THE UNIVERSITY OF BUCKINGHAM

Data-Independent Vs. Data-Dependent Dimension Reduction for Pattern Recognition in High Dimensional Spaces

By

Tahir Mohammed Hassan

Department of Applied Computing

The University of Buckingham

A thesis submitted for the Degree of MSc in Mathematics and Computation to the School of Science in the University of Buckingham

April 2017

Buckingham, United Kingdom

ABSTRACT

Data-Independent Vs. Data-Dependent Dimension Reduction for Pattern Recognition in High Dimensional Spaces

By Tahir Mohammed Hassan

There has been a rapid emergence of new pattern recognition/classification techniques in a variety of real world applications over the last few decades. In most of the pattern recognition/classification applications, the pattern of interest is modelled by a data vector/array of very high dimension. The main challenges in such applications are related to the efficiency of retrieval, analysis, and verifying/classifying the pattern/object of interest. The "Curse of Dimension" is a reference to these challenges and is commonly addressed by Dimension Reduction (DR) techniques. Several DR techniques has been developed and implemented in a variety of applications. The most common DR schemes are dependent on a dataset of "typical samples" (e.g. the Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA)). However, data-independent DR schemes (e.g. Discrete Wavelet Transform (DWT), and Random Projections (RP)) are becoming more desirable due to lack of density ratio of samples to dimension.

In this thesis, we critically review both types of techniques, and highlight advantages and disadvantages in terms of efficiency and impact on recognition accuracy. We shall study the theoretical justification for the existence of DR transforms that preserve, within tolerable error, distances between would be feature vectors modelling objects of interest. We observe that data-dependent DRs do not specifically attempts to preserve distances, and the problems of overfitting and biasness are consequences of low density ratio of samples to dimension.

Accordingly, the focus of our investigations is more on data-independent DR schemes and in particular on the different ways of generating RPs as an efficient DR tool. RPs suitable for pattern recognition applications are only restricted by a lower bound on the reduced dimension that depends on the tolerable error. Besides, the known RPs that are generated in accordance to some probability distributions, we investigate and test the performance of differently constructed over-complete Hadamard mxn (m<<n) submatrices, using the inductive Sylvester and Walsh-Paley methods. Our experimental work conducted for 2 case studies (Speech Emotion Recognition (SER)

and Gait-based Gender Classification (GBGC)) demonstrate that these matrices perform as well, if not better, than data-dependent DR schemes. Moreover, dictionaries obtained by sampling the top rows of Walsh Paley matrices outperform matrices constructed more randomly but this may be influenced by the type of biometric and/or recognition schemes. We shall, also propose the feature-block (FB) based DR as an innovative way to overcome the problem of low density ratio applications and demonstrate its success for the SER case study.

Dedicated to the memory of my father

MOHAMMED

who always believed in my ability to be successful in the academic arena. You are gone but your belief in me has made this journey possible.

ACKNOWLEDGEMENTS

Allah the most gracious and merciful: Who gave me energy, health and provided me with all the people to whom I am dedicating this hard work, which took a lot of time to finish.

First and foremost, I would like to express my sincere gratitude to my supervisor **Prof. Sabah Jassim** for the continuous support of my M.Sc. study and research, for his patience, motivation, enthusiasm, immense knowledge, and providing me with an excellent atmosphere for doing research. His guidance helped me in all the time of research and writing of this thesis.

I would like to thank **Dr. Abdulbasit Al-Talabani**, **Dr. Azhin Sabir**, and **Dr. Nadia Al-Hassan** (Visiting Postdoc at the university of Buckingham) for their advices, discussions, and fruitful collaborative work and I wish them all the best in the future.

All my love, gratitude, respect, and thanks go to my Mother **(Mena),** sisters, brothers, and friends. They were always supporting me and encouraging me with their best wishes.

I would like to express my sincere gratitude and appreciation to **The Higher Committee for Education Development –HCED** for sponsoring my MSc program of study.

ABBREVIATIONS

AGEnEI	Approximation sub-band Gait Entropy Energy Image
AVGEnEI	Approximation Vertical sub-band Gait Entropy Energy Image
C	Circulant Matrix
CB	City Block Distance
CN	Condition Number
CS	Compressive Sensing
CUR	CUR Decomposition
DIPCA	Data-Independent Principle Component Analysis
DR	Dimension Reduction
DS	Dimension Selection
DT	Dimension Transformation
DWT	Discrete Wavelet Transform
E	Euclidean Distance
FB	Feature Block
GBGC	Gait-Based Gender Classification
GEI	Gait Energy Image
GEnEI	Gait Entropy Energy Image
GEnI	Gait Entropy Image
GS	Gram-Schmidt Process
HH	Diagonal sub-band
HL	Vertical sub-band
JL	Johnson and Lindenstrauss Theorem
kNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LH	Horizontal sub-band
LL	Approximation sub-band
LLD	Low Level Descriptor
NSP	Null Space Property
PCA	Principal Component Analysis
RD	Random Over-complete Dictionary
RIC	Restricted Isometry Constant
RIP	Restricted Isometry Property

RP	Random Projection
SD	Structured Over-complete Dictionary
SER	Speech Emotion Recognition
SH	Sylvester-type Hadamard matrix
SMO	Sequential Minimal Optimization
SRD	Semi-Random Over-complete Dictionary
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TDA	Topological Data Analysis
TPCA	Topological Principal Component Analysis
VGEnEI	Vertical sub-band Gait Entropy Energy Image
W	Walsh Matrix
WP	Walsh-Paley Matrix

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DECLARATION

I hereby declare that my thesis entitled "***Data-Independent Vs. Data-Dependent Dimension Reduction for Pattern Recognition in High Dimensional Spaces***" is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text, and is not substantially the same as any that I have submitted, or, is concurrently submitted for a degree or diploma or other qualification at the University of Buckingham or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or is concurrently submitted for any such degree, diploma, or other qualification at the University of Buckingham or any other University or similar institution except as declared in the Preface and specified in the text.

Tahir Mohammed Hassan

CHAPTER ONE: INTRODUCTION

Advances in computational sciences and technologies over the last few decades have resulted in the emergence of new automatic pattern recognition/analysis techniques and systems in a variety of diverse applications/scenarios. These applications often involve a "large" dataset/database of records often representing multiple instances of a set of distinct objects/patterns. Each instance of the objects/patterns in many applications is modelled by a vector/array of a finite number of measurements/coordinates called the dimension of the vectors. In most interesting applications, the data vectors are of very high dimension. For example, in biometric systems the database may include records of Face images/videos, Fingerprints, Iris codes, handwritten text document, and/or speech recordings. The emerging field of Big Data analytics covers a variety of applications, including automatic medical diagnostic, that involve analysing large and complex types of data in order to discover known or hidden patterns and anomalies. The most common challenges that arise as a result of the high dimensionality of data for such applications relate to the efficiency of retrieving, analysing, and classifying the objects/patterns under investigation. Moreover, as the dimension of a dataset increases, the data points get further away from each other and some existing pattern of the points in a low dimensional space may disappear in high dimensional spaces. These challenges are often blamed on the so-called ***Curse of dimension***. Most common approaches to deal with this problem are based on reducing dimension using known samples of the objects of interest, but most such methods depend on the density of the samples within the modelling domain. This thesis is devoted to review and investigate the theoretical bases for dimension reduction techniques. We shall develop and test the performance of data/samples-independent dimension reduction schemes in two Pattern recognition applications.

In this introductory chapter, I shall primarily attempt to describe the research problem under investigation, our motivations, and the contributions made in this thesis.

1.1 High Dimensional Data and Curse of Dimension

A vector of dimension n is an array of a given type, and it is commonly used to mathematically model objects by incorporating their essential measurements/properties as coordinates. The use of such a model enables the use of computers to process, manipulate, and transform such objects using the wealth of knowledge inherent in the

algebraic structure of vector spaces $\mathbb{R}^n$ over the field $\mathbb{R}$ of real numbers (or other fields). A vector space is a set V of elements, called vectors, together with an operation of addition of vectors and an operation of multiplication satisfying a number of properties such as commutative and associative laws of vector addition, the distributive law of scalar multiplication, and the existence and uniqueness of zero vector and the negative of a vector. In this thesis, we only work with the n-dimensional vector space$\mathbb{R}^n$ whose elements are size n arrays of real numbers, and addition of two arrays is simply the usual addition of their corresponding entries while the scalar multiplication of an array $\underline{\boldsymbol{v}}$ by a scalar α is the product of α by every entry of $\underline{\boldsymbol{v}}$. It is essential to note that basic vector operations/functions defined on the $\mathbb{R}^2 and\ \mathbb{R}^3$ vector spaces, such as the Euclidean distance and the angle between two vectors, generalise naturally to high dimensional space but require proportionally more computational time. But this may not be true for more complex operations such as those commonly used in pattern analysis.

In many modern pattern recognition/analysis, classification and clustering applications, the dimension of vectors modelling the main objects of interest is tremendously high (hundreds, thousands or even in some applications are millions). High dimensional vector representation of objects presents several challenges to such modern computational problems. Among the well-known challenges one can list (1) the processing, analysis, and discovering discriminating features in such records; (2) facilitating building large databases of such objects allows very efficient or real time searching and retrieval; and (3) supporting essential datamining tasks.

Consideration of these challenges when computing and communication capabilities, at the early age of computing, was rather very modest by today's standards, the term (curse of dimension) became the common term to characterise the toughness of these challenges. Moreover, for most pattern recognition/classification applications, there is an added complexity associated with the fact that in such applications we may not have sufficiently large samples to be used for training purposes. Bellman (1961) who first coined the ***Curse of dimension*** term noted that "the sample size required to estimate a function of several variables grows exponentially with increasing number of variables" (Jolliffe, 2002). This means that, if we have more variables/dimensions, we need to have much more samples to fill the space. This comes from the fact that with increased dimension, the collected data samples get further away from each other and thus, the data sample density become very low. For instance, if S samples are enough to cover

$1D$ space with a good density, then we need S^2 samples to cover $2D$ space with same density, S^3 samples for $3D$ space, and so on.

The rapid increase of computing technology certainly led to the emergence of Big data applications and sophisticate machine learning schemes, but when the dimension gets truly high, analysis of the data becomes unstable/sensitive and greatly affects the efficiency of the applications. The low density of data samples of high dimensional vectors is one of the biggest issues for machine learning applications. In practice, more often we have a small number of samples compared to the number of dimensions of data sets. Moreover, for applications that use supervised learning, we need to divide the data set into two sets (training and testing) which makes the number of samples even much smaller compare to the number of dimensions. Simply, the curse of dimension problem makes analysis of high dimensional data oversensitive and intractable. In the case of supervised learning low sample density leads to overfitting and biasness. In general, there are a few ways to avoid curse of dimension, ***Dimension Reduction*** is one of them. This thesis is focused on reviewing, investigating and testing the performance of various dimension reduction techniques especially for low density scenarios.

1.2 Curse of Dimension – Face Recognition as an illustrating case.

A face recognition system uses some elaborate algorithm which on the input of a face image of a person, it will compare it with the records in a database of face images and returns/verifies the identity of the person only if the person has already been enrolled. The face database of the system, usually contains many face images of the people enrolled (or their digital representation). Each enrolled person may have several images which are taken under different recording conditions such as: light condition, face orientation angle, pose, emotion expression, etc. These images are either in greyscale or coloured, but for simplicity, we only consider greyscale with the same $m\times n$ resolution (number of pixels). Each image can be converted to an $N = (m\times n)$-dimensional vector by row or column concatenation. A popular face database that is regularly used for evaluating the performance of face recognition schemes, in this case, we use The Extended Yale B database which includes 2432 face images for 38 persons (Georghiades et al., 2000), figure 1-1, below displays 50 sample images of 5 people, each with 10 different images in this database. The size of each image is $m\times n = 192\times 168$. When converting each image by row concatenation, we get an $N = m\times n = 192\times 168 = 32256$ high dimensional vector of Bytes. For this database, the space

required to store the raw data for the images only is a matrix of size 2432×32256 Bytes while each row represents one image. Note that the above matrix representation of the database records is very convenient way to illustrate this case study but there are other ways to represent and process such databases. This database may not present a serious burden on storage, but for serious biometric systems where the database is expected to contain multiple images of tens of millions of persons, storage though a serious challenge it may not be the only or even the most difficult concerns. The working of the biometric system requires frequent searching through the database and retrieving images from the database, as well as implementing computationally expensive tasks of processing/analysing and classifying input fresh images. These some of the main challenges that are caused by the Curse of Dimension in this case study and many other applications.

Figure 1-1 Example of high dimensional data, 50 sample images of 5 people, each with 10 different images from The Extended Yale B database.

1.3 Approaches to mitigate the effect of Curse of Dimension

As we mentioned earlier, directly processing high dimensional data in many common applications may not be as easy as in low dimensional data spaces because their analysis is quite complicated and their efficiency may be beyond the available resources. It has long been recognised that appropriate dimension reduction transform becomes crucial to overcome various difficulties in relation to memory storage as well in conducting necessary analysis tasks. This is justified by the observation that the n-dimensional vectors representing the objects of interest in pattern analysis are unlikely to be scattered throughout the vast infinite space of $\mathbb{R}^n$. The 32256-dimensional vectors modelling face images in the previous example cannot be scattered densely throughout

the $\mathbb{R}^{32256}$ space no matter how many persons are enrolled on the face biometric database. In fact, in most pattern recognition/classification applications, the actual set of possible vectors in $\mathbb{R}^n$ modelling the objects of interest may be clustered within or very close to a subspace/manifold of much lower dimension. Figure 1-2 illustrates this concept, and determining the subspace of the low dimension may help reducing the effect of curse of dimension. In this figure, the points are almost close to a plane and therefore projecting the 3D points onto this plane reduces the dimension of the points from 3 to 2 and distances are reasonably preserved in the projected plane.

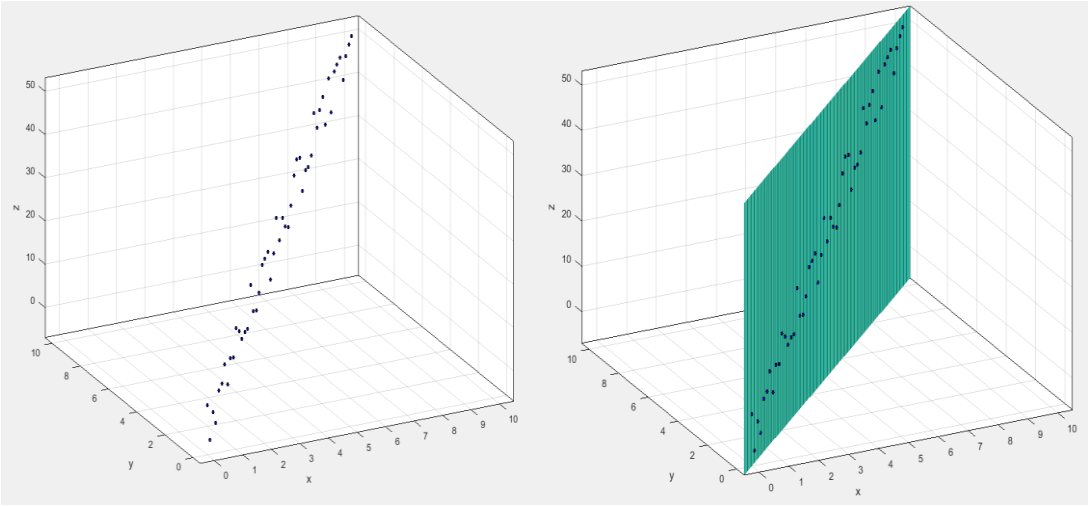


Figure 1-2 The concept of Dimension Reduction

In general, Dimension Reduction (DR) is the process of finding a lower dimensional data set X' that belongs to $\mathbb{R}^d$ from a high dimensional data set X belongs to $\mathbb{R}^N$ $(d \ll N)$ such that X' has nearly the same structure (approximation) of X and it retains almost all information that can be in X within a small relative error. In a mathematical term, this means that a DR is a projection of $\mathbb{R}^N$ onto a d-dimensional subspace of it so that the image X' of the set X in the projection subspace have similar geometric structure as X in the original vector space $\mathbb{R}^N$. Obviously, it is not possible always to get exactly the same structure of any data set after dimension reduction. However, good approximation is sufficient for some applications when a slight error in data accuracy is not a big issue and acceptable in practice. Moreover, with good DR projections means that X' can be processed more effectively than the original data X as the dimension of the reduced data is more manageable. Moreover, the density of the samples after projection will become higher with respect to the lower dimensional subspace, which is quite important for applications. Furthermore, the reduced version maintains all the essential information and it is a reliable approximation. In the

following chapters, we will explain that how DR techniques provide such approximation and how they preserve the sample dataset structure.

Suppose that A is a set of points that resides on a $d-dimensional$ space and it is embedded in an $N-dimensional$ space, $P: \mathbb{R}^N \to \mathbb{R}^d$. So, $\mathbb{R}^d$ is called ***intrinsic dimensional space*** for A and $\mathbb{R}^N$ is ***extrinsic dimensional space***. In general, any data set is easier to analyse in more efficient ways, and provides a better understanding in its intrinsic dimensional space rather than any other space. In the case of having a lower intrinsic dimension of a high dimensional data set ($d \ll N$), DR is quite meaningful and the goal of useful DR techniques is simply the process of moving from an extrinsic dimensional space to an intrinsic one. More precisely, Wang in (Wang, 2012) states "DR is a method of representing high dimensional data by their low-dimensional embedding", which means finding a linear transformation $P: \mathbb{R}^N \to \mathbb{R}^d$ which is usually called a projection. Theoretically, a high dimensional data set X can be projected onto different subspaces of different dimensions, however, the difficult challenge is to estimate the intrinsic dimension and find a subspace which is "best" fit to the dataset.

In any recognition/classification application, the coordinates of the vectors in $\mathbb{R}^N$ that model the objects of interest (e.g. face image), are referred to in the computing literature as features of the objects. In the case of a face image, these features are simply the pixel values in the image while facial features often refer to nose, eyes, mouth, eyebrow, or ear. The coordinates of the projected records in the lower dimension is usually referred to as ***meta-features***, each of which is a linear combination of the original features. In the literature, ***feature selection*** is a special simple type of dimension reduction, whereby a relatively small number of coordinates (i.e. features) in the n-dimensional vectors are retained and all other coordinates are replaced with zeros. This is useful, when there are evidences that many of the coordinates may have little or no relevance to the subject under consideration. In other words, the projection is done onto a *standard subspace* of $\mathbb{R}^N$ generated by a small subset of the *standard base* of the vector space $\mathbb{R}^N$. Note that, if $\underline{\mathbf{e}}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the unit vector in $\mathbb{R}^N$, whose i-th coordinate is 1 and all other coordinates are zeros, then the set $\{\underline{\mathbf{e}}_1, \ldots, \underline{\mathbf{e}}_i, \ldots, \underline{\mathbf{e}}_N\}$ is the standard basis for $\mathbb{R}^N$. In this thesis, we shall not use feature selection approach to dimension reduction, although we might use some simple types as a first step in some applications when the density is very low, see section (5.7).

Over the past few decades, mathematicians have investigated, and continue to do so, the curse of dimension problem for a variety of reasons and several DR techniques have been invented and developed many mathematical solutions for this problem and used these methods with a great deal of success in many applications. Commonly used DR techniques include Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Random Projections (RP).

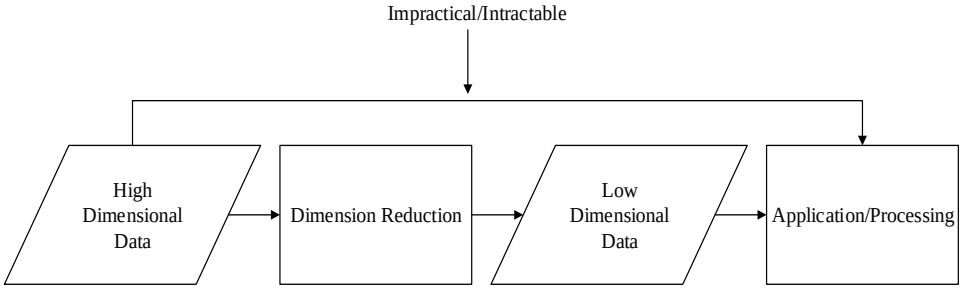


Figure 1-3 The goal of dimension reduction.

Due to issues like insufficient density of the original samples from a dataset or reliability of the model, uniqueness of output from DR is not guaranteed but for pattern recognition applications, it is essential that the distinguishing features of the pattern of interest are "preserved" with as little as possible loss in the lower dimension. This is a serious consideration through our investigations of this thesis.

1.4 Advantages of Dimension Reduction

The use of DR has been argued for in dealing with recognition/classification applications when the objects of interest are somewhat loosely modelled by very high dimensional vectors whose coordinates may involve a great deal of redundancies. The presence of redundancies in object representation reduces the discriminating power of the adopted high dimensional model.

(1) DR helps to overcome the challenge presented by the well-known problem (Curse of dimension) and provides a significant increase in data sample density, thus, it helps to boost performance of applications.

(2) Removes/reduces redundant and irrelevant features, and provides a compact representation of the objects of interest leading to efficient and more accurate object recognition/classification tasks.

(3) Ideal DR technique preserves data structure with a good probability. So, any solution of a data set after reduction is considered as reliable approximation solution of the original data.

(4) Another useful advantage of using DR is saving time and space/memory. Obviously, the dimension of a reduced data is significantly less than the original one especially in the case of Hard DR, so, less dimension means doing much less computation and the required memory to save the reduced data will be much smaller.

(5) The applicability of DR techniques is by no mean limited to very high dimensional models, and its use for relatively low dimensional cases enables a better visualization of datasets especially when the reduced dimension is ≤ 3.

However, these and other benefits, assumes the use of appropriate DR schemes that most likely be dependent on the domain of the application. In the rest of the thesis we shall attempt to review and investigate the various approaches to determine appropriate DR schemes.

1.5 Motivation

Advances in computer technology over the last few decades have led to the emergence of many successful algorithms to solve extremely difficult computational challenges in a variety of applications. Pattern recognition and classification is most widely researched applications and objects of interests are in most cases modelled by high dimension giving rise to curse of dimension. My ultimate motivation is to investigate mathematically inspired computations to deal with such challenges.

Having realised the benefits of using DR for dealing with the most common challenging pattern recognition and classification of objects modelled by very high dimensional arrays, I was therefore motivated was comprehend the mathematical justification for the existence of linear transformations that significantly reduce dimensions of the model without adversely influencing the performance of the recognition/classification scheme. Moreover, my initial study of the variety of applications revealed that in most cases we only have relatively small dataset of samples/instances of the objects of interest, which have added toughness to the curse of dimension challenge. This fact provided the motivation to search for dataset independent DR schemes. Having found that Random Projections provide a variety of data-independent DR techniques that have been shown to have the potential for success with a high probability (Johnson and Lindenstrauss, 1984).

1.6 Aims and Objectives of the thesis

(1) Investigate the mathematical theory of DR that is potentially applicable to pattern recognition/classification applications that adversely affected by curse of dimension.

(2) Review existing data-dependent DR schemes and investigate their implementation and advantages as well as limitations.

(3) Investigate the various Data independent DR schemes with focus on Random projections.

(4) Investigate the various approaches to generating Hadamard Random Projections as DR schemes and test the performance of these different approaches for 2 well-known pattern recognition/classification case studies.

The scope of the investigation in the last two chapters is influenced by the original experimental work conducted in these two-case studies: SER and GBGC. PCA was mainly used for DR rather than other schemes like LDA in the original work (Al-Talabani, 2015; Sabir, 2015). Therefore, we test the performance of Hadamard based RPs and compare their performance with PCA in both case studies.

1.7 Thesis Organization

- Chapter two provides relevant mathematical background for DR techniques with focusing on linear transformations and change of basis. It also presents the mathematical theory of existence of dimension reduction and the feasibility of such procedure, it also includes a general discussion on the classification of DR techniques.
- Chapter three is aimed at studying data-dependent DR techniques, it starts by reviewing the theory of eigenvalue problem followed by investigating several DR schemes that follow eigenvalue problem approach and its link to matrix factorization.
- Chapter four is aimed at studying data-independent DR techniques. It starts by studying wavelet-based DR approach. It also investigates different approaches to generating data-independent Hadamard based random projections for DR and the link to over-complete compressive sensing (CS) dictionaries.
- Chapter five presents the performance of various Hadamard based dictionaries for DR within the SER pattern recognition application. It also proposes the Feature Block (FB) based dimension reduction technique as an innovative solution to overcome the problem of low density ratio of samples to dimension.

- Chapter six presents the performance of various Hadamard based dictionaries for DR within the GBGC pattern recognition application.
- Chapter seven presents the conclusions of the thesis and potential directions for future research.

CHAPTER TWO: BACKGROUND

The first, and probably the most, serious computational challenge in pattern recognition/classification applications is the modelling of the complex objects of interest such as a face image, a speech signal, an MRI scan of the brain, … etc. Often such objects have obvious computer representation as high dimensional arrays of measurements while human, and in particular experts, can interpret and describe such objects with much less effort. Hence one can deduce that many types of real-life high dimensional models of data are not necessarily high dimensional (Wang, 2012). We all recognise that a large number of image pixels are redundant and it is the fact that is exploited in image compression techniques. In fact, any image contains smooth regions in that there are little variations in the pixel values and/or colours in such regions, i.e. there are lot of redundancies in such regions. While human can see and easily identify such regions, automatic identification of such facts by a computer requires a good model to represent the main characteristics of the pixel values in such regions.

For general object recognition/classification, even if the computer model of the investigated objects is genuinely high dimensional, then it is inconceivable that the full dataset of interest is scattered densely and uniformly across its model high dimensional vector space. Realising that most high dimensional datasets include some significantly large amounts of redundant features is an incentive to determine the "best" lower dimensional subspace of the whole space that capture all the discriminating features for the classification problem under consideration. Determining and locating the interesting and/or redundant features/entries in the high dimensional vector model of the objects under investigation require a good understanding of basic concepts in Linear Algebra and Matrix theory. Therefore, in this chapter we review the essential background in this field of Mathematics with focus on linear transformations and change of basis. We then describe the concept of dimension reduction in terms of matrix operations and discuss the mathematical theory of existence of dimension reduction procedures. We then end the chapter with a general discussion on the classification of Dimension reduction schemes.

2.1 Basic Terminology in Linear Algebra and Matrix theory

Traditionally, vectors are objects representing fixed length directed lines (i.e. has magnitude and direction/orientation), while scalars have only magnitudes. Linear Algebra is concerned with the study of spaces of vectors that could be scaled by a number system.

Definition (2.1): A ***vector space*** over a field F (elements are called scalars or numbers) is a triple $(V, +, .)$ where V is a set (whose elements are called vectors), satisfying the following properties for all vectors $u, v, w \in V$ and scalars $r, s \in F$:

1. (Closure of vector addition) $u + v \in V$
2. (Commutativity of addition) $u + v = v + u$
3. (Associativity of addition) $u + (v + w) = (u + v) + w$
4. (Additive identity) There exists and element $0 \in V$ such that $u + 0 = u = 0 + u$
5. (Additive inverse) There exists an element $-u \in V$ such that
$$u + (-u) = 0 = (-u) + u$$
6. (Closure of scalar multiplication) $r.u \in V$
7. (Distributive law) $r.(u + v) = r.u + r.v$
8. (Distributive law) $(r + s).u = r.u + s.u$
9. (Associative law) $(rs).u = r.(s.u)$
10. (Preservation of scale) $1.u = u$

The field of scalars could be the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, the rational numbers $\mathbb{Q}$, or even finite fields. In this thesis, we are only considering finite dimensional vector spaces over the field of real numbers, but most definitions and facts can be generalised to other fields. In other words, we will be working with the n-dimensional real vector space $\mathbb{R}^n$ whose vectors are size n-arrays of real numbers. In this vector space, vector addition and multiplication of a vector by a scalar are simply done coordinate by coordinate, i.e. for all $u, v \in \mathbb{R}^n$ and $s \in \mathbb{R}$:

$$u + v = [u_1, u_2, \cdots, u_n] + [v_1, v_2, \cdots, v_n] = [u_1 + v_1, u_2 + v_2, \cdots, u_n + v_n]$$

$$s.u = s.[u_1, u_2, \cdots, u_n] = [s.u_1, s.u_2, \cdots, s.u_n]$$

Definition (2.2): Let $v = [v_1, v_2, \cdots, v_n] \in \mathbb{R}^n$ be any vector. The ***Length/Magnitude*** of a is

$$\|v\| = \sqrt{{v_1}^2 + {v_2}^2 + \cdots {v_n}^2}$$

For example,

$$v = [2, 1, 2, 4] \rightarrow \|v\| = \sqrt{(2)^2 + (1)^2 + (2)^2 + (4)^2} = \sqrt{25} = 5$$

Definition (2.3): Let $x = [x_1, x_2, \cdots, x_n]$ and $y = [y_1, y_2, \cdots, y_n]$ be any two vectors in $\mathbb{R}^n$ and θ is the angle between them. The ***Dot/Inner*** product of $x\ and\ y$ is a scalar denoted by $x.y$ and given by

$$x.y = x_1y_1 + x_2y_2 + \cdots + x_ny_n = \|x\|\|y\|(\cos\theta)$$

From this formula, we get that $\cos\theta = \frac{x.y}{\|x\|\|y\|}$.

Simply, the ***angle between x and y*** while they are nonzero vectors is defined as:

$$\theta = \cos^{-1}\left(\frac{x.y}{\|x\|\|y\|}\right)$$

For example, $x = [2, 2, 1, 0]$ and $y = [1, 5, -3, 1]$

$$x.y = 2 * 1 + 2 * 5 + 1 * (-3) + 0 * 1 = 9$$

$$\|x\| = \sqrt{(2)^2 + (2)^2 + (1)^2 + (0)^2} = 3$$

$$\|x\| = \sqrt{(1)^2 + (5)^2 + (-3)^2 + (1)^2} = 6$$

$$\theta = \cos^{-1}\left(\frac{x.y}{\|x\|\|y\|}\right) = \cos^{-1}\left(\frac{9}{3 * 6}\right) = \cos^{-1}\left(\frac{9}{18}\right) = \cos^{-1}\left(\frac{1}{2}\right) \rightarrow \theta = 60^{\circ}$$

The above Euclidean norm/distance is just one example of an infinite number of norms/distances that can be defined on the vector space $\mathbb{R}^n$.

For any $p \in [1, \infty)$, the $l_p - norm$ is defined for any vector $x \in \mathbb{R}^n$ as follows:

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} & p \in [1, \infty) \\ max|x_i|, i = 1,2, \cdots, n & p = \infty \end{cases}$$

Note that $l_2 - norm$ stands for the Euclidean norm. Hence throughout this thesis $\|x\|_2 = \|x\|$. The most commonly used norms also include the $l_1 - norm,$ which is

called the city-block norm. Different norms define different geometries on the $\mathbb{R}^n$vector spaces, because distances and angles will have different meanings. Unless otherwise stated, we will only be working on vectors in the Euclidean space $\mathbb{R}^n$.

Definition (2.4): Let $x, y \in \mathbb{R}^n$ be any two vectors. $x\ and\ y$ are ***Orthogonal/Perpendicular*** if

$$x.y = 0$$

i.e. the angle between $x\ and\ y$, $\theta = 90^{\circ}$,

$$x.y = \|x\|\|y\|(\cos\theta) = \|x\|\|y\|(\cos 90) = \|x\|\|y\| * 0 = 0$$

One can think the way around, the ***angle between x and y*** is

$$\theta = \cos^{-1}\left(\frac{x.y}{\|x\|\|y\|}\right) = \cos^{-1}\left(\frac{0}{\|x\|\|y\|}\right) = \cos^{-1}(0) = 90^{\circ}, \text{ it means } \theta = 90^{\circ}$$

For example, $x = [-2,3,1], y = [4,1,5]$ be two vectors in $\mathbb{R}^3$.

$$x.y = (-2 * 4) + (3 * 1) = (1 * 5) = -8 + 3 + 5 = 0$$

$$\theta = \cos^{-1}\left(\frac{x.y}{\|x\|\|y\|}\right) = \cos^{-1}\left(\frac{0}{\|x\|\|y\|}\right) = \cos^{-1}(0) = 90^{\circ}$$

then we say $x\ and\ y$ are orthogonal/perpendicular.

Definition (2.5): A ***Unit*** vector is a vector of length 1. Any vector $x \in \mathbb{R}^n$ can be scaled by dividing all the coordinates in it by its magnitude and converted to a unit vector.

For example, $x = [4, 1, -2, 2]$

$$\|x\| = \sqrt{(4)^2 + (1)^2 + (-2)^2 + (2)^2} = \sqrt{25} = 5$$

Then, x is not a unit vector, we can scale it by dividing all the components by 5.

$$x = \left[\frac{4}{5}, \frac{1}{5}, \frac{-2}{5}, \frac{2}{5}\right]$$

$$\|x\| = \sqrt{\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{-2}{5}\right)^2 + \left(\frac{2}{5}\right)^2} = \sqrt{\frac{25}{25}} = 1$$

Now, x is a unit vector.

Definition (2.6): Let $x, y \in \mathbb{R}^n$ be any two vectors, then $x\ and\ y$ are two ***Orthonormal*** vectors if

$$x.y = 0 \quad and \quad \|x\| = \|y\| = 1$$

For example: We can easily show that $v_1 = \left[\frac{2}{3}, \frac{1}{3}, \frac{2}{3}\right]\ and\ v_2 = \left[\frac{-2}{3}, \frac{2}{3}, \frac{1}{3}\right] \in \mathbb{R}^3$ are two orthonormal vectors.

$$v_1.v_2 = \frac{2}{3} * \frac{-2}{3} + \frac{1}{3} * \frac{2}{3} + \frac{2}{3} * \frac{1}{3} = \frac{-4+4}{9} = 0$$

$$\|v_1\| = \sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2} = \sqrt{\frac{4}{9} + \frac{1}{9} + \frac{4}{9}} = 1$$

$$\|v_2\| = \sqrt{\left(\frac{-2}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2} = \sqrt{\frac{4}{9} + \frac{4}{9} + \frac{1}{9}} = 1$$

Definition (2.7): Let $V = \{v_1, v_2, \cdots, v_m\}$ be a set of vectors in $\mathbb{R}^n$ and $a_1, a_2, \cdots, a_m$ be m scalars in $\mathbb{R}$, then V is called ***linearly independent*** if the equation $v_1 a_1 + v_2 a_2 + \cdots + v_m a_m = 0$ holds with only $a_i = 0$ for all $i = 1,2, \cdots, m$. If this linear combination is equal to zero with at least one of $a_i \neq 0$, then, V is called ***linearly dependent***.

For example, $x = [0,2,4,-1], y = [0,4,8,-2]$ are two linearly dependent vectors as $2x = y$

$$2 * [0,2,4,-1] = 1 * [0,4,8,-2] \rightarrow 2 * [0,2,4,-1] - 1 * [0,4,8,-2] = 0$$

There are two scalars $2,-1$ such that $2x - y = 0$ while $2,-1 \neq 0$.

Definition (2.8): Let V be a vector space and $B = \{b_1, b_2, \cdots, b_n\}$ be a set of vectors in V. B forms a ***Basis*** of V if and only if it is linearly independent and spans/generates V, by span we mean for any vector $v \in V$, there is a linear combination of $b_1, b_2, \cdots, b_n$ such that $v = s_1 b_1 + s_2 b_2 + \cdots + s_n b_n$ where $s_i \in \mathbb{R}\ for\ i = 1,2 \cdots, n$. Such linear combination is unique for each vector v and the vector $[s_1, s_2, \cdots, s_n]$ is called coordinate vector of v relative to B.

For example, $B = \{[1,0,0,][0,1,0,][0,0,1]\}$ is called the standard basis of $\mathbb{R}^3$.

Definition (2.9): For a finitely generated vector space V, the ***Dimension*** of V is cardinality (number of elements) in a basis B of V which is denoted by $\dim(V)$.

In the above example, $B = \{[1,0,0,][0,1,0,][0,0,1]\}$ is standard basis of $\mathbb{R}^3$, So, $\dim(\mathbb{R}^3) = 3$, as the number of elements in B is equal to three. In fact, the set $\{\underline{\mathbf{e}}_1,\ldots, \underline{\mathbf{e}}_i, \ldots, \underline{\mathbf{e}}_n\}$ is the standard basis for $\mathbb{R}^n$ where $\underline{\mathbf{e}}_i = (0, \ldots, 0, 1,0,\ldots,0)$ is the unit vector in $\mathbb{R}^n$, whose i-th coordinate is 1 and all other coordinates are zeros, then $\dim(\mathbb{R}^n)$=n.

2.2 Linear Transformations

A function F that maps a vector space A into another vector space B, $F: A \to B$ is called linear transformation if it satisfies the following axioms for all $x, y \in A$ and any scalar $r \in \mathbb{R}$.

1. $F(x + y) = F(x) + F(y)$, [Addition preservation]

2. $F(rx) = rF(x)$, [Scalar multiplication preservation]

For example, a function $F: \mathbb{R}^3 \to \mathbb{R}^2$ where $F[x_1, x_2, x_3] = [x_1 + x_2, 2x_3]$ for any $x \in \mathbb{R}^3$ is a linear transformation. We can easily show that such map is a linear transformation, let $a, b \in \mathbb{R}^3$ and $r \in \mathbb{R}$.

1. $F(a + b) = F([a_1, a_2, a_3] + [b_1, b_2, b_3]) = [(a_1 + b_1) + (a_2 + b_2), 2(a_3 + b_3)]$

$$= [a_1 + a_2, 2a_3] + [b_1 + b_2, 2b_3] = F(a) + F(b)$$

2. $F(ra) = F(r[a_1, a_2, a_3]) = F([ra_1, ra_2, ra_3]) = [(ra_1 + ra_2), 2ra_3] = rF(a)$

2.3 Matrix-Terminology

Matrices play an important role in linear algebra. The set of all real valued matrices of a fixed size mxn forms a vector space of dimension N=mn over the field of real numbers. However, they also define linear transformations of vector spaces.

Definition (2.10): Let A *and* B be two matrices, B is called ***Transpose*** of A and denoted by $B = A^T$, if each entry b_{ij} *in* B is equal to a_{ij} *in* A. If $A = A^T$, then A is called ***Symmetric*** matrix.

For example, if $A = \begin{bmatrix} -2 & 0 & 5 \\ 1 & 1 & 2 \end{bmatrix}$, then $A^T = \begin{bmatrix} -2 & 1 \\ 0 & 1 \\ 5 & 2 \end{bmatrix}$

If $G = \begin{bmatrix} 0 & 2 & -3 \\ 2 & 1 & 6 \\ -3 & 6 & 0 \end{bmatrix}$, then $G^T = \begin{bmatrix} 0 & 2 & -3 \\ 2 & 1 & 6 \\ -3 & 6 & 0 \end{bmatrix}$, obviously $G = G^T$. So, G is a symmetric Matrix.

Definition (2.11): The ***Determinant*** of a square $n \times n$ matrix A is a scalar, denoted by $\det(A)$, defined as follows:

$$\det(A) = \sum_{j=1}^{n} a_{1j}\alpha_{1j}$$

Where the a_{ij} is an entry of A at position (i, j), and the coefficients $\alpha_{ij}$are given by:

$$\alpha_{ij} = (-1)^{i+j}\beta_{ij}$$

Where β_{ij} is the determinant of the $(n-1) \times (n-1)$ submatrix of A that obtained by deleting the i^{th} row and the j^{th} column of A.

Note: if $n = 1$, $A = [a_{11}]$, then $\det(A) = a_{11}$.

: if $n = 2$, $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, it is easy to show that:

$$\det(A) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

Example: the determinant of the matrix $B = \begin{bmatrix} 1 & 0 & 2 \\ 1 & -2 & 3 \\ 2 & 5 & 3 \end{bmatrix}$ is calculated as follows:

$$a_{11}\alpha_{11} = 1 * (-1)^{1+1} * \det\left(\begin{bmatrix} -2 & 3 \\ 5 & 3 \end{bmatrix}\right) = -21$$

$$a_{12}\alpha_{12} = 0 * (-1)^{1+2} * \det\left(\begin{bmatrix} 1 & 3 \\ 2 & 3 \end{bmatrix}\right) = 0$$

$$a_{13}\alpha_{13} = 2 * (-1)^{1+3} * \det\left(\begin{bmatrix} 1 & -2 \\ 2 & 5 \end{bmatrix}\right) = 18$$

$$\det(B) = -3$$

Definition (2.12): A square matrix A of size nxn is called ***Invertible,*** if there is a square matrix B of the same size such that $AB = BA = I$, where I is the identity matrix of size nxn. B is called the inverse of A and it is denoted by A^{-1}.

For example, $A = \begin{bmatrix} 1 & 2 \\ -2 & -5 \end{bmatrix}, B = \begin{bmatrix} 5 & 2 \\ -2 & -1 \end{bmatrix}$

$$AB = \begin{bmatrix} 1 & 2 \\ -2 & -5 \end{bmatrix}\begin{bmatrix} 5 & 2 \\ -2 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ -2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ -2 & -5 \end{bmatrix} = BA$$

So, B is the inverse matrix of A, ($B = A^{-1}$) and conversely.

Definition (2.13): A square matrix A of size nxn is ***Orthogonal matrix*** if it is invertible and $A^{-1} = A^T$, $(A^T A = I = AA^T)$. Or equivalently, if the rows/columns of A form an orthonormal basis of $\mathbb{R}^n$.

For example, $H = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$ is an orthogonal matrix,

$$H = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$H^T H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} * \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = HH^T$$

The previous linear transformation $F: \mathbb{R}^3 \to \mathbb{R}^2$ defined for any $x \in \mathbb{R}^3$ by the formula:

$$F[x_1, x_2, x_3] = [x_1 + x_2, 2x_3]$$

can be represented as a matrix, $F(x) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$.

In this format, the transformation becomes a matrix multiplication by a column vector $x \in \mathbb{R}^3$. In fact, all linear transformations in Euclidian space have a matrix representation.

On the other hand, it is obvious that any $(m\ x\ n)$ matrix A defines a linear transformation

$$T_A: \mathbb{R}^n \to \mathbb{R}^m$$

For any column vector $x \in \mathbb{R}^n$, the transformation defined by a matrix multiplication $T_A = Ax$.

2.4 Change of Basis and Coordinates

Let V be a finite dimensional vector space and $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ be two ordered bases of V. Any vector $v \in V$ can be presented uniquely as a linear combination of each basis, i.e.

$$v = s_1 a_1 + s_2 a_2 + \cdots + s_n a_n \,, \qquad s_i \in \mathbb{R} \; \forall i$$

This is the presentation of v in relative to A and the vector $v_A = [s_1, s_2, \cdots, s_n]^T$ is the coordinate vector of v in relative to A and it can be written in a matrix form

$$v = A v_A$$

The vector v has another unique representation in relative B in the same format

$$v = r_1 b_1 + r_2 b_2 + \cdots + r_n b_n \,, \qquad r_i \in \mathbb{R} \;\; \forall \, i$$

$$v = B v_B$$

From the above equations, we get that

$$A v_A = v = B v_B \rightarrow A v_A = B v_B$$

$$v_A = A^{-1} B v_B$$

So, in the above equation, the matrix $A^{-1}B$ transforms the coordinates of v in relative to B into its coordinates in relative to A. The inverse of this transformation is $B^{-1}A$.

The matrix $A^{-1}B$ is called transition matrix and it is a linear transformation which transform a vector in a coordinate system into a new one. This type of transformation is very important in the case linear dimensionality reduction techniques as some data dependent DR techniques reduce the dimension of a dataset after transforming it into a new space by changing its coordinate system. In the transformed space, some dimensions become very important and others become irrelevant and negligible, thus we can discard some of them and project the data set on only important dimensions.

2.5 Mathematical underpinning of DR

The feasibility of dimension reduction using random projections with providing a strong guarantee on "preserving" pairwise distances with high probability has been proved in 1984 by Johnson and Lindenstrauss in their well-known article (Extensions of Lipschitz maps into a Hilbert space).

Theorem (2.1) (Johnson and Lindenstrauss, 1984)

For any dataset A of n points in $\mathbb{R}^N$, and any $0 < \epsilon < 1$, there is a function $f: \mathbb{R}^N \to \mathbb{R}^k$, with

$$k \geq O(\epsilon^{-2} \log n)$$

such that for any two points $a, b \in A$

$$(1-\epsilon)\|a-b\|_2 \leq \|f(a)-f(b)\|_2 \leq (1+\epsilon)\|a-b\|_2$$

The JL theorem insures the existence of linear transformations that reduces dimensions of input vectors while the distances between the map of vectors are within any desired tolerance of the distance between the original vectors as long as the reduced dimension is bounded below by a number proportional to the tolerance level. Various modification of the JL theorem have been established that impose different restrictions on the value of k or on the nature of the vectors. More details will be discussed in Chapter 4.

2.6 Dimension Reduction and Preservation of Information

One of the most popular questions around dimension reduction process is that (is it possible to reduce the dimension of a dataset without a significant structure distortion? i.e. without losing too much information? To answer this question, first, we need to explain that what do we mean by dataset structure? The most natural characteristic of dataset structure, that is relevant to pattern recognition, is pairwise distances of data sample vectors and ideally DR techniques provide some guarantees on preserving these distances within a small tolerable error. We shall explain the concept of preserving dataset structure in a simple example. Consider the simple dataset in 2-dimensional space displayed in figure 2-1.

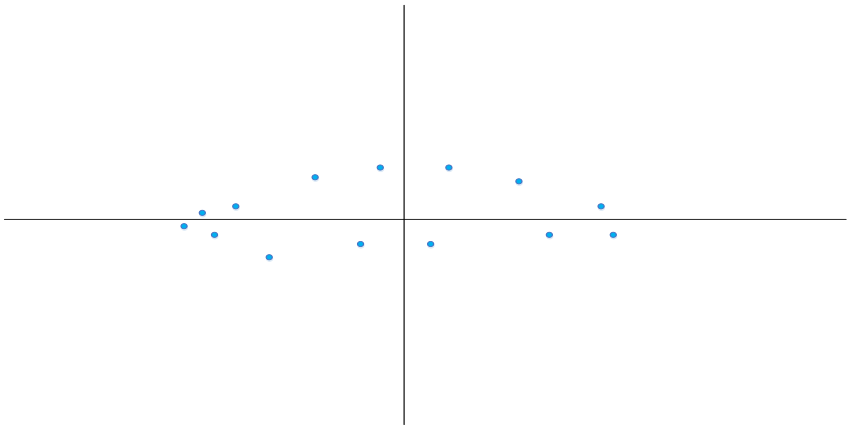


Figure 2-1 Two-dimensional dataset

The dimension of this data set can be reduced simply by discarding one of the dimensions and projecting it onto the other one. Clearly, the data set is more scattered along the X-axis rather than Y-axis and there is a large variance among the x-coordinates compare to the y-coordinates. This means that X-axis maintain more information of this data compare to the other axis. If we discard the Y-axis and project the data set onto X-axis, we will get the projected data in red points in figure 2-2. It can be clearly seen that the pairwise distances are preserved approximately, by looking to the distance between the two points $a\ and\ b$ in the original data set and the distance between their projection, it is fairly preserved. However, those points in the original data set that their x-coordinates are close, then their projections are also close to each other, nevertheless, the projected data is a proper approximation to the original one.

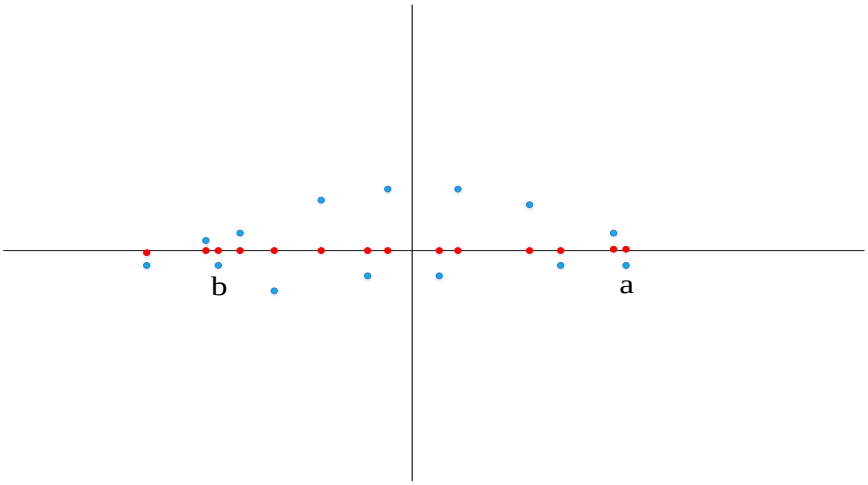


Figure 2-2 The projected data on X-axis is a good approximation

If we instead have projected our data set onto Y-axis by discarding X-axis as shown in figure 2-3, the projection causes a huge error in the pairwise distances. So, the pairwise distances are not preserved as the projected data is very inaccurate and it does not represent the original data properly. If we compare the distance between $a\ and\ b$ in the original data and the projected set, the distance between their projection is nearly zero while they are quite far from each other before projection. Thus, the projection on the X-axis is much better and provides more accuracy than projecting on Y-axis.

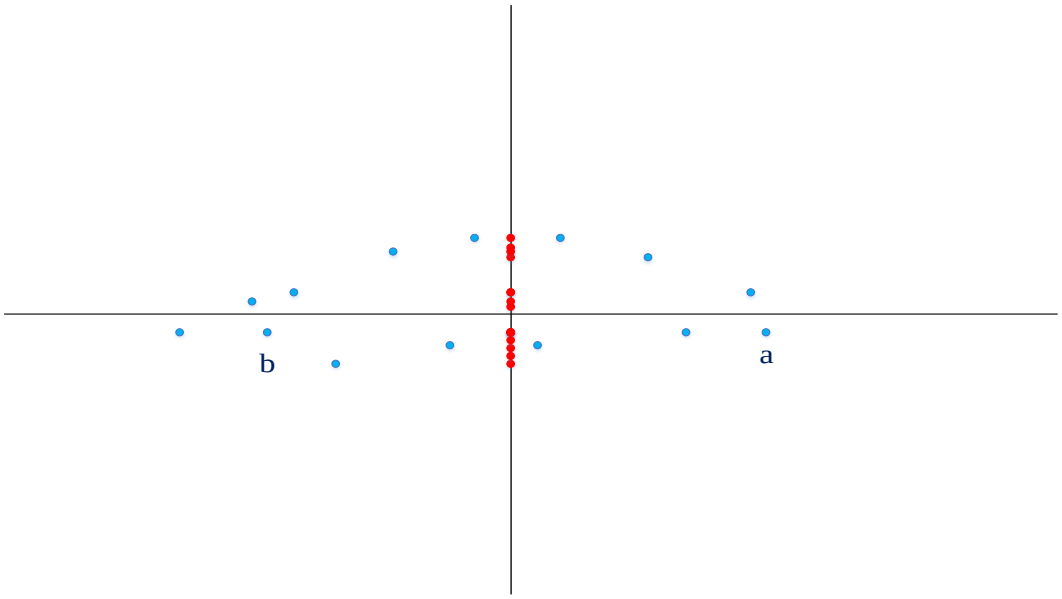


Figure 2-3 The projected data on Y-axis is inaccurate

Now, suppose that we have another dataset as shown in figure 2-4, which is simply a rotation of the previous dataset by 45º.

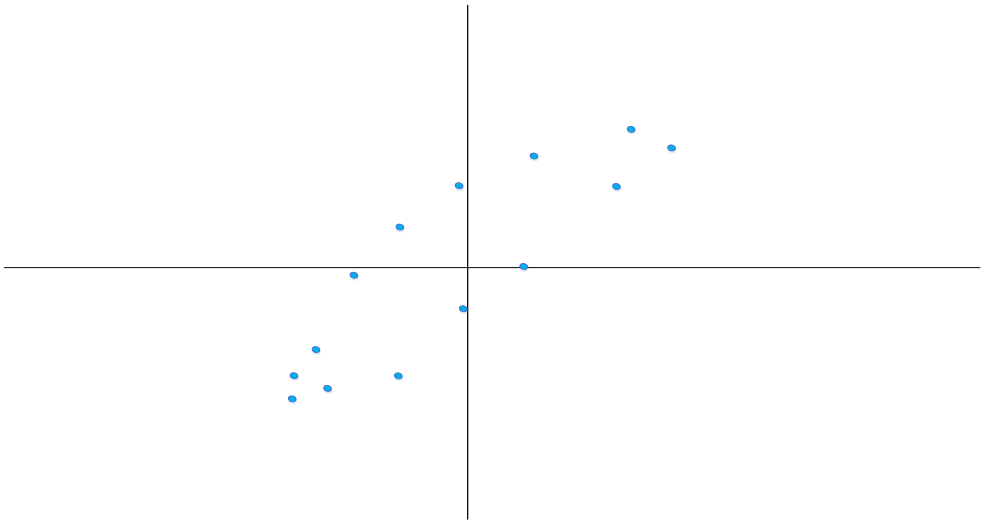


Figure 2-4 The rotated dataset by 45º

In this case, projecting our data onto X-axis or Y-axes does not provide a good data accuracy as the variance among both coordinates are nearly the same and discarding anyone of them will not preserve pairwise distances properly. We need to find another direction that is more suitable than these two axes for projection. This means discarding some axes is not the only way to project our data, in this case, the best direction/line for projection is the direction/line that captures maximum variance present in the dataset as shown in figure 2-5.

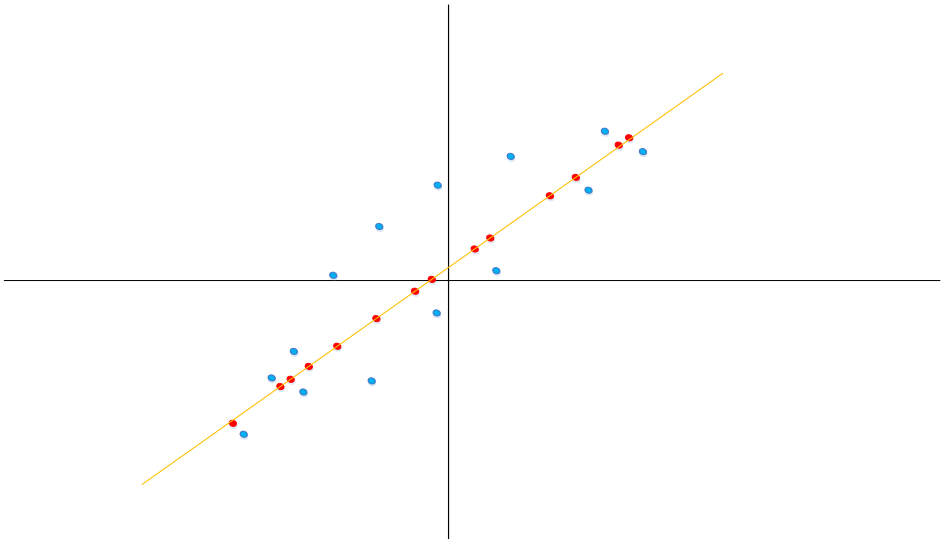


Figure 2-5 The optimal direction for projecting the rotated dataset

The above examples show that dimension reduction is not necessarily about discarding dimensions, it is a process of finding some dimensions that are important in some way for the dataset and maintain essential information (structure) after reduction. The projected subspace must be quite easy to compute for the reduced data while the structure is preserved with some tolerable errors. Based on the above explanation, we can say that; it is possible for DR techniques to reduce dimension without a significant distortion. These types of direction/basis that capture almost all the variance in a data set can be found by solving an eigenvalue problem of a matrix that model variation in the original data set. This is indeed one of the most common approaches in certain type of DR's which will be reviewed in the next chapter. So, dimension reduction and more precisely dimension transformation is a process of finding some new basis/directions that maintain almost all the information present in a dataset, such techniques reduce the dimension of a dataset carefully and having different DR techniques means that; there are different ways to provide such new basis and different ways to measure data distortion.

2.7 Classification of Dimensionality Reduction Methods

There are different ways of categorising dimensional reduction approaches. Dimension reduction problems have been classified in terms of the dataset feature model into two types: ***Hard*** and ***Soft*** DR depending on the number of extrinsic dimension of datasets (Wang, 2012). If the number of extrinsic dimensions is between hundreds and hundreds of thousands or above that, it is called hard DR problem and in this case an extreme reduction needs to be done. For instance, DR of facial images for recognition applications is one of the hard DR problems as the number of extrinsic dimension is

about hundreds of thousands. On the other hand, a DR problem is called soft if the number of extrinsic dimension is at most a few tens. In this case, the reduction is not drastic. The aim of soft DR problems is the analysis of the data rather than reduction as the original dimension of the data set is not truly high. In terms of the most challenging recognition/classification problems this classification is not of significant relevance because in general we are dealing with Hard DR problems. In general, DR methods can be classified as follows:

2.7.1 Dimension Selection (DS)

Dimension selection is one of the Dimensionality reduction methods which has been commonly used in the pre-processing stage of pattern recognition. This technique reduces the dimension of a data set by taking a proper dimension subset and discarding other dimensions out of the set of all dimensions based on a criterion. There are two major methods of Dimension Selection (DS): *Filter* and *Wrapper* (Kojadinovic and Wottka, 2000). The criterion of Filter method is measuring some properties of the dimensions using certain type of filters while the criterion for Wrapper method is finding a dimension subset which provides the "best" accuracy for the application. DS is a good alternative for other DR techniques where either the dimension of a data set is not very high or there are different types of dimension and some of them are highly correlated (Sabir, 2015). This method of DR is not the objective of this thesis while having knowledge about them will help to study other DR methods properly.

2.7.2 Dimension Transformation/Embedding (DT)

Dimension transformation method is a process of transforming a high dimensional dataset into a much lower subspace by using some mappings. Such method reduces dimension by providing a new linear or non-linear combination of the original data features, while in the case of dimension selection, we only choose a proper feature/dimension subset from the set of all features. In general, there are two types of feature transformation, either it is linear such as the Random Projections or non-linear like the Kernel PCA. Some of the techniques are unsupervised like the Principal Component Analysis (PCA) or supervised such as the Linear Discriminant Analysis (LDA). In this thesis, we focus on the Linear dimension transformation techniques. Simply, suppose that A_{mxN} is a data set consists of m −points in N −dimensional space (N is usually large). To reduce the dimension of this set using Linear DT, we only need to provide a suitable projection matrix (Linear Transformation) which is an overcomplete matrix/dictionary say $P_{d\times N}$ where $(d \ll N)$. Mathematically, this

projection matrix is a linear transformation which transforms the original high dimensional data into a much lower subspace. The original dataset is multiplied by the projection matrix to produce the low dimensional approximation $A'_{m\times d}$ using this formula

$$(A'_{m\times d} = A_{m\times N} P^T_{N\times d})$$

In the case of studying Linear DT techniques, all the questions are around the projection matrix. Some of the techniques extract the projection matrix from the data set itself, these techniques are called ***Data-Dependent*** such as PCA and LDA. We can also generate the projection matrix independently from the dataset by using ***Data-Independent*** techniques such as Random Projection (RP) Methods. It will be explained that how PCA, LDA, and RPs provide such projection matrix and how they guarantee information preservation in the following two chapters.

2.8 Summary

In this chapter, we reviewed some important basic concepts in Linear Algebra and Matrix Theory as the mathematical background of linear DR techniques. We first defined Vector Spaces and specified our work space, we then revised Linear Transformations and Change of Basis since all linear DR techniques whether it is data-dependent or data-independent are linear transformations that transform/imbed a high dimensional dataset into a much lower subspace by finding some new directions/basis that maintain essential information of the dataset in the transformed space. We stated the JL theorem as a mathematical underpinning of DR, and investigated the feasibility of dimension reduction without losing significant information in a simple example. We finally explained different classification methods of DR techniques. In the next chapter, we shall revise the theory of Eigenvalue problem and its computation. We will also investigate data-dependent DR techniques that use Eigenvalue problem approach.

CHAPTER THREE: DATA-DEPENDENT DIMENSION REDUCTION

In the last chapter, we noted that for a general dataset of high-dimensional vectors, modelling a recognition application, dimension reduction by feature selection may not be adequate. When the dataset of 2D points in section (2.6) where rotated by a certain angle, it became more susceptible to dimension reduction into one dimension with little loss in variation in the other direction. This meant that DR is about projecting the points onto a lower dimensional subspace that captures the maximum variation between the points in the direction of subspace basis vectors. The directions of the subspace basis vector, depend on the dataset, and can be found by solving the eigenvalue problem of a matrix that models variation in the original dataset. In fact, the further away the eigenvalue is from 0, the more important are the variations between the projected points onto the corresponding eigenvector(s). Data-dependent DR is dominated by different methods of finding bases built from eigenvectors for a carefully defined eigenvalue problem. This chapter is concerned with such approaches as well as modified techniques that have similar effect. We first review the theory of Eigenvalue problems and relevant computation with focus on high dimensions. This will be followed by investigating three different DR methods that follow the eigenvalue problem approach and its link to matrix factorisation.

3.1 Eigenvalues and Eigenvectors

Let A be an nxn square matrix of real numbers. A scalar number λ is said to be an eigenvalue of A if there exists a non-zero vector $v \in \mathbb{R}^n$ such that

$$Av = \lambda v$$

In this case, we say that v is corresponding eigenvector while λ is an eigenvalue of the matrix A. There are a few methods to calculate eigenvalues and eigenvectors, and ***Characteristic Equation*** method is one of them. Simply, the above equation can be written as follows:

$$Av = \lambda v \rightarrow Av - \lambda v = 0 \rightarrow (A - \lambda I)v = 0$$

Where I is the nxn identity matrix. The matrix equation $(A - \lambda I)v = 0$ is a homogeneous system, and has a non-trivial solution vector v whenever the matrix $(A - \lambda I)$ is not invertible, i.e.

$$\det(A - \lambda I) = 0$$

The left-hand side of this equation is a degree n polynomial $P(\lambda)$ in λ, and

$$P(\lambda) = 0$$

is called the ***characteristic equation*** of the matrix A.

The eigenvalues of a matrix A can be computed by solving its characteristics equation, and the corresponding eigenvectors can be computed by solving the matrix equation

$$Av - \lambda v = 0.$$

For example: if matrix $A = \begin{bmatrix} 0 & 3 \\ 5 & 2 \end{bmatrix}$ then

$$|A - \lambda I| = 0 \rightarrow \left| \begin{bmatrix} 0 & 3 \\ 5 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

Therefore,

$$\left| \begin{bmatrix} -\lambda & 3 \\ 5 & 2 - \lambda \end{bmatrix} \right| = (-\lambda)(2 - \lambda) - 15 = \lambda^2 - 2\lambda - 15 = (\lambda - 5)(\lambda + 3) = 0$$

The eigenvalues of A are $\lambda_1 = 5, \ \lambda_2 = -3$

For $\lambda_1 = 5$, $Av = \lambda v \rightarrow \begin{bmatrix} 0 & 3 \\ 5 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 5 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \rightarrow \begin{bmatrix} 3v_2 \\ 5v_1 + 2v_2 \end{bmatrix} = \begin{bmatrix} 5v_1 \\ 5v_2 \end{bmatrix}$

Which implies that $v_2 = \frac{5}{3} v_1$. Setting $v_1 = 1,$ yields the eigenvector $v = \begin{bmatrix} 1 \\ 5/3 \end{bmatrix}$.

Similarly, for $\lambda_2 = -3, \ v = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ is a corresponding eigenvector.

General remarks.

(1) if v is an eigenvector corresponding to an eigenvalue λ, then for any scalar value α the vector αv is also an eigenvector for λ. Hence v generates a 1-dimensional subspace of $\mathbb{R}^n$.

(2) The characteristic equation of A is a polynomial of degree n, and therefore it has n eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ not all distinct or even real numbers.

For example, $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ has 2 complex conjugate eigenvalues $\lambda_1 = i, and\ \lambda_2 = -i.$

And the matrix, $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ has 2 equal eigenvalues $\lambda_1 = \lambda_2 = 1.$

Theorem (3.1): Let A be a matrix of size $n \times n$. If $B = \{v_1, v_2, \cdots, v_n\}$ is the set of eigenvectors corresponding to the distinct eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ of A., then B is linearly independent.

Proof, see (Fraleigh et al., 1995).

Although linear independence of the eigenvectors is a useful property, but for computation purposes orthogonality of these vectors is more desirable. Note that, our interest in the eigenvalue problem is based on the observation that we need a change of basis so that a smaller number of the vectors in the new basis can capture the maximum amount of variation between the various vectors representing the given set of application objects. The variation in a dataset $U = \{u_1, u_2, \cdots, u_N\}$ of application records is represented by the covariance matrix

$$Cov(U) = \left[\langle u_i - \mu, u_j - \mu \rangle\right]$$

The covariance matrix is a symmetric matrix with real-valued entries. Here, μ is the average vector of the vectors in U, each coordinate of which is the mean of that coordinates of the vectors in U. In this case, the following is a very important property that has very useful implications for the DR process for the dataset U.

Theorem (3.2): Let A be a square real symmetric matrix, then the eigenvectors of A corresponding to different eigenvalues are orthogonal.

Proof: see (Fraleigh et al., 1995)

Definition (3.1): Let A_{nxn} be a matrix, then it is called ***orthogonally diagonalizable*** if there is a diagonal matrix D and an orthogonal matrix B such that $A = BDB^{-1} = BDB^{T}$.

Fundamental Theorem of Real symmetric matrix (3.3): Let A_{nxn} be a real symmetric matrix, then A is orthogonally diagonalizable and it has only real eigenvalues (Fraleigh et al., 1995).

Definition (3.2): Let A_{nxn} be a matrix and $\lambda_1, \lambda_2, \cdots, \lambda_n$ be the eigenvalues of A, then λ_1 is called the ***dominant eigenvalue*** of A if $|\lambda_1| > |\lambda_i| \, for\ i = 2,3, \cdots, n$ and the corresponding eigenvector say v_1 to the λ_1 is called the ***dominant eigenvector*** of A.

In the example (3.1), where $A = \begin{bmatrix} 0 & 3 \\ 5 & 2 \end{bmatrix}$ and $\lambda_1 = 5$, $\lambda_2 = -3$, clearly, $|\lambda_1| = 5, |\lambda_2| = 3$, and $|\lambda_1| > |\lambda_2|$. So, λ_1 is called the dominant eigenvalue of A and the corresponding eigenvector $v = \begin{bmatrix} 1 \\ 5/3 \end{bmatrix}$ is called the dominant eigenvector.

Sorting the eigenvalues of the covariance matrix of a dataset of objects in order of descending their magnitude/absolute value, plays an important role in the DR process, because more dominant eigenvalue is the more variation along its eigenvector is away from the average vector. In this way, we can find those directions that capture almost all variation present in a dataset.

In dimension reduction applications, computing the dominant Eigenpairs of the matrix that models the variation in the dataset of objects is an essential requirement. However, computing the eigenvalues and eigenvectors of large size matrices, which relates to the (curse of dimension) problem, by solving its characteristic equation is not stable because approximating the roots of high order polynomials is so sensitive and ill-condition (i.e. just a little inaccuracy in the variables can cause a significant error in the results) (Fraleigh et al., 1995)

The Power Method is an iterative procedure to approximately compute such Eigen pairs. For more details see (Fraleigh et al., 1995). The algorithm, repeatedly estimate the next eigenvector corresponding to the next dominant eigenvalue.

3.2 Computing Dominant/Top Eigenpairs of Real Symmetric Matrices

The reason of stating the above theorems and definitions is that, the Eigenpairs of any real symmetric matrix have some very nice properties and it is quite useful in the case of studying linear data-dependent DR techniques, because the linear transformations of some of these techniques are constructed by top eigenvectors of some real symmetric matrices. By top eigenvectors, we mean, after sorting the eigenvectors in the order of descending magnitude of the corresponding eigenvalues, those eigenvectors corresponding to the eigenvalues with high absolute value/magnitude.

Now, we are looking for a method to find a subset of top eigenpairs instead of calculating all by using the (Power Method) and Theorem (3.3).

Let A_{nxn} be a real symmetric matrix with its eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ such that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ and corresponding unit eigenvectors $v_1, v_2, \cdots, v_n$. Let D_{nxn} be a

diagonal matrix with the eigenvalues on its diagonal in the same order and B_{nxn} is another matrix which has the eigenvectors in its columns correspondingly. Using the definition of eigenpairs $Av = \lambda v$ for each eigenpair, then we can write the above equation in this way $AB = BD$, we know that B consists of $n-$orthogonal unit eigenvectors which implies that B is invertible and $B^{-1} = B^T$, then we get that

$$AB = BD \rightarrow A = BDB^{-1} = BDB^T$$

More precisely,

$$A = BDB^T = \lambda_1 b_1 b_1^T + \lambda_2 b_2 b_2^T + \cdots + \lambda_n b_n b_n^T$$

Clearly, (λ_1, b_1) is the dominant eigenpair of A and it can be found using the Power Method. Then, if we write the above equation in this form

$$A - \lambda_1 b_1 b_1^T = \lambda_2 b_2 b_2^T + \cdots + \lambda_n b_n b_n^T$$

Now, we get another matrix $(A - \lambda_1 b_1 b_1^T)$ with the dominant eigenpair (λ_2, b_2) and it can be found using the Power Method again. By repeating this procedure, we can compute the top eigenvalues successively in a descending order instead of computing all of them. Such procedure is quite interesting and we will explain that later why we are interested in computing top Eigen pairs. In the following sections, we shall study these data-dependent DR techniques that follows Eigen problem approach.

3.3 Principal Component Analysis (PCA)

Principal Component Analysis is perhaps the most popular linear DR technique which is data-dependent. The term PCA refers to the process of obtaining an orthonormal linear transformation that maps a high dimensional dataset into a lower subspace whose basis vectors correspond to the maximum variance directions in the original apace (Martinez and Kak, 2001). The matrix that represents this linear transformation is called the PCA projection matrix and such technique is data dependent as the projection matrix is extracted from the dataset of sample vectors representing objects of interest for a given application.

The key idea of this technique is that, coordinates of high dimensional data are usually highly correlated except the case that the data set is very small or it has a simple structure (Jolliffe, 2002). PCA finds a set of new uncorrelated directions/basis for the data set which are called principal components such that the data set has the maximum variance along the direction of the PCs in a descending order as shown in figure 3-1. More

precisely, it has the maximum variance along the direction of the first PC and it has the second maximum variance along the second direction and so on. By projecting the data set on the PCs, the variation present of the data set can be captured as much as possible. Interestingly, the first few PCs can capture almost all the variation present of the data set and the variation on other PCs is very small and they are negligible. This property makes PCA to be an effective dimension reduction technique and the reduction can be done by discarding these PCs with low variation/information. Moreover, PCA provides a better representation of any correlated dataset as the input data is possibly correlated while the projected data is uncorrelated. This transformation can be obtained for any dataset of records by computing the generalized Eigenproblem of the covariance matrix of the dataset as explained in the following section.

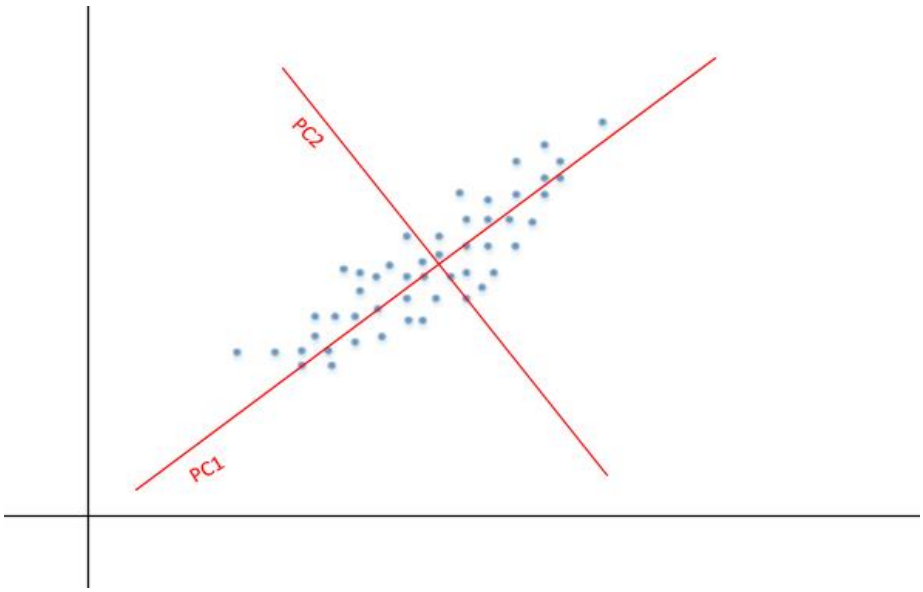


Figure 3-1 Principal Component Analysis captures variance present in a data set

3.3.1 PCA steps for Dimension Reduction.

The following steps explain that how to implement PCA technique on a dataset for DR purposes.

(1) Let $X = \{ x_{1,} x_2, \dots , x_m \}$ be a data set of m points in $\mathbb{R}^n$. Firstly, it is arranged in a matrix called data matrix X_{mxn}.

(2) Compute the centred data matrix A from the data matrix X in this way, compute the mean of each column/feature,

$$h[j] = \frac{1}{m}\sum_{i=1}^{m} x[i,j] \quad when \quad j = 1,2,\dots,n. \quad \rightarrow \quad H = [h_1, h_2, \dots, h_n]$$

Then A is computed using this formula

$$A_{mxn} = X_{mxn} - G_{mx1}H_{1xn} \quad when \quad G[i] = 1 \quad for \quad i = 1,2,\dots,m$$

(3) Compute the covariance matrix C which is a real symmetric matrix from the centred data matrix A using this formula,

$$C_{nxn} = A^T A$$

(4) Compute eigenvalues and eigenvectors of the covariance matrix C. Clearly, Eigenvectors of C are orthogonal and then normalize them by dividing each by its length (Normalization). Sort the orthonormal eigenvectors in the order of descending magnitude of the eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$.

(5) The orthonormal transformation matrix T is obtained by considering the k^{th} eigenvectors corresponding to the top k $(k \ll n)$ eigenvalues as its columns in the same order. To produce the projected/reduced data D_{mxk}, we do the projection using this formula

$$D_{mxk} = A_{mxn}T_{nxk}$$

In this way, we can reduce the dimension of a high dimensional data set A_{mxn} into a much lower dimensional subspace $(k \ll n)$ without losing too much information and providing a better representation

3.3.2 Covariance Matrix and Eigen Problem

The entries of covariance matrix are dot products and simply the dot product measures the similarity between two vectors. In this way, Covariance matrix measures correlation between data features/dimensions. We solve the generalized eigenproblem for covariance matrix and among the eigenvectors, we only choose those correspond to the top eigenvalues in the sense of their magnitude in order to choose the direction that maximizes the variance present in a dataset. Interestingly, Covariance matrix C is a real symmetric matrix, we can easily show that:

$$C^T = (A^T A)^T = A^T {A^T}^T = A^T A = C$$

Clearly, to create the projection matrix, we only use a subset of eigenvectors that corresponding top eigenvalues, this means that, we do not need to calculate all the eigenpairs and by using the steps in section (3.2), we can only calculate the top eigenpairs which is computationally cheaper than calculating all of them.

Furthermore, since the covariance matrix is a real symmetric matrix and according to Spectral theorem, it has only real eigenvalues. The matrices $A^T A$ and AA^T have the same eigenvalues and this follows from the fact that if $\lambda \neq 0$ is an eigenvalue of $A^T A$ and v is its eigenvector then:

$$(A^T A)v = \lambda v$$

$$A(A^T A)v = A\lambda v \rightarrow (AA^T)(Av) = \lambda(Av)$$

So, it shows that if λ is a non-zero eigenvalue of $A^T A$ with the eigenvector v, then it is also an eigenvalue of AA^T with the corresponding eigenvector Av. Consequently, if the number of data records is less than the number of dimensions $m \ll n$, there will be up to m useful eigenvector and the rest will have eigenvalues of zero. In this case, it is better to solve the eigenproblem for AA^T instead of $A^T A$ and in this way, we reduce the calculation significantly and it becomes more manageable since calculating eigenpairs is an expensive task (Turk and Pentland, 1991).

The following example illustrate, the use of PCA for face recognition.

Figure 3-2 A selection from a 200 training face images from the ORL database

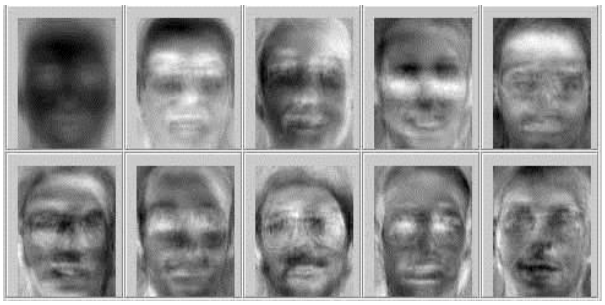


Figure 3-3 (10-most) significant Eigenfaces out of the 200 eigenfaces

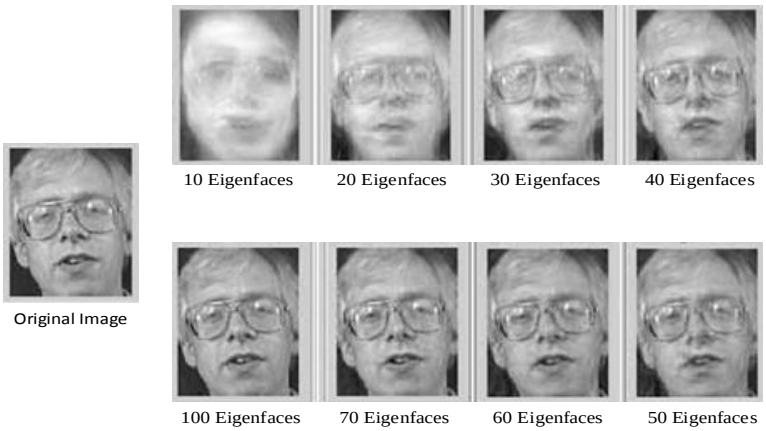


Figure 3-4 Incremental reconstruction of a face image from set of top Eigenfaces

The above example, illustrate the success of PCA in reconstructing high dimensional face image from a significantly reduced dimensional subspace in the PCA domain. The original image can be represented with a very good approximation, by 100 PCA coefficients using the 100 eigenvectors that corresponds to top 100 significant eigenvalues. This comes from the fact that PCA minimizes the reconstruction error.

3.3.3 Covariance matrix and Limitations of PCA

Despite the success of PCA as a dimension reduction scheme, some shortcomings of the scheme limit its use directly and modifications are essential. Here we list the 2 most important limitations:

(1) As we showed that if λ is a non-zero eigenvalue of $A^T A$ with the eigenvector v, then it is also an eigenvalue of AA^T with the corresponding eigenvector Av. Now, the point is that $A^T A$ and AA^T have the same set of non-zero eigenvalues. This property makes some limitation on PCA in practice by restricting the produced number of meaningful principal components, for instance, suppose we have a data set A consist of 50 samples in 5000-dimensional space resulting in a matrix of size (50×5000). In this case, PCA generates up to 50 useful eigenvectors and the rest will have associated eigenvalue of zero. In other words, it cannot produce more than 50 PCs and this restriction comes from the fact that PCA is a data dependent technique. To overcome with this limitation, ***Data Independent PCA*** (DIPCA) has been suggested in (Al-Talabani, 2015). DIPCA has the same projection matrix of PCA which is trained on another dataset and is used to project another given dataset which is independent from this projection matrix. In this way, we can pass such limitation and get benefit from the good characteristics of PCA projection matrix.

(2) The covariance matrix is an estimation of the variation between a dataset records away from their mean, but this estimation does not consider the topological relations among

them. The topological relationships provide information about the subspace that spanned by the data samples. This type of information and prior knowledges can be used to generate a more suitable subspace (projection matrix) and for this reason, ***Topological PCA*** (TPCA) has been suggested instead of the usual PCA in (Pujol et al., 2001). The covariance matrix of the TPCA is a linear combination of usual covariance matrix and a prior covariance matrix which contains topological relationships of the data samples. This method provides a more robust version of covariance matrix and thus it improves the general capabilities of PCA.

3.4 Linear Discriminant Analysis (LDA)

Although PCA is widely used as a dimension reduction prior to classification, in the construction of PCA and particularly covariance matrix, no consideration is given of class information/labels as PCA looks to the dataset as a global set. This makes PCA not to be an optimal dimensionality reduction technique for classification applications because it does not guarantee class discriminatory in the projected subspace and it might cause a huge overlapping between different classes. In fact, the performance of PCA based recognition could adversely influenced by the within class variation. In the case of face recognition variation in lighting, age and pose are examples of conditions leading to significant within class variation. For this reason, PCA has been modified and optimised for class discriminatory by the so called ***Linear Discriminant Analysis*** (LDA).

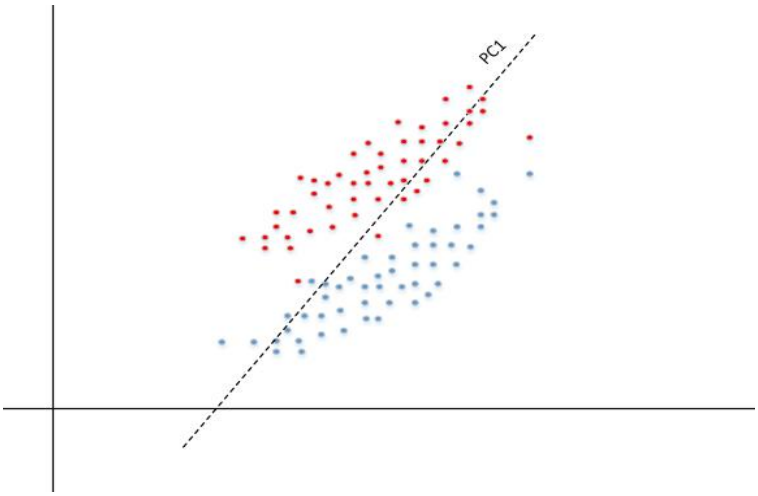


Figure 3-5 PCA does not consider class labels

LDA is also a data-dependent DR technique that has been commonly used in pattern-recognition and machine learning applications. It transforms a high dimensional dataset into a lower subspace that is optimal for class-discriminatory. It was first designed by Fisher (Fisher, 1936) which finds a new basis of a linear subspace of $\mathbb{R}^n$ along the

directions of which classes of the given dataset are well separated i.e. (a new basis that makes the distance between the mean of classes as far as possible and the variance of each class as small as possible). To obtain such basis, we compute two scatter matrices, between class scatter matrix S_B and within class scatter matrix S_W, and the goal is maximizing S_B and minimizing S_W. Equivalently, we aim to maximize the ratio $\frac{\det(S_B)}{\det(S_W)}$ which is called Fisher criterion and it yields by solving the generalized Eigen problem for the matrix $(S_W^{-1}S_B)$ (Martinez and Kak, 2001). Among all the eigenvectors of this matrix, we use some of the top eigenvectors corresponding top eigenvalues to generate LDA projection matrix which is an optimal subspace that separate different classes quite nicely.

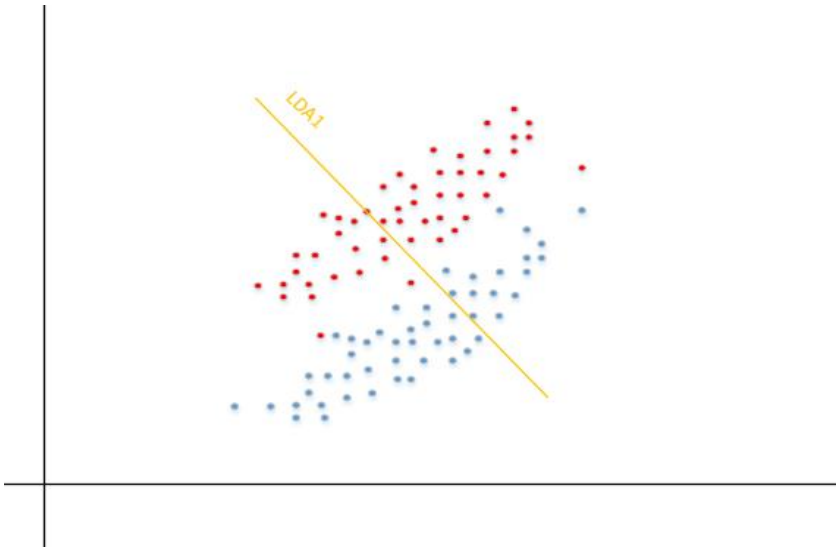


Figure 3-6 LDA provides a good class-discriminatory

3.4.1 LDA steps for Dimension reduction.

(1) Let $X = \{ x_1, x_2, \dots , x_m \}$ be a data set of m points in $\mathbb{R}^n$. Firstly, compute the centred data matrix A_{mxn}, see(step 1 and 2 PCA steps). Suppose there are C classes.

(2) Let n_i be the number of samples in the class i. Compute the mean vector m_i for each class the mean vector m for all the dataset.

(3) Compute the between and within scatter matrices S_B and S_W respectively.

$$S_B = \sum_{i=1}^{c} (m_i - m)(m_i - m)^T$$

$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_j - m_i)(x_j - m_i)^T$$

(4) Compute the eigenvalues and eigenvectors for the matrix $S_W^{-1}S_B$ and Sort the eigenvectors in order of descending magnitude of eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$.

(5) choose a d-eigenvectors ($d \ll n$) corresponding to the top d-eigenvalues to Create a projection matrix P by taking the d-eigenvectors as its columns in the same order. Do the projection using this formula $X' = XP$, where X' is an nxd-matrix representing the data set after reduction.

In this way, LDA provide some good class-discriminatory directions which is very important for classification/recognition applications. We cannot achieve this property with PCA as it considers the whole classes as a set globally. However, we might think LDA always outperforms PCA in classification and pattern recognition applications, but it is not always true. Especially, when there is a small (non-representative) training data set, PCA outperforms LDA and furthermore, LDA is more sensitive than PCA to different training set (Martinez and Kak, 2001).

3.5 Singular Value Decomposition (SVD)

The above linear DR schemes yields an approximate matrix factorisation of the data covariance matrix. Singular Value Decomposition or simply SVD is an exact matrix factorization methods which generalises the use of eigenvalue problem but for rectangular matrices of data.

Let A be a matrix of size mxn, and by SVD of A we mean that A can be uniquely represented as a product of three matrices $A = U\sum V^T$ where

A_{mxn}: the input data matrix.

U_{mxm}: matrix of left singular vectors, Orthogonal Matrix.

$\sum_{mxn}$: a rectangular diagonal matrix, there are r non-zero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ on its diagonal.

V_{nxn}: matrix of right singular vectors, Orthogonal Matrix.

The SVD is indirectly related to PCA, in the sense that solving the Eigenvalue problems for the covariance matrix, and its transpose, for the training set $X = \{ x_{1,} x_2, \dots , x_m \}$ of n-dimensional vectors is equivalent to decomposing the m×n matrix whose i-th row is $(x_i - \mu)$ where μ is mean vector of the elements in X. In this case, the singular values are the squares of the eigenvalues of the covariance matrix. Computing the SVM of a

rectangular mxn matrix A is based on solving the Eigenvalue problems of the 2 square matrices $(A^TA)\ and\ (AA^T)$ as described below.

The SVD decomposition of a matrix can be computed using the following Steps:

1. Compute the matrix $B = A^TA$, where A is the data matrix.
2. Calculate the eigenvalues and eigenvectors of B.
3. Calculate the singular values σ_i which are square roots of non-zero eigenvalues of B and sort them in a decreasing order.
4. Compute the three components of the factorization U, $\sum$ and V^T where
   $V = [v_1, v_2, \cdots, v_n]$ where $v_i$is a normalized eigenvector of $B = A^TA$.
   $\sum$ is a diagonal matrix, the diagonal entries are singular values σ_i in a decreasing order which are square roots of non-zero eigenvalues of $B = A^TA$. We assume for some index r, $(\sigma_1, \sigma_2, \cdots, \sigma_r)$ are non-zero and the rest are zero.
   $U = [u_1, u_2, \cdots, u_m]$ where $u_i = \frac{1}{\sigma_i}Av_i, \sigma_i \neq 0$. Now,

$$A = U\sum V^T = \sum_{i=1}^{r} u_i\sigma_i v_i^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

It is very important to sort the singular values in a decreasing order with corresponding left and right singular vectors especially in the case of using SVD for data reduction/compression. SVD as a matrix decomposition has some interesting applications, one of them is dimension reduction, although in the literature such procedure is referred to data compression which has the same meaning of dimension reduction. In the following section, we shall explain that how SVD can be used for this purpose.

Definition (3.3): Let A be a matrix of size $m\times n$ and $a_{i,j}$ is the entry of A at (i, j) position. the ***Frobenius norm*** of A is defined as

$$\|A\|_F = \sqrt{\sum_{i,j}\left(a_{i,j}\right)^2}$$

Theorem (3.4): Let A be a matrix of size $m \times n$ and the singular value decomposition of A is $A = U\sum V^T$, where $U\ and\ V$ are orthogonal matrices of left and right singular vectors respectively and $\sum$ is a diagonal matrix of singular values $(\sigma_1, \sigma_2, \cdots, \sigma_r)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $r = rank(A)$. Then for any $1 \leq k \leq r$, $\min\{\|A - B\|_F^2\ such\ that\ rank(B) = k\} = \sum_{i=k+1}^{r} \sigma_i^2$, where for any approximation matrix B of A, $\|A - B\|_F$ is called reconstruction error. The minimum of the equation is achieved with $B = A_K$, where $A_k = U_k \sum_k V_k^T$, $U_k\ and\ V_k$ are formed by the first k columns of $U\ and\ V$ and $\sum_k = diag(\sigma_1, \sigma_2, \cdots, \sigma_k)$ (Ye, 2005).

3.5.1 Dimension Reduction using SVD

Suppose we have a very large image/matrix A and we want to represent, save or send it as its SVD factors, the factors U, $\sum$ and V^T are also large matrices. Interestingly, SVD can be used to determine the most essential information in our data matrix, i.e. SVD can provide a compressed/reduced version of A without losing too much information. The key idea of using SVD as dimension reduction tool is the singular values. In general, some of singular values are very large and others are quite small. The ideal way to reduce the three components of SVD is to keep the most significant singular values and set others to zero. By setting most of small singular values to zero, we eliminate corresponding columns in the matrices $U\ and\ V$, see figure 3-7. In this way, we reduce the dimension of U, $\sum$ and V^T. At the same time, we retain almost all the information of our data matrix by keeping these top singular values because removing small singular values only cause losing a little information. Theorem (3.4) states that, retaining the top "k" singular values provide the optimal k-rank approximation of A.

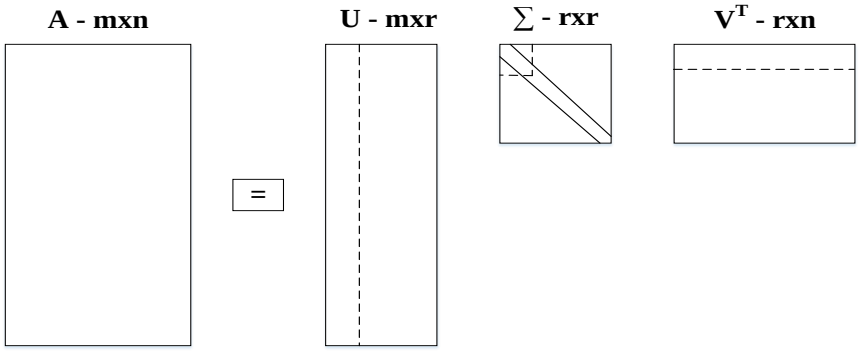


Figure 3-7 Using SVD for dimensionality reduction

3.5.2 Example: Image reduction/compression using SVD

In this example, we shall explain image compression using SVD. Let A be greyscale image of size 1000x1000. If we want to send or save this image, it contains $1000 \times 1000 = 1000000$ numbers (pixel intensity values) which is a huge number. The required memory can be reduced significantly by computing a good approximation of A using SVD which maintain all the essential information in A and reduce/remove redundant/irrelevant information. The reduced version of the image can be used more effectively instead of the original one with accepting that we lose little information (Kahu and Rahate, 2013). Firstly, we compute the SVD of A

$$A = \sum_{i=1}^{r} u_i \sigma_i v_i^T = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T + \cdots + u_r \sigma_r v_r^T$$

We will have some large singular values and others are very small while they are sorted in a decreasing order. Suppose A is the original image in figure 3-8, which is a 1000x1000 greyscale image. After computing SVD decomposition of A, if we keep only three singular values from top, it means we keep the first three terms of the summation in the SVD formula and the approximation is not good enough. By adding more singular values from the top, clearly, we obtain a better representation from 7, 10, to 15, by keeping only 15 terms, the approximation is not too bad and the picture is recognizable. Moving on to 25 and then 50 terms, when we retain only 50 singular values, the approximation image is fairly good. If we decide to use that image with only 50 terms, then the required memory for it is much smaller than the original as it consists of only $50 * (1000 + 1000 + 1) = 50 * 2001 = 100{,}050$ values, since in each term we have two vectors u_i, v_i^T of length 1000 with a scalar σ_i. If we compare the required space to save 100,050 values with 1,000,000 values, there is a huge difference while we retain almost all the information in this compressed version. In this way, we can use only 100,050 numbers instead of using 1,000,000 numbers.

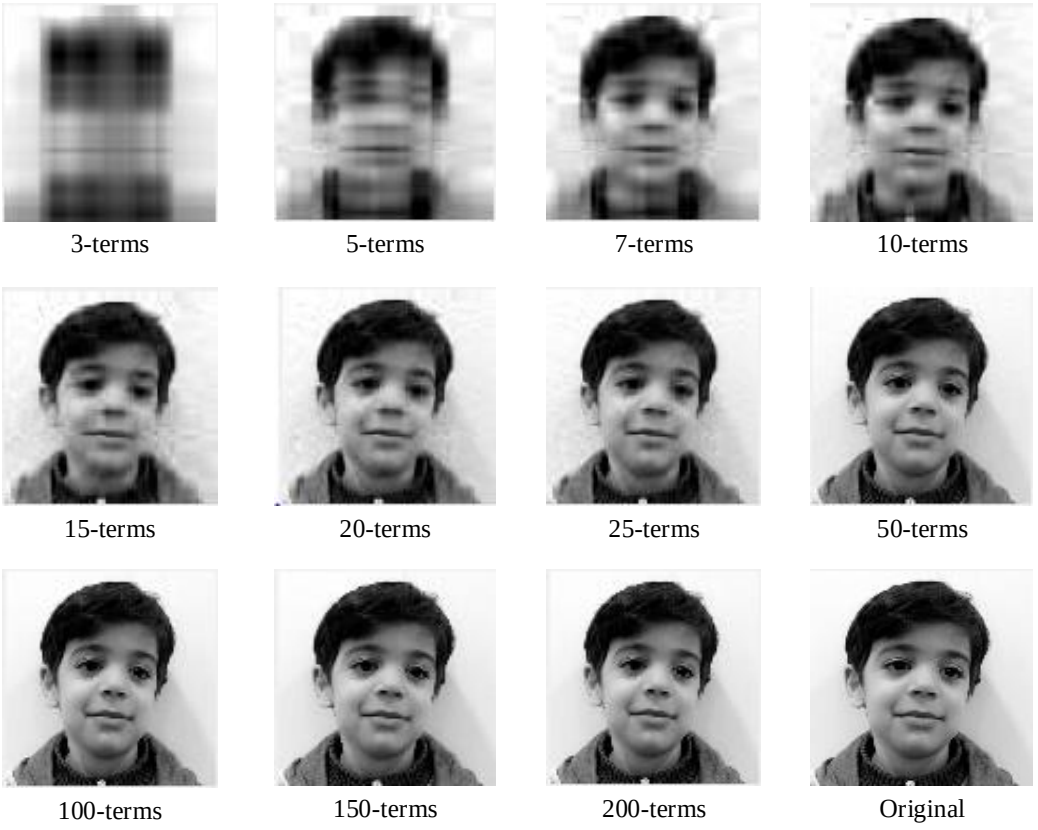


Figure 3-8 Image compression/reduction using SVD

3.6 CUR Decomposition

In matrix factorisation, the sparsity of the various factors is a useful property that help provide efficient computation. When we compute the SVD decomposition of a sparse matrix A, we obtain 2 components U and V^T that are generally dense and only the diagonal matrix $\sum$ is sparse. This lack of sparsity of the U and V^T factors is counted by some as one of the drawbacks of SVD. For this reason, another method of matrix decomposition has been developed (Drineas et al., 2008), which is much faster and easier than SVD to compute and it provides another simple DR technique. Here we shall briefly introduce this matrix factorisation.

CUR matrix Decomposition is another DR technique which is designed to maximize data sparsity in all its factor matrices. The goal of CUR is quite similar to SVD, for a given matrix A, CUR represents A as a product of three matrices $C, U \; and \; R$ while $C \; and \; R$ are alternatives to U and V^T, and CUR tries to make the reconstruction error $\|A - CUR\|_F$ as small as possible. Clearly, SVD provides optimal guarantee on the reconstruction error, so, CUR is expected to produce a greater error than SVD.

Definition (3.4) (Drineas et al., 2008)**:** Let A be a matrix of size $m \times n$, ***CUR-decomposition*** of A is an approximate matrix factorization $A' = CUR$, where C stands for column matrix and it is an actual subset of columns of A, R stands for row matrix and it is an actual subset of rows of A and U can be computed from $C\ and\ R$ as follows

C: a matrix of size $m \times c$, it consists of $(c < n)$-columns of A

R: a matrix of size $r \times n$, it consists of $(r < m)$-rows of A

U: is a matrix of size $c \times r$ and it is a pseudo inverse of the intersection of $C\ and\ R$.

Note: The pseudo inverse of any matrix B is denoted by $B^{\dagger}$ and computed as follows:

Given a matrix B of size $m \times n$ and compute its SVD decomposition

$$B = U\sum V^T = [u_1, u_2, \cdots, u_r] diag(\sigma_1, \sigma_2, \cdots, \sigma_r)[v_1, v_2, \cdots, v_r]^T$$

Then, the pseudo inverse of B which is denoted by $B^{\dagger}$ is defined using the orthogonality properties of U and V as follows:

$$B^{\dagger} = V\sum{}^{\dagger} U^T = [v_1, v_2, \cdots, v_r] diag(1/\sigma_1, 1/\sigma_2, \cdots, 1/\sigma_r)[u_1, u_2, \cdots, u_r]^T$$

Pseudo inverse is a special type of matrix inverse and it is one of the applications of SVD. Several Algorithms have been designed for this decomposition, some of the algorithms can be found in (Boutsidis and Woodruff, 2014) and one of the most compact versions of CUR is proposed in (Sun et al., 2007).

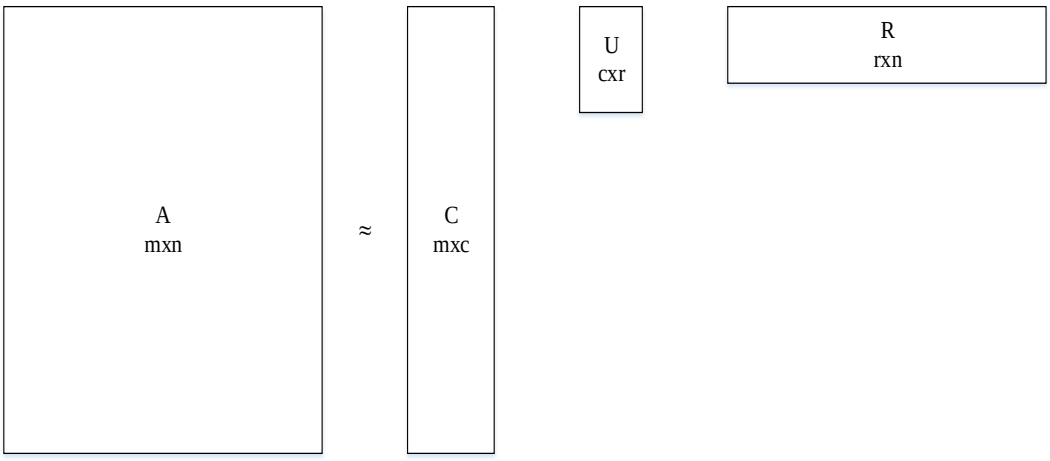


Figure 3-9 CUR Decomposition

3.7 Summary

In this chapter, we studied the theory of Eigenvalue problem with focus on important characteristics of eigenpairs of square real symmetric matrices and their role in data-dependent DR techniques. We critically investigated the two most widely used data-dependent DR schemes: PCA and LDA and highlighted their advantages and disadvantages. Having noted that the PCA has the effect of factorising the covariance matrix of the training dataset, we then studied the most relevant matrix factorisation that generalises and underpin the theory of PCA, namely the SVD matrix decomposition. We have studied data/image reduction/compression as an important application of the SVD method. Our investigation also covered the recently proposed CUR matrix decomposition which maintains sparsity in its factor and computationally cheaper than SVD.

All these methods, being defined in terms of a dataset of samples of digital representation of the objects under investigation do preserve the global information, relevant to the recognition task, that is conveyed by the training dataset. For example, the PCA captures the maximum variation between all pairs samples in the training set. In other word, there is no guarantee that the distances between every pair of samples are preserved before and after the projection. Accordingly, these data-dependent DR schemes are not JL compliant by design. This is the reason why recognition errors are dependent on the selected training dataset. In the next chapter, we shall focus on data-independent DR schemes that are designed to comply with JL theory. These include Discrete Wavelet transform (DWT) and Random Projections (RP). We shall study several examples of RP matrices and generate projection matrices from well-known Hadamard matrices.

CHAPTER FOUR: DATA-INDEPENDENT DIMENSION REDUCTION

The success of the data-dependent DR techniques, discussed in the previous chapter, for pattern recognition/classification applications depends on a number of factors. Firstly, the dataset of samples used for training the DR scheme need to be selected carefully to guarantee the widest representation of the typical objects of interest while taking into account all possible variants of vectors that model the same object/class and yet genuinely discriminates different classes. In many applications, the set of such objects is not easy to determine even when it is finite but large. Generally, one expect that the matching decision for any new sample would be more reliable when the sample is nearer to one of the training sample. Hence one may expect overfitting and biasness of the model to the training samples, so that a different training set may be less reliable. Moreover, the scalability of data-dependent DR techniques is not guaranteed when the population, of the objects of interest, expands by a large factor or by change of recording scenarios. A PCA system for face recognition that is trained for images captured in controlled scenarios and/or for a specific ethnic group may not perform as well when it is used to recognise people photographed in uncontrolled illumination/pose conditions. Therefore, DR techniques that are independent of training samples data are preferable. In this chapter, we review and investigate such dimension reduction schemes. Such techniques are to be based on reducing the dimension of individual biometric feature vectors independently of each other but using the same procedure. For example, down sampling/compressing face images by a fixed ratio can be considered as an independent DR that could be used for face recognition, where matching is done between down-sampled/compressed images. In fact, in this chapter we investigate DR schemes that are based on transforms that create or act on sparse biometric templates. For face images, frequency domain transforms such as wavelets and Discrete Cosine Transforms are suitable for these tasks. We shall first discuss the wavelet-based DR approach and will focus on using the emerging field of compressive sensing as a source of DR. In the last section, we focus on data-independent DR schemes that are based on the use of random submatrices of Hadamard matrices which are known to satisfy the Compressive sensing condition for unique recovery of sparse signals.

4.1 Discrete Wavelet Transform (DWT)

Wavelets are mathematical transformations that decompose a given signal/image hierarchically into its low and high frequency building blocks and DWT is a special case of wavelets (Al-Hassan, 2014). For an image I of size $m \times n$, there are exactly $m \times n$ wavelet coefficients that are divided into sub-bands of different frequency range. The original image can be reconstructed form Wavelet building blocks (sub-bands) without losing information as these transformations are invertible. In all but one sub-band, the majority of coefficients are very small, and hence these sub-bands can be made sparse by only considering a small number of significant coefficients to get a highly-compressed signal. Moreover, the inverse wavelet transformation of the compressed sub-band results in a very good quality signal that is almost indistinguishable from the original signal.

A wavelet function is a small waveform which unlike the trigonometric functions has its most energy concentrated in a small interval, called its support. However, like the trigonometric functions, any wavelet function W generates infinite versions (building blocks) of itself through a systematic scaling (usually by a factor of 2) and shifting by a fixed length (usually 1 unit). In this case, ψ is referred to as the mother wavelet. The process of scaling and shifting decomposes the space of all continuous square integrable signals $L^2(\mathbf{R})$, into a sequence of subspaces approaching $L^2(\mathbf{R})$. The initial subspace W_0 is generated by ψ and all its shifted copies. At the next stage, the scaled, by 2, version of ψ together with all its shifted copies generate a subspace W_{-1} of $L^2(\mathbf{R})$ and $W_{-1} = V_0 \oplus W_0$ where V_0 is the orthogonal complements of W_0. This process is repeated at infinitum and provides the Multi-resolution analysis of $L^2(\mathbf{R})$. Transforming a signal is done by repeatedly approximating the signal in terms of the generators of the V_i and the W_i using inner product of the output from previous stage. Therefore, we use the mother wavelet and its orthogonal complement as a filter bank.

There are various multi-resolution schemes that use Wavelet transforms to decompose a signal/image. The most commonly used scheme is the Pyramid scheme, which when applied on raw image I we get four wavelet sub-bands (LL, HL, LH, HH). At the second and other levels, we apply Wavelet again on the LL sub-band only to get the second level of decomposition. This process could continue as shown in figure (4.1). Therefore, the pyramid scheme decomposes the Image I at level q into $3q + 1$ sub-bands.

The Haar Wavelet is the simplest example of Discrete Wavelet transform (DWT) (Abdulla, 2007), and that is why we adopted it in this thesis. This transformation is linear and orthonormal and the filtering can be expressed in a matrix form:

$$H = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

Simply, it consists of two operators (sums and differences). For an image I and two adjacent pixel values ($p_1\ and\ p_2$), it is computed as follow:

$$[y_1, y_2] = H([p_1, p_2]) = \frac{1}{\sqrt{2}}[p_1, p_2] * \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

$$y_1 = \frac{1}{\sqrt{2}}(p_1 + p_2), \quad (Low\ Frequency)$$

$$y_2 = \frac{1}{\sqrt{2}}(p_1 - p_2), \quad (High\ Frequency)$$

For mxn images, it is customary to apply the wavelet transform in two stages of encoding (decomposing) at any level: In the first stage, we apply Haar Wavelet on horizontally adjacent pixels which will create 2 vertical sub-bands of size mx(n/2) coefficients. The left-hand block contains the low frequencies (sums) and it is called approximation sub-band/low sub-band, and the right-hand block contains high frequencies (differences) and it is called detailed sub-band/ high sub-band. We denote these two sub-bands as $L\ and\ H$. In the second stage, we apply the Haar wavelet on vertically adjacent pixel values within each of L and H sub-bands resulting in four sub-bands (LL, HL, LH, HH) at the first level of resolution. In the next level, the above process is repeated only on LL-sub-bands. At any resolution level, LL-sub-band approximates the original Image I. Other sub-bands, $HL, LH\ and\ HH$ maintain vertical, horizontal and diagonal texture components in the original Image I.

Back to our objective, Dimensionality reduction, as we explained that Wavelet is a multi-resolution signal analysis technique. It also can be considered as an effective DR technique. Each of the different wavelet sub-bands, at different levels, provide different feature vectors representation of the image with lower resolution, i.e. reduced dimension. In figure 4-1. The original Image is 512x512, if we converted to a vector by row concatenation, it becomes a vector in 262144-dimensional space which is very high. After applying Haar Wavelet at the first level, we obtain four Wavelet sub-bands (LL_1, HL_1, LH_1, HH_1) each of size 256x256. Each sub-band maintain important

information about the original image and it can be considered as a reduced version. In this level, each sub-band is a vector of 65536 dimension, the dimension is reduced from 262144 to 65536. At the second level, after applying Haar Wavelet on LL_1, (LL_2, HL_2, LH_2, HH_2) is obtained with each of size 128x128 and each sub-band can be represented as a 16384-dimensional vector. If we apply Haar Wavelet on LL_2 at third level. It will produce (LL_3, HL_3, LH_3, HH_3) each of size 64x64 which means a vector in 4096-dimensional space. In this way, a drastic dimension reduction can be done by moving from a level to the next one while each sub-band contains important features/information of the original input image. So, Wavelet reduce the dimension of image data with preserving different Features in different sub-band at different resolution.

Note: The amount of reduction in dimension achieved by the DWT at different resolution depth, increases the deeper one resolves the image. Finally, we observe that this process is applied to any image without depending on other images and the amount of reduction achieved by any of sub-band does not depend on what is in the image but on the image size. Moreover, there is another source to further reduce the dimension in each of the non-LL-sub-bands, because the majority of coefficients in such sub-bands are nearly 0, and if we set all these small coefficients then we get a sparse representation which is exploited in image compression. However, the positions of the significant non-zero coefficients in the non-LL sub-band are not easy to determine. The concept of compressive sensing, to be discussed later, is based on designing certain types of random matrices that can be used to project the image itself or its wavelet sub-bands directly onto to the significant coefficients only. The rest of the chapter, is devoted to investigating this kind of data-independent approach to DR.

Original Image

Figure 4-1 Pyramid Wavelet Transform for 1st, 2nd, and 3rd levels

4.2 Random Projections (RP)

Random projection, is a powerful linear DR tool which does not distort local properties significantly and at the same time, RP is totally independent from training data samples as the projection matrix is constructed independently. Papadimitriou says in his foreword in (Vempala, 2004) "if distance is all you care about, there is no reason to stay in high dimension". It is often remarked that, unlike the data-dependent DR's, random projections do not benefit from good dataset and it is not affected by bad datasets. Theoretical results demonstrate that, under certain conditions, there exist transformations on the Euclidian spaces whose range is a lower dimensional subspace that "preserve" pairwise distances within a relatively small error with *high probability* (Dasgupta and Gupta, 2003), (The term *high probability* here means that the chance of preserving pairwise distances is very high, i.e. it is highly possible to preserve the distances between almost all the points, and this term will be repeatedly used in the rest of the thesis). These conditions are related to the desirable value of the reduced dimension and on the error tolerance level. The JL lemma as stated in section (2.5) shows that for any set A of n points in any Euclidean space, there is a map to embed A into a Euclidean space of dimension $k \geq O(\epsilon^{-2} \log n)$ while it guarantees that this function does not distort pairwise distances by more than a factor $(1 \pm \epsilon)$ with a good probability (Johnson and Lindenstrauss, 1984). In this lemma, the original dimension of A is not directly involved but the value of k depends only on ϵ and the number n of points in set A. This means that for any n points in high dimensional space whether the number of dimension is hundreds or thousands, such a map exists. However, the stated lower bound on the value of k ensures that one is dealing with sufficiently dense set in the high dimension of the points.

Interestingly, the lemma says, such a map exists for any dataset, no matter how the data records are distributed or convoluted. This is one of the most important properties of random projections. However, it does not mean that a random projection matrix is suitable for every dataset and any application. Nonetheless, this property is quite useful while it is not easy to achieve with other data-dependent DR techniques like PCA and LDA, as mentioned before, the distribution of a given data set affect the performance of the PCA and it can make PCA success or fail. In fact, the practice of selecting the number of significant eigenvalue in the PCA scheme is normally linked with the given dataset which may not control the tolerable error. In fact, the number of significant eigenvectors are not linked to guaranteeing the preservation of distances between all pairs of point after the

projection, i.e. PCA is tolerance of existing anomalies representing some points that have large projection along the non-significant eigenvectors.

The JL theorem states that the lowest reduced dimension $k = O(\epsilon^{-2} \log n)$ in order to preserve pairwise distances within a small relative error $(1 \pm \epsilon)$. The number k must be sufficiently large to guarantee the above statement. In practice, we need to know what is the value of $O(\epsilon^{-2} \log n)$. There are some simplifications of the original proof of this theorem that also provide different lower bound and culminating in Dasgupta and Gupta's work that provides a more specific lower-bound in their theorem as follows:

Theorem (4.1) (Dasgupta and Gupta, 2003) : For any set A of n points in $\mathbb{R}^N$, and any $0 < \epsilon < 1$, there is a function $f: \mathbb{R}^N \to \mathbb{R}^k$, with

$$k \geq \frac{4}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \ln(n) = \frac{24}{\epsilon^2 \ (3 - 2 \ \epsilon)} \ln(n)$$

such that for any two points $a, b \in A$

$$(1 - \epsilon)\|a - b\|_2 \leq \|f(a) - f(b)\|_2 \leq (1 + \epsilon)\|a - b\|_2$$

A very interesting computational consequence of using random projections that are compatible with the conditions, stated in the above theorems, relates to efficiency. The construction of random projection matrices, in accordance with the JL theorem is totally independent from any training dataset in contrast of data-dependent techniques such as PCA, and LDA. Thus, this technique is computationally very efficient and its computational complexity is just $O(nNk)$ which is the cost of a matrix multiplication while the computational cost of PCA is $O(N^2 n) + O(N^3)$ (Bingham and Mannila, 2001). Here, N is the dimension of the original space, while k is the desired number of reduced dimension.

Finally, we observe that the distance inequality condition in the J-L theorem in the case when we are dealing with sparse high dimensional feature vectors have a significant relevance to the concept of compressive sensing discussed next.

4.3 Compressive Sensing (CS)

Traditional methods of signal acquisition follow Shannon's theorem which states that: to avoid losing information in the process of capturing a signal and guarantee a perfect signal recovery, one must sample the signal at a rate, which is known as Nyquist-rate, two times faster than the signal bandwidth, i.e. the sampling rate, must be greater than or equal to

the highest frequency of the signal (Candes and Wakin, 2008), (Baraniuk, 2007). In many applications, such as digital images and videos, the Nyquist-rate is very high and commercial devices cannot acquire samples at this rate. Compression techniques become necessary to obtain a sparse but very concise signal representation of the signal.

Transform coding techniques, such as JPEG and JPEG2000, find a proper basis for the signal that provides a sparse or a compressible representation for the signal, i.e. in terms of such basis, the signal has a few large coefficients and the rest are small and close to zero. Sparse representation is obtained by preserving the value and location of k largest coefficients and setting the rest $N - k$ coefficients to zero ($k \ll N$) without losing too much information. So, traditional protocols of signal acquisition sample at a rate which is very high to produce all the data and then most of it will be thrown away in the process of data compression. Now, the question is that, is it possible to directly acquire a compressed form of a signal which maintain the important part of the data without going through the above stages? This is the question that Compressive sensing tries to answer it.

The concept of ***compressive sensing*** (CS) was first introduced by David Donoho (Donoho, 2006) as a new paradigm of signal acquisition which relaxes significantly the Nyquist-Shannon sampling condition while facilitating the recovery of good quality digital signal. It relies on two fundamental premises: Sparsity and Incoherence. The first concept pertains to the signal of interest and the latter one is related with the sensing method. In fact, most of the signals involved in pattern recognition are sparse or compressible when represented in terms of proper bases. The CS paradigm states that in the case of sampling a sparse signal, the number of measurements needed to be collected can be extremely reduced compared to the number of required samples that suggested in the Shannon-Nyquist sampling theorem. The trick in this case is really to take measurements/meta-features that are linear combinations of the raw features. In short, compressive sensing provide a new source of dimension reduction by projecting onto a multiple basis rather than a single basis, using what mxN matrices (with m<<N) that satisfy similar properties to JL conditions but only when applied to sparse (or nearly sparse) signals. Such matrices are referred to as CS dictionaries. Here we shall give a brief introduction, but the reader is advised to consult with (Candes and Tao, 2005, and Donoho, 2006).

4.3.1 The sensing/sampling problem

Compressive Sensing (CS) simply correlate the signal of interest $x \in \mathbb{R}^N$ with a small number of non-adaptive linear measurements m $(m \ll N)$ which can be arranged as rows of an overcomplete matrix dictionary say $D_{m\times N}$, such a matrix is fixed and independent from the signal. The vector $y \in \mathbb{R}^m$ consists of the sampled values, (i.e. $(m \ll N)$ inner products between x and D), this process can be written as

$$y_{m\times 1} = D_{m\times N} * x_{N\times 1}$$

The main challenges in the sensing problem are: First, designing a stable dictionary that "preserves" the information/length of the $k-$sparse signals $(k \leq M)$ with a good probability while it reduces the dimension $(D: \mathbb{R}^N \rightarrow \mathbb{R}^m)$, such dictionary must allow us to reconstruct the full-length signal x from only m measurements y. Second, an algorithm to accurately recover the signal x.

4.3.2 Restricted Isometry Property (RIP)

An $m\times N$ $(m \ll N)$ dictionary D is said to have the ***Restricted Isometry Property*** (RIP) of order k if there exists a constant $0 < \delta_k < 1$, such that for any k-sparse vector/signal $x \in \mathbb{R}^N$:

$$(1 - \delta_k)\|x\|_2 \leq \|Dx\|_2 \leq (1 + \delta_k)\|x\|_2$$

Also, the smallest constant δ_k is defined as ***Restricted Isometry Constant*** (RIC) of order k. If D is a dictionary that satisfies RIP condition with the RIC δ_{2k}, and $x, y \in \mathbb{R}^N$ be any two $k-$sparse vectors then projection by this dictionary defines a Random projection that preserves, up to a tolerance error, the distances between pairs k-sparse vectors as follows:

$$(1 - \delta_{2k})\|x - y\|_2 \leq \|Dx - Dy\|_2 \leq (1 + \delta_{2k})\|x - y\|_2$$

Candes (Candès, 2008) proved that if a dictionary D satisfies RIP condition with the RIC $\delta_{2k} < \sqrt{2} - 1$, then the equation $y = Dx$ can be uniquely solved by $l_1 - minimization.$ for the sparsest solution.

Another criterion that guarantees unique recovery of the sparsest solution of $y = Dx$ by $l_1 - minimization$ is the Null Space Property (NSP) which imposes bounds on the $l_1 - norm$ of every set of k non-trivial vectors in the kernel of D. For details see (Candes and Tao, 2006) and (Rubinstein et al., 2010).

Both RIP and NSP are difficult to test directly for high dimensional matrices. However, a number of algebraic criteria have been developed that can be used to test the suitability

of an mxN matrix D (with m<<N) to act as a CS dictionary. These conditions are mostly associated with the ability to recovering the unique sparse solution based on L1-minimization. Sufficient, but not necessary, conditions include (1) coherence value of D defined as the largest absolute normalized inner product of pairs of columns of D. It has also been shown that if D satisfies RIP of order k, then every 2k-columns submatrix of D must be well-conditioned (i.e. the condition number CN = ratio of its maximum to its minimum singular values need to be < 2.5). Another indicator of RIP-compliance is the spark of D defined as the minimum number of linearly dependent columns. Clearly, spark(D)≤ m+1. Equality occurs when D has a full row rank. For more details see (Baraniuk et al., 2008, Rauhut, 2010, and Chen and Dongarra, 2005).

4.3.3 The Relation Between RIP and JL conditions

The JL theorem and its modified versions apply to any set A of n-points in $\mathbb{R}^N$, and guarantee the existence of a function f that transforms A into a lower dimensional subspace with at least $O(\epsilon^{-2}\log n)$ dimensions while approximately preserves distances, with a small relative error between any two vectors $x, y \in \mathbb{R}^N$. However, the RIP condition is very similar to that of the JL condition in preserving distances but only between vectors of k-sparsity. However, the RIP is independent of any training set of samples.

The JL condition applies whether the vectors are sparse or not, while with RIP condition are concerned with sparse vectors only. So, we can say that the RIP is a special case of the JL conditions. Linear transformations that satisfy the JL theorem are good candidates for CS applications and for data-independent DR. We shall next identify some classes of matrices, known for their suitability for CS dictionaries, as DR tools.

We first begin by investigating several examples of random sensing matrices generated by known probability density functions. These classes of matrices are predicted to satisfy RIP condition with high probability. The generation of these classes is guided by the Hecht-Nelson probabilistic assertion that as we go to high dimension, the number of nearly orthogonal directions increases, and hence the chance of picking a set of almost orthogonal is very high in high dimensional space (Hecht-Nielsen, 1994).

4.3.4 A Selection of CS dictionaries

1. Gaussian Random Matrix

Gaussian Random Matrix with normalized columns of size $(m\times n, m \ll n)$ is one of the most widely used RIP dictionaries in CS applications, entries of this dictionary are independently identically distributed from the Gaussian distribution with mean 0 and variance $1/m$ which is denoted by $x_{i,j}\sim N(0,1/m)$ (Al-Hassan, 2014).

2. Achlioptas Matrices

Achlioptas (Achlioptas, 2001) replaced the normal Gaussian entries $N(0,1)$ by either one of these two probability distributions: $\{1,0,-1\}$ with probabilities $\left\{\frac{1}{6},\frac{2}{3},\frac{1}{6}\right\}$ or $\{1,-1\}$ with probabilities $\left\{\frac{1}{2},\frac{1}{2}\right\}$

$$r_{ij}=\sqrt{3}\times\begin{cases}+1 & p=1/6\\ 0 & p=2/3\\ -1 & p=1/6\end{cases}$$

$$r_{ij}=\begin{cases}+1 & p=1/2\\ -1 & p=1/2\end{cases}$$

Li et al. (Li et al., 2006) proposed a generalized form of Achlioptas matrices as follows, where Achlioptas matrices are special cases for $s=1,3$

$$r_{ij}=\sqrt{s}\times\begin{cases}+1 & p=1/2s\\ 0 & p=1-1/s\\ -1 & p=1/2s\end{cases}$$

3. Bernoulli Random matrix

Bernoulli Random Matrix of size $(m\times n, m \ll n)$ is another example of random RIP matrix, its entries are identically independently distributed from the following distribution:

$$r_{ij}=\begin{cases}+1/\sqrt{m} & p=1/2\\ -1/\sqrt{m} & p=1/2\end{cases}$$

4. Semi-Structured Circulant (C) and Toeplitz matrices

Random dictionaries like Gaussian and Bernoulli random matrices guarantee sparse recovery using $l_1-minimization$ with near optimal required number of measurements. However, using structured dictionaries in some applications helps to raise the speed of sparse recovery algorithm significantly (Rauhut, 2010). The circulant matrix is one of the

widely-used matrices in CS applications, each row of this matrix is the right cyclic shift of the previous row and the Toeplitz matrix is known as a submatrix of circulant and each Toeplitz matrix can be embedded in a circulant matrix, every left to right descending diagonal is a fixed constant in both types as shown in figure 4-2.

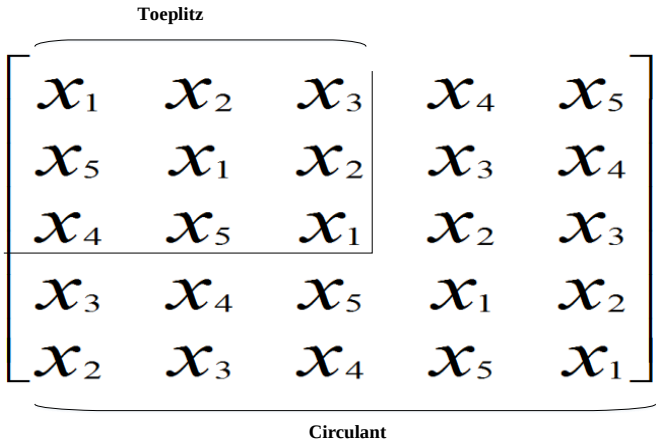


Figure 4-2 Circulant and Toeplitz Matrix

The overcomplete dictionary $(m \times n, m \ll n)$ is created ether by, creating the first row using a random distribution such as standard Gaussian $N(0,1)$ and generating other rows by shifting iteratively. Alternatively, one can generate the full size circulant matrix and select a subset of m −rows randomly.

Remarks: The simplicity of generating the above random matrices, comes at a price of less than adequate efficiency of the projection procedure. Moreover, the Hecht-Nielsen probabilistic assertion may not be valid for the lower range of high dimension and in any case random generation procedures may not be successful all the time. Therefore, many DR schemes that are based on Gaussian random matrices start by generating m pseudo random vectors in $\mathbb{R}^N$, then, these rows are converted to a set of orthonormal vectors using Gram-Schmidt process to obtain an orthonormal projection matrix.

4.4 Gram-Schmidt (GS) Process

Let $\{v_1, v_2, \cdots, v_k\}$ be a basis of a subspace V in $\mathbb{R}^n$, from this set, we can find an orthonormal basis $\{u_1, u_2, \cdots, u_k\}$ for V using the following algorithm which is called Gram-Schmidt process:

$$w_1 = v_1 \quad \rightarrow u_1 = \frac{w_1}{\|w_1\|}$$

$$w_2 = v_2 - \frac{v_2 . w_1}{w_1 . w_1} w_1 \quad \rightarrow u_2 = \frac{w_2}{\|w_2\|}$$

$$w_3 = v_3 - \frac{v_3 . w_1}{w_1 . w_1} w_1 - \frac{v_3 . w_2}{w_2 . w_2} w_2 \quad \rightarrow u_3 = \frac{w_3}{\|w_3\|}$$

$$\vdots$$

$$w_k = v_k - \sum_{i=1}^{k-1} \frac{v_k . w_i}{w_i . w_i} w_i \quad \rightarrow u_k = \frac{w_k}{\|w_k\|}$$

For example, let $v_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}$ and $V = span\{v_1, v_2\}$. We can find an orthonormal basis of V using Gram-Schmidt Process as follows:

$$w_1 = v_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \quad \rightarrow u_1 = \frac{w_1}{\|w_1\|} = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$w_2 = v_2 - \frac{v_2 . w_1}{w_1 . w_1} w_1 = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix} - \frac{4}{2} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} \rightarrow u_2 = \frac{w_2}{\|w_2\|} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Now, u_1, u_2 form an orthonormal basis of V.

Unfortunately, this procedure has some drawbacks and it is not an easy task in practice. Firstly, there is no absolute guarantee on the m pseudo random vectors to be linearly independent which is essential for GS. Secondly, GS algorithm is very high demanding and not stable (Jassim et al., 2009), (Bingham and Mannila, 2001). In fact, (Jassim et al., 2009) deviate from this naive RP generation and does not use the Gram-Schmidt orthogonalizing but adopt an efficient and stable method. This scheme, uses block diagonal matrices using a number of known small size orthonormal square matrices. It exploits the fact that small size orthonormal matrices can be generated from known

rotation/reflection matrices. For example, for any θ the following rotation matrices define orthonormal projections of the 2-dimensional plane (resp. the 3-dimensional Cartesian space):

$$R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \qquad R'_\theta = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(Jassim et al., 2009) scheme for RP generation of size 2n*x*2n, simply generates a random sequence of angles $\{\theta_1, \theta_2, \cdots, \theta_n\}$ and use the corresponding 2x2 rotation matrices $R_{\theta_i}{}'s$ to construct the following block diagonal matrix:

$$A = \begin{pmatrix} R_{\theta_1} & 0 & \cdots & 0 \\ 0 & R_{\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{\theta_n} \end{pmatrix}$$

For feature vector of odd size (2n+1) simply select an i and replace $R_{\theta_i}$with R'_{θ_i}.

This scheme is efficient and unlike the above naive scheme it has been shown to be computationally stable. Any user can change at will the randomly generated sequence $\{\theta_1, \theta_2, \cdots, \theta_n\}$, and thereby this procedure generates cancellable biometric templates (or biometric feature vectors). Moreover, these RP matrices being highly sparse make the process of transforming biometric templates extremely efficient.

The impact of using these random sensing matrices, whichever way generated, have been positive and yielding good performances in different applications. Next, we shall focus on constructing projection matrices from different types of Hadamard matrices and their impact will be tested in the remaining chapters of this thesis for different pattern recognition case studies.

4.5 Overcomplete Hadamard Submatrices

Hadamard Matrices are square and simple structured matrices, their entries are $+1$ or -1 and any two-distinct row/column vectors are mutually perpendicular (Agaian, 2011). Due to its simplicity and efficiency, it is found in several applications such as: Digital signal and image processing, combinatorial designs, quantum computing, physics, chemistry, etc. In relation to our objective, Random Projections and Dimensionality Reduction, these matrices provide a very interesting class of matrices consisting of orthogonal direction vectors, from which we can construct a variety of overcomplete Hadamard submatrices for dimensionality reduction simply by randomly selecting sufficient number of rows in terms of compatibility with J-L and RIP conditions.

Definition: A square matrix H of order N with entries $1\ and\ -1$ is called ***Hadamard*** matrix if it satisfies the following equation:

$$H_N H_N^T = N I_N = H_N^T H_N$$

Where H_N is a Hadamard matrix of size $N \times N$, H_N^T is the transpose of H_N and I_N is the NxN identity matrix. Obviously, the square matrix $\frac{1}{\sqrt{N}} H_N$ is an orthogonal matrix.

There are few different ways to construct such matrices, we shall explain three methods of constructing Hadamard matrices with examples and illustrative display format. We need to define a special type of matrix operation, the so called Kronecker/Tensor Product.

Definition: Let $A_{m\times n}\ and\ B_{p\times q}$ be two matrices, the ***kronecker product*** of $A\ and\ B$ is a matrix C of size $mp \times nq$ which is computed using this formula:

$$C = A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

For example, Let $A = \begin{bmatrix} 1 & -2 \\ 0 & 0.5 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 4 & -6 \\ 1 & 0 & 7 \end{bmatrix}$

$$C = A \otimes B = \begin{bmatrix} 1B & -2B \\ 0B & 0.5B \end{bmatrix} = \begin{bmatrix} 2 & 4 & -6 & -4 & -8 & 12 \\ 1 & 0 & 7 & -2 & 0 & -14 \\ 0 & 0 & 0 & 1 & 2 & -3 \\ 0 & 0 & 0 & 0.5 & 0 & 3.5 \end{bmatrix}$$

4.5.1 Sylvester-type Hadamard Matrices (SH)

The Sylvester construction method is a successive method of creating Sylvester-type Hadamard matrices of order N where $N = 2^n, n$ is a positive integer by using the following formula:

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} + & + \\ + & - \end{bmatrix}$$

Where $\pm\ stands\ for \pm 1$ respectively. This will be used later to display Hadamard matrices as a binary image, + and - replace with black and white pixel respectively.

$$H_N = H_{2^n} = H_2 \otimes H_2 \otimes \cdots \otimes H_2 = \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} + & + \\ + & - \end{bmatrix}$$

For example, by using this formula, we can create H_4 from H_2.

$$H_4 = H_2 \otimes H_2 = \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \begin{bmatrix} + & + \\ + & - \end{bmatrix} = \begin{bmatrix} + & + & + & + \\ + & - & + & - \\ + & + & - & - \\ + & - & - & + \end{bmatrix}$$

Recurrently, we can construct H_8 from H_2 and H_4.

$$H_8 = H_2 \otimes H_2 \otimes H_2 = H_2 \otimes H_4 = \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \begin{bmatrix} + & + & + & + \\ + & - & + & - \\ + & + & - & - \\ + & - & - & + \end{bmatrix}$$

$$= \begin{bmatrix} + & + & + & + & + & + & + & + \\ + & - & + & - & + & - & + & - \\ + & + & - & - & + & + & - & - \\ + & - & - & + & + & - & - & + \\ + & + & + & + & - & - & - & - \\ + & - & + & - & - & + & - & + \\ + & + & - & - & - & - & + & + \\ + & - & - & + & - & + & + & - \end{bmatrix}$$

Iterating this process yield Hadamard matrices of order 16, 32, and 64 or above.

Equivalently, the above formula can be written in this way,

$$H_N = H_{2^n} = H_2 \otimes H_2 \otimes \cdots \otimes H_2 = H_2 \otimes H_{2^{n-1}} = \begin{bmatrix} H_{2^{n-1}} & H_{2^{n-1}} \\ H_{2^{n-1}} & -H_{2^{n-1}} \end{bmatrix}$$

It means, a Hadamard matrix of order 2^n can be computed from a Hadamard matrix of order 2^{n-1} simply by collocating for copies of it in a 4×4 block and negating one of them. Figure 4-3 is the picture of Sylvester-type Hadamard matrices.

In generally, an entry of Sylvester-type Hadamard matrix at the position (j, k) can be computed individually by using this formula:

$$SH(j,k) = (-1)^{\sum_{i=0}^{n-1}(j_i k_i)}$$

Where $j_i \; and \; k_i$ are the i^{th} bits in the binary representations of $j \; and \; k$ respectively.

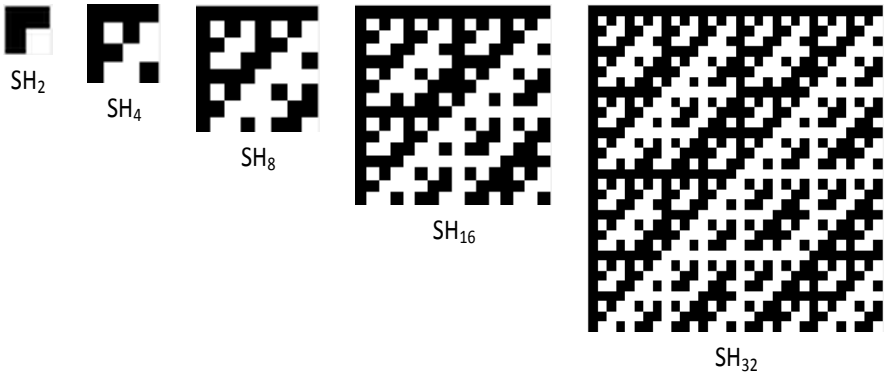


Figure 4-3 Binary display of Sylvester-type Hadamard Matrices of order 2, 4, 8, 16, and 32.

4.5.2 Walsh-Paley Matrices (WP)

Walsh-Paley construction method differs from the Sylvester construction by an iterative procedure that depend on a different tensor product of different constituent matrices. To create the Walsh-Paley Hadamard matrices of order $N\ where\ N = 2^n\ for\ n = 1,2,3,\cdots$ we use the following recursive formula:

$$WP_1 = [1], \qquad and\ \ for\ N = 2^n\ \ (n = 1,2,3,\cdots) \quad WP_N = \begin{bmatrix} WP_{\frac{N}{2}} \otimes [1 & 1] \\ WP_{\frac{N}{2}} \otimes [1 & -1] \end{bmatrix}$$

Hence,

$$WP_2 = \begin{bmatrix} WP_1 \otimes [1 & 1] \\ WP_1 \otimes [1 & -1] \end{bmatrix} = \begin{bmatrix} [1] \otimes [1 & 1] \\ [1] \otimes [1 & -1] \end{bmatrix} = \begin{bmatrix} + & + \\ + & - \end{bmatrix}$$

$$WP_4 = \begin{bmatrix} WP_2 \otimes [+ & +] \\ WP_2 \otimes [+ & -] \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes [+ & +] \\ \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes [+ & -] \end{bmatrix} = \begin{bmatrix} + & + & + & + \\ + & + & - & - \\ + & - & + & - \\ + & - & - & + \end{bmatrix}$$

$$WP_8 = \begin{bmatrix} WP_4 \otimes [+ & +] \\ WP_4 \otimes [+ & -] \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} + & + & + & + \\ + & + & - & - \\ + & - & + & - \\ + & - & - & + \end{bmatrix} \otimes [+ & +] \\ \begin{bmatrix} + & + & + & + \\ + & + & - & - \\ + & - & + & - \\ + & - & - & + \end{bmatrix} \otimes [+ & -] \end{bmatrix}$$

$$= \begin{bmatrix} + & + & + & + & + & + & + & + \\ + & + & + & + & - & - & - & - \\ + & + & - & - & + & + & - & - \\ + & + & - & - & - & - & + & + \\ + & - & + & - & + & - & + & - \\ + & - & + & - & - & + & - & + \\ + & - & - & + & + & - & - & + \\ + & - & - & + & - & + & + & - \end{bmatrix}$$

Generally, an entry at the position (j, k) of Walsh-Paley matrix can be expressed as

$$WP(j,k) = (-1)^{\sum_{i=0}^{n-1}(k_{n-i}+k_{n-i-1})j_i}$$

Where $j_i \; and \; k_i$ are the i^{th} bits in the binary representations of $j \; and \; k$ respectively.

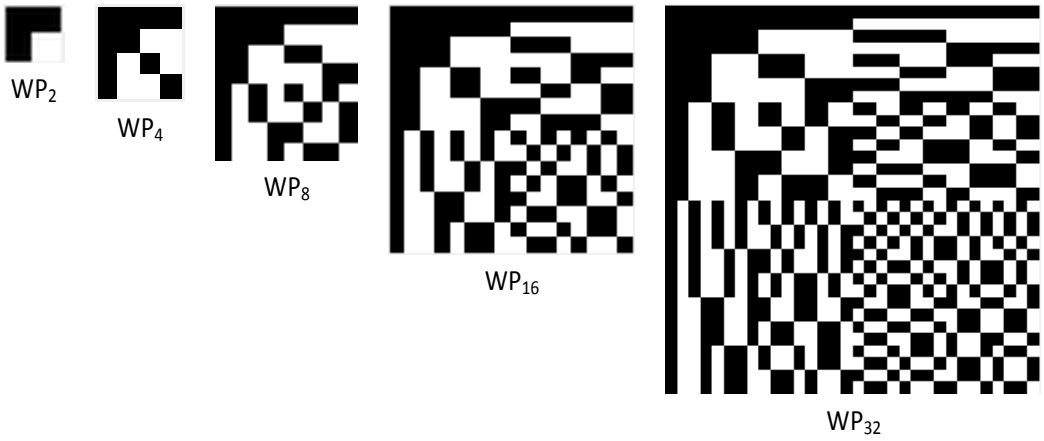


Figure 4-4 Binary display of Walsh-Paley Matrices of order 2, 4, 8, 16, and 32

4.5.3 Walsh Matrices (W)

The Walsh Hadamard matrices are constructed with yet a third recursive formula:

$$W_N = \left[W_2 \otimes A_1, Q \otimes A_2, \cdots, W_2 \otimes A_{\left(\frac{N}{2}\right)-1}, Q \otimes A_{\left(\frac{N}{2}\right)}\right]$$

$$Where \; N = 2^n \; for \; n = 1,2,3, \cdots and \; W_2 = \begin{bmatrix} + & + \\ + & - \end{bmatrix}, Q = \begin{bmatrix} + & + \\ - & + \end{bmatrix} \; and$$

$$A_i \; is \; the \; i^{th} \; column \; of \; a \; walsh \; matrix \; of \; order \frac{N}{2}$$

For example,

$$W_4 = [W_2 \otimes A_1, Q \otimes A_2] = \left[\begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \begin{bmatrix} + \\ + \end{bmatrix}, \begin{bmatrix} + & + \\ - & + \end{bmatrix} \otimes \begin{bmatrix} + \\ - \end{bmatrix}\right] = \begin{bmatrix} + & + & + & + \\ + & + & - & - \\ + & - & - & + \\ + & - & + & - \end{bmatrix}$$

$$W_8 = [W_2 \otimes A_1, Q \otimes A_2, W_2 \otimes A_3, Q \otimes A_4]$$

$$= \left[\begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \begin{bmatrix} + \\ + \\ + \\ + \end{bmatrix}, \begin{bmatrix} + & + \\ - & + \end{bmatrix} \otimes \begin{bmatrix} + \\ + \\ - \\ - \end{bmatrix}, \begin{bmatrix} + & + \\ + & - \end{bmatrix} \otimes \begin{bmatrix} + \\ - \\ - \\ + \end{bmatrix}, \begin{bmatrix} + & + \\ - & + \end{bmatrix} \otimes \begin{bmatrix} + \\ - \\ + \\ - \end{bmatrix},\right]$$

$$= \begin{bmatrix} + & + & + & + & + & + & + & + \\ + & + & + & + & - & - & - & - \\ + & + & - & - & - & - & + & + \\ + & + & - & - & + & + & - & - \\ + & - & - & + & + & - & - & + \\ + & - & - & + & - & + & + & - \\ + & - & + & - & - & + & - & + \\ + & - & + & - & + & - & + & - \end{bmatrix}$$

Generally, the (j, k) entry of a Walsh Hadamard matrix a can be computed as follows

$$W(j,k) = (-1)^{\sum_{i=0}^{n-1}(j_{n-i-1}+j_{n-i})k_i}$$

Where $j_i \; and \; k_i$ are the i^{th} bits in the binary representations of $j \; and \; k$ respectively.

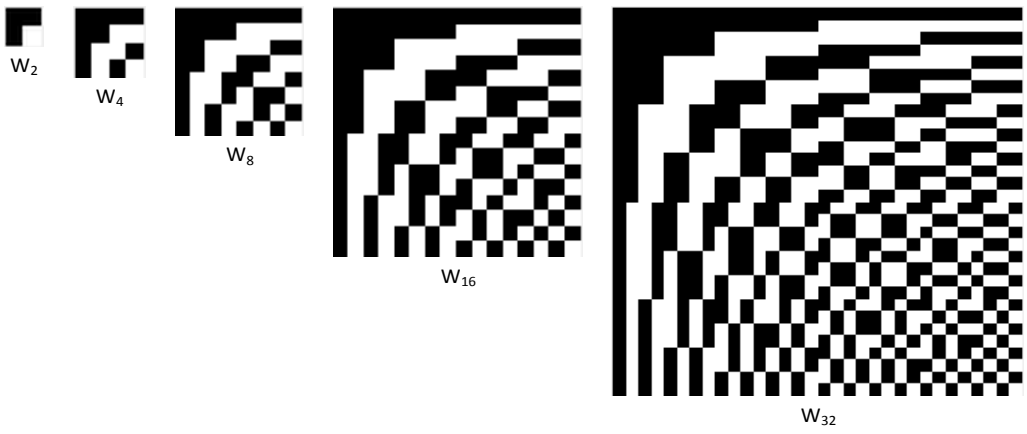


Figure 4-5 Binary display of Walsh Matrices of order 2, 4, 8, 16, and 32

Some Properties of Hadamard Matrices

Let H_N be a Hadamard matrix of order N

1. $H_N * H_N^T = NI_N = H_N^T * H_N$ where I_N is the identity matrix of the same order, this means, $\frac{1}{\sqrt{N}} H_N$ form an orthonormal matrix,

$$\frac{1}{\sqrt{N}} H_N * \frac{1}{\sqrt{N}} H_N^T = I_N = \frac{1}{\sqrt{N}} H_N^T * \frac{1}{\sqrt{N}} H_N$$

2. $|\det(H_N)| = N^{\frac{1}{2}N}$

3. Hadamard matrices can be changed into other Hadamard matrices by multiplying rows and column by -1 and by permuting rows and columns

4.5.4 Generating Over-complete Hadamard submatrices (Projection Matrices)

The above approaches are designed to construct different types of square Hadamard matrices. In this section, we explain how to generate different types of random overcomplete mxN dictionaries to be used as RP matrix for dimension reduction. In particular, we describe only three types: Fully Random, Semi Random and Structured over-complete dictionaries using the various Hadamard square matrices of order $N = 2^n$.

(1) Fully random over-complete dictionaries (RD): The m rows of these projection matrices are selected randomly without repetition from a selected Hadamard matrix of order N but we construct RD only from the SH matrices.

(2) Semi-random over-complete dictionaries (SRD): The m rows of these matrices are divided into two nearly equal size. The first part at the top of the projection matrix are the top rows of the selected Hadamard matrix while the other rows are randomly sampled without repetition from the rest of the same Hadamard matrix, in this thesis, we construct SRD from WP.

(3) Structured over-complete dictionaries(SD): The m rows of these projection matrices are the top m rows of the selected Hadamard matrix, but we construct SD only from the WP constructed matrices in this thesis.

Figure 4-6, below displays the binary illustration of some of the examples of the Hadamard over-complete dictionaries constructed in accordance to the above 3 choices. These are used for testing in the next section.

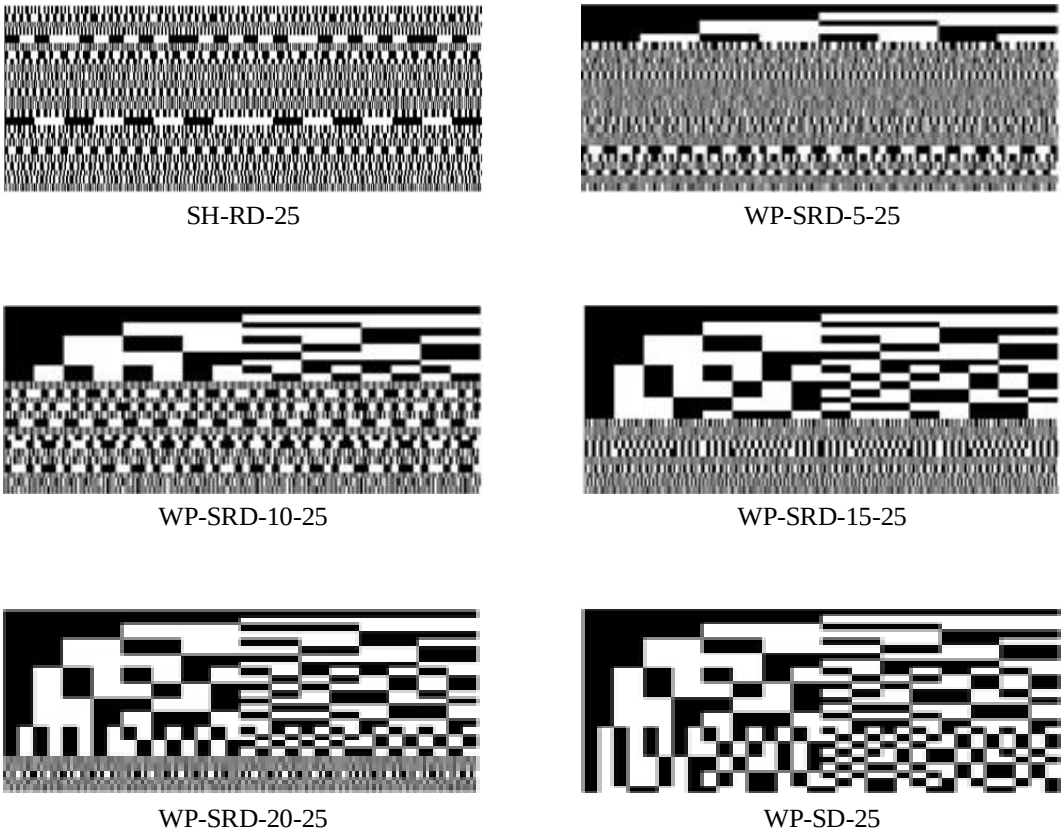


Figure 4-6 Binary display of SH-RD, WP-SRD, and WP-SD overcomplete matrices/dictionaries of size 25×512

Now, we give some explanation on the terminology of the dictionaries in figure (4-6), SH-RD-25 stands for a Fully random over-complete dictionary (RD) constructed from SH and the number of its rows ($m = 25$). WP-SRD-5-25 stands for a Semi-random over-complete dictionary (SRD) constructed form WP and the number of its rows ($m = 25$) while 5 rows at the top of the dictionary are the top 5 rows of selected WP matrix and the rest 20 rows are selected randomly from the rest of the matrix without repetition, this is true for all other WP-SRD. WP-SD-25 stands for a Structured over-complete dictionary (SD) constructed from WP and the number of its rows($m = 25$). The same construction method is used to generate over-complete Circulant matrices (C). In the rest of the thesis, we shall test the performance of the different types of Hadamard over-complete dictionaries when used for pattern recognition in different case studies.

4.6 Testing CS- compliance of different random Hadamard Matrices.

In this section, we report on the results of experiments conducted in collaboration with Dr. Nadia Al-Hassan (Visiting Postdoc at the university of Buckingham) to test the RIP characteristics of the various random overcomplete Hadamard matrices using the Sylvester and the Walsh-Paley constructions and compared the results with variant copies of Circulant dictionaries. These tests examine a random sample of 400 submatrices in terms of coherence, Condition number (CN), and the row ranks as measures of ability to recover sparse solutions.

The average coherence for 400 randomly selected submatrices of different sizes, ranging from 25×12 to the full size 25×512 to test the coherence CS property. Fig 4-7 shows, average coherence values only for submatrices of 25-columns. As can be seen, the coherence values for such submatrices of all Hadamard dictionaries and Circular matrices are comfortably within the bounding, $(1/\sqrt{m}) = 0.2 \leq$ coherence ≤ 1. Similar comments are true for the full-size matrices, displayed in Table 4.1 below.

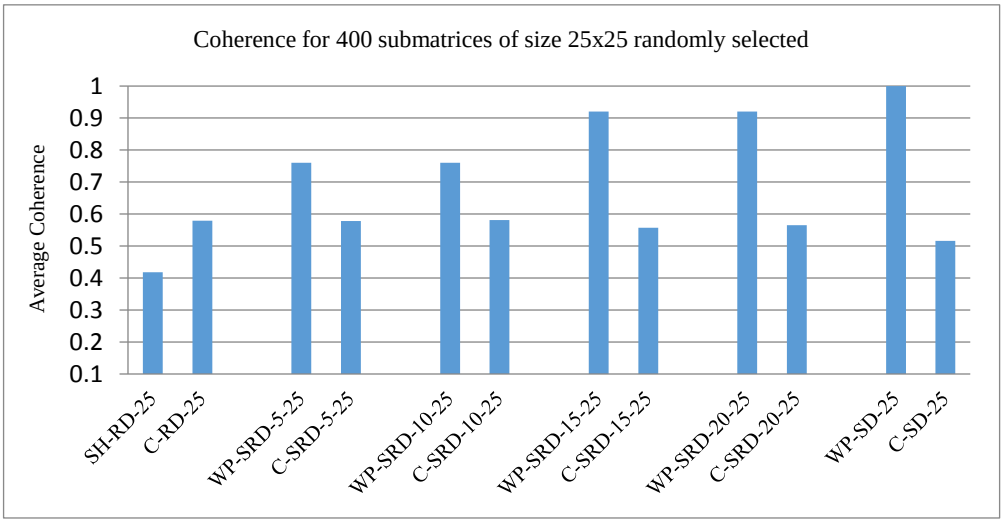


Figure 4-7 Average coherence for 400 randomly selected submatrices of size 25×25

In order to ensure sparse recovery, we also calculated condition number CN and row rank for the full size 25x512 dictionaries. The results, shown in Table 4-1, again confirm that all Hadamard matrices have CN=1 and markedly smaller than CN of Circulant matrices but still comfortably within the safe zone of CN < 2.5. Therefore, the different Hadamard and Circulant dictionaries are well-conditioned at full size. Moreover, all dictionaries attain maximum spark value with a row rank of 25.

Table 4-1 Coherence, Condition Number and Row Rank for the dictionaries

Dictionaries	RIP indicators		
	Coherence	**CN**	**Row Rank**
SH-RD-25	0.68	1	25
C-RD-25	0.75	1.44	25
WP-SRD-5-25	0.76	1	25
C-SRD-5-25	0.74	1.44	25
WP-SRD-10-25	0.76	1	25
C-SRD-10-25	0.79	1.45	25
WP-SRD-15-25	0.92	1	25
C-SRD-15-25	0.74	1.48	25
WP-SRD-20-25	0.92	1	25
C-SRD-20-25	0.76	1.69	25
WP-SD-25	1	1	25
C-SD-25	0.72	1.68	25

Finally, we conducted experiments to test the ability of these dictionaries to recover vectors of certain sparsity (a fraction to the size of the low-resolution image patch) specified by the definition of RIP. Therefore, the sparsity of vectors $\alpha \in \mathbb{R}^n$ for 400 5x5 patches were calculated for 10 images. Ideally, for such patches sparsity k must be < 6. The average k values, together with the standard deviations, for each dictionary are shown in figure 4-8, below. From the results, we noticed that the recovered coefficients are always sparse and the level of sparsity varies depending on the complexity of each tested patch. Notably, the WP-SD-25 dictionary gives a good sparse recovery with $k \in [1, 5.5]$. For the C-SD-25, $k \in [1, 6.85]$ and for the other dictionaries, the number $k > 8$ and reach nearly to 14.

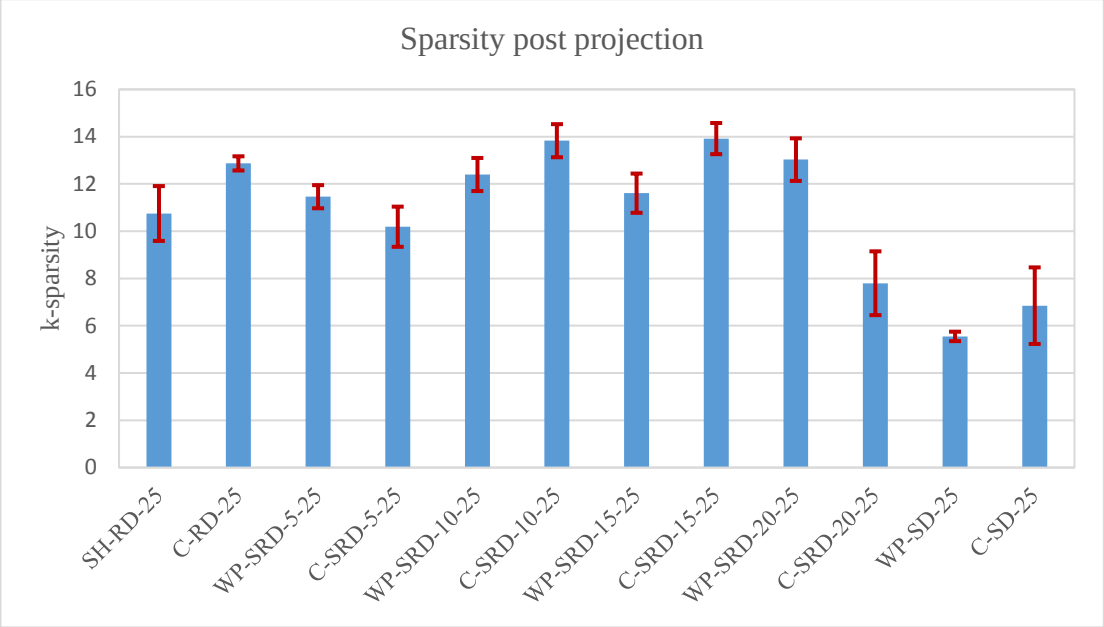


Figure 4-8 Mean and standard deviation for sparsity of 400 patches in 10images

From the above tests on coherence, CN, spark values, and k-sparsity one can conclude that the WP-SD-25 matrix is the only dictionary that satisfy the above necessary RIP conditions with high probability.

4.7 Summary

In this chapter, we investigated different JL compliant approaches to generating data independent RP's that can be used for linear dimension reduction, i.e. we were focused on designing dimension reducing projections that maintain distances between pairs of vectors within acceptable error tolerance before and after transformation. Having noted the relevance of JL condition to the recent emerging paradigm of compressive sensing (CS), we observed that the wealth of research conducted in the area of CS for designing a variety of CS dictionaries that facilitate significant reduction in the number of attributes (often referred to as meta-features) needed to model objects of interest in most interesting pattern recognition applications. Compliance of overcomplete dictionaries with the CS paradigm is dependent on a modified version of the JL condition. Instead of preserving distance between any pair of vectors, CS compliance is based on satisfying the Restricted Isometry Property (RIP) whereby the distance between sparsely represented vectors. We exploited these facts and investigated different classes of JL compliant DR matrices that are linked to over-complete CS dictionaries. These included various well investigated random matrices such as Gaussian and Bernoulli overcomplete dictionaries. However, we extended our investigation to include a large pool of the random selection of such dictionaries from the rows of the well-known class of Hadamard matrices constructed

using three different methods. We tested the compatibility of such dictionaries with RIP condition and found that random submatrices of Hadamard matrices form a very rich pool of RP tools for DR. In the rest of the thesis we shall test the performance of dictionaries in this pool in pattern recognition for different biometric case studies.

CHAPTER FIVE: CASE STUDY 1: SPEECH EMOTION RECOGNITION (SER)

5.1 Introduction

Speech is the most common form of interaction between people, and emotion may change the meaning of any uttered speech, perhaps conveying different meaning. For instance, a word like "really", could have a definitive, disbelief, admiration or even a query depending on the emotional expression of the speaker. Speech Emotion Recognition (SER) is concerned with identifying the emotional state of a speaker from the speech signal. SER is an important area of pattern recognition/classification research and could improve effectiveness and efficiency of many speech system/applications, e.g. it helps assess pilot's stressed-speech in aircraft cockpits. and is becoming very useful in many applications including in healthcare and human computer interaction (Al-Talabani, 2015).

In general, any pattern recognition problem, and in particular, SER can be summarized in three steps: (1) remove silence portions and extracting important speech features that discriminate different emotions from the raw speech samples; (2) pre-processing the extracted feature vectors by dimensionality reduction to remove redundancy and overcome the curse of dimension; and (3) use appropriate classifier(s). In the last step, usually the data set is divided into training and testing sets to build a model. Figure 5-1 shows SER steps.

Since the 2nd step is concerned with DR, then we take the SER as our first case study to investigate and compare the performance of the various DR approaches investigated in the past chapters. In particular, we consider the differently constructed Hadamard dictionaries as well as the PCA. We shall briefly describe the most commonly adopted feature extraction step in SER. We also, describe the selected testing database(s) and the main adopted classifier used for the SER. And then, we shall represent our results. As a benchmark, we adopted the various choices made in steps (1) and (3) from the work of Dr Al-Talabani in his PhD thesis, done at Buckingham University. I acknowledge and highly appreciate his guidance and help in the experimental work. We shall conduct different sets of experiments and propose the Feature-Block (FB) approach as an innovative approach to deal with lack of density ratio of samples to dimension. The FB approach is certainly useful for data-dependent DR schemes but we shall demonstrate its success in Data-independent Hadamard based DR schemes.

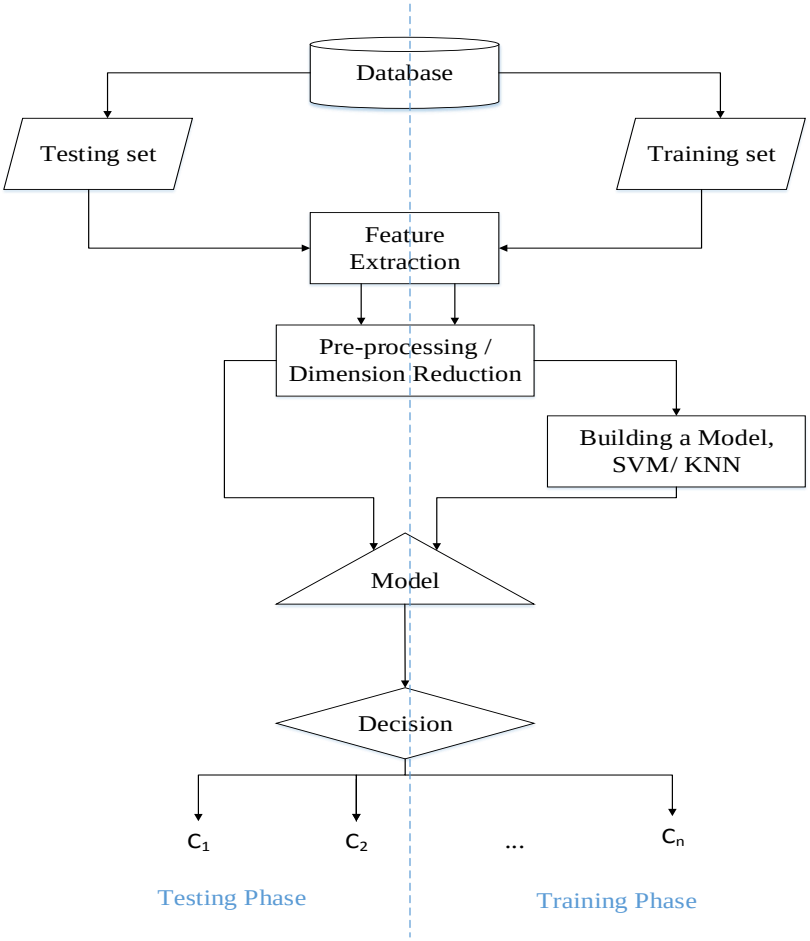


Figure 5-1 Pattern Recognition/SER steps

5.2 Feature Extraction: Low Level Descriptors (LLDs)

There are different approaches to extract emotion relevant speech features from speech signals, in this case study, we adopted the "brute force" approach. The openEAR toolkit/software (Eyben et al., 2009, and Schuller et al., 2009) was used by (Al-Talabani, 2015), to extract by brute force a total of 6552 features representing the Low Level Descriptors (LLD) baseline. For our performance testing experiments, we simply used these already extracted features for the FAU-Aibo database. The 6552 LLDs features are extracted as 39 functionals of 56 acoustic LLDs and corresponding first and second order delta regression coefficients, in total $(56 * 39 * 3 = 6552)$. The 56 acoustic LLDs are given in table 5-1, and the 39 statistical functionals are given in table 5-2.

Table 5-1 Low Level Descriptors (LLD) used in Acoustic analysis with openEAR

Feature Group	
Raw Signal	Zero-crossing-rate
Signal Energy Pitch	Logarithmic fundamental frequency F0 in Hz via cep- strum and autocorrelation (ACF). Exponentially smoothed F0 envelope.
Voice Quality	Probability of voicing $\left(\frac{\mathrm{ACF(T0)}}{\mathrm{ACF(0)}}\right)$
Spectral	Energy in bands 0-250Hz, 0-650Hz, 250- 650Hz, 1-4kHz 25%, 50 %, 75%, 90% roll-off point, centroid, flux, and rel. pos. of spectrum max. and min.
Mel-spectrum	Band 1-26
Cepstral	MFCC 0-12

Table 5-2 Functionals and their regressions coefficient applied to the LLD contour

Functionals	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. - arithmetic mean and Min. - arithmetic mean	2
Arithmetic mean, quadratic mean	2
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Centroid	1
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. Mean	4
Linear regression coefficients and corresp. approximation error	4
Quadratic regression coefficients and corresp. approximation error	5

5.3 The Support Vector Machine (SVM) Classifier.

The Support vector machine (SVM) is a well-known and commonly used supervised classifier. Given any training dataset of labelled samples of n-dimensional feature vectors, the learning process of SVM aims to find an optimal separating hyper-plane of the different classes of the training set. Optimality of the output SVM hyper-plane means that it has the maximum distance to the nearest training data samples on either side. The samples that are nearest to the SVM hyperplane are called the Support Vectors. Therefore, SVM maximize the width of the margin between the separating hyper-plane and support vectors. Such maximization assumes the existence of a unique solution for the problem which is expected to yield a better classification performance on the testing set.

Given a set of 2D-points $X = \{ x_{1,} x_2, \dots , x_m \}$ which consists of two linearly separable classes say class1 and class2, and let $Y = \{ y_{1,} y_2, \dots , y_m \}$ with $y_i \in \{+1, -1\}$ to stand for the class label of x_i. Figure 5-2, below, illustrates the SVM challenge of selecting the optimal hyperplane among all the possible class separating lines.

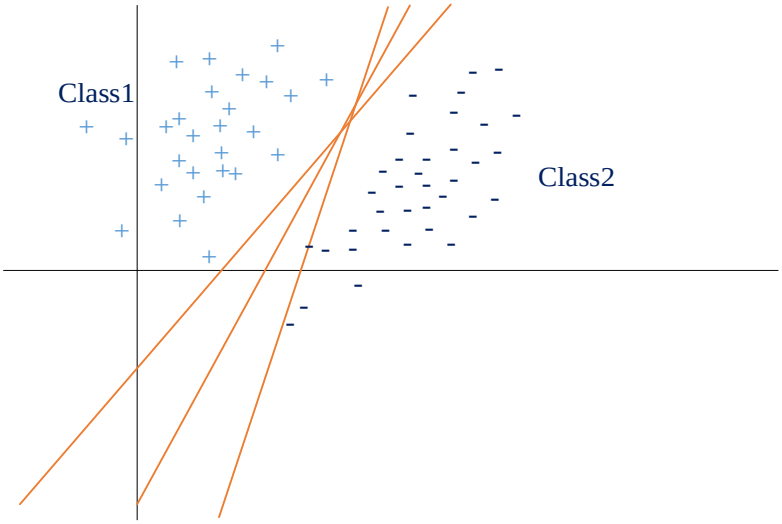


Figure 5-2 Separating hyper-planes

There may exist infinitely many lines separating the samples of the two classes. Some of the lines are very close to the training samples from one class or both, and the SVM classifier aims to find the line that lies as far as possible from the support vectors in the two classes as shown in figure 5-3.

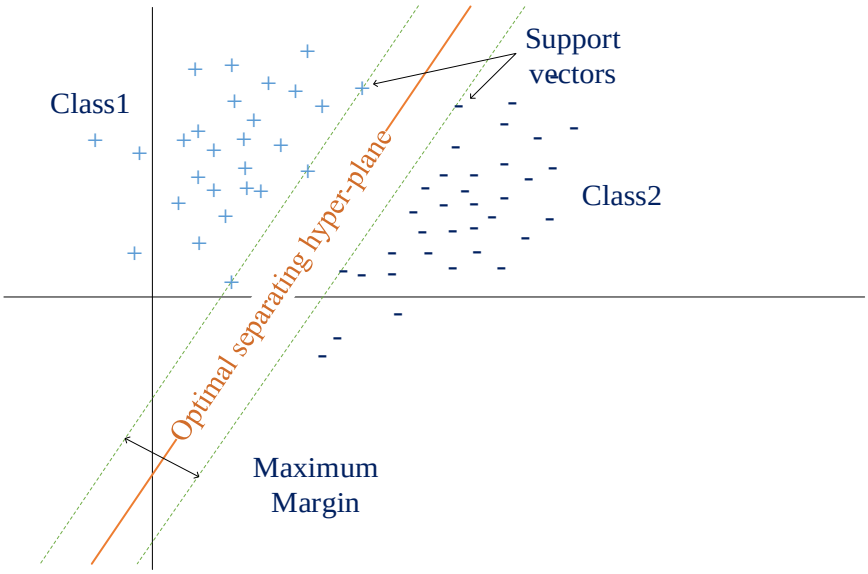


Figure 5-3 Optimal Separating Hyper-plane using SVM

Formally, a separating hyper-plane (decision boundary) can be determined by the unit vector w that is normal to the hyper-plane which determines the orientation and a scalar b (bias) which controls the displacement from origin, the equation of the hyper-plane can be written as follows:

$$f(x) = w^T x + b = 0$$

where x is any point in the hyperplane. In 2-dimensional space, this linear equation represents a line, it represents a plane in 3-dimensional space, and a hyper-plane in higher dimensional spaces. Also, the margin is represented by another two hyper-planes, and vectors on both sides of the hyperplane either belong to class1 or class2 depending weather,

$$w^T x + b = 1 \quad or \quad w^T x + b = -1$$

So, the distance between a support vector and the separating hyper-plane is $\left(\frac{1}{\|w\|}\right)$ and the width of the margin is twice of this distance,

$$M = \frac{2}{\|w\|}$$

In the optimization problem, we aim to maximize the width of the margin and this happen when the length of the normal vector w is minimized. We have two constraints for this optimization problem:

$$w^T x + b \geq 1 \quad \forall\, x \in class1$$

$$w^T x + b \leq -1 \quad \forall\, x \in class2$$

Equivalently, we can express the optimization problem in a simple way as follows:

$$\min \{\frac{1}{2}\|w\|\}$$

Subject to

$$y_i(w^T x + b) \geq 1 \quad for\ i = 1,2,\cdots m$$

This problem is known as Lagrangian optimization which can be solved using Lagrange multipliers to compute the weighting vector w and the bias b of the optimal hyper-plane, for more details see (Hastie et al., 2009).

5.4 Database used

In order to test the performance of our dictionaries within the adopted SER, and in line with any pattern recognition application, we need to use a benchmark database in the experiments we conducted to test the performance of the various dictionaries, we opted to use the very popular FAU-Aibo database. It was used in (Steidl, 2009), and consists of 51 children's sound samples while they interact with the Sony's pet robot Aibo. The children cohort at the time of recording were aged from 10 to 13. The dataset is divided into two parts 'OHM' and 'Mont' based on the data collection place and the number of speakers is 25 and 26 for each part respectively. In this database, there are five class of emotions that label the different samples: Anger, Neutral, Positive, Emphatic and rest. This database is of the non-prompted type, i.e. is recorded so that the participants are not aware of being monitored and are not instructed to express a specific emotion, but are expected to get into an emotional state, and then the produced emotion is recorded without their knowledge. The children believed that the robot is responding to their instruction. Five experts independently labelled each uttered word in the database with the emotion type. Unlike acted databases, this database present the most difficult challenge for automatic recognition. This is probably due to the fact that the way children express emotion is yet to mature or to control. This was a good reason for our choice to test the performance of our DR schemes, beside the fact that it has been used widely as a benchmark testing database.

5.5 Implementation and Experimental Setup

Speaker independent based experiment has been conducted using the 'OHM' part of the database to train the system while the 'MONT' part is used for testing. Linear kernel Support Vector Machine (SVM) and Sequential Minimal Optimization (SMO) method is adopted. It is well-known that SVM is more suitable than other classifiers for very high dimensional data applications and it was the natural choice for the SER application (Al-Talabani, 2015). Since SVM is a binary classifier i.e. it cannot classify more than two classes, methods to deal with multiclass data need to be adopted. In this work, we adopt 1 versus 1 class approach, i.e. an SVM is designed for each pair of emotions. And the final decision is made based on a majority voting among all the SVMs. The experiments conducted here are aimed at comparing the emotion recognition accuracy for the various SER schemes that only differ from each other in their dimension reduction steps. These schemes are therefore, are obtained by applying the differently constructed Hadamard submatrices and the PCA for DR. Recall that the feature vectors representing the dataset post extracting the emotion-relevant speech features (obtained from the speech samples using openEAR software) are of very high dimension 6552 and the number of available samples is $n = 9959$. In this case, dimensionality reduction is meant to be "preserving" pairwise distances, with respect to a tolerance error ϵ, but the success depends on the data sample density which in turn depend on the number of available samples. To boost the efficiency of the classifier and probably improve the recognition rate we need to set a sufficiently small value for ϵ. As a result, there will be a lower bound on the reduced dimension. In these experiments, we set $\epsilon = 0.5$, and according to the $k \geq \frac{24}{\epsilon^2(3-2\epsilon)}\ln(n)$, we need our dictionaries to reduce dimension not lower than 442. The dimension of the data set is reduced from 6552 to only 442 dimensions which is a significant reduction.

5.6 Results

To test the performance of our projection matrices as a DR tool, we first constructed the various types of overcomplete dictionaries that reduce the dimension from 6552 to 442 in all the cases. For the PCA scheme, we also selected the 442 most significant eigenvectors. The results are shown in figure 5-4. In the case of using no dimension reduction, the recognition accuracy is 38.2% while PCA with only 442 dimensions provides marginally higher accuracy of 38.5%. Differently constructed projection matrices provide different accuracy rate but with one exception (the WP-SRD-50-442) all outperform the PCA. The accuracy rate for SH-RD is 39% which is slightly higher

compare to original dimensions. All other WP-SRD projection matrices provide accuracy ranged from 38% - 40%, and the highest recognition accuracy of 41.1% is achieved by WP-SD. Note that all the features in the WP-SD are captured by first 442 rows at the top of the Full Walsh Paley Hadamard matrix, and this means that the top energy in the signal is captured by this dictionary which seem to explain its superior performance compare to other schemes.

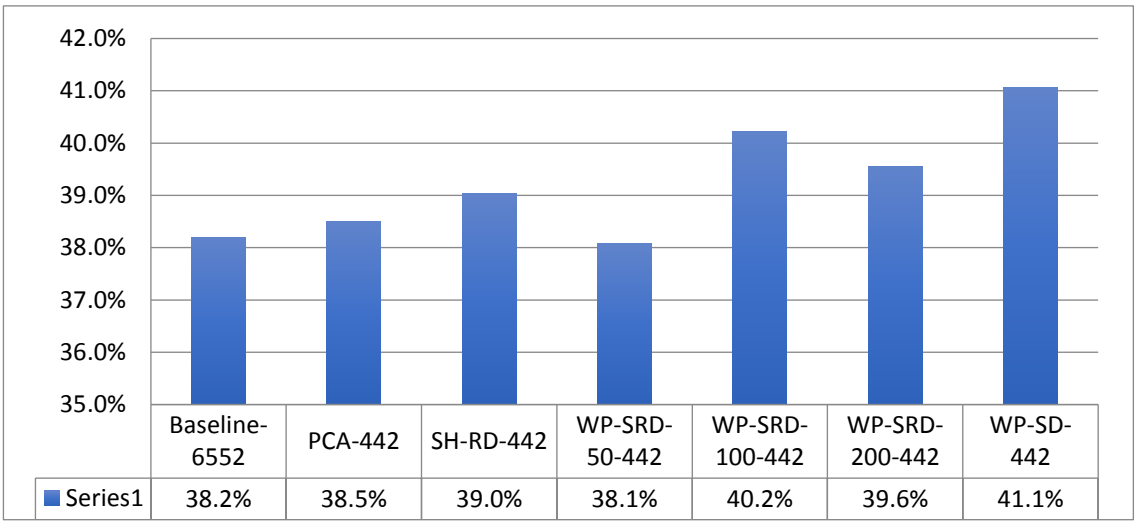


Figure 5-4 Performance of different overcomplete Hadamard Dictionaries over the FAU-Aibo database

The superior performance of the WP-SD, raises the question whether this can be improved by changing the tolerance error level ϵ. The set of experiments have been designed to use the WP-SD as a projection matrix and reduce the dimension of our dataset into different number of dimensions, by selecting different values of $0 < \epsilon < 1$. Taking into account some different values for ϵ, the corresponding lower bounds k for the reduced dimension will change accordingly as indicated in the set $\{(\epsilon, k) = (0.4,628), (0.5,442), (0.6,341), (0.7,282), (0.8,247), (0.9,228)\}$, from this set we can see that: by reducing the tolerance error level ϵ, the number of dimensions increases and vice versa, for $(\epsilon, k) = (0.4,628)$, the number of required dimensions is 628 which is the highest dimension that we tested as lower ϵ requires higher dimension and the reduced dimension will be very high. Also, the lowest dimension is 228 corresponding $\epsilon = 0.9$, furthermore, we reduced the dimension further to 25, 50, 75, 100, 221 without considering the lower bound. The results shown in Figure 5-5 reveal that reducing dimension to only 25-dimensions results in reduced accuracy to 40.3% than the above case when $\epsilon = 0.5\ and\ m = 442$. However, all other schemes outperform the $\epsilon = 0.5$ scheme by a minimum of 1.2% and the best performing is the scheme corresponds to $(\epsilon, m) =$

$(0.9\ , 228)$ which achieve accuracy of 43.4%, i.e. an increase of 2.3% over the (0.5, 442) scheme. The two previous experiments demonstrate that there is a direct relation between the combination pair (ϵ, m), and the selected type of Hadamard dictionary but the Walsh Paley construction and the choice of the top rows can produce the best performance. In the next set of experiments, we shall investigate a rather innovative way of reducing effect of low density ratio of samples to feature vector dimension.

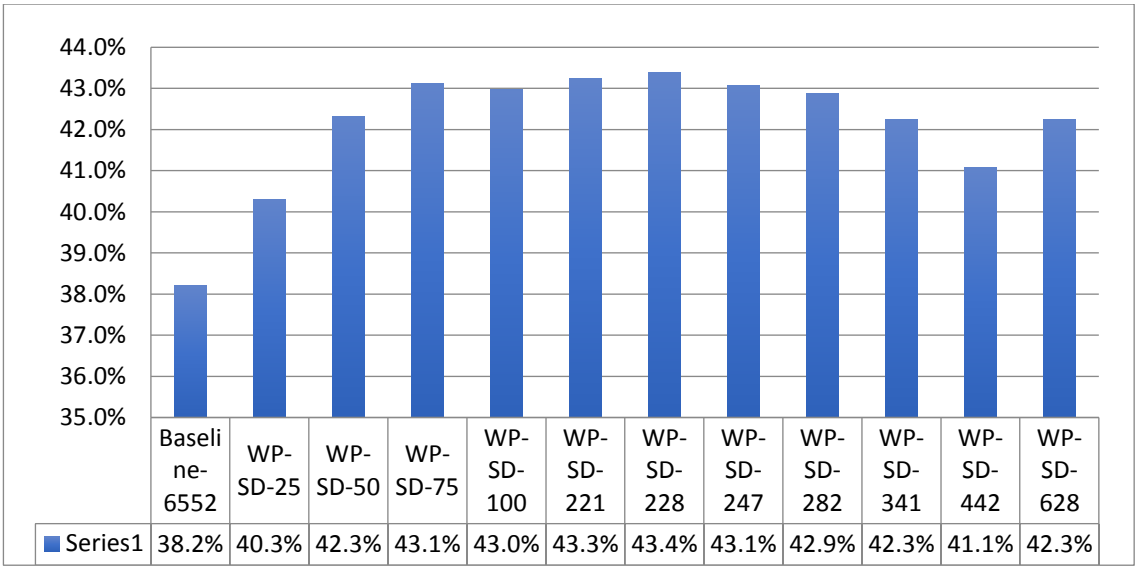


Figure 5-5 Recognition rate for FAU-Aibo database, post WP-SD Dictionaries for DR

5.7 Statistical-Based Feature Block (FB)

For any high dimensional data set, dimension reduction by data-dependent schemes is adversely impacted by the lack of density ratio of available samples to the dimension. In such cases one would ask whether all the individual, or groups, of features are emotion relevant in the same way. This often leads to considering feature selection or even to give different weight for different features. However, in some pattern recognition schemes, the feature vectors representing the objects of interest are made of several groups/blocks of different types. This is certainly the case with the SER because the 6552 coordinates represent feature groups of different types. In fact, the 6552-dimensional feature vector is made up of 39 statistical functionals of 56 acoustic LLD and their corresponding first and second ordered delta regression coefficients.

Here we propose feature-block approach to dimensionality reduction. Simply, it is a process of dividing features/dimensions into some blocks, which can be partitioned either randomly or based on some common properties of the features. Then, instead of processing the high dimensional data set, the blocks will be processed individually which have a lower dimension and higher sample density without discarding any dimension or

block. In this way, for each block the density ration increases dramatically when there are many different blocks.

We shall now conduct experiments to test the effectiveness of FB approach for the SER application. We first create 39 statistical-based feature block from FAU-Aibo database, each block contains only one statistical parameter, from each of sound frames, resulting in 168-dimensional vector for each of these parameters. Note that each statistical property repeated 168 times but of course not all the time for the same feature group. In these experiments, the dimension of FAU-Aibo database is reduced by constructing 39 feature blocks, instead of randomly constructing some blocks, we put all statistical parameters of the same type together in a block. Having reduced the dimension of each block to 168, the density ratio of the samples is increased by a factor of 39.

We now test the performance of our WP-SD schemes when the dimension of each block is further reduced to 100, 75, 50, 25 and compare with that of using entire 168-dimension. The results as shown in table 5-3. It can be clearly seen that; different feature block provides different accuracy rate due to fact that each statistical parameter contains different information of the data set. The recognition of a few blocks is between 30% - 40% and a fewer blocks under 30%. Interestingly, nearly half of the blocks provide a good recognition accuracy about 40% and above that. The maximum accuracy rate is achieved with the fifth block (minameandist) 44% with 168-dimensions. This rate provides a good improvement about 6% compare to accuracy of using the baseline features.

In the rightmost four columns of table 5-3, the dimension of the feature-blocks is reduced more. In general, the accuracy remains nearly the same or slightly lower, again, it shows that our dictionaries provide a very proper lower dimensional approximation and they fairly preserve the structure of high dimensional data in the transformed space. The maximum recognition accuracy is 44% with the feature-block (stddev) with only 100 dimensions, the same feature block provides the maximum accuracy 43% and 44% at 75 and 50 dimensions respectively which is significant compare to the case of using the 6552-baseline features. 43% accuracy as the maximum rate can be achieved with the (nzabsmean) feature-block with only 25 dimensions.

These experiments, not only demonstrate the effectiveness of the FB approach and opens the way to using sophisticated multi-blocks fusion schemes as well as ensemble of multiple classifiers for improved accuracy. However, this is outside the remit of this thesis but can be done in the future.

Table 5-3 Accuracy rates of Statistical-based Feature Blocks using WP-SD schemes for the FAU-Aibo database.

	Statistical Functionals, etc.	168 Dimensions	WP-SD-100	WP-SD-75	WP-SD-50	WP-SD-25
1	Range	41%	41%	41%	43%	41%
2	maxPos	29%	30%	29%	29%	27%
3	minPos	28%	28%	27%	27%	25%
4	maxameandist	41%	40%	40%	42%	41%
5	minameandist	44%	43%	42%	42%	40%
6	linregc1	35%	35%	35%	34%	30%
7	linregc2	40%	39%	38%	38%	37%
8	linregerrA	43%	42%	42%	42%	42%
9	linregerrQ	42%	42%	42%	41%	41%
10	qregc1	31%	30%	29%	32%	29%
11	qregc2	36%	35%	34%	34%	33%
12	qregc3	36%	35%	34%	33%	33%
13	qregerrA	42%	41%	41%	42%	42%
14	qregerrQ	41%	40%	42%	42%	41%
15	Centroid	31%	30%	30%	30%	27%
16	Variance	42%	42%	42%	42%	41%
17	Stddev	43%	44%	43%	44%	43%
18	Skewness	38%	36%	35%	34%	32%
19	Kurtosis	35%	33%	32%	31%	31%
20	quartile1	36%	35%	34%	31%	32%
21	quartile2	34%	32%	32%	29%	30%
22	quartile3	38%	37%	38%	38%	37%
23	iqr1-2	35%	33%	31%	27%	28%
24	iqr2-3	37%	37%	37%	37%	36%
25	iqr1-3	39%	37%	38%	38%	36%
26	percentile95.0	43%	42%	43%	42%	41%
27	percentile98.0	43%	41%	42%	42%	42%
28	Zcr	32%	33%	33%	34%	33%
29	numPeaks	34%	33%	34%	34%	32%
30	meanPeakDist	31%	32%	31%	32%	30%
31	PeakMean	41%	43%	42%	42%	42%
32	peakMeanMeanDist	42%	41%	42%	41%	41%
33	Amean	41%	41%	40%	39%	38%
34	Absmean	43%	43%	43%	42%	42%
35	Qmean	44%	41%	42%	41%	41%
36	Nzabsmean	43%	43%	43%	43%	43%
37	Nzqmean	43%	42%	42%	42%	42%
38	Nzgmean	39%	37%	37%	36%	37%
39	Nnz	30%	31%	30%	30%	30%
	Maximum Accuracy for each column	Minameandist	Stddev	Stddev	Stddev	Nzabsmean
		44%	44%	43%	44%	43%

5.8 Conclusion

In this chapter, we investigated the performance of various Hadamard based dictionaries for DR within the SER pattern recognition application. The experimental results demonstrated that these data-independent DR schemes, and with very few exceptions they outperform the PCA schemes not only using the same number of reduced dimension but with significantly lower dimensions. Among the various Hadamard based dictionaries, the WP-SD scheme, which uses the Walsh-Paley Hadamard construction and then select from the top rows only in their order, as the best performing scheme. We also proposed the Feature-Block based dimension reduction technique as an innovative solution to overcome the problem of low density ratio of samples to dimension. We have demonstrated the success of this approach in significantly increasing accuracy rates for the SER application with significant dimension reduction. Not only this approach can be extended to other pattern recognition schemes, where features can be naturally split into different groups, but opens the way for different approaches to fusion with promising results.

CHAPTER SIX: CASE STUDY 2: GAIT-BASED GENDER CLASSIFICATION (GBGC)

6.1 Introduction

Human Gait biometric refers to the profile of the way human walk, that may be detected unobtrusively from a distance with or without cooperation. It is not only useful for recognising a person from his/her style of walking but it could help classifying the gender, age group, and state of health of the person as well as in determining whether the person is carrying items/bags or wearing certain items of clothing. Gait analysis is one of the main challenging areas of biometric research that has an important role in authentication applications (access control) and in particular for security surveillance due to its efficiency and effectiveness.

Gait-Based Gender Classification (GBGC) as a special case of human gait analysis has many interesting applications: in smart surveillance and other recognition applications, it can be used as a pre-classification in human gait recognition to improve the recognition accuracy by restricting searching process on one gender. Demographic studies use such application for collecting gender based statistical information to support/improve customer service in stores(Sabir et al., 2014).

Like most other biometrics, dimension reduction may become a necessary step, in automatic gait recognition/classification. This case study is devoted to automatic gait-based gender classification only. The main aims of the experimental work, in this chapter, is similar to those followed in chapter 5 for the SER case study, is to conduct a comparative analysis the performance of the various data-independent DR Hadamard based dictionaries as well as the PCA scheme. Here, we shall use motion-based data features to represent human gait profile that was proposed in (Mawlood, 2016) and in particular we adopt the various feature vectors that were investigated intensively by Dr. Azhin Sabir in his DPhil thesis (Sabir, 2015) done at Buckingham University. I acknowledge and highly appreciate his guidance and valuable advice throughout this case study investigation as well as the experimental work.

6.2 Pre-processing and Feature Extraction

The process of GBGC like any other pattern recognition/classification system can be divided into three steps (see the description of this process given in the section (5.1). The

first step is pre-processing aims to segment the video frames, and subtract the background from the foreground (image of the person), and estimate the gait cycle (Sabir, 2015). The background subtraction generates a binary image/frame where 0 representing the background and 1 representing the foreground which refers to the person's silhouette. Variation in silhouettes size may significantly affect the recognition accuracy. The silhouette images are determined with respect to the size of a bounding box, and therefore some normalization will be done to make all the silhouettes have the same size. Necessarily, the silhouettes need to be horizontally aligned to be in the centre of the bounding boxes of all the frames. The final pre-processing step is gait cycle estimation, for more details see (Mawlood, 2016). Pre-processing may also include data/image quality improvement procedures to avoid undesired image distortion.

The pre-processing step is followed by human gait feature extraction. For our experiments on GBGC system, we use a gait feature called Gait Entropy Energy Image (GEnEI) proposed in (Mawlood, 2016). It is constructed using the two most widely used human gait features in the literature due to their effectiveness and simplicity known as Gait Energy Image (GEI) and Gait Entropy Image (GEnI). These three feature vectors are defined as follows:

$$GEI(x,y) = \frac{1}{T}\sum_{t=1}^{T} B(x,y,t)$$

Where t is the frame number and T is the total number of frames in a complete cycle in a sequence, x and y are coordinate values of the pixel (x,y) in the 2D image. B(x,y,t) stands for the pixel value (x,y) in frame t.

$$GEnI(x,y) = -\sum_{r=1}^{R} p_r\,(x,y)\,log_2\,p_r\,(x,y)$$

Again x and y are coordinate values of the pixel (x,y) and $p_r\,(x,y)$ is the probability that the pixel (x,y) takes on the r^{th} value along the whole frames. In this case, the silhouettes are binary images (1 or 0) and thus R = 2.

In order to provide a better human gate feature performance, GEnEI is defined as follows:

$$GEnEI(x,y) = \begin{cases} GEI(x,y) & if\ \ GEI(x,y) > 0\ and < 0.5 \\ GEnI(x,y) & Otherwise \end{cases}$$

In our experiments, Gait Entropy Energy Image feature (GEnEI) is used as a gait feature which is constructed based on Gait Energy Image (GEI) and Gait Entropy Image (GEnI). Then, three feature vector is constructed from GEnEI using wavelet transform at second level of decomposition. AGEnEI is the 1st feature vector based on Approximation coefficient LL sub-band, while VGEnEI is the 2nd feature vector based on Vertical coefficient sub-band. Due to the fact that when a person walks, his/her upper body part changes in a different way to the lower body part. We follow the approach taken by (Sabir, 2015), where the two parts of the human body are considered separately and determined according to the so-called *golden ratio* proportion (0.62 for the upper part and 0.38 for the lower part). Consequently, the 3rd type of feature vector model of Gait biometric template is the AVGEnEI which consists of two sub-vectors: the VGEnEI extracted from the upper body part and the AGEnEI from the lower body part. Each of these three feature vectors have dimension of 1500, and in our experiments, we shall test the performance of the various DR schemes using each separately.

6.3 k-Nearest Neighbours (kNN) Classifier

The k-Nearest Neighbour is a simple and efficient classification algorithm that classifies patterns/objects based on a similarity/distance function. The kNN store the feature vector templates of samples of the enrolled members each labelled with the class identity of the owner, and whenever a fresh sample is presented, kNN assigns a class label to it based on the majority vote of its k-Nearest Neighbours, (Theodoridis and Koutroumbas, 2003). In the case of k=1, the object is classified to the class of the nearest neighbour, and this would more useful when only one template per class is stored. In general, it is better to choose an odd value for k and more generally, k is better to be a value different from multiples of class numbers to avoid ties.

There are various distance functions to measure the similarities and the most widely used functions are Euclidean and City Block distance function. For any two vectors $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ in $\mathbb{R}^n$, the Euclidean and City Block Distance between $A\ and\ B$ are defined as follows respectively:

$$E(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

$$CB(A,B) = \sum_{i=1}^{n}|a_i - b_i|$$

In our experiments, City Block distance function is adopted. In reality, there are some inequalities relating these two different distance functions and their influence on accuracy for kNN is expected to be marginal, i.e. our choice does not have limiting effects on conclusions.

6.4 Database used

To test the performance of our DR schemes within the overall Gait-based Gender classification task, we use the well-known CASIA Gait Database, Dataset B, of videos. The videos were captured in an indoor environment with a simple background in order to facilitate the efficient detection and segmentation of the walking person's silhouette (Yu et al., 2006). There are three different recoding conditions covered in this database: Normal, clothing and carrying condition. In total, 124 people participated, 31 females and 93 males. There were 11 cameras with different view angle. For each view angle, each person has 10 gait video sequences; 6 normal, 2 with a coat, and 2 with a bag. In total, there are $124 \times 10 \times 11 = 13640$ gait sequences. Each video records an individual walking with one of the conditions (Normal, Coat wearing, Carrying bag) in a certain direction. Figure 6-1 shows example of different view angles, clothing and carrying conditions.

(a) Normal walking sequences

(b) Walking Normally (c) Walking with a coat (d) Walking with a bag

Figure 6-1 CASIA Gait Database, Dataset B

6.5 Implementation and Experimental Setup

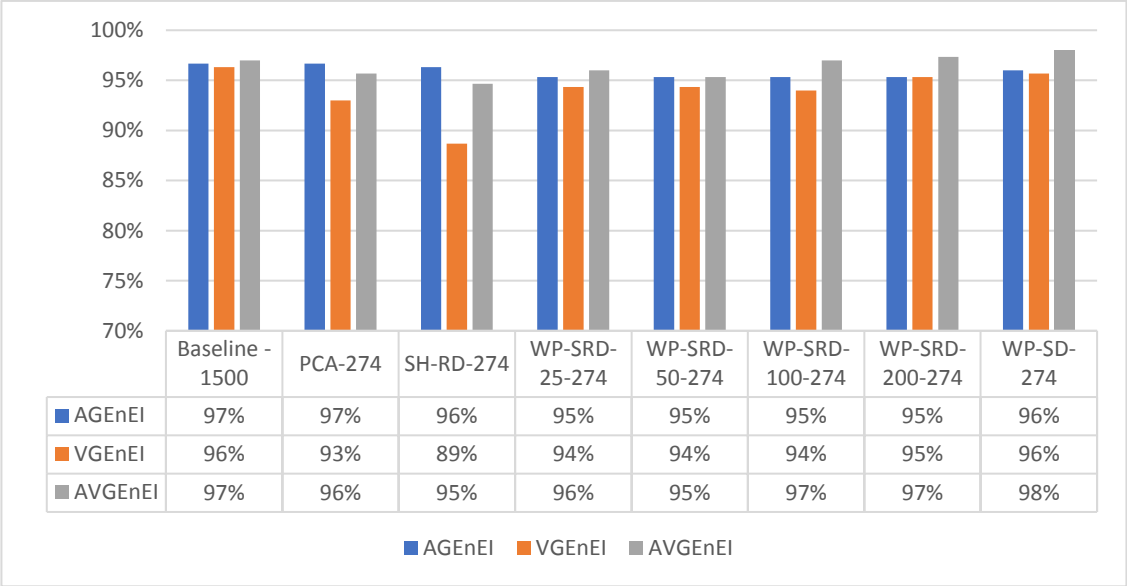
In this case study, we select the normal walking scenario (Where the person does not wear a coat or carry a bag) with only 90^o view angle between the camera and the line of walking. This scenario provides more dynamic information compare to other view angles. The restriction of the experiments to this scenario is due to the fact these tests are basically aimed as proof of concept. The CASIA-dataset B contains 31 females and 93 males which is imbalanced from the gender point of view and using all the samples for testing may raise doubts about the testing performance. Therefore, we select two random subsets of 25 female and 25 male participants from the normal gait sequences. For each of the 50 individuals, the dataset includes 6 normal gait sequences in the 90^o view angle. In total, there are 50×6×1=300 samples each represented by three feature vectors of dimensions 1500 each as discussed above. According to the $k \geq \frac{24}{\epsilon^2(3-2\epsilon)}\ln(n)$ lower bound condition, setting $\epsilon = 0.5$, sets the lower bound of 274 on the reduced dimension k for the three extracted feature vectors when using random, semi-random, and structured Hadamard submatrices and PCA to test and evaluate the performance of our differently constructed projection matrices in dimensionality reduction and compare it to the performance of using all the 1500 features.

The original experimental work conducted by Dr. Azhin on GBGC used kNN as the most common classifier used by the Gait Recognition community (Sabir, 2015). Consequently, in these experiments, k-Nearest Neighbour is used as classification method. It has been shown in (Mawlood, 2016) that for this proposed method, k=1 provides better results compare to k=3 or 5. So we test our system with only k=1 and adopt the 10-Fold Cross Validation protocol to determine the accuracy result for each of three feature vectors and the DR scheme.

6.6 Results

Figure 6-2 presents the performance in terms of the recognition accuracy rates for all combinations of feature vector and DR scheme. In the case of using no dimension reduction, the recognition accuracy is almost optimal for all the three feature vectors. For the AGEnEI feature vector, only the PCA with 274 dimensions matches accuracy of the full dimension, while all the Hadamard projection matrices provide nearly 96%. For the second VGEnEI feature vector, the full dimension scheme outperforms all but one of the DR reduction schemes. Only WP-SD scheme matches the performance of full dimension at 96%. The performance of the PCA is degraded by absolute 3% and SH-RD by absolute

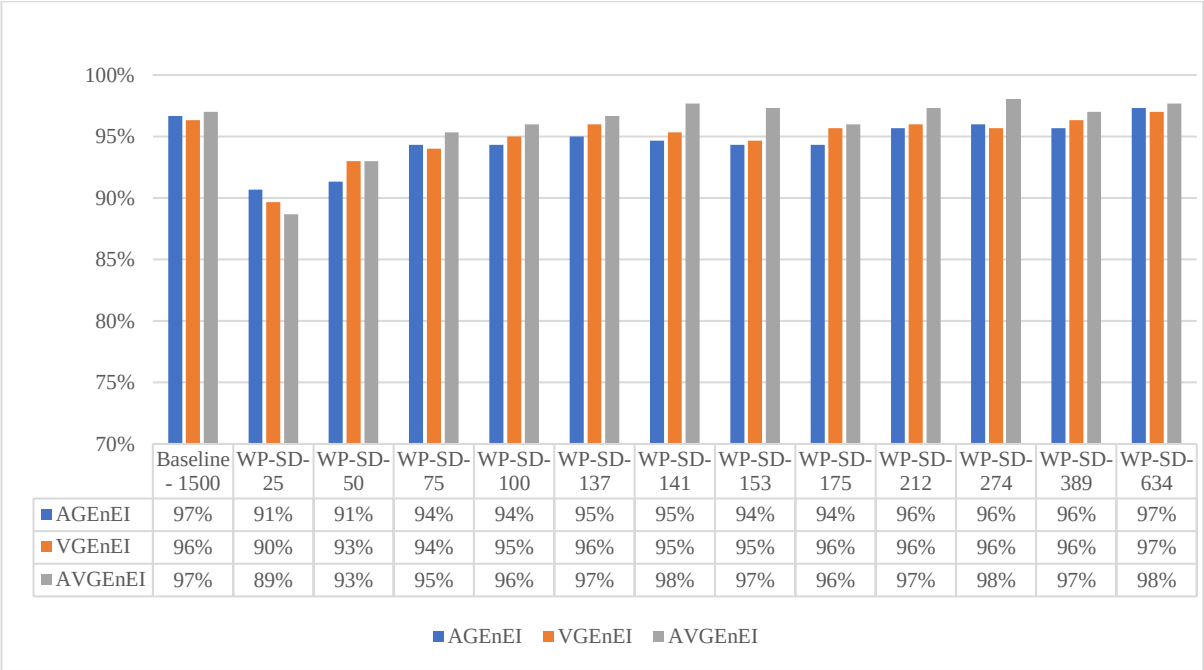
7%, while the remaining schemes are also degraded but by lower percentages. For the AVGEnEI feature vector, PCA provides 96% accuracy and in the case of SH-RD the accuracy 95%. The performance of the WP-SRD schemes increases proportionally to the number of rows from the top, and in fact when all the top 274 rows are selected the resulting WP-SD scheme outperform the full dimension scheme by attaining accuracy of 98%.



	Baseline - 1500	PCA-274	SH-RD-274	WP-SRD-25-274	WP-SRD-50-274	WP-SRD-100-274	WP-SRD-200-274	WP-SD-274
AGEnEI	97%	97%	96%	95%	95%	95%	95%	96%
VGEnEI	96%	93%	89%	94%	94%	94%	95%	96%
AVGEnEI	97%	96%	95%	96%	95%	97%	97%	98%

Figure 6-2 Gender accuracy rates for PCA and Hadamard sub-matrices schemes for DR.

Due to the relatively excellent performance of the WP-SD compared to all the other DR schemes, we conducted a new set of experiments to test the performance of the WP-SD with different number of lower dimensions, by selecting different values of $0 < \epsilon < 1$.The corresponding lower bounds for the reduced dimension will change accordingly as indicated in the set $\{(\epsilon, k) = (0.3,634), (0.4,389), (0.5,274), (0.6,212), (0.7,175), (0.8,153), (0.9,141)\}$, from this set we can see that: by reducing the tolerance error level ϵ, the number of dimensions increases and vice versa, for $(\epsilon, k) = (0.3,634)$, the number of required dimensions is 634 which is the highest dimension that we tested as lower ϵ requires higher dimension and the reduced dimension will be very high. Also, the lowest dimension is 141 corresponding $\epsilon = 0.9$, furthermore, we reduced the dimension further to 25, 50, 75, 100, 137 without considering the lower bound. The results shown in Figure 6-3 reveal that reducing dimension to only 25 results in the lowest accuracy across the 3 different feature vectors. But as the number of reduced dimension increases the very few exceptions the performance across the 3 feature vectors improves and the WP-SD 634 outperform or matched the performance of the full dimension schemes.



	Baseline - 1500	WP-SD-25	WP-SD-50	WP-SD-75	WP-SD-100	WP-SD-137	WP-SD-141	WP-SD-153	WP-SD-175	WP-SD-212	WP-SD-274	WP-SD-389	WP-SD-634
AGEnEI	97%	91%	91%	94%	94%	95%	95%	94%	94%	96%	96%	96%	97%
VGEnEI	96%	90%	93%	94%	95%	96%	95%	95%	96%	96%	96%	96%	97%
AVGEnEI	97%	89%	93%	95%	96%	97%	98%	97%	96%	97%	98%	97%	98%

Figure 6-3 Classification accuracy when using WP-SD for DR with different reduced dimension

6.7 Conclusion

In this second case study, we investigated the performance of various Hadamard based dictionaries for DR for GBGC problem. The experimental results demonstrated that the data-independent DR schemes perform very well with WP-SD schemes outperforming all other DR schemes. This performance may be attributed to the fact that the rows at the top of these dictionaries capture the highest energy in the input sample.

CHAPTER SEVEN: CONCLUSION AND FUTURE WORK

7.1 Review of the thesis

Curse of Dimension is challenging obstacle that occurs in an increasing number of pattern/object recognition/classification applications whereby the raw or pre-processed digital model of the patterns/objects are represented by vectors in linear high dimensional vector spaces. It has long been recognised that applying dimension reduction (DR) linear transformations to project the pattern/object vectors onto a significantly reduced subspace can provide mechanisms to deal with the adverse effects of curse of dimension without significant loss of information. Since recognition/classification of objects rely strongly on similarity/distance functions defined on the domains and codomains vector spaces of the deployed DR transformation, then loss of information is associated with the effect of the transform on the adopted similarity/distance between patterns/objects before and after transformation. The main aim of this thesis is to study and investigate various types of dimension reduction schemes that have been used, or suitable for use, to support efficient and reliable pattern recognition/classification applications. Naturally, we found that the most commonly practiced DR techniques for recognition/classification were obtained by a training process that works with a dataset of pattern/object samples that together hold most of the necessary variations to ensure a good performance. However, search for DR matrices as well as compressive sensing dictionaries have also been investigated by mathematicians who proposed useful random schemes that are independent of any training process. Accordingly, the work in this thesis was focused on investigating different data-independent DR schemes in contrast to data-dependent schemes.

In the first chapter, we described curse of dimension as the general reference to challenges associated with high extrinsic dimensionality of the modelled patterns/objects of interest in pattern recognition/classification applications. These challenges are mostly related to the efficiency of retrieval, analysis, and verifying/classifying the pattern/object of interest. Dimension reduction is a process of transforming the extrinsic high dimensional digital models into an intrinsic (much lower) dimensional subspace without losing relevant information. We noted that in the case of data-dependent DR, the adverse consequences of the "curse of dimension" intensify in the applications where the density ratio of available samples to the dimension of the feature space get smaller. This is most likely to be due to the difficulty in ensuring that the training data samples could form sufficient representation of the objects/patterns in the extrinsic (high) dimensional spaces.

In the second chapter, we studied some basic background concepts in Linear Algebra as the background of DR schemes, that are essential for developing data-dependent DR techniques as linear transformations (Change of basis) that transform a high dimensional dataset into a much lower subspace (coordinate system), in which the most recognition/classification related information hidden in a sample dataset are preserved with a sufficiently small tolerance error that is determined by the objectives of the application. We also presented the JL theorem, as the mathematical theory that underpins the existence of linear DR and determines the conditions that govern the relationship between the value of the reduced dimension and the tolerance error. Recognition/ classification relevant information relates to the extent of preservation of the distances between pairs of object digital samples before and after the DR transformation. The JL theorem ensures the existence of a function that maps any give dataset of high dimensional vectors into a much lower dimensional subspace without distorting pairwise distances significantly. We also discuss the classification of linear DR schemes into data-dependent and data-independent ones.

In the third chapter, we began by critically investigating the PCA and LDA as the most two widely used data-dependent DR schemes in pattern recognition/classification applications. PCA is the technique that project a high dimensional dataset on a subspace that captures almost all the variation present in the dataset while LDA provide a lower subspace that is optimal for class discrimination. We also covered SVD as a matrix decomposition technique and data compression/dimension reduction as one of its applications. These techniques successfully reduce the dimension of high dimensional datasets with a good accuracy, however, they are computationally high demanding as the system need to be trained on a suitable training set to extract the projection matrix (lower subspace). Furthermore, for any pattern recognition/classification application with data-dependent dimension reduction schemes, a sufficiently large number of training samples maybe required to model a robust system. Ideally the training set must include a wide range of possible variants of the pattern of interest and the scheme is ideally adaptive to the most relevant variation in capturing/recording samples. However, in most applications the available dataset is often small and for supervised learning schemes need to be divided into training and testing which makes the available training set even smaller and it leads to overfitting and biasness.

The most common characteristics of all these data-dependent linear DR schemes, is their reliance on the theory of matrix factorisation as well as their relevance to the question of

data/image compression. Despite their successes for many applications, we know that these schemes are not designed in accordance to the JL theory. Instead of preserving the information between every pair of objects, within tolerable errors, the investigated data-dependent linear DR methods minimise the global loss of information hidden in all the samples of a chosen training dataset, rather than each pair of samples in the training set. Moreover, the reduced dimension in these schemes are to set in accordance to the performance of the corresponding recognition scheme on a fixed testing dataset. This is the reason that the performance of these methods is influenced by the process of selecting the training set and scalability maybe in doubt, i.e. when the objects population increases significantly the scheme performance deteriorates. Consequently, removing dependence on observed samples is desirable in controlling the errors in the distances between any pair of samples within a chosen/fixed error tolerance. This is the main incentive to investigate data-independent DR schemes the existence of which are guaranteed by the JL condition.

In the fourth chapter, we first presented DWT as a well-known data-independent linear transformation of signals/images such that each sub-band of the transformed data is a linearly reduced dimensional representation of the original data. We then extended the investigation to giver a variety of RPs. Such schemes, unlike data-dependent schemes, are computationally very cheap and it costs just a matrix multiplication as the projection matrix is constructed independently of the dataset. Influenced by the JL theorem, we were focused on designing dimension reducing projections that maintain distances between pairs of vectors within acceptable error tolerance before and after transformation. Having noted the relevance of JL condition to the recent emerging paradigm of compressive sensing (CS), we observed that the wealth of research conducted in the area of CS for designing a variety of CS dictionaries that facilitate significant reduction in the number of attributes (often referred to as meta-features) needed to model objects of interest in most interesting pattern recognition applications. Compliance of overcomplete dictionaries with the CS paradigm is dependent on a modified version of the JL condition. Instead of preserving distance between any pair of vectors, CS compliance is based on satisfying the Restricted Isometry Property (RIP) whereby the distance between sparsely represented vectors. We exploited these facts and investigated different classes of JL compliant DR matrices that are linked to over-complete CS dictionaries. These included various well investigated random matrices such as Gaussian and Bernoulli overcomplete dictionaries. We extended the pool of RP schemes by using randomly constructed

overcomplete submatrices of Hadamard matrices which are known to be orthogonal. We attempted to adhere to the condition of JL theory in order to reduce the dimension of high dimensional data while maintain pairwise distances with high probability. The attraction of Hadamard matrices comes from the fact that they can be generated using different recursive procedures (Sylvester, Walsh, and Walsh-Paley). We also investigated the use of overcomplete circulant matrices that are known to satisfy RIP condition. Accordingly, we constructed pure random, semi-random, and structured random projections from well-known Hadamard matrices and circulant matrices. We have tested the compatibility of Hadamard-based overcomplete sub-matrices with the RIP conditions, and the results show that WP-SD is the only dictionary that satisfy RIP conditions with high probability. Such investigation motivated us to test the performance of these differently constructed Hadamard submatrices for the DR step in different biometric recognition as case studies.

In the fifth and sixth chapters, we have tested the performance of overcomplete Hadamard sub-matrices in two case studies SER and GBGC. In these two case studies, we first reduced the dimension of the corresponding original data models using differently constructed Hadamard submatrices. We then tested and compared the performances of the corresponding classification schemes over a well-known benchmark databases, see figure 5-4, 6-2.

7.2 Main findings of the study

The experimental results in the case studies show that among the various Hadamard based dictionaries, the WP-SD scheme, which selects rows from the top rows only of the Walsh-Paley Hadamard construction in their appearing order, is a very effective scheme as it outperforms semi-random and random schemes. In fact, the same pattern of performance of different Hadamard-based DR schemes was repeated for both case studies, i.e. the Walsh-Paley structured dictionary (WP-SD) outperforms all other random and semi-random dictionaries. These results are consistent with the results in Chapter four section (4.6) which showed that the WP-SD matrix is the only dictionary that satisfies RIP condition as a special case of JL condition.

The ease with which Hadamard based random projections can be constructed, provided an opportunity to investigate and develop a strategy to reduce the effect of low density ratio of available training samples to the dimension. We investigated and developed a novel Feature-Block based dimension reduction technique to overcome this problem by what might be considered as a feature selection approach. It works by splitting the features

into groups and applying usual dimension reduction on each group separately. Ideally these groups consist of similar data types and measure similar aspects of the signal. This approach results in increasing the samples to dimension ratio and reducing the effect of low density ratio problem especially if the original dimension can be subdivided naturally into a relatively large number of similar groups. The SER case study provided a perfect candidate to test this strategy, since the extracted feature vector consisted of 39 different types of features each is extracted from the different speech windows. We tested performance of this strategy for the SER application and we only confined the test to the use of the Walsh-Paley structured dictionary (WP-SD) for dimension reduction due to the fact that WP-SD was the best performing DR. The experimental work demonstrated the success of this strategy by the significant increase of accuracy rates for the SER application with significant dimension reduction. In some way this strategy can be seen as a hybrid of dimension reduction and dimension selection to deal with the curse of dimension.

The experimental work carried out for the two case studies confirmed some expected observation regarding the conditions, imposed by the JL theory, that govern the relation between the lower bound of the reduced dimension and the impact on performance of the resulting DR scheme and the recognition tasks. We know that the lower bound of the reduced dimension is dependent of the tolerance error, but we know that increasing the tolerance error increases false acceptances which means reduced accuracy. Consider the results in figure 5-5, which show the performance of SER when using of WP-SD for DR at different choices of reduced dimension. We observe that when the number of reduced dimension increases from 25 to 228, which represent lower bounds for decreased tolerance error, the SER accuracy increases reaching a peak at 228 after which the performance is significantly lower. In the GBGC case study, the results in figure 6-3 show that, for the first two feature vector GEI and GEnI increasing the number of dimensions from 25 all the way to 634 results in a marginal increase in the accuracy and peaking at 634, while for the third feature vector (GEnEI) the highest accuracy is achieved with 141-dimension and further increasing the dimensions does not help improve accuracy. These observations are pretty normal in most pattern recognition/classification since there is a direct link between classification performance and the reduced number of dimension, but the interest is in determining the optimal choice of reduced dimension.

Finally, we conclude that: Data-independent DR schemes perform as well, if not better, as data-dependent schemes, furthermore, among differently constructed Hadamard based

RPs, structured RP matrices deliver a better performance in terms of classification/ recognition accuracy than complete random and semi-random matrices. Interestingly, such schemes are very cheap and efficient compare to other schemes and they are not adaptive to any training set as adaptive schemes may lead to overfitting and biasness. In general, data-independent DR schemes are less sensitive to fresh samples in the classification/recognition systems. In fact, unlike the Data-dependent schemes, the data-independent DR schemes do not suffer from the scalability problem but could benefit from increased population size.

7.3 Future Work

Dimension reduction and especially for pattern recognition applications is an active area of research. Our investigations throughout this thesis revealed several areas that require more efforts. Here we only highlight few directions for our future work.

(1) In the future, the use of the feature-block strategy can be extended to other pattern recognition schemes, whenever features can be naturally split into different groups. This may also open the way for different approaches to fusion with promising results.

(2) In the last two chapters, we discovered that overcomplete Hadamard submatrices is a very rich pool for data-independent RPs for DR. We constructed three different types: SH-RD from Sylvester-type Hadamard matrices, WP-SRD, and WP-SD from Walsh Paley matrices, but we did not use the Walsh recursive constructions. Our work can naturally be extended to Walsh-based construction as well as other types of Hadamard matrices with a particular interest in constructing matrices that comply with CS requirements (i.e. RIP condition).

(3) The various Hadamard based random projection matrices, studied in this thesis, can be used and investigated in other pattern recognition/classification applications. Moreover, those satisfying the RIP condition (e.g. WP-SD dictionary) can be used as a good candidate in Compressive sensing applications. We also need to consider other biometrics to determine among other things if the observed pattern of performance by the different dictionaries in the two-case study are effected by the characteristics of recognition problem or not.

(4) In this thesis, we focused on comparing the different DR schemes by comparing the performance of pattern recognition schemes that deploy them as their DR step. In recent years, a new sophisticated approach to pattern and data analysis has emerged that argue for complementing the linear algebra aspects of the problem with

topological aspects. The emerging area of Topological Data Analysis (TDA) is based on taking into account the topology of the simplicial complexes of the digital high dimensional vector models of the object/pattern of interest. Rieck and Leitte, in (Rieck and Leitte, 2015), present a novel TDA-based evaluation scheme for DR techniques using the so-called Persistent Homology (from computational topology) which represent a well-known invariant of the sequence of simplicial complexes of the investigated dataset of high dimensional vectors. This invariant studies topological features of the given dataset in terms of the number of connected components, of the constructed simplicial complexes when one increases the connectivity distance threshold increases. They used this invariant to compare the quality of some data-dependent dimension reduction schemes. Such TDA approach can be expanded to assess data-independent DR techniques, and in particular it would be useful in providing a topological evaluation of the performance of our differently constructed Hadamard based overcomplete dictionaries.

REFERENCES

Abdulla, A., 2007. Wavelet Based Analysis for Video Packet Switching and Compression (M.Sc.). University of Buckingham, Buckingham.

Achlioptas, D., 2001. Database-friendly random projections. Presented at the ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems, ACM Press, pp. 274–281.

Agaian, S.S. (Ed.), 2011. Hadamard transforms. SPIE, Bellingham, Wash.

Al-Hassan, N., 2014. Mathematically inspired approaches to face recognition in uncontrolled conditions: super resolution and compressive sensing (doctoral). University of Buckingham.

Al-Talabani, A., 2015. Automatic Speech Emotion Recognition- Feature Space Dimensionality and Classification Challenges (doctoral). University of Buckingham.

Baraniuk, R., 2007. Compressive Sensing [Lecture Notes]. IEEE Signal Process. Mag. 24, 118–121.

Baraniuk, R., Davenport, M., DeVore, R., Wakin, M., 2008. A Simple Proof of the Restricted Isometry Property for Random Matrices. Constr. Approx. 28, 253–263.

Bingham, E., Mannila, H., 2001. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. Presented at the KDD, ACM, New York, NY, USA, pp. 245–250.

Boutsidis, C., Woodruff, D.P., 2014. Optimal CUR matrix decompositions. ACM Press, pp. 353–362.

Candès, E.J., 2008. The restricted isometry property and its implications for compressed sensing. Comptes Rendus Math. 346, 589–592.

Candes, E.J., Tao, T., 2006. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? IEEE Trans. Inf. Theory 52, 5406–5425.

Candes, E.J., Tao, T., 2005. Decoding by linear programming. IEEE Trans. Inf. Theory 51, 4203–4215.

Candes, E.J., Wakin, M.B., 2008. An Introduction To Compressive Sampling. IEEE Signal Process. Mag. 25, 21–30.

Chen, Z., Dongarra, J.J., 2005. Condition Numbers of Gaussian Random Matrices. SIAM J. Matrix Anal. Appl. 27, 603–620.

Dasgupta, S., Gupta, A., 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms 22, 60–65.

Donoho, D.L., 2006. Compressed sensing. IEEE Trans. Inf. Theory 52, 1289–1306.

Drineas, P., Mahoney, M.W., Muthukrishnan, S., 2008. Relative-Error CUR Matrix Decompositions. SIAM J. Matrix Anal. Appl. 30, 844–881.

Eyben, F., Wöllmer, M., Schuller, B., 2009. OpenEAR #x2014; Introducing the munich open-source emotion and affect recognition toolkit. Presented at the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6.

Fisher, R.A., 1936. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. Ann. Eugen. 7, 179–188.

Fraleigh, J.B., Beauregard, R.A., Katz, V.J., 1995. Linear algebra, 3rd ed. ed. Addison-Wesley, Reading, Massachusetts.

Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J., 2000. From few to many: generative models for recognition under variable pose and illumination. IEEE Comput. Soc, pp. 277–284.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. ed, Springer series in statistics. Springer, New York, NY.

Hecht-Nielsen, R., 1994. Context vectors: general purpose approximate meaning representations self-organized from raw data, in: Computational Intelligence: Imitating Life. IEEE Press, pp. 43–56.

Jassim, S., Al-Assam, H., Sellahewa, H., 2009. Improving performance and security of biometrics using efficient and stable random projection techniques. IEEE, pp. 556–561.

Johnson, W.B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space, in: Beals, R., Beck, A., Bellow, A., Hajian, A. (Eds.), Contemporary Mathematics. American Mathematical Society, Providence, Rhode Island, pp. 189–206.

Jolliffe, I.T., 2002. Principal component analysis, 2nd ed. ed, Springer series in statistics. Springer, New York.

Kahu, S., Rahate, R., 2013. Image Compression using Singular Value Decomposition. Int. J. Adv. Res. Technol. Volume 2, 244–248.

Kojadinovic, I., Wottka, T., 2000. Comparison between a filter and a wrapper approach to variable subset selection in regression problems. ESIT 2000.

Li, P., Hastie, T.J., Church, K.W., 2006. Very sparse random projections. Presented at the KDD, ACM Press, pp. 287–296.

Martinez, A.M., Kak, A.C., 2001. PCA versus LDA. IEEE Trans. Pattern Anal. Mach. Intell. 23, 228–233.

Mawlood, Z., 2016. Gait-Based Gender Classification Using Neutral and Non-Nuetral Gait Sequences (M.Sc.). Hasan Kalyoncu University, Turkey.

Pujol, A., Vitrià, J., Lumbreras, F., Villanueva, J.J., 2001. Topological principal component analysis for face encoding and recognition. Pattern Recognit. Lett. 22, 769–776.

Rauhut, H., 2010. Compressive Sensing and Structured Random Matrices, in: Theoretical Foundations and Numerical Methods for Sparse Recovery. pp. 1–92.

Rieck, B., Leitte, H., 2015. Persistent Homology for the Evaluation of Dimensionality Reduction Schemes. Comput. Graph. Forum 34, 431–440.

Rubinstein, R., Bruckstein, A.M., Elad, M., 2010. Dictionaries for Sparse Representation Modeling. Proc. IEEE 98, 1045–1057.

Sabir, A., Al-Jawad, N., Jassim, S., 2014. Feature selection gait-based gender classification under different circumstances. Presented at the SPIE 9139, Real-Time Image and Video Processing 2014.

Sabir, A.T., 2015. Human gait recognition under neutral and non-neutral gait sequences (doctoral). University of Buckingham.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A., 2009. Acoustic emotion recognition: A benchmark comparison of performances. Presented at the 2009 IEEE Workshop on Automatic Speech Recognition Understanding, pp. 552–557.

Steidl, S., 2009. Automatic classification of emotion-related user states in spontaneous childern's speech (PhD thesis). Depatrment of Computer Science, University of Erlangen-Nuremberg, Germany.

Sun, J., Xie, Y., Zhang, H., Faloutsos, C., 2007. Less is More: Compact Matrix Decomposition for Large Sparse Graphs, in: Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 366–377.

Theodoridis, S., Koutroumbas, K., 2003. Pattern recognition, 2. ed. ed. Academic Press, Amsterdam.

Turk, M., Pentland, A., 1991. Eigenfaces for Recognition. J. Cogn. Neurosci. 3, 71–86.

Vempala, S.S., 2004. The random projection method, DIMACS series in discrete mathematics and theoretical computer science. American Mathematical Society, Providence, R.I.

Wang, J., 2012. Geometric structure of high-dimensional data and dimensionality reduction. Springer, Beijing ; Higher Education Press ; Heidelberg ; New York.

Ye, J., 2005. Generalized Low Rank Approximations of Matrices. Mach. Learn. 61, 167–191.

Yu, S., Tan, D., Tan, T., 2006. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. Presented at the 18th International Conference on Pattern Recognition (ICPR'06), pp. 441–444.