GLOBAL-LOCAL PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS AND ITS EXTENSION IN REPRODUCING KERNEL HILBERT SPACE

AMINU MUHAMMAD

UNIVERSITI SAINS MALAYSIA

2021

GLOBAL-LOCAL PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS AND ITS EXTENSION IN REPRODUCING KERNEL HILBERT SPACE

by

AMINU MUHAMMAD

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

April 2021

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to you Dad, for the love, support and encouragement you gave me my entire life. I would also like to thank my supervisor, Assoc. Professor Dr. Noor Atinah Ahmad for her help and thoughtful advise. I am especially grateful to my co-supervisor Dr. Norhashidah Awang for her continued support and guidance during my graduate study.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

GLSP	Global Local Structure Preserving
KLPPLS-DA	Kernel Locality Preserving Partial Least Squares Discriminant Analysis
KNPPLS-DA	Kernel Neighborhood Preserving Partial Least Squares Discriminant Analysis
KNN	K-Nearest Neighbor
KUNPPLS-DA	Kernel Uncorrelated Neighborhood Preserving Partial Least Squares Discriminant Analysis
LE	Laplacian Eigenmap
LDA	Linear Discriminant Analysis
LLE	Locally Linear Embedding
LPP	Locality Preserving Projections
LPPLS-DA	Locality Preserving Partial Least Squares Discriminant Analysis
NIPALS	Nonlinear Iterative Partial Least Squares
NPE	Neighborhood Preserving Embedding
NPPLS-DA	Neighborhood Preserving Partial Least Squares Discriminant Analysis
PCA	Principal Component Analysis
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis

RBF	Radial Basis Function
RKHS	Reproducing Kernel Hilbert Space
SVM	Support Vector Machines
UNPPLS-DA	Uncorrelated Neighborhood Preserving Partial Least Squares Discriminant Analysis

LIST OF SYMBOLS

n	The number of data points
m	The number of features
d	The number of reduced features
C	The number of classes
x_i	The i-th data point
X	The data matrix
$\bar{X}$	The centred data matrix
S_b	The between class scatter matrix
S_w	The within class scatter matrix
S_t	The total scatter matrix
S	The affinity matrix
L	The graph Laplacian matrix
W	The transformation matrix
$\mathscr{H}$	The feature space
K	The kernel matrix
$\bar{K}$	The centred kernel matrix
ϕ, ψ	The nonlinear mapping functions
Φ, Ψ	The data matrices in feature space
$\bar{\Phi}, \bar{\Psi}$	The centred data matrices in feature space

S_b^{ϕ}, S_b^{ψ}	The between class scatter matrices in feature space
S_w^{ϕ}, S_w^{ψ}	The within class scatter matrices in feature space
S_t^{ϕ}, S_t^{ψ}	The total class scatter matrices in feature space

ANALISIS DISKRIMINAN KUASA DUA TERKECIL SEPARA GLOBAL-SETEMPAT DAN PELANJUTANNYA DALAM RUANG HILBERT KERNEL PENGHASILAN SEMULA

ABSTRAK

Pembelajaran subruang adalah satu pendekatan penting untuk mempelajari perwakilan dimensi rendah bagi suatu ruang dimensi tinggi. Apabila sampel data diwakili sebagai titik dalam ruang dimensi tinggi, pembelajaran dengan kedimensian tinggi menjadi mencabar kerana keberkesanan dan kecekapan algoritma pembelajaran turun dengan ketara apabila dimensi meningkat. Oleh itu, teknik pembelajaran subruang digunakan untuk mengurangkan kedimensian data sebelum menggunakan algoritma pembelajaran yang lain. Baru-baru ini, minat terhadap teknik pembelajaran subruang yang berdasarkan kerangka pemeliharaan struktur global dan tempatan (GLSP). Telah meningkat idea utama pendekatan GLSP adalah mencari transformasi data berdimensi tinggi kepada subruang berdimensi yang lebih rendah dengan maklumat struktur data global dan tempatan terpelihara dalam subruang berdimensi rendah. Tesis ini mempertimbangkan kes yang mana data daripada sampel dalam manifold dasar yang terbenam dalam ruang sekitar dimensi tinggi. Dua algoritma pembelajaran subruang baharu dipanggil 'locality preserving partial least squares discriminant analysis' (LPPLS-DA) dan 'neighborhood preserving partial least squares discriminant analysis' (NPPLS-DA) yang berdasarkan kerangka GLSP dicadangkan untuk pembelajaran subruang diskriminan. Tidak seperti analisis diskriminan kuasa dua terkecil separa konvensional (PLS-DA) yang bertujuan hanya untuk memelihara ruang data struktur Euclidan global ruang data, LPPLS-DA yang dicadangkan dan algoritma NPPLS-DA mencari pembenaman yang memelihara kedua-dua struktur global dan struktur manifold tempatan.

Hasilnya, kedua-dua LPPLS-DA dan NPPLS-DA mampu mengekstrak lebih banyak maklumat diskriminasi dalam data asal berbanding dengan PLS-DA dan sangat sesuai untuk pengurangan dimensi dan visualisasi kumpulan data yang kompleks. Selanjutnya, pelanjutan kernel LPPLS-DA dan NPPLS-DA dalam Ruang Hilbert (RKHS) kernel penghasilan semula dicadangkan untuk menangani situasi di mana wujud hubungan tak linear yang kuat di antara set-set data yang dicerap. Peningkatan prestasi algoritma yang dicadangkan berbanding dengan PLS-DA konvensional ditunjukkan melalui beberapa eksperimen. Ia telah menunjukkan bahawa LPPLS-DA dan NPPLS-DA sangat berkesan untuk analisis wajah (pengecaman dan perwakilan). Pelanjutan kernel kaedah-kaedah ini digunakan untuk tumor dan analisis data kimia dan ditunjukkan bahawa model dengan pelanjutan kernel mengatasi model linear apabila wujud hubungan tak linear kuat antara set data yang dicerap.

GLOBAL-LOCAL PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS AND ITS EXTENSION IN REPRODUCING KERNEL HILBERT SPACE

ABSTRACT

Subspace learning is an essential approach for learning a low dimensional representation of a high dimensional space. When data samples are represented as points in a high dimensional space, learning with the high dimensionality becomes challenging as the effectiveness and efficiency of the learning algorithms drops significantly as the dimensionality increases. Thus, subspace learning techniques are employed to reduce the dimensionality of the data prior to employing other learning algorithms. Recently, there has been a lot of interest in subspace learning techniques that are based on the global and local structure preserving (GLSP) framework. The main idea of the GLSP approach is to find a transformation of the high dimensional data into a lower dimensional subspace, where both the global and local structure information of the data are preserved in the lower dimensional subspace. This thesis consider the case where data is sampled from an underlying manifold embedded in a high dimensional ambient space. Two novel subspace learning algorithms called locality preserving partial least squares discriminant analysis (LPPLS-DA) and neighborhood preserving partial least squares discriminant analysis (NPPLS-DA) which are based on the GLSP framework are proposed for discriminant subspace learning. Unlike the conventional partial least squares discriminant analysis (PLS-DA) which aims at preserving only the global Euclidean structure of the data space, the proposed LPPLS-DA and NPPLS-DA algorithms find an embedding that preserves both the global and local manifold structure. As a result, both LPPLS-DA and NPPLS-DA can extract more discriminant information in the original data than PLS-DA and are well-suited for dimensionality reduction

and visualization of complex datasets. Furthermore, kernel extensions of LPPLS-DA and NPPLS-DA in reproducing kernel Hilbert space (RKHS) are proposed to handle situations where a strong nonlinear relation exist between the sets of observed data. Performance improvement of the proposed algorithms over the conventional PLS-DA is demonstrated through several experiments. It is shown that LPPPLS-DA and NPPLS-DA are very effective for face analysis (recognition and representation). Their kernel extensions are applied to tumor classification and chemical data analysis respectively and it was shown that the kernel extensions outperform their linear counterparts when a strong nonlinear relationship exist between the set of observed data.

CHAPTER 1

INTRODUCTION

1.1 Overview

In many research fields such as machine learning, computer vision, bioinformatics and pattern recognition, data points are represented as points in a high dimensional space. Researchers in such areas encounter difficulties working with such high dimensional data sets. The effectiveness and efficiency of many learning algorithms, such as clustering and classification algorithms, drop rapidly as the dimensionality increases (Guo and Dyer, 2005; Souza et al., 2016). A lot of techniques have been proposed in the past to reduce the dimensionality of the data by either selecting the most representative features from the original ones (feature selection) or by creating new features as linear combinations of the original features (feature extraction). These techniques include principal component analysis (PCA) (Yi et al., 2017; Zhao et al., 2019b), partial least squares (PLS) (Boulesteix and Strimmer, 2007; Rosipal and Krämer, 2005) and linear discriminant analysis (LDA) (Belhumeur et al., 1997). PCA is an unsupervised dimension reduction technique which captures most of the variance of a data, while LDA is a supervised dimension reduction technique which aim at discriminating the different classes in a data.

PLS is a statistical method that models the linear relationship between sets of observed variables X and Y by means of latent variables (components). The method was first developed by Herman Wold (Wold, 1966) and since then it gained wide acceptance in fields such as chemometrics, bioinformatics, social sciences, medicine etc.

The ability of PLS in handling high dimensionality and collinearity problems in spectral data makes it a powerful and standard tool for the analysis of chemical data in chemometrics (Aliakbarzadeh et al., 2016; Bai et al., 2017; Borràs et al., 2014; Hobro et al., 2010; Kemsley, 1996). Although PLS was not designed for discrimination and classification tasks, it has been successfully applied to these problems and its performance was outstanding (Barker and Rayens, 2003; Huang et al., 2005). PLS for discrimination or better known as partial least squares discriminant analysis (PLS-DA) was shown to have a statistical relationship with LDA and it was further suggested that PLS should be used instead of PCA when discrimination is the goal and dimension reduction is needed (Barker and Rayens, 2003). PLS-DA combines feature extraction and discriminant analysis into one algorithm and is well applicable for high dimensional data sets. Theoretically, PLS-DA finds a transformation of the high dimensional data into a lower dimensional subspace in which data samples of different classes are mapped far apart. The transformation is readily computed using the nonlinear iterative partial least squares (NIPALS) algorithm (Wold, 1966).

1.2 Problem Statement

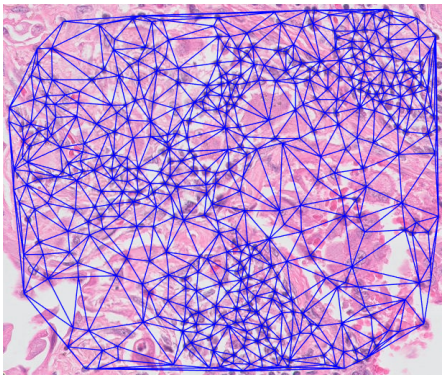
Although PLS-DA provide a principle way of dealing with high dimensional data, the method does not automatically lead to extraction of relevant features. Many studies (Brereton and Lloyd, 2014; Goodhue et al., 2012; Gromski et al., 2015; Mendez et al., 2020) have pointed out that when classification is the goal and dimension reduction is needed, PLS-DA should not be preferred over other traditional methods as it has no significant advantages over them. Some recent studies (Brereton and Lloyd, 2014; Lee et al., 2018b; Pomerantsev and Rodionova, 2018) also indicate the need to refine the

PLS-DA modeling practice strategies, especially in complex data sets such as multi-class, colossal and imbalanced data sets.

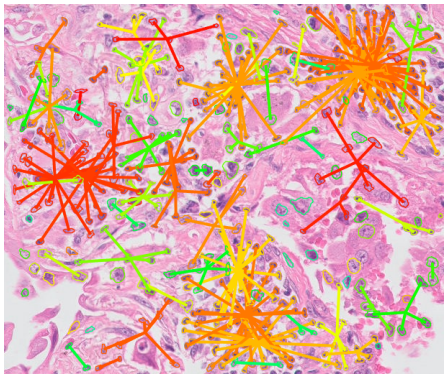
Another major drawback of the PLS-DA method is its lack of ability to preserve the local structure of data. PLS-DA sees only the global Euclidean structure of data. It fails to preserve the local structure of data point if the data points lie on a nonlinear manifold hidden in the high dimensional Euclidean space. Fortunately, several techniques that can effectively preserve the local structure of data point have been proposed (He et al., 2005a; He and Niyogi, 2004; Roweis and Saul, 2000; Shikkenawis and Mitra, 2016). However, when treating multi-clustered data as in the case of appearance-based face recognition and cancer classification, there is a need to treat simultaneously, both the global clustering structure as well as the local clustering structure (Cai, 2017; Liu et al., 2013).

In digital pathology image analysis, spatial arrangement of nuclei in histopathological images has been shown to be able to predict patient outcomes (Lu et al., 2021; Nguyen et al., 2014; Zhou et al., 2019). Cell graphs have been proposed to model the relationship between different cell nuclei and the tissue micro-environment using graph features. The graph can be constructed via global approaches such as Voronoi or Delaunay triangulation methods (Basavanhally et al., 2009) or via local approach such as the FLock method (Lu et al., 2021). After cell graph construction, features related to edge length and node density are then extracted to predict disease outcome. Figure 1.1 show graphs constructed using both the global and local approaches.The PLS-DA method is design to capture only the global information uncover from the global graph approaches, important information involving local spatial interaction may be left un-

(a) Voronoi



(b) Delaunay



(c) FLock

Figure 1.1: Cell nuclei graphs

exploited. Since both the global and local features are useful in the context of cancer grading, a method that capture both the global and local information is highly desirable. Also, capturing both the global and local information will lead to a better approach for modeling the tissue micro-environment.

1.3 Motivation of Study

Since global and local structures are both important for many pattern classification problems, a method that can reduce the dimensionality of a data set while preserving both of its global and local structures is thus highly desirable. The most current trends in feature extraction methods in general are seeing more and more approaches embracing the global and local structure preserving (GLSP) framework and are reporting very

promising results (Abeo et al., 2019; Lee, 2018; Song and Shi, 2018; Wan et al., 2018; Yao et al., 2018; Zhao et al., 2019a; Zhao and Jia, 2018). Their approaches are similar which involves developing an objective function that includes both the global and local structure features of a data set. The optimal direction for projection is thus the optimal solution of the objective. The resulting optimization problem is also equivalent to a generalized eigenvalue problem and can be solved using existing high performance computational methods for eigenvalue problems.

1.4 Research Objectives

The overall aim of this study is to develop an effective subspace learning (feature extraction) technique. Specifically, this study aims to enhance the overall performance of PLS-DA in feature extraction for complex high dimensional datasets. The objectives of this study include:

1. To show the effectiveness of global local PLS-DA method in discrimination and classification of face and chemical datasets.

2. To propose variants of the PLS-DA method which are more effective for discrimination and classification problems.

3. To propose new nonlinear subspace learning techniques based on the improved techniques to handle situations where strong nonlinear relationship exist between sets of observed data.

4. To show the effectiveness of the nonlinear extensions of the global local PLS-DA method in discrimination and classification of chemical and gene expression datasets.

1.5 Research Contribution

In an effort to make PLS-DA more effective for feature extraction and discrimination, we propose modifications to PLS-DA termed locality preserving PLS-DA (LPPLS-DA) and neighborhood preserving PLS-DA (NPPLS-DA), which are based on the GLSP framework. The local geometric structure of data is integrated into PLS-DA, which was almost never considered in the literature. We employed two different criteria to model the local geometric structure and obtain two feature extraction algorithms. Two efficient algorithms are developed to solve the resulting optimization problems, and their computational complexities are carefully discussed. The new algorithms are interesting in a number of perspectives.

1. The proposed LPPLS-DA and NPPLS-DA algorithms are fundamentally based on discriminant and spectral graph analysis. They are designed to solve different criteria from the PLS-DA method.

2. LPPLS-DA shares some similar properties with NPPLS-DA. Both techniques aim to discover both the global and the local geometric structure of data. However, their objective functions are entirely different.

3. Both LPPLS-DA and NPPLS-DA construct graphs over labeled data points to uncover the intrinsic discriminant structure in the data.

4. Both LPPLS-DA and NPPLS-DA are linear techniques which makes them suitable for practical applications. They may be conducted in the original space or in the reproducing kernel Hilbert space (RKHS) into which data points are mapped. This approach gives rise to nonlinear variants of LPPLS-DA and NPPLS-DA

called kernel LPPLS-DA (KLPPLS-DA) and kernel NPPLS-DA (KNPPLS-DA) respectively.

1.6 Scope of Study

This study focuses on improving the performance of PLS-DA in dealing with complex high dimensional datasets. Several approaches are proposed to refine the PLS-DA modeling practice strategies, especially in complex datasets such as multi-class datasets, imbalanced datasets, highly nonlinear datasets and manifold learning. All of the proposed approaches need to construct a graph to capture the local geometric structure of the data. Since, PLS-DA is already a supervised dimension reduction technique, we used the knowledge of the class labels while constructing the graph in the newly proposed variants of PLS-DA. This way, the discriminating ability of the PLS-DA technique can be enhanced to a higher extent.

The newly proposed approaches are applied to a large number of complex high dimensional datasets including face datasets, chemical datasets and biomedical datasets. Note that, all the datasets used in this study are publicly available as benchmark datasets for evaluating the performance of newly proposed machine learning techniques.

1.7 Thesis Organization

The organization of this thesis is as follows. Chapter 2 provides a detailed review of some of the most popular subspace learning techniques. Specifically, we give a detailed review of classical and manifold based subspace learning techniques. Chapter 3 intro-

duces the newly proposed locality preserving PLS-DA (LPPLS-DA) algorithm, plus a detailed computational analysis of LPPLS-DA. Extensive experimental results were also carried out on face databases to demonstrate the effectiveness of the LPPLS-DA method. The LPPLS-DA algorithm is also applied to complex chemical datasets and the experimental results are presented in Chapter 4. In chapter 5, a nonlinear extension of the LPPLS-DA algorithm in reproducing kernel Hilbert space (RKHS) is introduced to handle situations in which data are highly nonlinear. A detailed computational analysis of the nonlinear version of LPPLS-DA as well as extensive experimental results on gene expression datasets are also presented in Chapter 5. In Chapter 6, we introduce two new algorithms called neighborhood preserving PLS-DA (NPPLS-DA) and uncorrelated neighborhood preserving PLS-DA (UNPPLS-DA) for discriminant feature extraction. The extensive experimental results on face databases are also presented in Chapter 6. The kernel extensions of NPPLS-DA and UNPPLS-DA in reproducing kernel Hilbert space (RKHS), and the computational analysis of the algorithms are presented in Chapter 7. The extensive experimental results on some spectra datasets are also presented in Chapter 7. Finally, we provide some concluding remarks and suggestions for future research direction in Chapter 8.

CHAPTER 2

SUBSPACE LEARNING

Subspace learning is a framework applicable in research areas such as machine learning and pattern recognition where data samples are represented as points in high-dimensional spaces. Learning in high dimensional spaces becomes challenging because the performance of learning algorithms drops drastically as the number of dimensions increases. This phenomenon is known as *"curse of dimensionality"*. Thus, subspace learning techniques are first employed to reduce the dimensionality of the data before other learning techniques are applied. Subspace learning techniques consist of classical dimensionality reduction techniques and manifold learning techniques (Li and Allinson, 2009). In this chapter, a detailed review of some of the most popular subspace learning techniques is provided.

2.1 Classical Dimensionality Reduction Techniques

With the recent advances in computer technologies, there has been an explosion in the amount of data generated, stored and analyze. Most of this data are high dimensional in nature, ranging from several hundreds to thousands. Clustering or classification of such high dimensional datasets is almost infeasible. Thus, classical dimensionality reduction algorithms are used to map the high dimensional data into a lower dimensional subspace prior to the application of the conventional clustering or classification algorithms. The most popular algorithms for this purpose are principal component analysis (PCA) (Bartenhagen et al., 2010; Ma and Dai, 2011; Ma and Kosorok, 2009;

Yeung and Ruzzo, 2001), linear discriminant analysis (LDA) (Brereton, 2009; Cai et al., 2007; Hastie et al., 1995) and partial least squares discriminant analysis (PLS-DA) (Boulesteix and Strimmer, 2007; Nguyen and Rocke, 2002a,0; Pérez-Enciso and Tenenhaus, 2003; Tan et al., 2004). PCA is an unsupervised dimension reduction algorithm which aims at maximizing the variance of the new representations in the lower dimensional subspace. While LDA and PLS-DA are supervised dimensionality reduction algorithms which attempts to maximize between class covariance in the projected space. In addition to maximizing the between class covariance, LDA also attempts to minimize within class covariance in the projected space. These algorithms have proved successful for dimension reduction in many fields of research (Bahreini et al., 2019; Lee et al., 2018a; Li et al., 2020; Mas et al., 2020; Sitnikova et al., 2020; Xie et al., 2019; Yang et al., 2018; Zhao et al., 2018). The classical PCA, LDA and PLS-DA algorithms are linear dimensionality reduction algorithms and their performance can be restrictive when handling highly nonlinear datasets (Cao et al., 2011; Mika et al., 1998). To overcome this limitation, nonlinear extensions of PCA (Bartenhagen et al., 2010; Liu et al., 2005; Schölkopf et al., 1997), LDA (Baudat and Anouar, 2000; Cai et al., 2011) and PLS-DA (Song et al., 2018; Srinivasan et al., 2013; Štruc and Pavešić, 2009) through "kernel trick" have been proposed. The main idea of the kernel based techniques is to map the data into a feature space using a nonlinear mapping function. For a properly chosen nonlinear mapping function, an inner product can be defined in the feature space by a kernel function without defining the nonlinear mapping explicitly. The nonlinear extensions of the PCA and PLS-DA algorithms usually outperforms the linear PCA and PLS-DA algorithms when the data is highly nonlinear. In what follows, we give a brief review of the classical PCA, LDA and PLS-DA algorithms.

2.1.1 Principal Component Analysis

PCA is a well-known feature extraction technique in machine learning and pattern recognition. PCA seeks directions on which the data points are distributed with maximum variance. Given a set of n data points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \in R^m$. Let the vector $\boldsymbol{z} = \{z_1,\ldots,z_n\}$ represent the m-dimensional data points such that $z_i = \boldsymbol{w}^T\boldsymbol{x}_i$ is the one-dimensional map (representation) of $\boldsymbol{x}_i$ $(i = 1,2,\ldots,n)$, and $\boldsymbol{w} \in R^m$ denotes the transformation vector. The variance of the data points in the one-dimensional space can be calculated as follows (Jolliffe, 1986):

$$\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-\bar{z}_i)^2 \tag{2.1}$$

where $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Equation (2.1) can be reduced to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-\frac{1}{n}\sum_{i=1}^{n}z_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}^T\boldsymbol{x}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-\boldsymbol{w}^T\bar{\boldsymbol{x}})^2 \\
&= \boldsymbol{w}^T\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right)w \\
&= \boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}
\end{aligned} \tag{2.2}$$

where $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$ denotes the variance of the data points in the original space. Consequently, the objective function of PCA is given as follows:

$$\max_{\boldsymbol{w}^T\boldsymbol{w}=1} \boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w} \tag{2.3}$$

Using Lagrange multiplier method, the objective function (2.3) can be converted into an eigenproblem. Let

$$L(\boldsymbol{w},\lambda) = \boldsymbol{w}^T \boldsymbol{S}\boldsymbol{w} - \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1) \tag{2.4}$$

where λ is the Lagrange multiplier. Differentiating $L(\boldsymbol{w},\lambda)$ and setting the result to zero, one can get

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}} = \boldsymbol{S}\boldsymbol{w} - \lambda\boldsymbol{w} = 0 \tag{2.5}$$

or,

$$\boldsymbol{S}\boldsymbol{w} = \lambda\boldsymbol{w} \tag{2.6}$$

where $\boldsymbol{w}$ is the eigenvector of $\boldsymbol{S}$ and λ is the corresponding eigenvalue.

2.1.2 Linear Discriminant Analysis

LDA aims at finding a lower dimensional subspace in which data samples from the same class remain close to each other while data samples from different class are mapped far apart. Given a set of n data points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \in R^m$ belonging to C different classes. The LDA method solve the following objective function (Cai et al., 2007):

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}, \tag{2.7}$$

$$\boldsymbol{S}_b = \sum_{c=1}^{C} n_c(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu})^T, \tag{2.8}$$

$$\boldsymbol{S}_w = \sum_{c=1}^{C} \sum_{j=1}^{n_c} (\boldsymbol{x}_j^{(c)} - \boldsymbol{\mu}^{(c)})(\boldsymbol{x}_j^{(c)} - \boldsymbol{\mu}^{(c)})^T, \tag{2.9}$$

where $\boldsymbol{\mu}$ denotes the global centroid, $\boldsymbol{\mu}^{(c)}$ denotes the centroid of the cth class, n_c denotes the size of data samples in the cth class and $\boldsymbol{x}_j^{(c)}$ denotes the jth sample in the cth class. The matrices $\boldsymbol{S}_b \in R^{m\times m}$ and $\boldsymbol{S}_w \in R^{m\times m}$ are called the between-class and the within-class matrices, respectively. Let

$$L(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}} \tag{2.10}$$

To optimize the objective function (2.7), LDA requires the derivative of $L(\boldsymbol{w})$ and sets it to zero (Fukunaga, 2013):

$$\begin{aligned} \frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w})(2\boldsymbol{S}_b \boldsymbol{w}) - (\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w})(2\boldsymbol{S}_w \boldsymbol{w})}{(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w})^2} \\ &= \frac{\boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}} - \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w})^2} \boldsymbol{S}_w \boldsymbol{w} \\ &= 0 \end{aligned} \tag{2.11}$$

Equation (2.11) can be reduced to

$$\boldsymbol{S}_b \boldsymbol{w} = \lambda \boldsymbol{S}_w \boldsymbol{w}. \tag{2.12}$$

where $\lambda = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}$. Thus, the optimal $\boldsymbol{w}$ can be computed as the generalized eigenvector of $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ corresponding to the eigenvalue λ.

2.1.3 Partial Least Squares

PLS is a well known method for modeling the linear relationship between two sets of observed variables. The method is widely used as a feature extraction method to deal with undersampled and multi-collinearity issues usually encountered in high dimensional data (Ahmad et al., 2006; Jia et al., 2016). Given two set of observed variables $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in R^{n \times m}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^T \in R^{n \times N}$. PLS decomposes the zero mean data matrices $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$ into the following form (Rosipal and Krämer, 2005):

$$\begin{aligned} \bar{\boldsymbol{X}} &= \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \\ \bar{\boldsymbol{Y}} &= \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F} \end{aligned} \tag{2.13}$$

where the $n \times d$ matrices $\boldsymbol{T}$ and $\boldsymbol{U}$ represents the score matrices of the d extracted components, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the $m \times d$ and $N \times d$ loading matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, and the $n \times m$ matrix $\boldsymbol{E}$ and the $n \times N$ matrix $\boldsymbol{F}$ correspond to residual matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. The PLS method is based on the nonlinear iterative partial least squares (NIPALS) algorithm (Geladi and Kowalski, 1986; Wold, 1975) which find weight vectors $\boldsymbol{w}$ and $\boldsymbol{c}$ such that

$$[cov(\boldsymbol{t}, \boldsymbol{u})]^2 = \max_{\boldsymbol{w}^T\boldsymbol{w}=\boldsymbol{c}^T\boldsymbol{c}=1} [cov(\bar{\boldsymbol{X}}\boldsymbol{w}, \bar{\boldsymbol{Y}}\boldsymbol{c})]^2 \tag{2.14}$$

where $cov()$ denotes the sample covariance between variables. The outline of the NIPALS algorithm can be summarized as follows:

Step 1: Randomly initializes $\boldsymbol{u}$, usually $\boldsymbol{u}$ is set to be one of the columns of $\bar{\boldsymbol{Y}}$

Step 2: Compute the $\bar{\boldsymbol{X}}$ weights $\boldsymbol{w}$:

$$\boldsymbol{w} = \bar{\boldsymbol{X}}^T \boldsymbol{u} / \boldsymbol{u}^T \boldsymbol{u} \tag{2.15}$$

$\boldsymbol{w}$ can be normalized, i.e. $\|\boldsymbol{w}_1\| = 1$.

Step 3: Compute the $\bar{\boldsymbol{X}}$ scores $\boldsymbol{t}$:

$$\boldsymbol{t} = \bar{\boldsymbol{X}} \boldsymbol{w} \tag{2.16}$$

Step 4: Compute the $\bar{\boldsymbol{Y}}$ weights $\boldsymbol{c}$:

$$\boldsymbol{c} = \bar{\boldsymbol{Y}}^T \boldsymbol{t} / \boldsymbol{t}^T \boldsymbol{t} \tag{2.17}$$

Step 5: Compute an updated set of $\bar{\boldsymbol{Y}}$ scores $\boldsymbol{u}$:

$$\boldsymbol{u} = \bar{\boldsymbol{Y}} \boldsymbol{c} \tag{2.18}$$

Step 6: Test for convergence on the change in $\boldsymbol{t}$, if $\boldsymbol{t}$ converges proceeds to the next step, else return to Step 2.

Step 7: Deflate the data matrices $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$:

$$\begin{aligned} \bar{\boldsymbol{X}} &= \bar{\boldsymbol{X}} - \boldsymbol{t}(\boldsymbol{t}^T \bar{\boldsymbol{X}}) / \boldsymbol{t}^T \boldsymbol{t} \\ \bar{\boldsymbol{Y}} &= \bar{\boldsymbol{Y}} - \boldsymbol{t}(\boldsymbol{t}^T \bar{\boldsymbol{Y}}) / \boldsymbol{t}^T \boldsymbol{t} \end{aligned} \tag{2.19}$$

The PLS method has been used in discrimination problems (i.e., separating distinct data samples) and classification problems (i.e., assigning new data samples to predefined groups) (Bai et al., 2017; Nguyen and Rocke, 2002b). In this case, the input data

matrix $\boldsymbol{Y}$ is replaced by a dummy (class membership) matrix containing class information and the procedure is called partial least squares discriminant analysis (PLS-DA). The objective function of PLS-DA is as follows:

$$\boldsymbol{w}^* = \max_{\boldsymbol{w}^T\boldsymbol{w}=1} [cov(\bar{\boldsymbol{X}}\boldsymbol{w}, \boldsymbol{Y})]^2 \tag{2.20}$$

where $\boldsymbol{Y}$ denotes the class membership matrix defined as:

$$\boldsymbol{Y} = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \dots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \dots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_C} & 0_{n_C} & \dots & 1_{n_C} \end{pmatrix} \tag{2.21}$$

where n_i (for $i = 1, 2, \dots, C$) represents the number of samples in the i-th class, $\sum_{i=1}^{C} n_i = n$ (total number of samples), and 0_{n_i} and 1_{n_i} are $n_i \times 1$ vectors of zeros and ones respectively. For example, if the data set contains two classes, then the matrix $\boldsymbol{Y}$ is designed as a single-column vector with entries of 1 for all samples in the first class and 0 for

samples in the second class, i.e.

$$\boldsymbol{Y} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

Further, if the data have three classes, then the $\boldsymbol{Y}$ matrix is encoded with three columns as follows,

$$\boldsymbol{Y} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix},$$

One may choose to centre the class membership matrix $\boldsymbol{Y}$ to have zero mean. Since

$$\begin{aligned} [cov(\bar{\boldsymbol{X}}\boldsymbol{w}, \boldsymbol{Y})]^2 &= \frac{1}{(n-1)^2} (\boldsymbol{Y}^T \bar{\boldsymbol{X}} \boldsymbol{w})^T (\boldsymbol{Y}^T \bar{\boldsymbol{X}} \boldsymbol{w}) \\ &= \frac{1}{(n-1)^2} \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{Y}^T \boldsymbol{X} \boldsymbol{w}, \end{aligned} \tag{2.22}$$

the objective function (2.20) can be rewritten in the following equivalent form:

$$\max_{\boldsymbol{w}^T \boldsymbol{w}=1} \boldsymbol{w}^T \bar{\boldsymbol{X}}^T \boldsymbol{Y} \boldsymbol{Y}^T \bar{\boldsymbol{X}} \boldsymbol{w} \tag{2.23}$$

Using Lagrange multiplier method, the objective function (2.23) can be reduced to an eigenproblem of the form:

$$\bar{\boldsymbol{X}}^T \boldsymbol{Y}\boldsymbol{Y}^T \bar{\boldsymbol{X}} \boldsymbol{w} = \lambda \boldsymbol{w} \tag{2.24}$$

Thus, the optimal weight (projection) vector $\boldsymbol{w}$ in (2.23) can be obtained as the eigenvector of $\bar{\boldsymbol{X}}^T \boldsymbol{Y}\boldsymbol{Y}^T \bar{\boldsymbol{X}}$ corresponding to the eigenvalue λ in (2.24). It was shown that the eigenstructure (2.24) is basically that of a slightly altered version of the between class scatter matrix in LDA (Aminu and Ahmad, 2019; Barker and Rayens, 2003). Therefore, what PLS-DA does is basically maximizing between class separation.

2.2 Manifold Learning Techniques

The classical PCA, LDA and PLS-DA methods aim at preserving the global Euclidean structure of the data space; if data points happen to reside on nonlinear submanifold embedded in the high dimensional ambient space, this can pose a problem for these methods. Thus, several manifold based dimension reduction algorithms have been proposed to discover the local manifold structure. These algorithms include Laplacian eigenmaps (LE) (Belkin and Niyogi, 2002,0), locally linear embedding (LLE) (Roweis and Saul, 2000), neighborhood preserving embedding (NPE) (He et al., 2005a) and locality preserving projections (LPP) (He and Niyogi, 2004). These methods are designed to determine a subspace where local structure of data points are well preserved, but how well they are able to capture the global structure of a dataset is still not well understood. In what follows, we give a brief review of the LE, LLE, LPP and NPE algorithms.

2.2.1 Laplacian Eigenmap

Laplacian eigenmaps (LE) (Belkin and Niyogi, 2002) is a local nonlinear subspace learning technique based on spectral graph theory. The method attempts to preserve the local geometrical structure of data after dimension reduction. Specifically, LE seek an embedding such that nearby points on the manifold are mapped close to each other in the low dimensional subspace. Suppose $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ denotes the set of n data points sampled from an underlying manifold $\mathscr{M}$ embedded in a high dimensional ambient space R^m. LE first construct a graph G with the data points $\boldsymbol{x}_i$ $(i=1,\ldots,n)$ as nodes and a weight matrix $\boldsymbol{S}$ that assign weights to the edges between the nodes. The weight matrix $\boldsymbol{S}$ can be computed as follows:

$$S_{ij}=\begin{cases}1; & \text{if } \boldsymbol{x}_i\in N_p(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j\in N_p(\boldsymbol{x}_i)\\ 0; & \text{otherwise.}\end{cases} \tag{2.25}$$

where $N_p(\boldsymbol{x}_i)$ denotes the set of p nearest neighbors of $\boldsymbol{x}_i$. Let $\boldsymbol{z}=(z_1,\ldots,z_n)^T$ denotes the map determined by LE where z_i is the one-dimensional map of $\boldsymbol{x}_i$ $(i=1,\ldots,n)$. LE realize the optimal map by solving the following minimization problem:

$$\min_{\boldsymbol{z}}\sum_{i,j=1}^{n}(z_i-z_j)^2S_{ij} \tag{2.26}$$

The minimization problem (2.26) can be reduced to

$$\begin{aligned}
\min_{\boldsymbol{z}} \sum_{i,j=1}^{n} (z_i - z_j)^2 S_{ij} &= \min_{\boldsymbol{z}} \sum_{i=1}^{n} (z_i^2 + z_j^2 - 2 z_i z_j) S_{ij} \\
&= \min_{\boldsymbol{z}} (\sum_{i=1}^{n} z_i^2 D_{ii} + \sum_{j=1}^{n} z_j^2 D_{jj} - 2 \sum_{ij=1}^{n} z_i z_j S_{ij}) \\
&= 2 \min_{\boldsymbol{z}} (\sum_{i=1}^{n} z_i^2 D_{ii} - \sum_{i=1}^{n} z_i^2 S_{ij}) \\
&= 2 \min_{\boldsymbol{z}} (\boldsymbol{z}^T \boldsymbol{D} \boldsymbol{z} - \boldsymbol{z}^T \boldsymbol{S} \boldsymbol{z}) \\
&= 2 \min_{\boldsymbol{z}} \boldsymbol{z}^T (\boldsymbol{D} - \boldsymbol{S}) \boldsymbol{z} \\
&= \min_{\boldsymbol{z}} 2 \boldsymbol{z}^T \boldsymbol{L} \boldsymbol{z} \qquad (2.27)
\end{aligned}$$

where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$ is the graph Laplacian (Chung and Graham, 1997) and $\boldsymbol{D}$ is a diagonal matrix whose entries are column (or row) sum of $\boldsymbol{S}$, $D_{ii} = \sum_j S_{ij}$. In order to remove an arbitrary scaling factor in the embedding, the following constraint is impose:

$$\boldsymbol{z}^T \boldsymbol{D} \boldsymbol{z} = 1$$

Finally, the minimization problem (2.26) reduces to

$$\arg \min_{\boldsymbol{z}^T \boldsymbol{D} \boldsymbol{z} = 1} \boldsymbol{z}^T \boldsymbol{L} \boldsymbol{z} \qquad (2.28)$$

Minimizing the objective function (2.28) is an attempt to ensure that if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close on the manifold, then their low dimensional representations z_i and z_j are close as well.

Solving the minimization problem (2.28) is equivalent to finding the eigenvector

corresponding to the smallest eigenvalue of the following generalized eigen-problem

$$\boldsymbol{L}\boldsymbol{z} = \lambda \boldsymbol{D}\boldsymbol{z} \tag{2.29}$$

Equivalently, the embedding $\boldsymbol{z}$ can be obtained as the eigenvector corresponding to the largest eigenvalue of the following generalized eigen-problem

$$\boldsymbol{S}\boldsymbol{z} = \lambda \boldsymbol{D}\boldsymbol{z} \tag{2.30}$$

2.2.2 Locally Linear Embedding

Locally linear embedding (LLE) (Roweis and Saul, 2000) is another nonlinear subspace learning technique which is also based on spectral graph theory. The main idea in LLE is that each data point resides on a local linear patch of a manifold and can be reconstructed by a linear combination of its nearest neighbors. Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the set of n data points sampled from an underlying manifold $\mathscr{M}$ embedded in a high dimensional ambient space R^m. The reconstruction errors are then measured by the cost function:

$$\varepsilon(\boldsymbol{S}) = \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \sum_{j=1}^{n} S_{ij}\boldsymbol{x}_j \right\|^2 . \tag{2.31}$$

The weights S_{ij} characterized the contribution of the data point $\boldsymbol{x}_j$ to the reconstruction of $\boldsymbol{x}_i$. To compute the weights S_{ij}, the cost function (2.31) is minimized subject to two constraints: 1) the rows of the weight matrix sum to one, i.e., $\sum_j S_{ij} = 1$, 2) each data point $\boldsymbol{x}_i$ is reconstructed only from its nearest neighbors, enforcing $S_{ij} = 0$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are not neighbors. Let $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ denotes the map of the original data points to a line where z_i represents $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$. The LLE algorithm determines a neighbor-

hood preserving mapping by minimizing the following embedding cost function:

$$\varphi(\boldsymbol{z}) = \sum_{i=1}^{n}(z_i - \sum_{j=1}^{n} S_{ij}z_j)^2. \tag{2.32}$$

Let

$$p_i = z_i - \sum_{j=1}^{n} S_{ij}z_j \qquad (i = 1,\ldots,n)$$

which can be written in vector form as

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{z} - \boldsymbol{S}\boldsymbol{z} \\ &= (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{z} \end{aligned}$$

The embedding cost function (2.32) can be reduced to

$$\begin{aligned} \varphi(\boldsymbol{z}) &= \sum_{i=1}^{n}(z_i - \sum_{j=1}^{n} S_{ij}z_j)^2. \\ &= \sum_{i=1}^{n}(p_i)^2 \\ &= \boldsymbol{p}^T\boldsymbol{p} \\ &= \boldsymbol{z}^T(\boldsymbol{I} - \boldsymbol{S})^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{z} \end{aligned} \tag{2.33}$$

Thus, the embedding $\boldsymbol{z}$ that minimizes the cost function (2.32) is given by the eigenvector corresponding to the smallest eigenvalue of the following eigen-problem:

$$(\boldsymbol{I} - \boldsymbol{S})^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{z} = \lambda\boldsymbol{z} \tag{2.34}$$

2.2.3 Locality Preserving Projections

Locality preserving projections (LPP) (He and Niyogi, 2004; He et al., 2005b) is basically a linear approximation of the nonlinear LE technique. Similar to the LE technique, LPP seek an embedding that preserves the local geometrical structure of the original data. Given a set of n data points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \in \mathscr{M}$ and $\mathscr{M}$ is a nonlinear manifold embedded in R^m, LPP models the local geometrical structure of the data by an adjacency graph G with the data points $\boldsymbol{x}_i$ $(i=1,\ldots,n)$ as nodes and a weight matrix $\boldsymbol{S}$ that assign weights to the edges between the nodes. The weight matrix $\boldsymbol{S}$ can be define as follows:

$$S_{ij} = \begin{cases} \exp\Big(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{t}\Big); & \text{if } \boldsymbol{x}_i \in N_p(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_p(\boldsymbol{x}_i) \\ 0; & \text{otherwise.} \end{cases} \tag{2.35}$$

where $N_p(\boldsymbol{x}_i)$ denotes the set of p nearest neighbors of $\boldsymbol{x}_i$. Let z_i denotes a one dimensional map of $\boldsymbol{x}_i$ $(i=1,\ldots,n)$. LPP minimizes the following objective function:

$$\sum_{i,j=1}^{n} (z_i - z_j)^2 S_{ij}. \tag{2.36}$$