta \mathbf{w}_{s_1} = \varepsilon_b^{s_1} \cdot \sigma_{s_1}(iter) \cdot \frac{(\mathbf{x} - \mathbf{w}_{s_1})}{\|\mathbf{x} - \mathbf{w}_{s_1}\|} \quad (5.10)$$

$$\Delta \mathbf{w}_i = \varepsilon_n^i \cdot \sigma_i(iter) \cdot \frac{(\mathbf{x} - \mathbf{w}_i)}{\|\mathbf{x} - \mathbf{w}_i\|} + \exp\left(-\frac{d_{s_1 i}}{\varsigma}\right) \cdot \beta \cdot \frac{\sum_{i=1}^{|N_{s_1}|} d_{s_1 i}}{|N_{s_1}|} \cdot \frac{(\mathbf{w}_i - \mathbf{w}_{s_1})}{\|\mathbf{w}_i - \mathbf{w}_{s_1}\|}, \forall i \in N_{s_1}$$

where β , $|N_{s_1}|$ and $d_{s_1 i}$ denote the multiplier in the definition of the repulsive force, the number of the direct topological neighbors of the winning prototype $\mathbf{w}_{s_1}$ and distance from the winning prototype $\mathbf{w}_{s_1}$ to its topological neighbor $\mathbf{w}_i$, $\forall i \in N_{s_1}$, respectively. The parameter ς controls the decreasing effect of the repulsive force based on the proximity of two prototypes. In our following experiments, we choose values 0.1 and 2 for ς and β , respectively. According to the preliminary trials, clustering performance is not very sensitive to the values of these two parameters.

5.3.3 Determination of the optimal number of clusters

In cluster analysis tasks, automatic determination of the optimal number of clusters remains an open problem. Most conventional clustering algorithms need to pre-specify the number of prototypes, which is assumed to equal the number of underlying clusters. However, a subjectively presumed number may possibly lead to unsatisfactory clustering results. In the past few decades, many cluster validity measures have been proposed to evaluate the quality of the clustering results obtained under different algorithmic configurations, e.g. different pre-specified numbers of prototypes. Among them, statistics or information theory based model selection criteria, such as Akaike's Information Criterion (AIC), Bayesian Information Criterion

(BIC), Minimum Description Length (MDL) and Minimum Message Length (MML) have received much attention [30], [236], [263], [277]. Although these model selection criteria are commonly subject to the Occam's Razor principle, they differ in the originations and formulations. The AIC [278] explicitly subtracts the dimension of the model from the model log likelihood with respect to the available feature vectors so as to penalize the model with high goodness-of-fit but large complexity. The BIC [279] was derived within the Bayesian framework to approximate the posterior probability of the model, up to normalization. In comparison with the AIC, the BIC tends to choose more parsimonious models. Their relative pros and cons have been argued extensively in the literature [280]. Both MML [281] and MDL [282] conform to the minimum encoding length principle, which means that an optimal model to describe the available feature vectors should have the shortest code length to encode the model complexity and the goodness-of-fit between the available feature vectors and the model. In contrast, the MML takes into consideration the structure of the model parameter space (e.g. the priors of parameters and the Fisher information matrix) when encoding the model complexity. It is usually used to simultaneously choose the model structure and estimate the associated model parameters. The MDL generally employs some intuitive and simple ways to encode the model complexity and can be used to choose the optimal model structure with the associated model parameters under such structure estimated by other techniques. The MDL and MML can incorporate various encoding schemes [283], [284], which may result in many variants suitable for different applications. It is worth noting that one MDL formulation proposed by Rissanen is formally but not conceptually equivalent to the BIC. All the above-mentioned model selection criteria can be coupled with a suitable clustering algorithm to detect the optimal number of clusters. Some successful examples have been reported in literatures [30], [79], [236], [263], [277]. Our proposed RGNG algorithm follows the MDL framework suggested in [236], which can distinguish between inliers and outliers and encode them in different manners.

Specifically, we formulate the task of determination of the optimal number of clusters as seeking for the extreme MDL value calculated according to the clustering results over different numbers of prototypes on a given a dataset $\mathbf{V}$ consisting of N feature

vectors. Due to the existence of outliers, we divide $\mathbf{V}$ into two subsets $\mathbf{I}$ and $\mathbf{O}$, which are composed of inliers and outliers, respectively. Accordingly, the MDL measure can be expressed by:

$$MDL(\mathbf{V}, \mathbf{W}) = modL(\mathbf{I}, \mathbf{W}) + errorL(\mathbf{I}, \mathbf{W}) + modL(\mathbf{O}) \quad (5.11)$$

where $\mathbf{W}$ represents a set of C prototypes, i.e. $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C\}$. We evaluate the complexity of the entire clustering model by the term $modL(\mathbf{I}, \mathbf{W})$. Its encoding length is defined by the summation of two terms: (1) the length of encoding all the prototypes, denoted by $L(\mathbf{W})$; (2) The length of encoding the prototype indices for all the available feature vectors, denoted by $L(\mathbf{V}(\mathbf{W}))$. The term $errorL(\mathbf{I}, \mathbf{W})$ denotes the residual error when describing all inliers $\mathbf{I}$ by the prototype set $\mathbf{W}$. The description length of the outlier set $\mathbf{O}$, denoted by $modL(\mathbf{O})$, is usually encoded in the same way as the prototype set. The model capacity of describing the whole dataset $\mathbf{V} = \mathbf{I} + \mathbf{O}$ is reflected by the last two terms in Eq. (5.11). According to [236], we implement the MDL measure as follows.

$$MDL(\mathbf{V}, \mathbf{W}) = C \cdot K + N \cdot \log_2 C + \sum_{i=1}^C \sum_{\mathbf{x} \in S_i} \sum_{k=1}^n \max \left(\log_2 \left(\frac{|\mathbf{x}_k - \mathbf{w}_{ik}|}{\eta} \right), 1 \right) + |\mathbf{O}| \cdot K \quad (5.12)$$

where C , n , N , η , S_i and $|\mathbf{O}|$ represent the number of prototypes, the feature space dimensionality, the total number of available feature vectors in the given dataset, the resolution of encoding, the receptive field of prototype $\mathbf{w}_i$ to contain inliers and the cardinality of the outlier set, respectively. Here, K is the number of bits to encode one prototype, which can be calculated in terms of the maximum feature-wise value difference and the encoding resolution, i.e.

$$K = \left\lceil \log_2 \left(\frac{\max_{1 \leq k \leq n} (\mathbf{x}_k^{\max} - \mathbf{x}_k^{\min})}{\eta} \right) \right\rceil \quad (5.13)$$

For the term $L(\mathbf{V}(\mathbf{W}))$, we use the fixed length encoding scheme [236], in which each available feature vector is encoded by $\log_2 C$ bits and accordingly $L(\mathbf{V}(\mathbf{W})) = N \cdot \log_2 C$. If the residual error with respect to feature vector $\mathbf{x}$ and the prototype $\mathbf{w}_i$ is feature-wise independent, the encoding length of $\mathbf{x}$ with respect to $\mathbf{w}_i$ can be calculated by $L(\mathbf{x} - \mathbf{w}_i) = \sum_{k=1}^n \max\left(\log_2\left(\frac{|\mathbf{x}_k - \mathbf{w}_{ik}|}{\eta}\right), 1\right)$. Accordingly, the term $errorL(\mathbf{I}, \mathbf{W})$ can be defined by:

$$errorL(\mathbf{I}, \mathbf{W}) = \sum_{i=1}^C \sum_{\mathbf{x} \in S_i} L(\mathbf{x} - \mathbf{w}_i) = \sum_{i=1}^C \sum_{\mathbf{x} \in S_i} \sum_{k=1}^n \max\left(\log_2\left(\frac{|\mathbf{x}_k - \mathbf{w}_{ik}|}{\eta}\right), 1\right) \quad (5.14)$$

To calculate the MDL value by Eq. (5.12) at the completion of each growing stage, outliers with respect to the current prototypes need to be determined. According to [236], if a feature vector $\mathbf{x}$ is an indeed outlier, the length of encoding it by the index of its nearest prototype and the corresponding residual error will be larger than directly encoding it like a prototype. Therefore, by checking the change of the MDL value defined as follows when $\mathbf{x}$ is artificially moved from the inlier set to the outlier set, we can judge whether $\mathbf{x}$ is an outlier or not.

$$\begin{aligned} \Delta MDL &= \{L((\mathbf{I} - \{\mathbf{x}\})(\mathbf{W})) + K\} - \{L(\mathbf{I}(\mathbf{W})) + L(\mathbf{x} - \mathbf{w}_i)\} - \varphi_{S_i} \cdot K \\ &= \{(N-1) \cdot \log_2(C - \varphi_{S_i}) + K\} \\ &\quad - \left\{N \cdot \log_2 C + \sum_{k=1}^n \max\left(\log_2\left(\frac{\|\mathbf{x}_k - \mathbf{w}_{ik}\|}{\eta}\right), 1\right)\right\} - \varphi_{S_i} \cdot K \end{aligned} \quad (5.15)$$

where φ_{S_i} equals “1” if $\mathbf{x}$ is the only feature vector in the receptive field S_i before the movement, otherwise it equals “0”. If ΔMDL is negative, feature vector $\mathbf{x}$ is regarded as an outlier since encoding it as like a prototype results in a smaller code length than encoding it as an inlier. In this way, outliers can be detected by Eq. (5.15) at the completion of each growing stage. It is worth noting that during the training process, when the number of prototypes grows to the actual number of underlying clusters in

the given dataset, if the positions of these prototypes are attracted far away from the actual cluster centers by the outliers, the resulting MDL value can be very large. Accordingly, the minimum value of MDL may not be obtained at such number of prototypes even though it actually equals the number of underlying clusters. This explains why the GNG-M algorithm may not successfully find the optimal number of clusters and their centers, while our proposed RGNG algorithm can find them.

5.3.4 Implementation of the RGNG algorithm

The RGNG algorithm starts with a few prototypes, usually two, and then a certain number of training epochs are spent to train the current number of prototypes by the updating rule in Eq. (5.10). Due to the incorporation of the outlier resistant strategy, the adaptive learning rates and the cluster repulsion scheme, the RGNG algorithm is robust to the existence of outliers. At the completion of training the current prototypes, a new prototype is inserted around the prototype with the largest local accumulated error. At the present of outliers, it may not be reasonable to simply define this accumulated error by the sum of residual errors since the outliers may also undesirably contribute to the calculation of the local accumulated error. In our proposed GNG algorithm, the local accumulated error with respect to prototype $\mathbf{w}_i$ is defined by:

$$\sum_{\mathbf{x} \in S_i} \exp\left(-\frac{\|\mathbf{x} - \mathbf{w}_i\|}{\text{harmdist}^i}\right) \cdot \|\mathbf{x} - \mathbf{w}_i\| \quad \text{with} \quad \text{harmdist}^i = |S_i| \left(\sum_{\mathbf{x} \in S_i} (\|\mathbf{x} - \mathbf{w}_i\|)^{-1} \right)^{-1} \quad (5.16)$$

where harmdist^i denotes the harmonic average of distances from prototype $\mathbf{w}_i$ to all the feature vectors in its receptive field S_i . Therefore, excessively large accumulated error caused by the outliers can be suppressed. As training proceeds, the MDL value is calculated by Eq. (5.12) at the completion of each growing stage until the pre-specified maximum number of prototypes is reached. Finally, the number of prototypes and their positions corresponding to the smallest MDL value over the whole growing process will be detected as the optimal number of clusters and their centers.

The pseudo-code of the RGNG algorithm is given in Table 5.2. Note that we add a small positive number $\varepsilon = 1e-8$ to any calculated distance so as to avoid trivial distance values.

Table 5.2: Pseudo-code of the Robust Growing Neural Gas (RGNG) algorithm

Initialize a set of prototype vectors (usually 2) $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2\}$ randomly, the learning rates $\varepsilon_{bi}, \varepsilon_{bf}, \varepsilon_{ni}, \varepsilon_{nf}$, the ranking counter pre_node^l of each current prototype to 0, a connection set $\mathbf{C}$, $\mathbf{C} \subset \mathbf{W} \times \mathbf{W}$ to an empty set: $\mathbf{C} = \emptyset$ and κ, η used to calculate the MDL value. Set the maximum number of prototypes to grow to as $pre_numnode$ and the maximum training epoch Max_iter at each growing stage. Set the initial training epoch number: $m = 0$ and the iteration step in the training epoch m : $t=0$. Accordingly, the overall iteration step $iter$ at one growing stage is: $iter = m \cdot N + t$. The set of available feature vectors is denoted by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

(a) While the current number of prototypes is smaller than or equal to $pre_numnode$ and some predefined performance requirement, if exists, is not met

(b) For $m = 0$ to $Max_iter - 1$

- Calculate the harmonic average distance $d_m^i(0)$ for each prototype $\mathbf{w}_i, \forall i \in \mathbf{W}$, with respect to its current position by Eq. (5.8)
- Calculate the learning rates for the current prototype l , when it serves as a winner or its topological neighbors, respectively, by Eq. (5.9). The learning rates remain same in the following training for the prototype l , before a new prototype is inserted
- Set $t=0$, and accordingly $iter = m \cdot N + t$.
- Set $trainingset = \mathbf{X}$, i.e. include all the available feature vectors into $trainingset$

(c) If $trainingset$ is not empty

- Draw randomly a feature vector $\mathbf{x}_t^m$, at the iteration step t in the training epoch m , from the $trainingset$

- Determine the nearest s_1 and the second nearest prototype s_2 ($s_1, s_2 \in \mathbf{W}$) by $s_1 = \arg \min_{i \in \mathbf{W}} \|\mathbf{x}_t^m - \mathbf{w}_i^{iter}\|$, $s_2 = \arg \min_{i \in \mathbf{W} \setminus \{s_1\}} \|\mathbf{x}_t^m - \mathbf{w}_i^{iter}\|$ with $iter = m \cdot N + t$
- If a connection between s_1 and s_2 does not exist already, create it: $\mathbf{C} = \mathbf{C} \cup \{(s_1, s_2)\}$. Set the *age* of the connection between s_1 and s_2 to 0, i.e. $age_{(s_1, s_2)} = 0$
- Update the winning prototype and its direct topological neighbors according to Eqs. (5.6)-(5.10)
- Increment the *age* of all edges emanating from s_1 by 1: $age_{(s_1, i)} = age_{(s_1, i)} + 1$, $\forall i \in N_{s_1}$
- Remove edges with *age* values larger than $\alpha_{\max}$. If this results in some prototypes with no emanating edges, remove those prototypes as well
- Increment the iteration step t by 1 and remove the used feature vectors $\mathbf{x}_t^m$ from the set *trainingset*

(c) End If

(b) End For

- Calculate the MDL value with respect to the current prototype by Eqs. (5.12)-(5.15)
- Save the current prototypes, if the calculated MDL value is smaller than the previous value
- Calculate the local accumulated error for each current prototype by Eq. (5.16)
- Determine the prototype q with the maximum accumulated error:

$$q = \arg \max_{i \in \mathbf{W}} \mathbf{E}_i.$$

- Determine, among the neighbors of q , the prototype f with the maximum accumulated error: $f = \arg \max_{i \in N_q} \mathbf{E}_i$.
- Add a new prototype r to the network and interpolate its position from q and f :

$$\mathbf{W} = \mathbf{W} \cup \{\mathbf{w}_r\}, \quad \mathbf{w}_r = (2 \cdot \mathbf{w}_q + \mathbf{w}_f) / 3$$

- Insert edges connecting the new prototype r with prototypes q and f , and remove the original edge between q and f .

$$\mathbf{C} = \mathbf{C} \cup \{(r, q), (r, f)\}, \quad \mathbf{C} = \mathbf{C} \setminus \{(q, f)\}$$

- Set the ranking counter pre_node^r of the newly inserted prototype r to 0 and for other existing prototypes, increment their ranking counters by 1, i.e.

$$pre_node^r=0, \quad pre_node^l=pre_node^l+1, \quad \forall l \in \mathbf{W} \setminus \{r\}.$$

(a) End While

Output the number of prototypes and their final positions corresponding to the minimum MDL value over all growing stages as the optimal number of clusters and their centers

5.4 Experimental results

In this section, we evaluate the performance of our proposed RGNG algorithm on 5 artificial and UCI benchmark datasets. Compared with the GNG-M algorithm, which incorporates the MDL measure into the conventional GNG algorithm in the same way as the RGNG does, the RGNG algorithm shows superior performance.

5.4.1 Experimental setup

In general, a robust clustering algorithm should be less sensitive to different parameter configurations and therefore we use the same parameter settings for different experimental datasets. In our following experiments, the parameters of the GNG-M algorithm are set as their typical values suggested by Fritzke [274], i.e. $\varepsilon_b = 0.05$, $\varepsilon_n = 0.006$, $\alpha_{max} = 100$, $\alpha = 0.5$ and $\beta = 0.001$. The parameter set of our RGNG algorithm is set as: $\varepsilon_{bi} = 0.1$, $\varepsilon_{bf} = 0.01$, $\varepsilon_{ni} = 0.005$, $\varepsilon_{nf} = 0.0005$, $\alpha_{max} = 100$, $\eta = 1e-6$. The maximum number of prototypes $pre_numnode$ is specified according to the scale of the given dataset, and kept same for both GNG-M and RGNG algorithms.

We independently run each of these two clustering algorithms 10 times. The initial states of the two starting prototypes are differently configured for each of the 10 runs.

The detected optimal number of clusters and some other performance measures are used to evaluate the cluster quality. Here, we apply the two-sample t -test [81] to measure the statistical difference between the clustering results obtained by the RGNG and GNG-M algorithms, which is reported in Tables 5.3-5.5.

5.4.2 Experiments on artificial data

To examine the clustering performance of the RGNG and GNG-M algorithms when dealing with the datasets containing many outliers, we designed two 2D datasets and applied the two clustering algorithms on them.

The first dataset D_1 is shown in Figure 5.1a. Five clusters S_1 , S_2 , S_3 , S_4 and S_5 are generated according to different Gaussian distributions with their mean vectors located at $[0,1]$, $[0,2]$, $[-1,0]$, $[1,0]$ and $[0,-1]$, respectively. Among them, S_1 , S_2 and S_3 have spherical shapes; S_4 and S_5 have elliptical shapes. The number of feature vectors in these five clusters is 150, 150, 200, 100 and 100, respectively. Figure 5.1b shows the contaminated version of dataset D_1 , in which 84 outliers are added. Figure 5.2a shows the second dataset D_2 , which consists of $k=25$ clusters arranged in a 5×5 2D grid. Each cluster includes 50 feature vectors sampled from the corresponding Gaussian distribution with a total number of $N=1,250$ feature vectors. We add 104 outliers to contaminate this dataset as shown in Figure 5.2b. In Figures 5.1a, 5.1b, 5.2a and 5.2b, the solid black crosses represent the actual cluster centers. Note that, in the contaminated versions of datasets D_1 and D_2 , four outliers with their coordinates at $(10, 0)$, $(10, 0)$, $(20, 0)$ and $(20, 0)$ are not displayed in Figures 5.1b and 5.2b.

The following measures are used to evaluate the performance of a clustering algorithm. Since the underlying cluster structures are completely known a priori, we use the average optimal number of clusters detected by the clustering algorithm over 10 runs, denoted by $OptN$, as the first performance measure. The second performance measure we employed is a class of external cluster validity indices [285], which evaluate the distinction between the obtained clustering results and the actual cluster structures known a priori. Specifically, we employ three famous external cluster

validity indices, namely Rand Statistic (RS), Jaccard Coefficient (JC), Folkes and Mallows index (FM). Assume that a , b , c and d denote the number of feature vectors that belongs to the same underlying cluster and the same detected cluster, belongs to the same underlying cluster and the different detected cluster, belongs to the different underlying cluster and the same detected cluster, belongs to the different underlying cluster and the different detected cluster, respectively. Obviously, the maximum number of pairs of feature vectors in the dataset $M = N(N-1)/2$ equals $a+b+c+d$. The three external validity indices are defined by: $RS = (a+d)/(a+b+c+d)$, $JC = a/(a+b+c)$, $FM = \sqrt{a^2/((a+b)(a+c))}$, respectively. It has been proven that large values of these indices indicate high similarity between the clustering results obtained by applying the clustering algorithms and the already known cluster structures. To demonstrate the robustness of our proposed RGNG algorithm, two indices are used to evaluate the intermediate clustering results when the number of prototypes grows to the actual number of underlying clusters: Founded Clusters ($FouC$) denotes the number of underlying clusters which is detected by at least one prototype. Mean Square Error (MSE) is the average distance between the actual cluster centers and the positions of the prototypes. These two measures are averaged over 10 runs to evaluate the performance of the clustering algorithm when applied on the datasets containing many outliers with the number of prototypes growing to the actual number of underlying clusters. The calculation of these performance measures on the contaminated datasets will only rely on the inliers regardless of the added outliers to highlight the performance of the algorithm in finding the actual clusters.

Figures 5.1c-5.1f plot the MDL values averaged over 10 runs, calculated at the completion of each growing stage by using the GNG-M and RGNG algorithms on the clean and contaminated dataset D_1 , respectively. The maximum number of prototypes to grow is specified as 20. It is obvious that our proposed RGNG algorithm can successfully find the actual number of underlying clusters and is less sensitive to initializations, input sequence orders and the presence of outliers, while the GNG-M algorithm ends with the wrong cluster numbers on both clean and contaminated datasets. Other performance measures used to evaluate these two clustering algorithms on dataset D_1 are reported in Table 5.3. We can observe that the RGNG algorithm

yields higher RS , JC and FM values than the GNG-M algorithm in both clean and contaminated datasets, which indicate the better partitioning quality of the RGNG. Meanwhile, at the growing stage with the number of prototypes equal to the actual number of underlying clusters, the RGNG algorithm can successfully detect all the underlying clusters in both clean and contaminated datasets over each of the 10 runs and achieve much lower MSE values. It is worth noting that the calculation of the MDL value at the completion of each growing stage heavily depends on the current positions of the prototypes, since the code lengths of the quantization errors and the detected outliers will differ significantly with respect to the prototype positions. For the GNG-M algorithm, we can observe from Table 5.3 that when the number of prototypes grows to “5”, it cannot detect all the actual clusters and thus yield a higher MSE value on the clean dataset. On the contaminated dataset, the finally obtained prototype positions will deviate significantly from the actual cluster centers when the number of prototype equals the actual number of underlying clusters, which can be observed from the last two measures in Table 5.3 for the GNG-M. However, after some prototypes have been attracted near the outliers, the newly inserted prototypes may find the actual clusters. These facts indicate that the GNG-M algorithm is highly sensitive to the presence of outliers. This explains why the GNG-M algorithm detects the larger number of clusters.

Figure 5.1a

Figure 5.1b

Figure 5.1c

Figure 5.1d

Figure 5.1e

Figure 5.1f

Figure 5.1: (a) Plot of the clean dataset D_1 ; (b) Plot of the contaminated dataset D_1 ; (c) and (d) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the clean D_1 ; (e) and (f) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the contaminated D_1 .

Table 5.3: Clustering results of GNG-M and RGNG algorithms on datasets D_1 over 10 runs (standard deviation in bracket).

Datasets	Methods	$OptN$	$FouC$	MSE	RS	JC	FM
Clean D_1	GNG-M	15.2 (4.4)	3.6 (0.7)	0.1712 (0.0263)	0.8940 (0.0576)	0.5052 (0.2696)	0.6888 (0.1781)
	RGNG	5.0 (0.0)	5.0 (0.0)	0.0156 (0.0072)	0.9937 (0.0020)	0.9711 (0.0090)	0.9853 (0.0047)
	p -value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
Contaminated D_1	GNG-M	9.7 (1.8)	1.7 (0.8)	1.6072 (0.9778)	0.8877 (0.0066)	0.5676 (0.0277)	0.7250 (0.0219)

Chapter 5: Robust Growing Neural Gas Algorithm with Application in Cluster Analysis

	RGNG	5.0 (0.0)	5.0 (0.0)	0.0197 (0.0084)	0.8943 (0.0017)	0.6029 (0.0050)	0.7522 (0.0039)
	<i>p</i> -value	0.0000	0.0000	0.0001	0.0066	0.0009	0.0011

Table 5.4: Clustering results of GNG-M and RGNG algorithms on multimodal datasets D_2 over 10 runs (standard deviation in bracket).

Datasets	Methods	<i>OptN</i>	<i>FouC</i>	<i>MSE</i>	<i>RS</i>	<i>JC</i>	<i>FM</i>
Clean D_2	GNG-M	26.7 (1.3)	21.8 (1.1)	0.1483 (0.2047)	0.9990 (0.0005)	0.9751 (0.0125)	0.9874 (0.0063)
	RGNG	25.0 (0.0)	25.0 (0.0)	0.0041 (0.0007)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	<i>p</i> -value	0.0004	0.0000	0.0390	0.0000	0.0000	0.0000
Contaminated D_2	GNG-M	30.3 (5.2)	19.9 (1.3)	2.6348 (1.0885)	0.9888 (0.0060)	0.7668 (0.0838)	0.8673 (0.0544)
	RGNG	25.0 (0.0)	25.0 (0.0)	0.0051 (0.0015)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	<i>p</i> -value	0.0045	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 5.2a

Figure 5.2b

Figure 5.2c

Figure 5.2d

Figure 5.2e

Figure 5.2f

Figure 5.2: (a) Plot of the clean multimodal data set D_2 ; (b) Plot of the contaminated multimodal data set D_2 ; (c) and (d) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the clean D_2 ; (e) and (f) Plot of the MDL value versus the number of prototypes during the growing process of GNG-M and RGNG algorithm on the contaminated D_2 .

Figures 5.2c-5.2f plot the MDL values averaged over 10 runs, calculated at the completion of each growing stage by applying the GNG-M and RGNG algorithms on the clean and contaminated dataset D_2 , respectively. The maximum number of prototypes to grow is specified as 50. Some performance indices of these two algorithms are reported in Table 5.4. Compared with our proposed RGNG algorithm, the GNG-M algorithm shows inferior performance as it fails to find the actual cluster number 25 on both clean and contaminated dataset D_2 and accordingly yields much lower RS , JC and FM values than the RGNG algorithm does. Moreover, when the number of prototypes grows to equal the actual number of underlying clusters, the average number of clusters $OptN$ detected by the GNG-M over 10 runs significantly deviates from the actual value 25 and consequently the corresponding MSE value is much higher.

In comparison with the GNG-M algorithm, our proposed RGNG algorithm demonstrates superior performance in both clean and contaminated datasets, as evidenced by the small p -values shown in Tables 5.3 and 5.4.

5.4.3 Experiments on UCI data

To evaluate the clustering performance of the RGNG and GNG-M algorithms in real-world applications, we choose one benchmark dataset “Wine” from the UCI machine learning repository [78], and compare the clustering performance of the two algorithms. In real-world datasets, the number of underlying clusters may not equal the number of classes since some classes may contain several sub-clusters. In such case, we may have little prior knowledge about the actual number of underlying clusters. However, the explicit underlying cluster structures in “Wine” dataset can be verified by the particularly high classification accuracy obtained from applying the QDC on the whole dataset. In order to demonstrate the robust performance of our proposed RGNG algorithm, we also apply the two clustering algorithms on a contaminated version of “Wine” dataset generated by adding two outliers faraway from the existing feature vectors in the original dataset. To perform cluster analysis, we first remove the class labels in the original dataset and then apply the RGNG and

the GNG-M algorithms on the unlabeled clean and contaminated “Wine” datasets, which is normalized to $[-1, 1]$. Here, we employ the RS , JC and FM measures to evaluate the clustering results. By comparing these values obtained by executing the GNG-M and RGNG algorithms, the RGNG algorithm shows better partitioning quality in both noisy and contaminated datasets. The performance measures reported in Table 5.5 show that the two clustering algorithms actually detect different number of clusters. With their respective numbers of clusters, the RGNG algorithm achieves higher RS , JC and FM values than those achieved by the GNG-M algorithm, which means that the RGNG algorithm can yield better clustering results.

Table 5.5: Clustering results of GNG-M and RGNG algorithms on “Wine” datasets over 10 runs (standard deviation in bracket).

Datasets		Methods	$OptN$	RS	JC	FM
WINE	Clean	GNG-M	2.5 (0.5)	0.7906 (0.1191)	0.5972 (0.1603)	0.7426 (0.1198)
		RGNG	3.0 (0.0)	0.9202 (0.0146)	0.7875 (0.0343)	0.8808 (0.0217)
		p -value	0.0077	0.0031	0.0017	0.0021
	Contaminated	GNG-M	2.6 (1.0)	0.7979 (0.0963)	0.6252 (0.1411)	0.7743 (0.0917)
		RGNG	3.0 (0.0)	0.9317 (0.0054)	0.8165 (0.0131)	0.8989 (0.0079)
		p -value	0.2069	0.0004	0.0005	0.0005

5.5 Conclusions

In this chapter, many robust variants of the conventional prototype-based clustering algorithms and several prototype-based clustering algorithms that can automatically determine the optimal number of clusters were reviewed. We developed a RGNG algorithm for offline cluster analysis, which is less sensitive to initializations, input sequence orders and the presence of outliers due to the incorporation of the outlier resistant scheme, the adaptive modulation of learning rates and cluster repulsion method into the conventional GNG framework. By seeking the extreme value of the Minimum Description Length (MDL) measure calculated at the completion of each

growing stage, the optimal number of clusters can be automatically determined. Furthermore, topology relationships among prototypes can be established. Experiments over several artificial and UCI datasets demonstrated the superior performance of the proposed RGNG algorithm over the GNG-M.

Chapter 6: Conclusions and Future Work

6.1 Conclusions

This thesis investigated respectively dimensionality and prototype reduction techniques for pattern analysis and specifically focused on feature extraction techniques for supervised classification and clustering techniques for unsupervised classification. The major research work covered in the thesis can be summarized as follows.

- **Dimensionality reduction for supervised classification**

We conducted a thorough literature survey on two categories of DLDR techniques: (1) the BODLDR techniques, in which the criterion functions are related to the Bayes error formulated parametrically or non-parametrically under certain conditions and the suboptimal solutions can be numerically derived at the expense of high computational costs; (2) The SMDLDR techniques, in which the criterion functions are directly formulated based on the geometric class structures of the training samples and the globally optimal solutions in the closed form can be analytically derived at the expense of losing the link with the Bayes error. The FLDA technique as well as its heteroscedastic extension, the CLDA technique proposed recently, was revisited in-depth. The Bayes optimality condition for the FLDA solutions was recapitulated. The potential deficiencies inherent in the conventional Fisher and Chernoff criteria, which destroy their link with the Bayes error in general sense, were unveiled and discussed.

Two categories of enhanced parametric BODLDR techniques were developed under the Gaussian homoscedastic and heteroscedastic assumptions, respectively. In the first category, a direct BODLDR technique and a class of 8 CWFLDA techniques were developed under the Gaussian homoscedastic assumption. The former one directly used the union of the pairwise Bayes errors in the transformed subspace as the

criterion function, which was analytically formulated under the Gaussian homoscedastic assumption. A suboptimal orthonormal transformation matrix can be obtained by numerically optimizing the criterion function via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme. The later class of techniques employed the newly devised class-wise weighted Fisher criteria in quotient and subtraction forms, respectively, with discriminant vectors subject to different conjugated orthogonality constraints being extracted in the batch, sequential forward or sequential backward manner. In the second category, a Gaussian heteroscedastic direct BODLDR technique and a class of 8 CWCLDA techniques motivated by the Bayes optimality under the Gaussian heteroscedastic assumption were developed. The former one directly used the union of the pairwise Chernoff upper bounds of the Bayes error in the transformed subspace as the criterion function, which was analytically formulated under the Gaussian heteroscedastic assumption. A suboptimal orthonormal transformation matrix can be obtained by numerically optimizing the criterion function via the gradient descent coupled with the Gram-Schmidt orthogonalization scheme. The later class of techniques employed the newly devised class-wise weighted Chernoff criteria in quotient and subtraction forms, respectively, with discriminant vectors subject to different conjugated orthogonality constraints being extracted in the batch, sequential forward or sequential backward manner. The class-wise weighted Chernoff criteria can be regarded as the heteroscedastic extension of the class-wise weighted Fisher criteria so as to capture the discriminatory information present in the difference between both class means and class covariance matrices. The effectiveness of the between-class and within-class weighting terms involved in the class-wise weighted Fisher and Chernoff criteria was validated via three artificial datasets subject to the Gaussian homoscedastic or heteroscedastic assumption. The performance of the proposed techniques to address real-world problems was demonstrated on several UCI multi-class datasets and compared with their archetypes in terms of both optimal and average dimensionality reduction performance with regard to three classifiers, i.e. LDC, QDC, k NN. We proposed three relative performance indices, namely RBPI, RAPI and ROPI, to comprehensively evaluate each experimental technique. A general observation was that many proposed

techniques could demonstrate better performance than their existing archetypes in terms of the three relative performance indices.

We conducted an extensive overview of the existing discriminative dimensionality reduction techniques capable of handling high-dimensional and undersampled problems along two main avenues: discriminative subspace analysis and discriminative manifold learning. We developed two classes of subspace based CWFLDA techniques, namely 2 CS-CWFLDA techniques and 4 RS-CWFLDA techniques, which can take advantage of the discriminatory information present in different subspaces while retaining the promising performance in the well-sampled cases. Specifically, the discriminant vectors subject to different conjugated orthogonality constraints were either extracted in the sequential forward manner from the two complementary subspaces, i.e. the null and column spaces of the average within-class scatter matrix, or extracted in the batch or sequential backward fashion from a single subspace, i.e. the column space of the regularized average within-class scatter matrix, based on the quotient class-wise weighted Fisher criterion. The performance of the proposed techniques was extensively evaluated on various types of high-dimensional data sources including handwritten digit images, facial images, text documents and microarray gene expression profiles, and compared with that of several state-of-the-art FLDA variants. Two of the proposed techniques demonstrated consistently better performance than the other experimental techniques in terms of the three relative performance indices.

- **Prototype reduction for unsupervised classification**

Robust variants of the conventional prototype-based clustering algorithms and several prototype-based clustering algorithms that can automatically determine the optimal number of clusters were reviewed. We developed a RGNG algorithm for cluster analysis, which is less sensitive to initializations, input sequence orders and the presence of outliers due to the incorporation of the outlier resistant scheme, the adaptive modulation of learning rates and cluster repulsion method into the conventional GNG framework. Moreover, by seeking the extreme value of the

Minimum Description Length (MDL) measure calculated at the completion of each growing stage, the optimal number of clusters can be automatically determined. Furthermore, topology relationships among prototypes can also be established. Experiments demonstrated the superior performance of the proposed RGNG algorithm over the GNG-M.

6.2 Recommendations for future work

Based on the research work conducted in this thesis, some research directions deserving future investigations are recommended as follows.

- In this thesis, we have demonstrated the promising performance of the proposed enhanced parametric BODDLDR techniques in comparison with their existing archetypes. However, we did not provide a detailed analysis on the performance distinctions among different types of DLDR techniques, e.g. the DLDR techniques subject to different conjugated orthogonality constraints, the DLDR techniques based on quotient and subtraction criteria, the DLDR techniques with the discriminant vectors extracted in different manners, etc. As one future work, we attempt to investigate different scenarios which may specifically suit certain types of the DLDR techniques. Moreover, we attempt to further analyze and improve the subtraction class-wise weighted Fisher and Chernoff criteria instead of the merely direct extensions from the quotient criteria as we did in the current thesis.
- In this thesis, we have only investigated the offline DLDR techniques. However, the core idea behind our proposed class-wise weighted criteria can also be incorporated into some existing online DLDR techniques to enhance their performance. For example, in the existing online FLDA technique [36], [37], the class means can be updated in the sequential or chunk manner when new training samples are presented, which enable the online updating of the pairwise between-class scatter matrix. The average within-class scatter matrix and the class prior probabilities can also be updated in the online manner. Consequently, we can

construct the online class-wise weighted Fisher criterion so as to establish the online CWFLDA techniques. It is worth noting that when the sample size is particularly small at the initial training stage, the online CWFLDA techniques will encounter the small sample size problem, which may be handled by the generalized CWFLDA techniques described in Chapter 4. The online Chernoff criteria based techniques also deserve further investigations.

- Theoretically speaking, the proposed enhanced parametric BODLDR techniques are particularly suitable for the well-sampled problems since the accurate pairwise Bayes error estimates require sufficient training samples being available. Although several CWFLDA techniques have been generalized to handle the high-dimensional and undersampled problems in Chapter 4, they actually comes down to the weighted FLDA with weighting terms mainly determined by the reciprocal of the squared Mahalanobis distances between class pairs. This is because that the pairwise two-class Bayes accuracies of most class pairs may achieve the common maximum values in the high-dimensional and undersampled scenario and consequently no longer influence the between-class and within-class weights. However, such degenerated weighting terms can still take effect to suppress the undesirable influences from those classes that are far away from the remaining classes. In the high-dimensional and undersampled scenario, almost all class covariance matrices are singular. Consequently, some regularization schemes must be used to enable the feasibility of the Chernoff criteria based DLDR techniques. Different types of regularized schemes specifically suitable for the Chernoff criteria based DLDR techniques deserve future investigations. However, the Chernoff criteria based techniques may still become invalid since the discriminatory information present in the difference between the regularized class covariance matrices may possibly make little sense in terms of the generalization power. That's why we did not generalize the CWCLDA techniques in this thesis. However, in real-world applications, some challenging tasks may involve a considerable number of training samples with high dimensionality, e.g. video-based face recognition, image retrieval, text categorization, protein fold classification, etc. Such kind of problems will not suffer from the small sample size problem due to

the relatively high ratio between the training sample size and the feature dimensionality. The proposed enhanced parametric BODLDR techniques may possibly find promising applications to handle such kind of problems, which deserves extensive investigations in future.

- We can observe that some FLDA variants extract the discriminatory information from distinct subspaces and thus demonstrate different discrimination power, while others extract the discriminatory information from the same subspace and thus are essentially equivalent to each other. Therefore, it is desirable to conduct a unified analysis of several state-of-the-arts FLDA variants theoretically and empirically so as to unveil the intrinsic equivalence and difference between different variants. Such analysis may also suggest us assembling the discriminative features extracted from many partially or fully disjoint subspaces so as to boost the performance.
- In this thesis, we did not address the problem of automatic determination of the optimal dimensionality to reduce to by applying the DLDR techniques. Although the determination of the optimal dimensionality remains an open problem, we attempt to try some typical solutions in future. For example, we attempt to estimate the leave-one-out error on training samples at each reduced dimension to choose the optimal dimensionality. Moreover, we can employ feature selection techniques after the DLDR techniques so as to further condense the already extracted features.
- Due to use of reference vectors and the Euclidean distance metric to represent prototypes, the proposed RGNG algorithm are mainly suitable for datasets having hyper-spherical clusters. Extensions to other distance metrics such as the Mahalanobis distance involving the covariance matrix as the parameter, and more complex prototypes like shell prototypes [286] in which the similarity is measured with respect to specific shapes instead of reference vectors deserve future study. Moreover, the proposed RGNG algorithm cannot scale well with the feature space dimensionality. As one future work, we attempt to apply the RGNG in a lower-dimensional space obtained by applying the manifold learning techniques to reduce the dimensionality of the original feature space while maximally retaining the

useful data structure information. Consequently, the RGNG algorithm can be applied to handle high-dimensional datasets. It is worth noting that spectral clustering intrinsically belongs to such type of methods. In fact, multiple lower-dimensional spaces with different dimensions or attained by different manifold learning techniques allow further applying the cluster ensemble [287] techniques, by which a more coherent clustering result can be eventually obtained.

- In this thesis, we have employed the MDL measure to determine the optimal number of clusters, which was specifically formulated to consider the existence of outliers. As one future work, we attempt to extend other popular cluster validity measures to explicitly take into account the possibly existing outliers, incorporate them into the RGNG algorithm, and compare the performance of the resulting algorithms with that of the currently proposed RGNG algorithm.
- In this thesis, our research work is only focused on the offline clustering techniques. As one future work, we attempt to devise an online RGNG algorithm. Specifically, when a new input sample is available, we can determine whether it is an outlier with regard to the current positions of prototypes via the MDL principle. Different from the offline RGNG algorithm, the detected outliers will be treated as new cluster centers since they cannot be reasonably assigned to any existing clusters. After every certain number of training iterations, all existing clusters may undergo the merging and splitting process based on some already presented representative training samples stored in the memory. The resulting prototypes will be used in the subsequent clustering procedure. If all the input samples are available, the offline RGNG can be applied to refine the clustering results obtained by the online algorithm with the prototypes obtained by the online algorithm as the starting point. It is worth noting that in the online RGNG, the initial restricting distance associated with each prototype will be pre-specified by users instead of being calculated based on all the available samples as done in the offline RGNG.
- The integration of dimensionality reduction and prototype reduction techniques for solving offline or online supervised or unsupervised learning problems may

Chapter 6: Conclusions and Future Work

constitute promising research directions deserving future investigations. Some existing works [14], [15], [16], [17] have verified its feasibility.

Author's Publication List

Journal Publications

1. A. K. Qin and P. N. Suganthan, "Robust growing neural gas algorithm with application in cluster analysis," *Neural Networks special issue on Recent Developments in Self-Organizing Systems*, 17 (8-9): 1135-1148, 2004.
2. E. K. Tang, P. N. Suganthan, X. Yao and A. K. Qin, "Linear dimensionality reduction using relevance weighted LDA," 38(4): 485-493, *Pattern Recognition*, 2005.
3. A. K. Qin, P. N. Suganthan and M. Loog, "Uncorrelated heteroscedastic LDA algorithm based on the weighted pairwise Chernoff criterion," 38(4): 613-616, *Pattern Recognition*, 2005.
4. A. K. Qin and P. N. Suganthan, "Initialization Insensitive LVQ algorithm based on dynamic cost function adaptation," 38(4): 613-616, *Pattern Recognition*, 2005.
5. A. K. Qin and P. N. Suganthan, "Enhanced neural gas network for prototype based clustering," 38(8): 1275-1288, *Pattern Recognition*, 2005.
6. Qin-Yu Zhu, A. K. Qin, P. N. Suganthan, Guang-Bin Huang, "Evolutionary extreme learning machine," 38(10): 1759-1763, *Pattern Recognition*, 2005.
7. J. J. Liang, A. K. Qin, P. N. Suganthan and S. Baskar, "Comprehensive learning particle swarm optimizer for global optimization of multimodal functions," *IEEE Transactions on Evolutionary Computation*, 10(3): 281-295, 2006.
8. J. J. Liang, S. Baskar, P. N. Suganthan and A. K. Qin, "Performance evaluation of multiagent genetic algorithm," *Natural Computing*, 5(9): 83-96, 2006.
9. A. K. Qin and P. N. Suganthan and M. Loog, "Generalized null space uncorrelated fisher discriminant analysis for linear dimensionality reduction," *Pattern Recognition*, 39(9): 1805-1808, 2006.
10. A. K. Qin, V. L. Huang and P. N. Suganthan, "Differential evolution with strategy and parameters self-adaptation," submitted to *IEEE Transactions on Evolutionary Computation*, 2007.

Conference Publications

1. A. K. Qin, P. N. Suganthan, C. H. Tay and H. S. Pa, "Personal identification system based on multiple palmprint features," *Proc. of the 9th International Conference on Control, Automation, Robotics and Vision (ICARCV'06)*, Singapore, December, 2006.
2. A. K. Qin and P. N. Suganthan, "Efficient Feature Extraction Based on Regularized Uncorrelated Chernoff Discriminant Analysis," *Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06)*, Hong Kong, August 2006.
3. V. L. Huang, A. K. Qin and P. N. Suganthan, "Self-adaptive Differential Evolution Algorithm for Constrained Real-Parameter Optimization," *Proc. of the 2006 IEEE Congress on Evolutionary Computation (CEC'06)*, Vancouver, BC, Canada, July 2006.
4. A. K. Qin and P. N. Suganthan, "Self-adaptive Differential Evolution Algorithm for Numerical Optimization," *Proc. of IEEE Congress on Evolutionary Computation (CEC05)*, Edinburgh, UK, Sept. 2005.
5. A. K. Qin, S. Y. M. Shi, P. N. Suganthan and M. Loog, "Enhanced Direct Linear Discriminant Analysis for Feature Extraction on High Dimensional Data," *Proc. of the 20th National Conference on Artificial Intelligence (AAAI-05)*, Pittsburgh, Pennsylvania, USA, July 2005.
6. J. J. Liang, A. K. Qin, P. N. Suganthan and S. Baskar, "Evaluation of comprehensive learning particle swarm optimizer," *Springer's Lecture Notes in Computer Science, ICONIP'04*, vol. 3316, pp. 230-235, 2004.
7. J. J. Liang, A. K. Qin, P. N. Suganthan and S. Baskar, "Particle swarm optimization algorithms with novel learning strategies," *Proc. of IEEE International Conference on systems, man and cybernetics 2004 (SMC 2004)*, Hague, Netherlands, Oct. 2004.
8. A. K. Qin, P. N. Suganthan, and J. J. Liang, "A new generalized LVQ algorithm via harmonic to minimum distance measure transition," *Proc. of IEEE Int. Conf. on systems, man and cybernetics (SMC 2004)*, Hague, Netherlands, Oct. 2004.
9. A. K. Qin and P. N. Suganthan, "A novel kernel prototype-based learning algorithm," *Proc. of the 17th Int. Conf. on Pattern Recognition (ICPR'04)*, Cambridge, UK, Aug. 2004.

AUTHORS' PUBLICATION LIST

10. A. K. Qin and P. N. Suganthan, "Kernel neural gas algorithms with application to cluster analysis," *Proc. of 17th Int. Conf. on Pattern Recognition (ICPR'04)*, Cambridge, UK, Aug. 2004.
11. A. K. Qin and P. N. Suganthan, "Growing generalized learning vector quantization with local neighborhood adaptation rule," *Proc. of IEEE Int. Conf. on Intelligent Systems*, Sofia, Bulgaria, June 2004.
12. A. K. Qin and P. N. Suganthan, "A robust neural gas algorithm for clustering analysis," *Proc. of International Conference on Intelligent Sensing and Information Processing (ICISIP2004)*, Chennai, India, Jan. 2004.

Bibliography

- [1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd edition, New York: Wiley Interscience, 2000.
- [2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, Academic Press, 1990.
- [3] H. Bunke, A. Sanfeliu, *Syntactic and Structural Pattern Recognition - Theory and Applications*, Singapore: World Scientific series in Computer Science, 1990.
- [4] 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.), 2: 35-855. Amsterdam: North-Holland, 1982.
- [5] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(3): 43-252, 1980.
- [6] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Norwell, MA, USA: Kluwer Academic Publishers, 1998.
- [7] M. A. Carreira-Perpinan, "A review of dimension reduction techniques," *Technical Report CS-96-09*, Department of Computer Science, University of Sheffield, January 1997.
- [8] I. K. Fodor, "A survey of dimension reduction techniques," *LLNL technical report*, UCRL-ID-148494, 2002.
- [9] I. T. Jolliffe, *Principal component analysis*, 2nd edition, New York: Springer, 2002.
- [10] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, 9(7): 1493-1516, 1997.
- [11] M. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society Series B*, 61(3): 611-622, 1999.
- [12] N. Cristianini and J. S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge, UK: Cambridge University Press, 2000.

BIBLIOGRAPHY

- [13] T. Hastie and W. Stuetzle, "Principle curves," *Journal of the American Statistical Association*, 84(406): 502-516, 1989.
- [14] S. Ozawa, S. N. Pang, and N. Kasabov: A modified incremental principal component analysis for on-line learning of feature space and classifier, *Lecture Notes in Artificial Intelligence*, 3157: 231-240, 2004.
- [15] S. Ozawa, S.-L. Toh, S. Abe, S. N. Pan and N. Kasabov, "Incremental learning of feature space and classifier for face recognition," *Neural Networks*, 18(5-6): 575-584, 2005.
- [16] S. Ozawa, S. N. Pang and N. Kasabov, "Incremental learning of feature space and classifier for on-line pattern recognition," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 10(1): 57-65, 2006.
- [17] S. Ozawa, S. N. Pang and N. Kasabov, "On-line feature selection for adaptive evolving connectionist systems," *International Journal of Innovative Computing, Information and Control*, 2(1): 181-192, 2006.
- [18] T. Kohonen, *Self-Organizing Maps*, 3rd edition. Berlin, Germany: Springer-Verlag, 2001.
- [19] C. M. Bishop, M. Svensén, and C. K. I. Williams, "GTM: the generative topographic mapping," *Neural Computation*, 10(1): 215-234, 1998.
- [20] J. H. Friedman, "Exploratory projection pursuit," *Journal of the American Statistics Association*, 82(397): 249-266, 1987.
- [21] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Singapore: John Wiley and Sons, 2001.
- [22] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, California, USA: Sage Publications, Beverly Hills, 1978.
- [23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by local linear embedding," *Science*, 290(5500): 2323-2326, 2000.
- [24] D. L. Donoho, C. Grimes, "Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data," *Proc. of the National Academy of Sciences of the United States of American*, 100(10): 5591-5596, 2003.
- [25] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290(5500): 2319-2323, 2000.

BIBLIOGRAPHY

- [26] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems applied and computational harmonic analysis," (21): 113-127, 2006.
- [27] Z. Y. Zhang and H. Y. Zha, "Principal manifolds and nonlinear dimension reduction via tangent space alignment," *SIAM Journal of Scientific Computing*, 26(1): 313-338, 2004.
- [28] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, 15(6): 1373-1396, 2003.
- [29] P. Mitra, C. A. Murthy and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3): 301-312, 2002.
- [30] M. H. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 26(9): 1154-1166, 2004.
- [31] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, 5: 845-889, 2004.
- [32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, 97(12): 273--324, 1997.
- [33] H. Liu and L. Yu. "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, 17(4): 491-502, 2005.
- [34] L. Yu and H. Liu. "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, 5: 1205-1224, 2004.
- [35] M. Dash and H. Liu, "Feature selection for classification," *International Journal of Intelligent Data Analysis*, 1(3): 131-156, 1997.
- [36] S. N. Pang, S. Ozawa, and N. Kasabov, "Chunk incremental LDA computing on data streams," *Lecture Notes in Computer Science*, 3497: 51-56, 2005.
- [37] S. Pang, S. Ozawa, N. Kasabov, "Incremental linear discriminant analysis for classification of data streams," *IEEE Transactions on Systems, Man and Cybernetics B*, 35(5): 905-914, 2005.
- [38] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," *Proc. of the 7th ACM SIGKDD International*

BIBLIOGRAPHY

- Conference on Knowledge Discovery and Data Mining (KDD'01)*, pp. 245-250, San Francisco, USA, 2004.
- [39] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1): 4-37, 2000.
- [40] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, Singapore: World Scientific, 2002.
- [41] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall, 1998.
- [42] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," *Proc. of the International Joint Conference on Neural Networks (IJCNN'04)*, pp. 985- 990, Budapest, Hungary, 2004.
- [43] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, Belmont, CA: Wadsworth & Brooks Press, 1984.
- [44] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, 31(3):264-323, 1999.
- [45] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: John Wiley and Sons, 1990.
- [46] P. Berkhin, "Survey of clustering data mining techniques," *Accrue Software, Inc.*, http://www accrue.com/products/rp_cluster_review.pdf, 2002.
- [47] C. Borgelt, "Prototype-based classification and clustering," Habilitation Report, <http://fuzzy.cs.uni-magdeburg.de/~borgelt/habil.html>, 2006.
- [48] J. W. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher, 2001.
- [49] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Prof. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, Portland, Oregon, 1996.
- [50] J. G. Augustson and J. Minker, "An analysis of some graph theoretical cluster techniques," *Journal of the ACM*, 17(4): 517-588, 1970.

BIBLIOGRAPHY

- [51] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *Proc. of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 94-105, Seattle, Washington, 1998.
- [52] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: a hierarchical clustering algorithm using dynamic modeling," *IEEE Computer: Special Issue on Data Analysis and Mining*, 32(8):68-75, 1999.
- [53] S. Savaresi, D. L. Boley, S. Bittanti, and G. Gazzaniga, "Cluster selection in divisive clustering algorithms," *Proc. of the 2nd SIAM International Conference on Data Mining*, pp. 299-314, Arlington, VA, USA, 2002.
- [54] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten, "Neural gas network for vector quantization and its application to time-series prediction," *IEEE Transactions on Neural Networks*, 4(4): 558-569, 1993.
- [55] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks: I, II," *IEEE Transaction on Neural Networks*, 13(3): 645-677, 2002.
- [56] A. Ben-Hur, D. Horn, H. T. Siegelmann, V. Vapnik, "Support vector clustering," *Journal on Machine Learning Research (JMLR)*, 2:125-137, 2001.
- [57] M. Girolami, "Mercer kernel based clustering in feature space," *IEEE Transaction on Neural Networks*, 13(3): 780-784, 2002.
- [58] Q. Song, "A robust information clustering algorithm", *Neural Computation*, 17(12): 2672-2698, 2005.
- [59] A. Y. Ng, M.I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm, *Proc. of 16th Annual Conference on Neural Information Processing Systems (NIPS'02)*, pp. 849-856, Vancouver, Canada, 2001.
- [60] A. Strehl, *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, The University of Texas at Austin, 2002.
- [61] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE Transactions on Evolutionary Computation*, 11(1): 56-76, 2007.
- [62] Q. Song and N. Kasabov, "ECM, a novel on-line, evolving clustering method and its applications," *Proc. of the 5th Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES'01)*, pp. 87-92, Dunedin: University of Otago, 2001.

BIBLIOGRAPHY

- [63] N. Kasabov and Q. Song, "DENFIS: Dynamic evolving neural-fuzzy inference system and its application for time-series prediction," *IEEE Transactions on Fuzzy Systems*, 10(2): 144-154, 2002.
- [64] J. Liu, J. P. Y. Lee, L. J. Li, Z.-Q. Luo, and K. M. Wong, "Online clustering algorithms for radar emitter classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8): 1185-1196, 2005.
- [65] S.-W. Kim, B. J. Oommen, "A brief taxonomy and ranking of creative prototype reduction schemes," *Pattern Analysis Application*, 6(3): 232-244, 2003.
- [66] S. W. Kim, B. J. Oommen, "Enhancing prototype reduction schemes with recursion: a method applicable for "large" data sets," *IEEE Transactions System Man and Cybernetics, Part B*, 34(3): 1384-97, 2004.
- [67] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, New York: J. Wiley, 2004.
- [68] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, 3:1415-1438, 2003.
- [69] J. S. Ma, "Function replacement vs. kernel trick," *Neurocomputing*, 50: 479-483, 2003.
- [70] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data," *Applied Statistics*, 44(1): 101-115, 1995.
- [71] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, Berkeley, CA, 1967.
- [72] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.
- [73] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.
- [74] R. N. Dave and R. Krishnapuram "Robust clustering methods: A unified view," *IEEE Transactions on Fuzzy Systems*, 5(2): 270-293, 1997.
- [75] B. Fritzke, "A growing neural gas network learns topologies," *Proc. of 8th Annual Conference on Neural Information Processing Systems (NIPS'95)*, pp. 625-632, Denver, CO, USA, 1995.

BIBLIOGRAPHY

- [76] M. Loog and R. P. W. Duin, "Non-iterative heteroscedastic linear dimension reduction for two-class data: From Fisher to Chernoff," *Proc. of the 4th Joint IAPR International Workshops SSPR 2002 and SPR 2002 (S+SSPR'02)*, pp. 508–517, Windsor, Ontario, Canada, 2002.
- [77] M. Loog and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6): 732–739, 2004.
- [78] C. Blake, E. Keogh, and C. J. Metz, "UCI repository of machine learning databases," Dept. of Information and Computer Science, University of California at Irvine, CA, <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998.
- [79] R. S. Zemel, *A Minimum Description Length Framework for Unsupervised Learning*. Ph.D. thesis, University of Toronto, 1994.
- [80] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, New York, N.Y.: John Wiley and Sons, Inc., 1991.
- [81] O. Kallenberg, *Foundations of Modern Probability*, New York: Springer-Verlag, 1997.
- [82] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, New York: John Wiley & Sons, 1992.
- [83] T. Lissack and K. S. Fu, "Error estimation in pattern recognition via $-L^\alpha$ distance between posterior density functions," *IEEE Transactions on Information Theory*, 22(1): 34-45, 1976.
- [84] M. Ben-Bassat, "Use of distance measures, information measures and error bounds in feature evaluation," P. R. Krishnaiah and L. N. Kanal (Eds.), *Handbook of Statistics*, volume 2, pp. 773-791, North-Holland Publishing Company, 1982.
- [85] M. E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Transactions on Information Theory*, IT-16(4): 368-372, 1970.
- [86] E. A. Patrick and F. P. Fisher, II, "Non-parametric feature selection," *IEEE Transactions on Information Theory*, IT-15: 577-584, 1969.
- [87] G. T. Toussaint, "Feature selection criteria and contextual decoding algorithms in statistical pattern recognition," *Ph.D. dissertation*, University of British Columbia, Vancouver, Canada, 1972.

BIBLIOGRAPHY

- [88] M. Röhl and C. Weihs, "Optimal vs. Classical Linear Dimension. Reduction," Proc. of the 22nd Ann. GfKI Conf., pp. 252-259, 1998.
- [89] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, 220(4598): 671-680, 1983.
- [90] K. Demuynck, J. Duchateau and D. V. Compernelle, "Optimal feature sub-space selection based on discriminant analysis," *Proc. of EUROSPEECH'99*, pp. 1311-1314, Budapest, Hungary, 1999.
- [91] R. Lotlikar and R. Kothari, "Adaptive linear dimensionality reduction for classification," *Pattern Recognition*, 33(2): 185-194, 2000.
- [92] R. Irving, *Introduction to Matrix Theory and Linear Algebra*, Holt, Rinehart and Winston, 1971.
- [93] G. Saon and M. Padmanabhan, "Minimum Bayes error feature selection for continuous speech recognition," *Proc. of 13th Neural Information Processing Systems (NIPS'01)*, pp. 800-806, Denver, USA, 2001.
- [94] J. Nocedal and S. J. Wright, *Numerical Optimization. Springer Series in Operations Research*, New York, NY: Springer-Verlag, 1999.
- [95] N. Campbell, "Canonical variate analysis—a general formulation," *Australian Journal of Statistics*, 26: 86-96, 1984.
- [96] N. Kumar and A. G. Andreou, "A generalization of linear discriminant analysis in maximum likelihood framework," *Proc. of the Joint Statistical Meeting, Statistical Computing Section*, Chichago, IL, USA, 1996.
- [97] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, 26(4): 283-297, 1998.
- [98] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern recognition*, 37(5): 857-874, 2004.
- [99] J. T. Tou and R. P. Heydorn, *Computer and Information Sciences, Volume 2*, New York: Academic Press, 1967.
- [100] C. C. Babu and R. N. Kalra, "On feature extraction in multiclass pattern recognition," *International Journal of Control*, 15: 595-601, 1972.

BIBLIOGRAPHY

- [101] H. P. Decell and J. A. Quirein, "An iterative approach to the feature selection problem," *Purdue University Conference on Machine Processing of Remotely Sensed Data*, 1973.
- [102] E. G. Schukat-Talamazzini, J. Hornegger, and H. Niemann, "Optimal linear feature transformations for semi-continuous hidden markov models," *Proc. of the 20th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, pp. 369-372, Detroit, USA, 1995.
- [103] L. J. Buturović, "Toward Bayes-optimal linear dimension reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(4): 420-424, 1994.
- [104] M. E. Aladjem, "Non-parametric discriminant analysis via recursive optimization of Patrick-Fisher distance," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 28(2): 292-299, 1998.
- [105] R. Lotlikar and R. Kothari, "Bayes-optimality motivated linear and multilayered perceptron-based dimensionality reduction," *IEEE Transactions on Neural Networks*, 11(2): 452-463, 2000.
- [106] J. N. Kapur, *Measures of information and their applications*, New Delhi, India: Wiley, 1994.
- [107] J. C. Principe, J. W. Fisher III, and D. Xu, "Information theoretic learning," S. Haykin (Ed.), *Unsupervised Adaptive Filtering*. New York, NY: Wiley, 2000.
- [108] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval," *Technical Report CSPL-328*, University of Michigan Ann Arbor, Communications and Signal Processing Laboratory, 2001.
- [109] M. Zhu and T. Hastie "Feature extraction for non-parametric discriminant analysis," *Journal of Computational & Graphical Statistics*, 12(1): 101-120, 2003.
- [110] J. H. Friedman, W. Stuetzle, and A. Schroeder, "Projection pursuit density estimation," *Journal of the American Statistical Association*, 79(387): 599-608, 1984.
- [111] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4): 388-400, 1993.
- [112] J. Y. Zhang and Y. X. Liu, "SVM decision boundary based discriminative subspace induction," *Pattern Recognition*, 38(10): 1746-1758, 2005.

BIBLIOGRAPHY

- [113] D. C. Tao and X. O. Tang, "Non-parametric discriminant analysis in relevance feedback for content-based image retrieval," *Proc. of the 17th International Conference on Pattern Recognition (ICPR'04)*, pp. 1013-1016, Cambridge, UK, 2004.
- [114] K.-P. Chung, C. C. Fun, "A hierarchical non-parametric discriminant analysis approach for a content-based image retrieval system," *IEEE International Conference on e-Business Engineering (ICEBE'05)*, pp. 346-351, Beijing, China, 2005.
- [115] Z. F. Li, W. Liu, D. H. Lin, X. O. Tang, "Non-parametric subspace analysis for face recognition," *Proc. of 2005 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 961-966, San Diego, CA, USA, 2005.
- [116] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, 8: 376-- 386, 1938.
- [117] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society. Series B*, 10: 159-203, 1948.
- [118] J. P. Ye and Q. Li, "LDA/QR: An efficient and effective dimension reduction algorithm and its theoretical foundation," *Pattern Recognition*, 37(4): 851-854, 2004.
- [119] F. X. Song and D. Zhang, and J. Y. Yang, "A novel dimensionality reduction approach for face recognition," *Neurocomputing*, 69: 1683-1687, 2006.
- [120] D. Foley and J. Sammon, "An optimal set of discriminant vectors," *IEEE Transactions on Computers*, 24(3): 281-289, 1975.
- [121] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6): 978 - 983, 1988.
- [122] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition* 34(7): 1405-1416, 2001.
- [123] F. X. Song, S. H. Liu, and J. Y. Yang, "Orthogonalized Fisher discriminant," *Pattern Recognition*, 38(2): 311-313, 2005.

BIBLIOGRAPHY

- [124] J. P. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research*, 6: 483-502, 2005.
- [125] Z. Jin, J. Y. Yang, Z. M. Tang, and Z. S. Hu, "A theorem on the uncorrelated optimal discriminant vectors," *Pattern Recognition* 34(10): 2041-2047, 2001.
- [126] Y. Xu, J. Y. Yang, and Z. Jin, "Theory analysis on FSLDA and ULDA," *Pattern Recognition*, 36(12): 3031-3033, 2003.
- [127] H.-C. Kim, D.-J. Kim, and S.-Y. Bang, "Extensions of LDA by PCA mixture model and class-wise features," *Pattern Recognition*, 36(5): 1095-1105, 2003.
- [128] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8): 1274-1286, 2006.
- [129] H. Li, T. Jiang, K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, 17(1): 157-165, 2006.
- [130] X. P. Qiu and L. D. Wu, "Non-parametric maximum margin criterion for face recognition," *Proc. of 2005 IEEE International Conference on Image Processing (ICIP'05)*, pp. 918-921, Genova, Italy, 2005.
- [131] J. Yan, B. Y. Zhang, S. C. Yan, Q. Yang, H. Li, Z. Chen, W. S. Xi, W. G. Fan, W. Y. Ma, and Q. S. Cheng, "IMMC: incremental maximum margin criterion," *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pp. 725-730, Seattle, Washington, USA, 2004.
- [132] P. F. Hsieh, D. S. Wang, C. W. Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information," *IEEE Transactions on Pattern Analysis Machine Intelligence*, 28(2): 223-235, 2006.
- [133] J. D. Tubbs, W. A. Coberly, and D. M. Young, "Linear dimension reduction and Bayes classification with unknown population parameters," *Pattern Recognition*, 15(3): 167-172, 1982.
- [134] H. Brunzell and J. Eriksson, "Feature reduction for classification of multidimensional data," *Pattern Recognition*, 33(10): 1741-1748, 2000.

BIBLIOGRAPHY

- [135] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," *Proc. the 25th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, pp. 1747-1750, Istanbul, Turkey, 2000.
- [136] Z. Q. Hong and J. Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," *Pattern Recognition*, 24(4): 317-324, 1991.
- [137] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, 84(405): 165-175, 1989.
- [138] I. Pima and M. Aladjem, "Regularized discriminant analysis for face recognition," *Pattern Recognition*, 37(9): 1945-1948, 2004.
- [139] W. Zhao, R. Chellappa, and P. Phillips, "Subspace linear discriminant analysis for face recognition," *Technical Report CAR-TR-914*, Center for Automation Research, University of Maryland, College Park, 1999.
- [140] J. W. Lu, K. N. Plataniotis and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, 26(2): 181-191, 2005.
- [141] D. Q. Dai and P. C. Yuen, "Regularized discriminant analysis and its application to face recognition," *Pattern Recognition*, 36(3): 845-847, 2003
- [142] 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*, 14(2): 214-223, 2004.
- [143] R. P. W. Duin, M. Loog, and R. Haeb-Umbach, "Multi-class linear feature extraction by nonlinear PCA," *Proc. of the 15th International Conference on Pattern Recognition (ICPR'00)*, pp. 398-401, Barcelona, Spain, 2000.
- [144] M. Loog, R. R. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7): 762-766, 2001.
- [145] M. Loog, *Approximate pairwise accuracy criteria for multiclass linear dimension reduction: generalisations of the Fisher criterion*, Number 44 in WBBM Report Series. Delft University Press, Delft, 1999.

BIBLIOGRAPHY

- [146] Y. Li, Y. Gao, and H. Erdogan, "Weighted pairwise scatter to improve linear discriminant analysis," *Proc. of the 6th International Conference of Spoken Language Processing (ICSLP'00)*, Beijing, China, 2000.
- [147] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6): 623-627, 2000.
- [148] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1): 1-12, 2000.
- [149] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press and J. Wiley, 1983.
- [150] Q. Yang and X. Tang, "Recent advances in subspace analysis for face recognition," *Invited Paper, Advances in Biometric Person Authentication*, S. Z. Li *et al.* (Eds.), Springer-Verlag Berlin Heidelberg, pp. 275-287, 2004.
- [151] G. Shakhnarovich, B. Moghaddam, "Face recognition in subspaces," *Handbook of Face Recognition*, S. Z. Li and A. K. Jain (Eds.), New York: Springer, 2004.
- [152] Q. Tian, Y. Fainman, and S. H. Lee, "Comparison of statistical pattern-recognition algorithms for hybrid processing. II. Eigenvector-based algorithm," *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 5(10): 1670-1682, 1988.
- [153] S. Raudys and R. P. W. Duin, "On expected classification error of the Fisher linear classifier with pseudo-inverse of the covariance matrix," *Pattern Recognition Letters*, 19(5-6): 385-392, 1998.
- [154] M. Skurichina and R. P. W. Duin, "Stabilizing classifiers for very small sample size," *Proc. the 13th International Conference on Pattern Recognition (ICPR'96)*, pp. 891-896, Vienna, Austria, 1996.
- [155] M. Skurichina and R. P. W. Duin, "Regularization of linear classifiers by adding redundant features," *Pattern Analysis and Applications*, 2(1): 44-52, 1999.
- [156] J. P. Ye, R. Janardan, C.-H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8): 982-94, 2004.

BIBLIOGRAPHY

- [157] Q. Li and J. P. Ye, "A two-stage linear discriminant analysis via QR-decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6): 929-941, 2005.
- [158] Y. Q. Cheng, Y. M. Zhuang, and J. Y. Yang, "Optimal Fisher discriminant analysis using the rank decomposition," *Pattern Recognition*, 25(1): 101-111, 1992.
- [159] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8): 831-836, 1996.
- [160] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng, "Discriminant analysis of principal components for face recognition," Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang (Eds.), *Face Recognition: From Theory to Applications*, pp. 73-85, Springer-Verlag, 1998.
- [161] P. N. Belhumeur, J. Hespanha, and D. Kiregeman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 711-720, 1997.
- [162] J. Yang and J. Yang, "Why can LDA be performed in PCA transformed space," *Pattern Recognition*, 36(2): 563-566, 2003.
- [163] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2): 230-244, 2005.
- [164] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, 10(33): 1713-1726, 2000.
- [165] R. Huang, Q. S. Liu, H. Q. Lu, and S. D. Ma, "Solving the small sample size problem of LDA", *Proc. of the 16th International Conference on Pattern Recognition*, pp. 29-32, Quebec, Canada, Aug. 2002.
- [166] W. M. Zheng, L. Zhao and C. R. Zou, "An efficient algorithm to solve the small sample size problem for LDA," *Pattern Recognition*, 37(5): 1077-1079, 2004.

BIBLIOGRAPHY

- [167] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(1): 4-13, 2005.
- [168] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data-with application to face recognition," *Pattern Recognition*, 34(10): 2067-2070, 2000.
- [169] J. W. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Transactions on Neural Networks*, 14(1): 117-126, 2003.
- [170] J. W. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face Recognition Using LDA Based Algorithms," *IEEE Transactions on Neural Networks*, 14(1): 195-200, 2003.
- [171] J. R. Price and F. G. Timothy, "Face recognition using direct, weighted linear discriminant analysis and modular subspaces," *Pattern Recognition*, 38(1): 209-219, 2005.
- [172] H. Gao, J. W. Davis, "Why direct LDA is not equivalent to LDA," *Pattern Recognition*, 39(5): 1002-1006, 2006.
- [173] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228--233, 2001.
- [174] C. C. Paige and M. A. Saunders, "Toward a generalized singular value decomposition," *SIAM Journal on Numerical Analysis*, 18: 398-405, 1981.
- [175] J. P. Ye, T. Li, T. Xiong, and R. Janardan, "Using uncorrelated discriminant analysis for tissue classification with gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(4):181-190, 2004.
- [176] J. P. Ye and T. Xiong, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis," *Journal of Machine Learning Research*, 7: 1183-1204, 2006.
- [177] X. G. Wang and X. O. Tang, "Dual-space linear discriminant analysis for face recognition," *Proc. of 2004 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 564-569, Washington, DC, USA, 2004.
- [178] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 696-710, 1997.

BIBLIOGRAPHY

- [179] X. G. Wang and X. O. Tang, "Random sampling LDA for face recognition," *Proc. of 2004 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 259-265, Washington, DC, USA, 2004.
- [180] X. G. Wang and X. O. Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9): 1222-1228, 2004.
- [181] M. Bressan and J. Vitria, "Non-parametric discriminant analysis and nearest neighbor classification," *Pattern Recognition Letters*, 24(15): 2743-2749, 2003.
- [182] B.-C. Kuo and D. Landgrebe, "Non-parametric weighted feature extraction for classification," *IEEE Transactions on Geoscience and Remote Sensing*, 42(5): 1096-1105, 2004.
- [183] J. P. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Proc. of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04)*, pp. 1569-1576, Vancouver, Canada, 2004.
- [184] M. Li and B. Z. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, 26(5): 527-532, 2005.
- [185] X. F. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," *Proc. of the 19th Annual Conference on Neural Information Processing Systems (NIPS'05)*, pp. 499-506, Vancouver, Canada, 2005.
- [186] J. Yang, D. Zhang, X. Yong and J. Y. Yang, "Two-dimensional discriminant transform for face recognition," *Pattern Recognition* 38(7): 1125-1129, 2005.
- [187] X.-Y. Jing, H.-S. Wong, D. Zhang, "Face recognition based on 2D Fisherface approach," *Pattern Recognition*, 39(4): 707-710, 2006.
- [188] X. F. He and P. Niyogi, "Locality preserving projections," *Proc. of the 17th Neural Information Processing Systems (NIPS'03)*, Vancouver, Canada, 2003.
- [189] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," *Proc. of 2005 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 846-853, San Diego, CA, USA, 2005.
- [190] H. Zhao, S. Sun, Z. Jing, and J. Yang, "Local structure based supervised feature extraction," *Pattern Recognition*, 39(8): 1546-1550, 2006.
- [191] W. W. Yu, X. L. Teng, and C. Q. Liu, "Face recognition using discriminant locality preserving," *Image and Vision Computing*, 24(3): 239-248, 2006.

BIBLIOGRAPHY

- [192] X. F. He S. C. Yan, Y. X. Hu, P. Niyogi, H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3): 328-340, 2005.
- [193] L. Jimenez and D. Landgrebe, *High dimensional feature reduction via projection pursuit*, PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 96-5, 1996.
- [194] C. Santa Cruz and J. R. Dorronsoro, "A nonlinear discriminant algorithm for feature extraction and data classification," *IEEE Transaction on Neural Networks*, 9(6): 1370-1376, 1998.
- [195] E.-K. Lee, D. Cook, S. Klinke, and T. Lumley, "Projection pursuit for exploratory supervised classification," *SFB 649 Discussion Papers SFB649DP2005-026*, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.
- [196] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, 40(12): 3043-3054, 1992.
- [197] A. Biem, S. Katagiri and B.-H. Juang, "Pattern recognition using discriminative feature extraction," *IEEE Transactions on Signal Processing*, 45(2): 500-504, 1997.
- [198] S. Katagiri, C.-H. Lee, and B.-H. Juang, "A generalized probabilistic descent method," *Proc. of the Acoustic Society of Japan, Fall Meeting*, pp. 141-142, 1990.
- [199] M. Ordowski and G. G. L. Meyer, "Geometric linear discriminant analysis for pattern recognition," *Pattern Recognition*, 37(3): 421-428, 2004.
- [200] S. Kaski and J. Peltonen, "Informative discriminant analysis," *Proc. of the 20th International Conference on Machine Learning (ICML'03)*, pp. 329-336, Washington DC, USA, August 21-24, 2003.
- [201] T. E. Flick, L. K. Jones, R. G. Priest and C. Herman, "Pattern classification using projection pursuit," *Pattern Recognition*, 23(12): 1367-1376, 1990.
- [202] X. C. Wang and K. K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognition*, 36(10): 2429-2439, 2003.

BIBLIOGRAPHY

- [203] E. Choi and C. Lee, "Optimizing feature extraction for multiclass problems," *IEEE Transactions on Geoscience and Remote Sensing*, 39(3): 521-528, 2001.
- [204] X. W. Liu, A. Srivastava and K. Gallivan, "Optimal linear representations of images for object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5): 662-666, 2004.
- [205] A. Sato, "Discriminative dimensionality reduction based on generalized LVQ," *Proc. of 2001 International Conference on Artificial Neural Networks (ICANN'01)*, pp. 65-72, Vienna, Austria, August 21-25, 2001.
- [206] M.-K. Tasy, K.-H. Shyu and P.-C. Chang, "Feature transformation with generalized learning vector quantization for handwritten Chinese character recognition," *IEICE Transactions on Information & Systems*, E82-D(3): 687-692, 1999.
- [207] G. Faddi, A. Kocsor, and L. Tóth, "Initialization of directions in projection pursuit learning," *Proc. of 2004 International Joint Conference on Neural Networks (IJCNN'04)*, pp. 1-4, Budapest, Hungary, 2004.
- [208] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, 4(2): 164-171, 2000.
- [209] C. J. Liu and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6): 570-582, 2000.
- [210] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, 4: 119-155, 2003.
- [211] D. de Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, and R. P. W. Duin, "Supervised locally linear embedding," *Proc. of 2003 Joint International Conference on Artificial Neural Networks and Neural Information Processing (ICANN'03/ICONIP'03)*, pp. 333-341, Istanbul, Turkey, 2003.
- [212] X. M. Bai, B. C. Yin, Q. Shi, Y. F. Sun, "Face recognition based on supervised locally linear embedding method," *Journal of Information and Computational Science*, 2(4): 641-646, 2006.

BIBLIOGRAPHY

- [213] D. de Ridder, M. Loog, M. J. T. Reinders, "Local Fisher embedding," *Proc. of the 17th International Conference on Pattern Recognition (ICPR'04)*, pp. 295-298, Cambridge, UK, 2004.
- [214] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Non-linear dimensionality reduction techniques for classification and visualization," *Proc. of the 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 645-651, Edmonton, AB, Canada, 2002.
- [215] X. Geng, D. C. Zhan, and Z. H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on System, Man, Cybernetics, Part B*, 35(6): 1098:1107, 2005.
- [216] J. P. Zhang, S. Z. Li, and J. Wang, "Manifold learning and applications in recognition," *Intelligent Multimedia Processing with Soft Computing*, Yap Peng Tan, Kim Hui Yap, Lipo Wang (Eds.), Springer-Verlag, Heidelberg, 2004.
- [217] J. P. Zhang, H. X. Shen, and Z. H. Zhou, "Unified locally linear embedding and linear discriminant analysis algorithm (ULLELDA) for face recognition," *Proc. of the 5th Chinese Conference on Biometric Recognition: Advances in Biometric Person Authentication (SINOBIOMETRICS)*, pp. 296-304, Guangzhou, China, 2004.
- [218] S. C. Yan, H. J. Zhang, Y. X. Hu, B. Y. Zhang, and Q. S. Cheng, "Discriminant analysis on embedded manifold," *Proc. of the 8th European Conference on Computer Vision (ECCV'04)*, pp. 121-132, Prague, Czech Republic, 2004.
- [219] M.-H. Yang, "Extended isomap for pattern classification," *Proc. of the 18th national conference on Artificial intelligence*, pp. 224-229, Edmonton, Alberta, Canada, 2002.
- [220] Y. M. Wu, K. L. Chan, and L. Wang, "Face recognition based on discriminative manifold learning," *Proc. of the 17th International Conference on Pattern Recognition (ICPR'04)*, pp. 171-174, Cambridge, UK, 2004.
- [221] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance function using equivalence relations," *Proc. of the 20th International Conference on Machine Learning (ICML'03)*, pp. 11-18, Washington DC, USA, 2003.

BIBLIOGRAPHY

- [222] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition*, 34(7): 1405-1416, 2001.
- [223] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. of IEEE*, 70: 939-952, 1982.
- [224] C. E. Thomaz, D. F. Gillies, "A maximum uncertainty LDA-based approach for limited sample size problems - with application to face recognition," *Proc. of the 18th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'05)*, pp. 89-96, Natal, 2005.
- [225] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahrouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and [rediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, 1(2): 133-143, 2002.
- [226] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286(5439): 531-537, 1999.
- [227] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and R. T. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. of the National Academy of Sciences of the United States of America*, 98(26): 15149-15154, 2001.
- [228] C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. Rifkin, M. Angelo, M. Reich, E. S. Lander, J. P. Mesirov, and T. R. Golub, "Molecular classification of multiple tumor types," *Bioinformatics*, 17(s1), s316-s322, 2001.
- [229] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van. de Rijn, M. Waltham, A. Pergamenschikov, Jeffrey C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Pharmacogenomics*, 24(3):227-235, 2000.

BIBLIOGRAPHY

- [230] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. Meltzer, "Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks," *Nature Medicine*, 7(6): 673-679, 2001.
- [231] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, 21(5):631-43, 2005.
- [232] TREC. *Text Retrieval conference*. <http://trec.nist.gov>, 1999.
- [233] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms For Pattern Recognition and Image Processing*. Norwell, MA: Kluwer, 1999.
- [234] J. W. Han, M. Kamber, and A. K. H. Tung, "Spatial clustering methods in data mining: a survey," H. Miller, and J. W. Han (Eds.), *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis.
- [235] S. C. Ahalt, A. K. Krishnamurty, P. Chen, and D. E. Melton, "Competitive learning algorithms for vector quantization," *Neural Networks*, 3(3): 277-291, 1990.
- [236] H. Bischof, A. Leonardis, and A. Selb, "MDL principle for robust vector quantization," *Pattern Analysis and Application*, 2(1): 59-72, 1999.
- [237] S. K. Bhatia and J. S. Deogun, "Conceptual clustering in information retrieval," *IEEE Transactions on System, Man, Cybernetics, Part B*, 28(3): 427-436, 1998
- [238] D. Jiang, C. Tang, and A. Zhang. "Cluster analysis for gene expression data: A survey," *IEEE Transaction on Knowledge and Data Engineering*, 16(11):1370-1386, 2004.
- [239] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 29(6): 778-785, 1999.
- [240] F. L. Chung, and T. Lee, "Fuzzy competitive learning," *Neural Networks*, 7(3): 539-551, 1994.

BIBLIOGRAPHY

- [241] T. M. Martinetz, "Competitive Hebbian learning rule forms perfectly topology preserving maps," *Proc. of 1993 International Conference on Artificial Neural Networks (ICANN'93)*, pp. 427-434 Amsterdam, Netherlands, 1993.
- [242] P. S. Bradley and U. M. Fayyad, "Refining initial points for k -means clustering," *Proc. 15th Int. Conf. on Machine Learning (ICML'98)*, pp. 91-99, Morgan Kaufmann, San Francisco, CA, 1998.
- [243] B. Zhang and M. Hsu, "K-harmonic means-a data clustering algorithm," *Technical Report HPL-1999-124*, Hewlett-Packard Labs, <http://www.hpl.hp.com/techreports/1999/HPL-1999-124.html>.
- [244] T.-N. Yang, S.-D. Wang, and S.-J. Yen, "Fuzzy algorithms for robust clustering," *Proceeding of International Computer Symposium*, Houliian, Taiwan, 2002.
- [245] K. K. Chintalapudi and M. Kam, "The credibilistic fuzzy c -means algorithm," *Proc. of 1998 IEEE International Conference on Systems, Man and Cybernetics*, pp. 2034-2040, San Diego, California, USA, 1998.
- [246] S. C. Chu, J. F. Roddick and J. S. Pan, "A comparative study and extension to k -medoides algorithms," *Proc. of 5th International Conference on Optimization: Techniques and applications (ICOTA'01)*, pp. 1708-1717, Hong Kong, China, 2001.
- [247] Y. Ohashi, "Fuzzy clustering and robust estimation," *Proc. of 9th Meet. SAS Users Grp. Int.*, Hollywood Beach, FL, 1984.
- [248] R. N. Dave, "Characterization and detection of noise in clustering," *Pattern Recognition Letter*, 12(11): 657-664, 1991.
- [249] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy System*, 1(2): 98-110, May 1993.
- [250] G. Beni and X. Liu, "A least biased fuzzy clustering method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9): 954-960, 1994.
- [251] H. Timm, C. Borgelt, C., Döring, and R. Kruse, "Fuzzy Cluster Analysis with Cluster Repulsion," *Proc. of European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Germany, 2001.

BIBLIOGRAPHY

- [252] H. Timm and R. Kruse, "A modification to improve possibilistic fuzzy cluster analysis," *Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'02)*, pp. 1460 – 1465, Piscataway, NJ, USA 2002.
- [253] H. Timm, C. Borgelt, C. Döring, and R. Kruse, "An extension to possibilistic fuzzy cluster analysis," *Fuzzy Sets and Systems*, 147(1): 3-16, 2004.
- [254] N. R. Pal, K. Pal and J. C. Bezdek, "A mixed c -means clustering model," *IEEE International Conference on Fuzzy System*, pp. 11-21, Spain, 1997.
- [255] N. R. Pal, K. Pal, J. M. Keller, J. C. Bezdek, "A Possibilistic fuzzy c -means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, 13(4): 517-530, 2005.
- [256] J. L. Chen and J. H. Wang, "A new robust clustering algorithm – density-weighted fuzzy c -means," *Proc. of Conf. IEEE SMC '99*, Japan, vol. 3, pp. 90-94, 1999.
- [257] K.-L. Wu, M.-S. Yang, "Alternative c -means clustering algorithms," *Pattern Recognition*, 35(10): 2267-2278, 2002.
- [258] R. J. Hathaway and J. C. Bezdek, "Generalized fuzzy c -means clustering strategies using L_p norm distances," *IEEE Transactions on Fuzzy System*, 8(5): 576-582, 2000.
- [259] J. S., Zhang and Y. W. Leung, "Robust clustering by pruning outliers," *IEEE Transactions on System, Man and Cybernetics, Part. B*, 33(6): 983-999, 2003.
- [260] J. Leski, "Towards a robust fuzzy clustering," *Fuzzy Sets and Systems*, 137(2): 215-233, 2003.
- [261] H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," *Pattern Recognition*, 30(7): 1109-1119, 1997.
- [262] H. Frigui and R. Krishnapuram, "A robust competitive clustering algorithm with applications in computer vision," *IEEE Transactions on Pattern Analysis Machine Intelligence*, 21(5): 450-465, 1999.
- [263] D. Pelleg and A. Moore, "X-means: Extending k -means with efficient estimation of the number of clusters," *Proc. of the 17th International Conference on Machine Learning (ICML'00)*, pp. 727-734, San Francisco, USA, 2000.
- [264] G. Hamerly and C. Elkan, "Learning the k in k -means," *Proc. of 17th Annual Conference on Neural Information Processing Systems (NIPS'03)*, pp. 281-288, Vancouver and Whistler, British Columbia, Canada, 2003.

BIBLIOGRAPHY

- [265] T. M. Martinetz and K. Schulten, "Topology representing networks," *Neural Networks*, 7(3): 507-522, 1994.
- [266] M. Winter, G. Metta, and G. Sandini, "Neural-gas for function approximation: a heuristic for minimizing the local estimation error," *Proc. of IEEE-INNS-ENNS International Joint Conference on Neural Networks, (IJCNN'00)*, pp. 535-538, Italy, 2000.
- [267] A. S. Atukorale and P. N. Suganthan, "Hierarchical overlapped neural gas network with application to pattern classification," *Neurocomputing*, 35(1-4): 165-176, 2000.
- [268] B. L. Zhang, M. Y. Fu and H. Yan, "A nonlinear neural network model of mixture of local principle component analysis: application to handwritten digits recognition," *Pattern Recognition*, 34(2): 203-214, 2001.
- [269] J. G. Wang and H. Yan, "A hybrid method for unconstrained handwritten numeral recognition by combining structural and neural "gas" classifiers," *Pattern Recognition Letter*, 21(6-7): 625-635, 2000.
- [270] B. Hammer, M. Strickert and T. Villmann, "Supervised neural gas with general similarity measure," *Neural Processing Letters*, 21(1): 21-44, 2005.
- [271] K. Rose, E. Gurewitz, and G. Fox, "Statistical mechanics and phase transitions in clustering," *Physical Review Letters*, 65(8): 945-848, 1990.
- [272] A. Carlevarino, R. Martinotti, G. Metta, and G. Sandini, "Incremental growing neural network and its application to robot control. *Proceeding of the 2000 International Joint Conference on neural networks (IJCNN'00)*, pp. 323-328, Como, Italy, 2000.
- [273] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. of the IEEE*, 86(11): 2210-2239, 1998.
- [274] B. Fritzke, "Some competitive learning methods," *Technique Report*, Institute for Neural Computation, Ruhr-University, Bochum, <http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/JavaPaper/>, 1997.
- [275] B. Fritzke, "Growing cells structures - a self-organizing network for unsupervised and supervised learning," *Neural Networks*, 7(9):1441-1460, 1994.

BIBLIOGRAPHY

- [276] S. Ray and R. H. Turi, "Determination of number of clusters in k -means clustering and application in color image segmentation," *Proc. of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*, pp. 137-143, Calcutta, India, 1999.
- [277] H. Yu, "Automatically determining number of clusters," *Information Retrieval* (CMU CS11-741) Final Report, 1998.
- [278] H. Akaike, "A new look at the statistical identification model," *IEEE Transactions on Automatic Control*, AC-19: 716-723, 1974.
- [279] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, 5(2): 461-464, 1978.
- [280] Y. H. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation," *Biometrika*, 92(4): 937-950, 2005.
- [281] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer Journal*, 11(2): 195-209, 1968.
- [282] J. Rissanen, "Modeling by the shortest data description," *Automatica*, 14: 465-471, 1978.
- [283] P. Grünwald, "A tutorial introduction to the minimum description length principle," In: *Advances in Minimum Description Length: Theory and Applications* (edited by P. Grünwald, I.J. Myung, and M. Pitt), MIT Press, 2005.
- [284] R. Baxter, *Minimum Message Length Inference: Theory and Applications*, PhD thesis, Department of Computer Science, Monash University, Clayton, Australia, 1996.
- [285] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: Part I," *SIGMOD Record*, 31(2): 40-45, 2002.
- [286] R. Krishnapuram and H. O. Nasraoui, "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation - part 1&2," *IEEE Transactions on Fuzzy Systems*, 3(1): 29--43, 1995.
- [287] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal on Machine Learning Research (JMLR)*, 3:583-617, 2002.