BIAN, XIAO. Sparse and Low-Rank Modeling on High Dimensional Data: A Geometric Perspective. (Under the direction of Dr. Hamid Krim.)

High dimensional data exhibits distinct properties compared to its low dimensional counterpart, which causes a common performance decrease and a formidable computational cost increase of traditional approaches. Novel methodologies are therefore needed to characterize data in high dimensional spaces. In this thesis, we study the representation of high dimensional data by different low dimensional structures embedded in high dimensional spaces, and focus on novel paradigms for general machine learning purpose, such as clustering, classification and inference.

To address the nonlinearity of imagery sequential data, we map images into a circulant operator space with a proper Riemannian metric associated to data samples. A low rank model to achieve the optimal circulant operator subspace for the given dataset is proposed with an efficient algorithm on video-based human activity classification.

In order to further characterize general high dimensional data, we study the union-of-subspaces (UoS) model, as a generalization of the subspace model. The UoS model maintains the simplicity of the subspace model, and additionally has the ability to address nonlinear data. We show a sufficient condition to use $l_1$ minimization to reveal the underlying UoS structure, and propose a bi-sparsity model to recovery data from errors/corruptions. An effective algorithm, robust subspace recovery via bi-sparsity pursuit, is proposed and applied to data recovery and clustering.

Furthermore, in order to learn latent variables for a given dataset, we consider the analysis dictionary learning problem for the UoS model, and relate it to the sparse null
space problem. We thereby develop an algorithm, sparse null space basis pursuit, to learn the analysis dictionary, and show its efficacy in both synthetic data and real-world data.
Sparse and Low-Rank Modeling on High Dimensional Data: A Geometric Perspective

by
Xiao Bian

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina
2014

APPROVED BY:

__________________________
Dr. Huaiyu Dai

__________________________
Dr. Larry K. Norris

__________________________
Dr. Edgar Lobaton

__________________________
Dr. Hamid Krim
Chair of Advisory Committee
DEDICATION

To my mom and my dad.

Xiao’s research interests span the area of machine learning, signal processing and computer vision. He currently focus on unsupervised learning, particularly sparse and low-rank models of high dimensional data analysis, and potential applications on image/video inference, clustering and classification.
ACKNOWLEDGEMENTS

Everyone is shaped by their culture, their environment, and especially by the people around. By all means, I cannot emphasize enough the importance of the people I met to me, to this work, during my PhD study. This piece of work is not just about my research, but also a tip of iceberg that how much help and guidance I have received during these years.

I would like to first thank my advisor Dr. Hamid Krim for his instruction and accompany on my journey in high dimensional spaces. He has guided me through my exploration using his extraordinarily broad knowledge and sharp intuition in different areas, such as mathematics, signal processing, machine learning, computer vision, etc. Moreover, his great patience and trust on students has no doubt encouraged me to overcome difficult problems, and has created a very active and healthy research atmosphere in VISSTA group.

I would also like to thank my other committee members, in alphabetical order, Dr. Huaiyu Dai, Dr. Liyi Dai, Dr. Edgar Lobaton and Dr. Larry Norris, for their invaluable advice and help during my PhD study. I have been extremely fortunate to have such a diversified committee, from engineering to mathematics, to guide me through any potential problems. Moreover, I want to especially acknowledge Dr. Alex Bronstein for all the fascinating discussions, which help me formulate some of the ideas of my research work.

Additionally, I want to express my gratitude to all my lab mates, for the time shared with them, for their support and all helpful discussions.

I have my special thanks, in alphabetical order, to Dr. Yu Du, Lu Han, Shun Miao, Dr. Ye Tian, Dr. Tian Wang, Ziteng Wang and Dr. Sheng Yi.
Finally, I want to sincerely thank my parents and my wife. None of this would ever happen or even mean anything to me, without their heartwarming understanding and support.
# TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................. x

LIST OF FIGURES ........................................................................................................... xi

Chapter 1: Introduction ................................................................................................. 1
    1.1 Outline .................................................................................................................. 2
    1.2 Notation ............................................................................................................... 3

Chapter 2: Review of Previous Research ................................................................... 4
    2.1 The Curse of Dimensionality .............................................................................. 4
    2.2 Low Dimensional Structures in High Dimensional Spaces ............................... 6

Chapter 3: An Optimal Circulant Operator Subspace: Mending Nonlinearity .......... 14
    3.1 Introduction ......................................................................................................... 14
    3.2 Background ........................................................................................................ 16
    3.3 Image representation: circulant operator space ............................................... 17
        3.3.1 Approximated shift invariance .................................................................. 17
        3.3.2 Circulant operator space ......................................................................... 19
        3.3.3 Metric on an image operator space .......................................................... 20
    3.4 Image Operator Space on Human Activity Video Sequences ......................... 21
        3.4.1 Geometric Space of Human Activity Video Sequences ............................ 22
        3.4.2 Operator Space Construction ................................................................... 23
    3.5 Optimal Operator Subspace Pursuit .................................................................... 24
LIST OF TABLES

Table 5.1  Clustering error (%) on the Extended Yale Face Database B compared
to state-of-the-art methods [45] [68] [102] . . . . . . . . . . . . . . . . 73

x
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Euclidean distance become indiscernible in a high dimensional space</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>The histogram of Euclidean distances in different dimensions between random points under Gaussian distribution</td>
<td>7</td>
</tr>
<tr>
<td>2.3</td>
<td>$E{d_{max}/d_{min}}$ vs Dimensions</td>
<td>8</td>
</tr>
<tr>
<td>3.1</td>
<td>Graph of nonlinear function $g(\alpha, \beta)$ compared to a 2-D gaussian</td>
<td>18</td>
</tr>
<tr>
<td>3.2</td>
<td>Sample frames of human activity video clips (Upper: running; Lower: Siding)</td>
<td>22</td>
</tr>
<tr>
<td>3.3</td>
<td>Fiber bundle structure for data space</td>
<td>23</td>
</tr>
<tr>
<td>3.4</td>
<td>Homeomorphism between $\mathcal{H}$ and $\mathcal{B}$</td>
<td>25</td>
</tr>
<tr>
<td>3.5</td>
<td>Examples human activity video sequence and difference images. (The upper case is running, and the lower case is walking)</td>
<td>29</td>
</tr>
<tr>
<td>3.6</td>
<td>Optimal operator space intends to obtain a more compact representation for each sample video sequence, upon which a classifier is built.</td>
<td>31</td>
</tr>
<tr>
<td>3.7</td>
<td>Classification rate vs percentage of data using as training set</td>
<td>32</td>
</tr>
<tr>
<td>3.8</td>
<td>(a) A sample walking sequence (b) The walking sequence embedded in 1D and 2D (c) A sample bending sequence (d) Bending sequence embedded in 1D and 2D (x-axis is the index of frames)</td>
<td>35</td>
</tr>
<tr>
<td>3.9</td>
<td>(a) A sample bending sequence (b) Bending sequence embedded in 1D and 2D (x-axis is the index of frames)</td>
<td>36</td>
</tr>
</tbody>
</table>
Figure 3.10  (a) Embedded sequence of walking. Left: before correction. Right: after correction. (b) Embedding sequence of bending. Left: before correction. Right: after correction.

Figure 3.11  (a) Embedded sequence of walking. (b) Project another sequence of walking into the same space (c) Project a sequence of bending into the same space.

Figure 3.12  Classification rate compared to results without using continuous HMM. (Red: with continuous HMM. Blue: without continuous HMM)

Figure 4.1  Data in a union of subspaces

Figure 5.1  Illustration of Problem 5.1.1

Figure 5.2  An example of robust subspace exact recovery.

Figure 5.3  Comparison with Robust PCA.

Figure 5.4  Overall recovery results of RoSuRe and Robust PCA. [0 0.2] is mapped to [1 0] of grayscale image.

Figure 5.5  Background subtraction on traffic videos (static camera)

Figure 5.6  Background subtraction on traffic videos (panning camera)

Figure 5.7  Coefficient matrix $W$ (a) without rearrangement according to the position of the camera (b) with rearrangement according to the position of the camera.

Figure 5.8  Sample face images in Extended Yale face database B

Figure 5.9  Clustering Accuracy vs The value of $\lambda$

Figure 5.10  Affinity matrix for face images from different subjects

Figure 5.11  Recovery results of human face images. The three rows from top to bottom are original images, the components $E$, and the recovered images, respectively.

Figure 6.1  An example of the result of SNS-BP

Figure 6.2  $\|\hat{N}\|_0/\|N\|_0$ vs Sparsity

Figure 6.3  Sample synthetic data matrix and its underlying structure

Figure 6.4  Sparse coefficients using learned analysis dictionary

xii
Figure 6.5  Textures and the corresponding patches for training  . . . . . . . . . 91
Chapter 1

Introduction

High dimensional data analysis has recently become a central topic in both academia and industry. The fast progress of techniques related to data acquiring, storage and processing has lead the researchers, from applied mathematicians to electrical engineers, to face problems of data of increasing size and dimensions.

Many well-established problems, such as data classification and clustering, therefore encounter novel challenges in the context of high dimensionality. Generally, the increasing dimensionality comes with redundant and irrelevant information, which may actually hinder the discovery of key information without carefully designed methods. The combinatorial nature of finding the suitable features essentially makes the problem of feature selection even more challenging with high dimensional data. Additionally, insufficient samples in the vast high dimensional space make the probabilistic estimation unreliable. Furthermore, the distinct properties of high dimensional space, such as discernible Euclidean distance [2] and “the hubness” [82], also impact the performance of traditional data analysis algorithms on high dimensional data.

In this thesis, we seek to understand the high dimensional space from a geometric perspective, and to explore the low rank and sparse models for high dimensional data analysis. Based on these geometric models, we further propose novel algorithms to overcome the challenges inherent to high dimensional data. As we later elaborate, we achieve superior performances on the classification and clustering problems in various high dimensional data types.
1.1 Outline

The rest of the thesis is organized as follows. In Chapter 2, we first give a review of the research on the properties of high dimensional spaces, and the state-of-the-art on the high dimensional data models. In fact, these works shed light on our framework of high dimensional data analysis.

In Chapter 3, we propose a low dimensional model, an optimal circulant operator subspace, to analyze imagery data. Natural images exhibit high nonlinearity, which usually causes the failure of linear low dimensional models in finding a low dimensional embedding. Instead of directly working on the high curvature unknown image space, we utilize the circulant operator space, with a properly defined Riemannian metric based on given image samples, to find an appropriate low dimensional embedding of the given image set. To further analyze the sequential imagery data, we exploit a fibre bundle formalism to model various realizations of each trajectory, and characterize these high dimensional data sequences by an optimal operator space. The low dimensional structure intrinsic to the data is further explored, by minimizing the dimension of the operator subspace under data-driven constraints.

In light of the parsimonious degrees of freedom of real-world data, we introduce a more advanced model, the union of subspaces (UoS), to characterize high dimensional data in Chapter 4. As a generalization of the subspace model, this model preserves the simplicity of the subspace model, and approximate the nonlinearity of the data distribution using a union of subspaces. We demonstrate the formulation of the UoS model as a $l_1$ minimization problem, and show a sufficient condition for the minimizer to reveal the underlying UoS structure of the given dataset. We further discuss the state-of-the-art algorithms to solve this problem.

We consider the problem of recovering data from errors/corruptions in high dimensional spaces using the UoS model in Chapter 5. We propose a bi-sparse model as a framework to analyze this problem and provide a novel algorithm, named robust subspace recovery (RoSuRe), to recover the union of subspaces in presence of sparse corruptions. We show that, this algorithm not only recover data from errors but also provide the information about the UoS structure of the given dataset. We further demonstrate
the effectiveness of our method by experiments on both synthetic data and real-world vision data, and illustrate the potential applications on data recovery and data clustering problem.

Towards the discovery of latent variables, we study sparse models in dictionary learning in Chapter 6, and relate this problem with another interesting topic, sparse null space problem (SNS). We investigate the relation between the SNS problem and the analysis dictionary learning problem, and show that the SNS problem plays a central role, and may be utilized to solve dictionary learning problems. Moreover, we propose an efficient algorithm of sparse null space basis pursuit, and extend it to a solution of analysis dictionary learning. Experimental results on numerical synthetic data and real-world data are further presented to validate the performance of our method.

Finally, in Chapter 7, we conclude our journey in high dimensional spaces with sparsely distributed, structured data.

1.2 Notation

A brief summary of notations used throughout this thesis is follows: The dimension of a $m \times n$ matrix $X$ is denoted as $\text{dim}(X) = (m,n)$. $\|X\|_0$ denotes the number of nonzero elements in $X$, while $\|X\|_1$ denotes the vector $l_1$ norm. For a matrix $X$ and an index set $J$, we let $X_J$ be the submatrix containing only the columns of indices in $J$. $\text{col}(X)$ denotes the column space of matrix $X$. We write $P_{\Omega_A}X$ as the orthogonal projection of matrix $X$ on the support of $A$, and $P_{\Omega_A}X = X - P_{\Omega_A}X$. The sparsity of a $m \times n$ matrix $X$ is denoted by $\rho(X) = \frac{\|X\|_0}{mn}$. 
Chapter 2

Review of Previous Research

2.1 The Curse of Dimensionality

The term “the curse of dimensionality” was first proposed by Richard Bellman in the context of dynamic programming [11]. With more and more exposure to high dimensional data in both academia and industry, it has been widely used to represent the challenges of high dimensional data analysis. In this section, we provide a brief review of recent and current research on exploring the properties of high dimensional space. Indeed, these distinct properties of high dimensional space unveil the reason why popular algorithms like $k$-nearest neighbors (KNN) fail, and further shed light on possible approaches on high dimensional data analysis.

One obvious yet important property of high dimensional space is the insufficiency of data samples. In particular, the volume among samples goes exponentially with the dimensionality, which essentially implies that, to maintain the same level of sample density, we also need the sample size to increase exponentially. Take images/videos for instance, each data point has typically tens of thousands dimension. It is hence very difficult to estimate the data distribution in a high dimensional space, with a limited number of samples [16].

From the perspective of pattern recognition, the power of classifiers degrade with a fixed sample size as the number of features increases, which is referred to as the Hughes effect [53]. One potential reason is that, with an increasing number of features, we
also introduce more noise components to the data \[107\]. Additionally, the number of irrelevant features also increases in a higher dimensional feature space. In fact, in most real-world problems, an effective data model is usually dominated by only a few variables. It would rather lead to ambiguity with many more features than the intrinsic degrees of freedom of a model. Considering human activity classification for example, the pixels of background and those of a person’s clothes are typically not related to the class identity of each presented image. It is hence unwise to treat each pixel as one feature and to address the activity classification problem in this high dimensional feature space, since the same activities may show in completely different backgrounds and contexts \[16\].

Furthermore, the Euclidean distance in high dimensional spaces becomes indiscernible \[51\] \[2\], which essentially shakes the foundation of KNN-based approaches on high dimensional data.

Specifically, \[51\] shows that, assume the data distribution in all dimensions are i.i.d and all the appropriate moments are finite, then for every \(\epsilon > 0\), the maximum distance \(d_{\text{max}}\) and the minimum distance \(d_{\text{min}}\) of \(N\) points to the query point satisfy the following relation,

\[
\lim_{d \to 0} P(d_{\text{max}} \leq (1 + \epsilon)d_{\text{min}}) = 1. \tag{2.1}
\]

Under this circumstance, the nearest neighbor of a query point becomes unstable.
Since a lot of distances are similar, a small perturbation may lead to an entirely different nearest neighbor.

In Fig. 2.2, we show the distribution of pair-wise distances of randomly generated points in different dimensions. In particular, the data distribution in each dimension is i.i.d Gaussian. To better visualize the results, we further normalize the distance such that the maximal distance in each case is 1. We can see the trend that, from $d = 2$ to $d = 200$, the distribution of distance becomes increasingly concentrated and inclines to 1. In the 200 dimensional space, all pair-wise distance become similar to each other and close to the maximum.

We further show the mean ratio of $d_{\text{max}}$ to $d_{\text{min}}$ in spaces with various dimensions in Fig. 2.3. We use the same method to generate data as in Fig. 2.2. When the dimensionality of the space increases, the ratio of $d_{\text{max}}$ to $d_{\text{min}}$ is getting closer to 1. All these numerical results illustrate that the maximum distance and minimum distance of a dataset become indiscernible in high dimensional spaces. These phenomena essentially imply the instability of kNN-based approaches in high dimensional spaces.

### 2.2 Low Dimensional Structures in High Dimensional Spaces

A natural approach to high dimensional data analysis is dimension reduction. If we can indeed find a more compact space of the given data, then the original challenging high dimensional problem would degenerate to its low dimensional counterpart. Extensive research has been carried out to model and to analyze the distribution of a given dataset as some low dimensional structure embedded in a high dimensional space [85] [35] [96]. Yet ahead of reviewing the algorithms of dimension reduction, it is of fundamental interest to validate the existence of such low dimensional structures in high dimensional data.

To answer this question from a theoretical perspective, Johnson-Lindenstrauss lemma (JL lemma) [54] [31] shows that, for a $n$-point set $S$, there exists a linear mapping $f$ that map these $n$ points into a Euclidean space of dimensionality $O(\log(n)/\epsilon^2)$, with the
Figure 2.2: The histogram of Euclidean distances in different dimensions between random points under Gaussian distribution
Figure 2.3: \( E\{d_{\text{max}}/d_{\text{min}}\} \) vs Dimensions
pairwise distance between these points preserved as

\[(1 - \epsilon)\|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1 + \epsilon)\|x - y\|^2, \forall x, y \in S. \tag{2.2}\]

Given a high dimensional dataset, especially when the dimensionality is larger than the number of data points, JL lemma guarantees that we can always find a linear mapping to map the high dimensional data into a low dimensional space, without much loss of information.

Furthermore, from a practical perspective, extensive works on different dataset show that low dimensional structures exist ubiquitously in real-world datasets, such as human faces, natural images and video sequences [96] [35] [85] [25]. Indeed, the degrees of freedom of a real-world dataset is typically much lower than the ambient space dimension. The challenge, therefore, is to recovery the underlying low dimensional structure in a given high dimensional dataset.

In particular, when the data is concentrated in a subspace, and is contaminated by mild noise, principal component analysis (PCA) is introduced to find the optimal subspace for the dataset [76] [37]. PCA essentially finds the directions that maximize the variance of a given dataset. The distribution of the data is therefore preserved in a low dimensional space to a maximum extent. Equivalently, we can also formulate this problem as an constrained optimization problem as follows,

\[
\min \|X - Y\|_F \quad \text{s.t. } \text{rank}(X) \leq k, \tag{2.3}
\]

where \(X\) is the data matrix, and \(Y\) is the corresponding low dimensional embedding. Note that (2.3) has a closed-form solution, by keeping the top \(k\) singular values and their associated singular vectors of \(X\) [37].

However, PCA is not robust to outliers, and often fails to recover the intrinsic low dimensional structure with the presence of outliers. To further solve this issue, robust PCA was proposed to recover the low dimensional structure with the presence of errors/outliers [25]. Specifically, given a dataset \(X = [x_1, \cdots, x_n]\), \(x_i \in \mathbb{R}^n\), the goal is to decompose \(X\) into two parts: a low rank matrix \(L\) representing the low dimensional subspace, and a sparse matrix \(E\) representing the potential errors/outliers. The decom-
position can be done by simultaneously minimizing the nuclear norm of $L$, which can be seen as a convex relaxation of the rank of $L$ [22], and the $l_1$-norm of $E$, which is the convex relaxation of the sparsity of $E$ [65],

$$\min_{L,E} \|L\|_* + \lambda \|E\|_1 \text{ s.t. } X = L + E. \quad (2.4)$$

This model can recover the low dimensional subspace for a given dataset more accurately with the presence of outliers and sparse errors [25].

Besides the subspace model, data may exhibit nonlinear relations among themselves, and are therefore distributed in a manifold embedded in a high dimensional space from a more general perspective. Imagine the points distributed in a 2-sphere $S^2$ embedded in $R^3$. Finding an optimal subspace for this dataset, which is essentially $R^3$ in this case, will make us lose insight of the degree of freedom being actually 2. Various manifold learning algorithms are proposed to learn the underlying data manifold embedded in high dimensional spaces [85] [9] [35] [95]. One effective way is to preserve the geodesic distance instead of Euclidean distance in the original high dimensional space, upon which the algorithm Isomap is built [95]. Specifically, a graph based on $k$ nearest neighbors is constructed. Isomap then pursues a low dimensional embedding by preserving the pairwise geodesic distances on that graph. The resulting low dimensional embedding reflects more of the intrinsic geometry of the given data compared to linear embeddings as the subspace model. Yet one potential problem for Isomap is that, miss-calculated $k$ nearest neighbors can cause a dramatical change on geodesic distances, the so called “short-circuit errors” [5]. The whole embedding may change, as a result.

In contrast to Isomap that emphasizes longer distances more than the shorter ones, another way to preserve the intrinsic geometry of high dimensional data, named locally linear embedding (LLE), is to focus on shorter distances. Specifically, LLE utilizes $k$ nearest neighbors to estimate the tangent space of each point, and preserves the local neighborhood relations in each tangent plane in the low dimensional embedding space. Instead of preserving geodesic distances, LLE are more robust to “short-circuit errors” by concentrating on local neighborhoods.

Isomap and LLE all start from constructing a kNN-based graph, and focus on preserv-
ing some properties of this graph in the low dimensional embedding space. Essentially, this weighted neighborhood graph is considered to be an approximation of the low dimensional data manifold. It is therefore natural to consider approaches that preserves the graph structure rather than only distances. Laplacian eigenmaps is one of the representative state-of-the-art in this category [9]. This algorithm reconstruct the graph in the low dimensional embedding space by preserving the corresponding Laplacian operator. Laplacian eigenmaps focus on intrinsic geometric structures, and is hence insensitive to noise [9].

As we discussed above, manifold learning algorithms rely on the approximation of the tangent space of the data manifold using $k$ nearest neighbors. When the $k$ nearest neighbors become unreliable, or the local sample density is not sufficient to estimate the tangent space, then the accuracy of the low dimensional embedding degrades dramatically. Unfortunately, these are the typical issues we are facing in high dimensional data analysis [39] [51] [82]. Without knowing reliable relations among data samples, embedding data into a low dimensional space still suffers from “the curse of dimensionality” from the beginning. Recent research on sparse representation shows that we can utilize the property of high dimensional data to conquer them [42] [34] [44] [101].

The story of sparse models starts from the notion that the number of degrees of freedom of high dimensional data is much smaller than the dimensionality of the ambient space [42]. In fact, consider the model behind a given high dimensional data sample, the underlying dominant variables are typically very few, and hence can be represented in a low dimensional space. Specifically, assume that each data sample is a linear combination of $k$ atoms from an atom set or dictionary $D$, we then have

$$ x = \sum_{d_i \in D} w_i d_i, i = 1, \cdots, k. \tag{2.5} $$

To formulate this problem into a matrix form, (2.5) is then written as

$$ x = Dw, \|w\|_0 \leq k. \tag{2.6} $$

The problem (2.6) essentially implies choosing $k$ atoms from an $n$-atom set, which
seems fail to escape from “the curse of dimensionality”. However, recent research shows that we can efficiently find $w$, when $D$ approximately preserves the length of sparse vectors, by considering the convex relaxation of (2.6) [42] [23], and thereby solving the following optimization problem,

$$\min_w \|w\|_1 \text{ s.t. } Dw = x. \quad (2.7)$$

We can then use a $k$-sparsity vector $w$ to represent each $x$. Consider data samples that share the same set of underlying variables and reside in the same subspace, we therefore have a union of subspaces (UoS) for different subsets of data that are dominant and represented by different subsets of variables. The UoS model can be seen as a generalization of the subspace model, but is able to address the nonlinear data distribution of different topologies. In particular, the data samples in the same subspace preserve the local linearity. However, data from different subspaces exhibit highly nonlinear relations. The addition of two points from different subspaces is not able to produce any valid data. The vast volume among the union of subspaces coincides with the fact that real-world data only reside in a tiny fraction of the high dimensional ambient space. Moreover, the UoS model relies less on the local relations of each data, and is hence insensitive to different topologies of data distributions [93] compared to kNN-based manifold learning algorithms [5].

Research has also explored the UoS structure of a high dimensional dataset when a well-formulated dictionary is not available [44] [68] [93]. Specifically, the data matrix $X$ is utilized as a dictionary $D$, and the problem of (2.6) is formulated as follows,

$$\min_W \|W\|_1 \text{ s.t. } XW = X, \text{diag}(W) = 0. \quad (2.8)$$

In [44], the UoS model for data under mild noise is studied, and it has been successfully applied to the motion segmentation problem. [93] further considers the $l_1$ norm of $w$ for each data to distinguish very few outliers from the majority of the data, and provide a thorough theoretical analysis on it. [68] follows a different model that, rather than modeling the coefficient matrix $W$ as a sparse matrix, they consider the low rank structure of $W$. They further design a column wise-sparse matrix $E$ to represent the few outliers.
in the data matrix, and the problem is formulated as

$$\min_{W,E} \|W\|_* + \lambda \|E\|_1 \ s.t. \ X = XW + E.$$  \hspace{1cm} (2.9)

This problem can be solved efficiently by linearized alternating direction method[67].
Chapter 3

An Optimal Circulant Operator Subspace: Mending Nonlinearity

3.1 Introduction

One of the major challenges of high dimensional data analysis is that, nonlinearity often accompanies high dimensionality. The high dimensionality, representing the context of the problem, becomes even more puzzling when the curved data manifold exploits the vast space of a high dimensional world. In particular, when the data reside in a manifold embedded in a Euclidean space, traditional dimension reduction techniques, such as PCA, can only find the minimum subspace that contains the data manifold. Therefore, if the data manifold itself is highly curved, i.e. of high curvatures, pursuing a linear subspace just provides a very coarse measure of the data space.

Unfortunately, it is practically very common that we have a curved data space rather than a linear one. Take images for example, when represented in the popular matrix form, even a simple translation motion causes nonlinearity in the matrix space. In particular, consider an image $I_1 = I(u, v)$ and its translation $I_2 = I(u + x, v + y)$, generally there is no linear relation such as $I_1 = \alpha I_2$ between them. Moreover, if we further consider the three consecutive points, $I(u, v), I(u + x, v + y), I(u + 2x, v + 2y)$, using $(I(u, v) + I(u + 2x, v + 2y))/2$ to approximate $I(u + x, v + y)$ would be a terrible idea. This essentially implies that the image space is full of high curvature points, that it is rather difficult to
do any interpolation.

The first step to analyzing a large high dimensional dataset therefore always involves addressing the nonlinearity. On one hand, we can directly learn the data manifold from the given samples, and therefore have a nonlinear dimension reduction on the dataset [18] [1] [35] [10] [104]. This approach utilizes the relations among various images to estimate the underlying manifold structure. It, however, requires dense samples to cover the high curvature area, which is extremely difficult to satisfy when the data dimension is high. On the other hand, we can mitigate the nonlinearity of the data by extracting appropriate features by exploring the structures inside each sample [48]. Yet it is difficult and rather heuristic to design feature extractors that are adaptive to a specific dataset. Moreover, the lack of a framework of feature extractor design impacts the generalizability of this class of method. The problems of these existing methods naturally raise a question about the integration of these two classes of approaches, i.e., is there a general framework for feature extraction to mitigate the nonlinearity, or equivalently mapping the given data to another space, and further learn the manifold structure in the feature space? If so, would it provide us a more accurate description of the given high dimensional data?

We discuss these questions in this chapter, and propose a novel framework for high dimensional data analysis. We describe image sequences using the formalism of fiber bundles, and construct an operator space $H$ which is homeomorphic to the manifold of hidden states of sequences. The operators on the data space have a specific targeted output to facilitate the study of the geometry of the space. As a result, instead of working on an a priori unknown data space (space of images), we exploit the corresponding smoother operator space to categorize and classify different data sequences, by first developing an algorithm to find the optimal low dimensional operator space where the discriminating information is compactly stored. It is worth pointing that, fundamentally different from current manifold learning techniques, this operator space pursuit method allows us to emphasize structures within each data point on the manifold, as such information is often crucial in high dimensional data such as images. In addition, we propose an operator-based manifold learning algorithm to explore the nonlinear dynamics of each video sequence data and build hidden Markov models on the embedding sequences to
explicitly characterize the temporal relations among frames. We illustrate the viability of the framework and validate the models by human activity video sequence data analysis.

3.2 Background

Of keen interest to artificial intelligence and computer vision, research activity in video sequence data analysis has witnessed an explosive growth in recent years. Tools from areas, such as machine learning, computer vision, optimization and statistical analysis have been collectively utilized to address this problem [59] [78] [61] [99] [18] [1] [104].

In contrast to classical low dimensional signal processing, where random noise is the major challenge, the key factor in video sequence analysis is the discovery of useful information hidden in high dimensional data. More specifically, natural images are widely recognized to reside in a low dimensional space given the fact that pixels are strongly correlated, even though they generally occupy a high dimensional space. It is hence natural to seek and apply algorithms of nonlinear dimension reduction to estimate and recover the intrinsic low dimensional structure. In [18][104], different manifold learning techniques are applied to analyze human activity video sequences as curves on the given data manifold. By embedding video sequences as high dimensional curves into a low dimensional space, temporal dynamics of video sequences are assumed to be preserved, and subsequent classification is based on the similarities of low dimensional curves. On account of the highly curved property of the image space, these algorithms heavily rely on extracting the critical features in the preprocessing step. In [18], silhouettes of gestures in each frame are extracted such as each video is treated as a curve on the assumed manifold. In[104], they utilize the contour of a human body as a shape so that sequences are seen as flows on the shape manifold, and statistical analysis can be carried out via flat connections on the well-defined manifold.

While this type of methods manages to preserve dynamic features of video clips in a low dimensional space, the preprocessing cost to extract relevant information is heavy. The preprocessing cost further increases dramatically when handling more complex scenarios as multiple objects of interest in a video clip are being presented.

Another promising methodology to analyze video sequences relies on extracting dis-
criminant spatial features among frames [59] [78] [61] [99]. Each video sequence is hence considered as a set of feature vectors, and the distance between two videos is determined by the two sets. Coupled with sparse modeling and deep learning, these methods show superior ability to analyze large data set. An over-complete dictionary is automatically learned from the unlabeled data, and with a small set of labeled data point, different class of samples are represented by different combinations of words in the dictionary. One drawback of these techniques is their failure to account for temporal information in video sequences. While their classification ability is generally strong (e.g. videos of cars vs. human), their performance at more refined scenarios, such as recognizing different type of human actions, leaves more room for improvement.

3.3 Image representation: circulant operator space

3.3.1 Approximated shift invariance

Denote the $L^2$ space of functions $x(u, v)$ such that $\int |x(u, v)|^2 du dv < \infty$ by $E$. Each Image, represented by a $m \times n$ matrix, may be seen as a discrete two dimensional function $x(u_i, v_i)$ sampled from $x(u, v) \in E$. However, $E$ and its natural metric are not sufficient to describe natural images due to numerous group transform operations for natural images beyond the $L_2$ metric. Indeed, it is critical to consider invariance under different operations, such as shift and rotation, to define an appropriate metric for image space.

Among all general operations on images, shift is one of the most fundamental ones, and can be seen as a basic building block to constructing other more elaborate operations. On the other hand, human vision shows an acute ability to identifying shift invariance rather than other operations, such as rotation and deformation, etc. However, strict shift invariance can only be achieved by using the power spectrum of the corresponding Fourier transform. A loss in phase information impacts classification of different types of images [79][56]. In this section, we focus on introducing another space defined on $E$ to obtain an approximated shift invariance, such that under the associated metric, only images under small translation are mapped to be close to each other.

Specifically, we define the notion of approximated shift invariance as follows,
**Definition** (Approximated shift invariance) For an image \( x(u,v) \in \mathcal{E} \), consider a functional \( h \) of \( \mathcal{E} \), and let \( g(\alpha, \beta) = \| h \circ x - h \circ x(u - \alpha, v - \beta) \| \), we say \( h \circ x \) satisfies the approximated shift invariance with parameter \((c, K)\) if

\[
g(\alpha, \beta) = \begin{cases} 
\leq K_1 & |\alpha| < c_1, |\beta| < c_1 \\
\geq K_2 & |\alpha| > c_2, |\beta| > c_2
\end{cases}, \tag{3.1}
\]

where \( 0 < c_1 < c_2 \) and \( 0 < K_1 < K_2 \).

Towards a closed form of \( h \) for each \( x \in \mathcal{E} \), we first construct a closed form of \( g(\alpha, \beta) \) as follows,

\[
g(\alpha, \beta) = 1 - e^{-(\frac{\alpha^2}{2c_1^2} + \frac{\beta^2}{2c_2^2})}. \tag{3.2}
\]

We can see that \( g(\alpha, \beta) \) in (3.2) has the same shape as a negative gaussian, and satisfies \( \min g(\alpha, \beta) = f(0,0) = 0 \). Specifically, when \( \alpha \ll c_1 \) and \( \beta \ll c_2 \), \( g(\alpha, \beta) \approx 0 \) and when \( \alpha \gg c_1 \) and \( \beta \gg c_2 \), \( g(\alpha, \beta) \approx 1 \).

With \( f(\alpha, \beta) \) as in (3.2), we can calculate the corresponding functional \( h \) for each \( x \), given that

\[
g(\alpha, \beta) = h \circ x(u + \alpha, v + \beta) = h \ast x. \tag{3.3}
\]
The explicit form of $h$ is then pursued by the following convex optimization problem,

$$h_x = \arg\min \{ \| g - h \ast x \| \}. \quad (3.4)$$

### 3.3.2 Circulant operator space

Note that $h \ast x$ can be replaced by $H \circ x$, where $H$ is a circulant operator defined on $\mathcal{E}$. A circulant operator defined on $\mathbb{R}^n$ can be represented in the form of

$$H = \begin{pmatrix} c_0 & c_{n-1} & \cdots & c_1 \\ c_1 & c_0 & \cdots & c_2 \\ \vdots & \vdots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & \cdots & c_0 \end{pmatrix}, \quad (3.5)$$

where $h$ is the first row of $H$ as $[c_0, c_{n-1}, \cdots, c_1]^T$.

In fact, all circulant matrices form a commutative algebra [33]. Consider a circulant matrix space is $\mathcal{H}$, for any $A, B \in \mathcal{H}$, $A + B \in \mathcal{H}$ and $AB \in \mathcal{H}$. Moreover, a pairwise commuting of a set of circulant matrices leads to a simultaneous diagonalization [36] [33]. They hence share the same eigenvectors, which are given by $v_j = (1, \omega_j, \omega_j^2, \cdots, \omega_j^{n-1}), \omega_j = e^{2\pi ij/n}$. All the above features lead to the property that the topology of $\mathcal{H}$ is an $n$-dimensional vector space. This, in turn, affords us to define an inner product and hence of distance in the vector space. Furthermore, given that all circulant matrices share the same eigenvectors as the Fourier transform matrix, the computation may further be simplified if performed in the spectral domain.

Specifically, let $F(\cdot)$ denote the Fourier transform, we have

$$H \circ x = h \ast x = F^{-1}\{F(h) \cdot F(x)\} = g, \quad (3.6)$$

where $g$ is a fixed 2D gaussian function, and $A \cdot B$ is the Hadamard product of $A$ and $B$. 

It follows that

\[ G = F(g) = F(h) \cdot F(x). \] (3.7)

Finally, \( h \) is uniquely determined by the following convex optimization problem,

\[ h = \arg \min \{ \| G - F(h) \cdot F(x) \| \}, \] (3.8)

### 3.3.3 Metric on an image operator space

Defining an appropriate metric on a circulant operator space, so as to measure distances among different elements in the space, is crucial for distinguishing different images for our problem. It seems arbitrary to define a metric on an \( n \)-dimensional vector space, and one simple way is to inherit the metric from Euclidean space \( \mathbb{R}^n \). However, in order to characterize the image sequences behind the circulant operator space, we may need further constraints to define a more appropriate metric for the operator space. To comply with the nature of the image representation, as well as to simplify the expression, we use the name “image operator space” instead of the metric circulant operator space hereafter.

Since \( H \) can be parameterized by \( h \in \mathbb{R}^n \), we therefore use \( h \) instead of \( H \) to represent an image operator. In particular, consider \( h \in \mathcal{H} \) and \( x \) its associated image, it is natural to define a quotient space on \( \mathcal{H} \), such that operators with the same associated image form an equivalence class,

\[ \tilde{\mathcal{H}} = \mathcal{H} / \sim = \{ [h] : h \in \mathcal{H} \} = \{ \{ f \in \mathcal{H}, f \sim h \} : h \in \mathcal{H} \}, \]

\[ f \sim h \text{ if and only if } h \circ x = f \circ x. \]

Let \( \Delta h \) be in the tangent space \( T_h \tilde{\mathcal{H}} \) of \( h \), it follows that

\[ \Delta h \sim 0 \text{ if } \Delta h \circ x = 0. \]
Consequently, for \( u, v \in T_h\tilde{\mathcal{H}} \), the Riemannian metric on \( \tilde{\mathcal{H}} \) can be defined as

\[
\langle u, v \rangle_h = \langle u \circ x, v \circ x \rangle,
\] (3.9)

where \( h \) is the associated image operator of \( x \). It is trivial to show that \( \langle u, u \rangle_h = 0 \) iff \( u \sim 0 \), and (3.9) is hence positive definite.

The geometry of \( \tilde{\mathcal{H}} \) is therefore determined by the associated image space \( X \) as in (3.9) for all \( x \in X \). We can then theoretically calculate the distance between any two points \( a \) and \( b \) on \( \tilde{\mathcal{H}} \) by measuring the length of the geodesic between them. However, the metric (3.9) needs information from the corresponding points in the image space \( \mathcal{E} \). Generally, very limited images/video frames on the entire trajectory are available, which makes the estimation of the complete metric on the tangent bundle impractical.

To overcome the difficult due to insufficient samples, we estimate the distance by projecting one image onto the others’s tangent space, and then carrying out the evaluation in that tangent space. Specifically,

\[
L(a, b) = \|a \circ x_b - b \circ x_a\|_2
\] (3.10)

\[
dist(a, b) = \frac{L(a, b) + L(b, a)}{2}
\] (3.11)

Note that to ensure the symmetry property of distance, we use the mean of the two measurements as the final estimated distance between \( a \) and \( b \).

### 3.4 Image Operator Space on Human Activity Video Sequences

The geometry of the image space and its corresponding operator space are capable of addressing various scenarios. Finer structures can be defined on the image space based on the needs of applications, and subsequently can be analyzed in the operator space. In this section, we introduce the necessary geometric structures to analyze formulate human activity video sequences, and formulate the problem within this framework.
3.4.1 Geometric Space of Human Activity Video Sequences

Let the space of all images be $\mathcal{E}$, then each image sequence may be viewed as a sampled curve in $\mathcal{E}$. Consider $\mathcal{E}$ as a metric space, implying that there is no guarantee that curves from the same class will be close to each other in the original image space, since for a given class of image sequences, different realizations exist.

Let each frame in a video sequence correspond to a hidden state which lies on a manifold $\mathcal{B}$ embedded in $\mathcal{E}$. These hidden states may be seen as control variables of a certain activity. We may then invoke the fiber bundle formalism to describe the data set (Fig.3.3), $\mathcal{E}$ being the global space for all image frames. In the case of $m \times n$ gray-level images, $\mathcal{E}$ is the space of $m \times n$ matrices; $\mathcal{B}$ is the base space of the bundle for all control variables; fiber $\mathcal{F}$ over $b \in \mathcal{B}$ is the space for different realizations of a control variable. $\pi : \mathcal{E} \rightarrow \mathcal{B}$ is a continuous surjection such that for a neighborhood $U \in \mathcal{B}$, $\pi^{-1}(U)$ is homeomorphic to the product space $U \times \mathcal{F}$.

Then each image sequence is a high dimensional curve in $\mathcal{E}$ corresponding to a low dimensional curve which lies on the base manifold $\mathcal{B}$. Under this setting, different realizations of the same activity, may vary in $\mathcal{E}$, but should follow the same trajectory in $\mathcal{B}$ up to some noise term. It is hence natural to focus on finding the base manifold $\mathcal{B}$, and extracting important features from $\mathcal{B}$ to categorize video sequences.
3.4.2 Operator Space Construction

In human activity analysis, or in video sequence analysis in general, there is no explicit form of the base manifold $B$. We then utilize the image operator space $\mathcal{H}$, and build the homeomorphism between $\mathcal{H}$ and $B$. We can subsequently investigate the space $\mathcal{H}$ instead of the unknown manifold $B$.

In particular, to establish a homeomorphism between $B$ and $\mathcal{H}$, it is equivalent to construct a 1-1 and onto mapping in between. Considering the data as sampled curves in $\mathcal{E}$, we see $B$ as the set of equivalent classes of elements of fiber $\pi^{-1}(p) = \{F \text{ over point } p \in \mathcal{E}\}$: $B = \{[x] : x \in \pi^{-1}(p)\}$. To map all points in one fiber $\mathcal{F}_p$ to a unique element in $\mathcal{H}$, we can empirically establish a homeomorphism between the operator space $\mathcal{H}$ and $B$ as follows:
\[ h_x = \arg \min_h \sum_i \| x_i * h - g \|_2, \]  
(3.12)

where \( x_i, i = 1, \ldots, m \) are samples on fiber \( \mathcal{F}[p] = \pi^{-1}(p_x) \), and \( g \) is a fixed 2-D Gaussian function.

Since the objective functional is convex, we have a unique solution \( h \{ \mathcal{F}[p] \} \) for each fiber \( \mathcal{F}[p] \). Similar to (3.8), the optimal operator has a closed form solution as

\[
\begin{align*}
X_i &= F(x_i), \quad G = F(g), \\
H_x &= \frac{\sum_i G \cdot X_i^*}{\sum_i X_i \cdot X_i^*}, \\
h_x &= F^{-1}(H_x). 
\end{align*}
\]
(3.13) (3.14) (3.15)

Consequently there exists a 1-1 and onto mapping between the operator space \( \mathcal{H} = \{ h \{ \mathcal{F}[p] \} \} \) and \( \mathcal{B} \). Therefore, instead of working on the unknown structures of \( \mathcal{B} \), we use sequences of operators in \( \mathcal{H} \) to categorize and analyze image sequences.

### 3.5 Optimal Operator Subspace Pursuit

#### 3.5.1 Problem formulation

Since the operator subspace in \( \mathcal{H} \) for a video sequence \( X = [x_1, x_2, \ldots, x_m] \) lies in a low dimensional subspace as a base manifold \( \mathcal{B} \), it is natural to find the optimal subspace by solving the following constrained dimension minimization problem,

\[
\begin{align*}
\min & \ dim(\mathcal{H}) \\
\text{s.t.} & \| h_i(x_i) - g \|_2 \leq c, \quad h_i \in \mathcal{H},
\end{align*}
\]
(3.16)

where \( \{ x_i \}, i = 1, \ldots, m \) are frames of a given image sequence.

Practically, minimizing the dimension of an operator space amounts to finding the least rank matrix \( \mathbf{H} \) under the constraints in Eq. (3.16). The rank minimization problem
is generally NP-hard [22], this problem may, however, be treated as a constrained nuclear
norm minimization, which may be seen as a tightest convex relaxation of Eq. (3.16).
Therefore, we can replace the objective function in Eq. (3.16) by \( \|H\|_* \), yielding the
following nuclear norm minimization problem,

\[
\min \|H\|_* \\
\text{s.t.} \|X_i h_i - g\|_2 \leq c, H = [h_1 | \cdots | h_m],
\]

(3.17)

where \( \{X_i\}, i = 1, \cdots, m \) are diagonally structured matrices with Fourier coefficients of
each frame on the diagonal, and \( h_i \) the corresponding operator.
3.5.2 Solution by singular value thresholding

Eq. (3.17) can be formally rewritten as,

\[
\min \|H\|_* \\
\text{s.t.} \|A_i(H) - g\|_2 \leq c, \text{ for } i = 1, \cdots, m,
\]

(3.18)

where \(A(\cdot) : \mathbb{R}^n \to \mathbb{R}^n\) is a linear operator.

Since singular value thresholding operator has been successfully used for large scale nuclear norm minimization problem, building upon [22], we also develop a modified version of the singular value thresholding algorithm adapted to our problem in (3.18).

For a matrix \(X\), the singular value thresholding operator is stated as:

\[
D_\tau(X) = U\tau(\Sigma)V^*, \tau(\Sigma) = diag(\sigma_i - \tau_+),
\]

(3.19)

with the following theorem [22]

**Theorem 3.5.1** For each \(\tau \leq 0, Y \in \mathbb{R}^{m \times n}\), the singular value thresholding operator is the solution of the following problem,

\[
D_\tau(Y) = \arg \min_X \frac{1}{2}\|X - Y\|^2_F + \tau\|X\|_*.
\]

(3.20)

From Theorem 3.5.1, we can see that the singular value thresholding operator is closely connected to nuclear norm minimization problem. By selecting a large \(\tau\), we proceed with the following approximate optimization problem,

\[
\min \tau\|H\|_* + \frac{1}{2}\|H\|^2_F \\
\text{s.t.} \|A_i(H) - g\|_2 \leq c, i = 1, \cdots, m.
\]

(3.21)

To further solve (3.21), we consider the generalized Lagrangian of (3.21) as

\[
L(H, y, s) = \tau\|H\|_* + \frac{1}{2}\|H\|^2_F + \sum_i (\langle y_i, g - A_i(H) \rangle - s_i c),
\]

(3.22)
where \( y_i \) and \( s_i \) are the Lagrangian multipliers.

We have the following optimization functional,

\[
\arg\min_H L(H, y, s) = \arg\min_H \{\tau\|H\|_* + \frac{1}{2}\|H\|_F^2 - \sum_i \langle y_i, A_i(H) \rangle \}.
\] (3.23)

Consider the dual of \( A_i \), we have \( \langle y_i, A_i(H) \rangle = \langle A_i^*(y_i), H \rangle \). Then (3.23) may be rewritten as

\[
\arg\min_H L(H, y, s) = \arg\min_H \{\tau\|H\|_* + \frac{1}{2}\|H - \sum_i A_i^*(y_i)\|_F^2 \}.
\] (3.24)

According to Theorem (3.5.1), also letting \( P_K \) be the orthogonal projection onto a second order cone \( K \), we have the following iteration,

\[
\begin{aligned}
H^k &= \mathcal{D}_\tau(\sum_i A_i^*(y_i^{k-1})) \\
\begin{bmatrix} y^k \\ s^k \end{bmatrix} &= P_K \left( \begin{bmatrix} y^{k-1} \\ s^{k-1} \end{bmatrix} + \delta_k \begin{bmatrix} g - A(H^k) \\ -c \end{bmatrix} \right).
\end{aligned}
\] (3.25)

In particular, the projection \( P_K \) is given by [22]

\[
P_K : (y, s) = \begin{cases} 
(y, s), & \|y\| \leq s, \\
\frac{\|y\|+s}{2\|y\|}(y, s), & -\|y\| \leq s\|y\|, \\
(0, 0), & s \leq -\|y\|.
\end{cases}
\] (3.26)

### 3.6 Algorithm for Human Activity Classification and Experimental Results

Video-based human activity analysis is a good validating candidate for its intrinsic low dimensional structure [18] represented in a very high dimensional data space. The high dimensionality is due to the intrinsic complexity of variability among different individuals (appearance, gestures, etc) and highly nonlinear deformation among frames. We next carry out experiments of human activity classification to demonstrate the performance of
our algorithm. To compare to the state-of-art, we use the database from [48] (also used by [18] [15]), which are 188 × 144, 25fps low resolution video sequences of different human activities, such as walking, bending, running, jumping, etc, by 9 individuals. Sample frames are shown in Fig. 3.5.

3.6.1 Preprocessing: difference image

Most state-of-art work [18] [48] [104], if not all, include a background subtraction step as a preprocessing to only use shapes or silhouettes of human body in each frame. The argument is that the textural and background information are irrelevant to activities. Under this setting, the results of background subtraction will affect the performance of human activities classification due to its imperfection. The computational complexity of background subtraction may further limit the potential of these algorithms to carry out the analysis in real time. In our work, the variability of textural information is already considered under the formalism of fibre bundles, and the potential for an operator-based approach to cope with additive noise terms [90], allow us to perform a coarser preprocessing without affecting the classification performance.

Considering rather the gestures in each frame, the variation across frames is more pronounced, and hence more intrinsic to human activities. we use the centered difference image of two neighboring frames as the input to the optimal operator space pursuit without any other preprocessing. This dramatically reduces the burden of preprocessing, and as we note in the next section, gives us the potential to classify high dimensional data sequences in real time.

3.6.2 Video sequence similarity measure

Consider two video sequences $X_1$ and $X_2$, the corresponding optimal operator spaces are $H_1$ and $H_2$, respectively. Instead of measuring the distance between $X_1$ and $X_2$, we consider the distance between $H_1$ and $H_2$ in the image operator space, and then utilize the metric defined in (3.11).
Figure 3.5: Examples human activity video sequence and difference images. (The upper case is running, and the lower case is walking)
Specifically, we have the frame-to-frame distance based on (3.11),
\[
d(h_q, h_p) = \frac{\|h_p(x_q) - h_p(x_p)\|_2 + \|h_q(x_p) - h_q(x_q)\|_2}{2}
= \frac{\|h_p(x_q) - g\|_2 + \|h_q(x_p) - g\|_2}{2}.
\]
(3.27)

We can subsequently define the frame-to-sequence distance as a generalization of point-to-set distance as follows,

**Definition** (Frame-to-sequence distance) For a sequence \(X = [x_1, \ldots, x_n]\) and its corresponding operator sequence \(H = [h_{x_1}, \ldots, h_{x_n}]\), and a frame \(z\) with its operator \(h_z\), the distance between \(H\) and \(h_z\), \(d(h_z, H)\), is defined as
\[
d(h_z, H) = \min_{h_x \in H} d(h_x, h_z).
\]
(3.28)

Intuitively, we search for the element in the optimal operator space to yield the minimum deviation from the ideal output, which essentially implies the most appropriate operator in the operator space to characterize the given image. Built upon the definition of frame-to-sequence distance, we further define the sequence-to-sequence distance as the Hausdorff distance of two operator sequences as follows,

**Definition** (Sequence-to-sequence distance)

For two sequences \(H^1 = [h^1_1, \ldots, h^1_p]\) and \(H^2 = [h^2_1, \ldots, h^2_q]\), the distance between \(H^1\) and \(H^2\), \(D(H^1, H^2)\) is defined as
\[
D(H^1, H^2) = \max(\text{mean}\{d(h^1, H^2)\}, \text{mean}\{d(h^2, H^1)\}), \forall h^1 \in H^1 \text{ and } h^2 \in H^2.
\]
(3.29)

### 3.6.3 Experimental results

Instead of using the entire video sequences of each sample, we further segment every sequence into 10-frame segments. This setting, similar to what has been used in [15] [48], has two advantages over using whole human activity sequences as in [104] [1] [18]. First,
Figure 3.6: Optimal operator space intends to obtain a more compact representation for each sample video sequence, upon which a classifier is built.

for real applications, it is unrealistic to wait until one target finishes its activity to do the analysis. Second, using a small interval of input data, will dramatically reduce the computation complexity [22], since the computation cost here is \( O(N \times L^2) \) [66], where \( N \) is the resolution of each frame and \( L \) is the length of the video sequences.

In the Weizmann human activity dataset [48], human activities of 9 people are collected. To demonstrate the performance of our framework, we randomly pick 1, 3, 5, 7 people’s samples respectively as a training set, and the others as a testing set. And for each input data (10-frame segments from testing set), we assign it to the same class of its nearest neighbor under the measure of Def. (3.6.2). The results are shown in Fig. 3.7. Notice that we also intentionally use a smaller data set to illustrate the generalizability of our algorithm. For 7 people using as a training set, which is about 77.78% of activity sequences, we can get a classification rate of 97.92%, which is higher than most of the previous results [15] [104] [18] using the same data base, and comparable to [48], which uses the leave-one-out test for classification. Moreover, in our algorithm, there is no need to do alignment as [104] or to use the entire video sequence as [18]. In either case, the processing time will be increased for buffering and preprocessing the entire input sequence.

Additionally, from Fig. 3.7, we can see that the classification rate for a smaller training data set is not severely decreased. In the case of 5 people, about 55.55% data, as a training set, the classification rate is 95.57%, and 91.53% in the case of 3 people, 92.08% in the case of 1. Very few works have shown the robustness of their method in the presence of a
Figure 3.7: Classification rate vs percentage of data using as training set
small training set, this result proves the superior generalizability and efficiency compared to other state-of-the-art.

### 3.7 Visualizing Operator Evolution

#### 3.7.1 Image operator local embedding

In order to model and visualize an image operator sequence, it is important to pursue a low dimensional representation of a given operator sequence while preserving important geometric information. Essentially, this problem can be stated as follows: Given a set of points \( X = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R}^n \), find out a low dimensional representation of \( X \) as \( Y = \{y_1, \ldots, y_n\}, y_i \in \mathbb{R}^d, d \ll n \), such that certain geometric structures among the original data set are well preserved.

In the image operator space, as we discussed in Section 3.3.3, the geometric information is largely captured by the distance measure between pairs of points. Additionally, since points are projected on each other’s tangent space during the measurement, the distance between close points provides a better approximation and is hence more important. The problem of pursuing a low dimensional representation for an operator sequence can therefore be restated as finding the optimal low dimensional global coordinates, while preserving neighborhood relations for all points with minimal discrepancy.

To be more precise, the pairwise relation can be represented as a matrix \( W \), where \( w_{ij} \in [0, 1], i, j = 1, \ldots, n \). We then have an optimization problem to find a low dimensional representation of an image operator sequence as follows,

\[
\Phi(Y) = \sum_{i,j} w_{ij} (y_i - y_j)^2 = \|Y - WY\|_2^2, \tag{3.30}
\]

\[s.t. \ Y = [y_1 | \ldots | y_n], Y^TY = I. \tag{3.31}\]

Intuitively, by minimizing the objective function, we force the embedding coordinates of points similar to each other in the original space to still be close in the embedding space. While for the points far away from each other, the corresponding coordinates can freely vary. In particular, this problem has a closed form solution by finding the lowest eigen-
vectors (except the one corresponding to eigenvalue 0) of $I - W$ [86]. Specifically, if we embed an operator sequence into a $q$-dimensional space, then the embedding coordinates $Y$ is a $n \times q$ matrix as follows,

$$Y = [v_{n-1} | \ldots | v_{n-q}],$$

$v_i$ is the $i$-th eigenvector for $L = I - W$.

**Algorithm 1** Image operator local embedding

Input: Image operator sequence $H = [h_1, \ldots, h_n]$, Image sequence $X = [x_1, \ldots, x_n]$, dimension of the data manifold $d$

Output: Embedding matrix $Y$

1: Find distance matrix $D$ such that $d_{ij} = dist(h_i, h_j)$
2: Calculate the adjacency matrix $A$ such that $a_{ij} = \exp \left( -\frac{d_{ij}^2}{2\sigma^2} \right)$
3: Determine the local weights for each $h_i$ such that

$$\hat{w}_{ij} = \begin{cases} a_{ij} & \text{if } h_j \in N(h_i) \\ 0 & \text{otherwise} \end{cases},$$

$W$ is then achieved by normalizing each row of $\hat{W}$
4: Solve the minimization problem for the embedding coordinates,

$$Y = \arg \min \| Y - WY \|_F^2, \text{ s.t. } Y^T Y = I$$

To proceed, we need the matrix $W$ that records pairwise relations for all samples. As the notion of distance on the image operator space defined in Section 3.3.3, we choose the diffusion distance between pairs of points such as

$$\hat{w}_{ij} = \begin{cases} \exp \left( -\frac{dist^2(h_i, h_j)}{2\sigma^2} \right) & \text{if } h_j \in N(h_i) \\ 0 & \text{otherwise} \end{cases}$$

(3.32)

where $N(h_i)$ denotes the neighborhood of $h_i$. Then $W$ is finalized by normalizing each row of $\hat{W}$.

Algorithm 1 illustrates the procedure for operator-based local embedding in detail.
Figure 3.8: (a) A sample walking sequence (b) The walking sequence embedded in 1D and 2D (c) A sample bending sequence (d) Bending sequence embedded in 1D and 2D (x-axis is the index of frames)
Figure 3.9: (a) A sample bending sequence (b) Bending sequence embedded in 1D and 2D (x-axis is the index of frames)
Fig.(3.8) and Fig.(3.9) show examples of embedding a human activity video sequence into a low dimensional space. Interestingly, clear cyclic structures are both shown in the embedding curves of walking in 1D and 2D (Fig.(3.8)), due to the periodic properties of walking activity. On the other hand, the first and the last part of bending sequence, which are both dramatically different from the middle part, are almost the same (Fig.(3.9)), since for bending, the first few frames and the last few frames are very similar.

### 3.7.2 Image operator correction

Practically, input video sequences include noise as well as variations, which inevitably will affect the corresponding operator sequences. Consider human activity video sequences: a certain gesture may vary from time to time even in one sample video. One approach to alleviating this situation is by averaging the neighborhood of given samples. With matrix $W$ recording the local relations of each point, we can use the following equation to update the operator coefficients,

$$\hat{H} = WH. \quad (3.33)$$

More specifically, for each $h_i$, we have

$$\hat{h}_i = \sum_{h_j \in N(h_i)} w_{ij} h_j, \quad (3.34)$$

This approach is equivalent to finding the weighted mean of each sample’s neighborhood. The weights are calculated per Equation (3.32). An improved algorithm for an operator sequence local embedding is hence achieved by alternatively updating each operator’s value and their neighborhoods, as shown in Algorithm 2.

As Fig.(3.10) shows, the temporal correlations are more pronounced after the correction procedure. Additionally, the ”averaging” effect reduces the noise level, and results in similar frames being closely clustered together. This outcome not only provides a more robust low dimensional representation, but also affords us advantages to exploring the temporal dynamics using hidden Markov model, as we elaborate in Section 3.8.

In Fig.(3.11), two video clips from different classes are projected onto the embedding
Algorithm 2 Improved Image operator local embedding

Input: Image operator sequence $H = [h_1, \cdots, h_n]$, Image sequence $X = [x_1, \cdots, x_n]$, dimension of the data manifold $d$

Output: Embedding matrix $Y$

1: for $k = 1, \ldots, \text{MaxIter}$ do
2:    Find distance matrix $D$ such that $d_{ij} = \text{dist}(h_i, h_j)$
3:    Calculate the adjacency matrix $A$ such that $a_{ij} = \exp(-\frac{D_{ij}}{2\sigma^2})$
4:    Determine the local weights for each $h_i$ such that
5:    $\hat{W}_{ij} = \begin{cases} a_{ij} & \text{if } h_j \in N(h_i) \\ 0 & \text{otherwise} \end{cases}$
6:    Then $W$ is achieved by normalizing each row of $\hat{W}$
7:    Update operator sequence $H$ by $H = WH$
8: end for
9: Solve the minimization problem for the embedding coordinates $Y = \text{arg min} \|Y - WY\|_F^2$, s.t. $Y^TY = I$

Figure 3.10: (a) Embedded sequence of walking. Left: before correction. Right: after correction. (b) Embedding sequence of bending. Left: before correction. Right: after correction.
space of a video sequence of walking. It shows that two curves from the same type of activity are quite close to each other in the low dimensional embedding space. However, for different type of human activity sequences, the embedded curve can be very different, as shown in Fig.(3.11)(c). These results essentially illustrate that the temporal relations are critical to distinguishing different human activity videos.

3.8 Exploring temporal relations: hidden Markov model for image operator sequence

Most video sequences are temporarily strongly correlated. This clear fact constitutes the foundation for techniques in video compression and video retrieval. There is hence fundamental interest in characterizing the temporal dynamics of video sequences for the benefit of recognition and classification. In the previous sections, our effort concentrated on finding robust as well as compact spatial representations of images. In particular, we proposed a nonlinear dimension reduction algorithm for image operator sequences in Section 3.7, and more importantly, we can see that strong correlations are well preserved among neighboring frames in the low dimensional embedding space. From another perspective, each video sequence may be seen as a observation sequence of the low dimensional stochastic process in the embedding space. Instead of processing rather curly high dimensional curves, we can characterize the dynamic features of given video sequences in the low dimensional embedding space, while capturing the spatial features by image operators. This allows for a more comprehensive tool to evaluate video sequences.

In the case of human activity video sequence, it is particularly easy for a human to predict the following potential gesture given the current state. This implies a Markov property for human activity. We hence adopt continuous hidden Markov model (HMM) to characterize the temporal dynamics of a sequence. Specifically, we view the low dimensional embedding space as the state space, given the experimental results that similar images are well clustered together. The original video sequence is seen as the observation sequence.

We next introduce the mathematical background of continuous HMM, which is based
Figure 3.11: (a) Embedded sequence of walking. (b) Project another sequence of walking into the same space. (c) Project a sequence of bending into the same space.
on [81], and then will subsequently focus on the construction of a state space and its transition probability distribution as well as the observation probability distribution. To conclude, experimental results are shown when the temporal dynamics characterized by continuous HMM are injected in our video sequence model.

### 3.8.1 Continuous hidden Markov model

The hidden Markov model has been widely applied to sequential data modeling, and quite successfully in speech recognition [81]. In particular, continuous HMM is applied to the case when the observations, instead of being described as a finite discrete set, are characterized by probability distributions attached to each state of the model.

Mathematically, assume that a system has $N$ states as $S_1, S_2, \ldots, S_N$, and in each time instant $t = 1, 2, \ldots, T$, the state of the system can change from $q_{t-1} = S_i$ to $q_t = S_j$ as a Markov chain. $q_t$ is the actual state at time $t$. The transition probability is defined as

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i), \quad 1 \leq i, j \leq N,$$

(3.35)

where

$$a_{ij} \geq 0, \quad \sum_{j=1}^{N} a_{ij} = 1.$$  

(3.36)

Additionally, the initial state probabilities are

$$\pi_i = P(q_1 = S_i), \quad 1 \leq i, j \leq N.$$  

(3.37)

Observations under continuous HMM can be modeled as

$$b_{jO} = P(O | q_t = S_j) = \sum_{m=1}^{M} c_{jm} \mathcal{F}(O, \mu_{jm}, U_{jm}),$$

(3.38)

where $c_{jm}$ is the mixture coefficient for the $m$th mixture in state $j$ such that $\sum_{m=1}^{M} c_{jm} =$
1, \( c_{jm} \geq 0 \) so that the pdf \( b_{jO} \) is normalized. Usually \( \mathcal{F} \) is chosen as Gaussian density [81]. A hidden Markov model can therefore be represented as \( \lambda = (A, B, \pi) \) such that \( A = \{a_{ij}\}, B = \{b_{ik}\}, \pi = \{\pi_i\} \).

For a hidden Markov model, evaluation and training are seen as basic problems that share general interest. The two problems may be formulated as follows:

1. Evaluation

   Given the observation sequence \( O = O_1O_2\ldots O_T \) and a model \( \lambda = (A, B, \pi) \), find \( P(O|\lambda) \).

2. Training

   Given the observation sequence \( O = O_1O_2\ldots O_T \), find model \( \lambda = (A, B, \pi) \) to maximize \( P(O|\lambda) \).

In the context of human activity analysis, evaluation means that given the models of every class of activity and an input video sequence, calculating the probability of this sequence generated by each model. We can then assign the input sequence to the one with maximal probability. Backward-forward algorithms provide the solution with time complexity \( O(N^2T) \) for an evaluation problem [81]. For the training problem, given a set of human activity video sequences, we want to find out the optimal model so that we can use it in an evaluation problem. In fact, not only is there no known way to analytically solve this problem, but usually multiple local maximum solutions also exist in the problem. Therefore, we use the Baum-Welch method [81] to find a solution that locally maximizes \( P(O|\lambda) \).

### 3.8.2 Experimental results

As we discussed in Section 3.7, image operator sequences embedded in a low dimensional space preserve temporal correlations, and after correction, neighboring points are well clustered. We therefore use the low dimensional embedding space \( Y \) as the state space of continuous HMM. In particular, the frames of embedded sequence \( Z = \{z_1, z_2, \ldots, z_T\} \) is clustered into \( M \) states \( \{q_1, q_2, \ldots, q_m\} \) in \( Y \), which form the state space of a continuous HMM model.
Assume that the distance between a given observation and a state obeys a Gaussian distribution, the observation is modeled as

\[ b_{jO} = P(O|q_t = S_j) = N(d(h_O, S_j), \mu_j, \sigma_j). \]  

(3.39)

Similar to the definition of frame-to-sequence distance in Section 3.6.2, the distance between \( h_O \) and \( S_j \) is also defined as a point-to-set distance via the metric defined in Section 3.3.3. More specifically, state \( S_j \) represents a cluster of operators as \( H_j = \{h_{1j}, h_{2j}, \cdots, h_{N_j} \} \). Define the distance between \( h_O \) and \( S_j \) as the minimal distance from \( h_O \) to points in set \( H_j \), written as

\[ d(h_O, S_j) = \min_{i=1, \ldots, N_j} \{dist(h_O, h_{ij})\}. \]  

(3.40)

Under the above setting, we can estimate the parameters of a continuous HMM of given human activity sequences by the Baum-Welch method [81]. After obtaining each HMM model of training sequences, each test sequence is matched to the class with maximal probability in the evaluation step. The classification performance is compared to a recent approach for the same data set [15] shown in Fig.(3.12). We can see that accounting for temporal information into our model, the classification rate generally improves.

### 3.9 Conclusion

We proposed in this chapter, a novel geometry-based framework for high dimensional data sequence analysis. Instead of exploring the unknown high dimensional data space, we utilize the optimal operator spaces to compactly represent the information of the high dimensional trajectories. Based on the image operator space, we pursued the optimal subspace of operator sequences by reformulating it as a rank-optimization problem, and an associated algorithm is developed. Furthermore, the nonlinear dynamics of each sequence is characterized by embedding the operator sequence into a low dimensional space, in which low dimensional states are modeled by continuous HMM. Finally, we applied the algorithms to video-based human activity classification, and cast a series of
Figure 3.12: Classification rate compared to results without using continuous HMM. (Red: with continuous HMM. Blue: without continuous HMM)
experiments to show its high classification rate and robustness to the size of training data.

Future research will include analysis of more complicated scenarios, such as multiple object interaction, and applications involving video retrieval and index.
Chapter 4

A Generalization of Linear Subspace Model: A Union of Subspaces

4.1 Introduction

A successful data model always reach a delicate balance between the complexity of the model itself and the fidelity of the model to the given data. With limited samples, we can always fit the given data perfectly with an arbitrarily complicated model. However, it would not provide much prediction power, and would easily fall into the overfitting trap. Moreover, the difficulty to find the optimal parameter(s) of a model may also enhance with the increase of the model complexity.

The widely used linear subspace model is an elegant example of this principle. The model itself is simple, yet extremely powerful when dealing with real-world high dimensional data [96] [25] [60]. On one hand, the low dimensional subspace structure exists in most real-world dataset. On the other hand, the simplicity of the model prevents overfitting. Furthermore, we have algorithms like PCA, to systemically pursue an optimal subspace model that “best*” fits a given dataset.

However, there are finer structures hidden from the prevailed subspace model. In particular, instead of one subspace, a union of subspaces (UoS) may better reflect the

*The measure of “best” in PCA is in $l_2$ norm, which is also a very popular measure of the fitness of a model.
relations among data samples, especially for high dimensional data. For instance, in [44] [68], the UoS model fits well the face images of different subjects under different lighting conditions and motion trajectories in a video sequence. In [87] [3], image patches are modeled well using the UoS model and the image denoising problem are further addressed.

From a geometric perspective, the UoS model can better approximate the nonlinear structures of the data space. We may interpret this model as using a set of tangent planes to fit the whole nonlinear data manifold, since after subtracting the mean of each sample, the original tangent spaces as a union of affine subspaces degenerate to a union of subspaces. As the number of subspaces increasing, we can achieve an more accurate approximation of the data manifold.

The key problem is then about how to construct a union of subspaces that fit the given dataset. This problem is equivalent to finding the clusters of data points are from the same subspace. In fact, when the clustering structure of data samples is unveiled, we can then follow the traditional wisdom, like PCA, to construct the subspace for each cluster. In particular, we introduce the concept “self-representation”, upon which a sparse model is built to recover the relations among data samples. When the UoS structure fits the given dataset, the clustering structure also emerges by exploring the relations discovered
from the self-representation model.

The remainder of this chapter is organized as follows. We first provide the prerequisite concepts and describe the self-representation (SR) model in Section 4.2. The relation between UoS and the SR model are further elaborated from a geometric perspective in Section 4.3. In Section 4.4, we discuss the algorithms to efficiently pursue the optimal solution of the SR model.

4.2 Sparsity meets self-representation

Consider a set of data points $\mathbf{x} \in \mathbb{R}^d$ uniformly sampled from a union of subspaces $S = \bigcup_{i=1}^J S_i$, then assumed sufficient sample density, each sample can be represented by the others from the same subspace with probability 1.† Mathematically, we represent the data matrix by $\mathbf{X} = [\mathbf{x}_1 | \mathbf{x}_2 | \ldots | \mathbf{x}_n]$, yielding

$$\mathbf{X} = \mathbf{XW},$$

where $\mathbf{W}$ is $n \times n$ block-diagonal matrix with zeros diagonals.

Intuitively, we use dataset $\mathbf{X}$ to represent itself. The important information about the relations among data samples is then recorded in the coefficient matrix $\mathbf{W}$. However, multiple solutions of $\mathbf{W}$ may exist, and even worse, a give data point may not necessarily use the points in the same subspace to represent itself. We hence need an additional constraint that all samples are represented only by other samples from the same subspace, to find a $\mathbf{W}$ that correctly reflects the UoS structure in $\mathbf{X}$. Specifically, the space of $\mathbf{W}$ can be then defined as follows,

**Definition (k-block-diagonal matrix)** We say that a $n \times n$ matrix $\mathbf{M}$ is k-block-diagonal if and only if

1. There exists a permutation matrix $\mathbf{P}$, such that $\tilde{\mathbf{M}} = \mathbf{PMP}^{-1}$ is a block-diagonal matrix

---

† All hyperplanes of a subspace here are of measure 0. Therefore the distribution of samples will span the entire subspace with probability 1.
2. the maximum dimension of each block of $\tilde{M}$ is less or equal than $k$.

The space of all such matrices is denoted as $BM_k$.

We next define the space of $X$ based on the space $BM_k$ of $W$.

**Definition (k-self-representative matrix).** We say that a $d \times n$ matrix $X$ with no zero column is $k$-self-representative if and only if

$$X = XW, W \in BM_k, W_{ii} = 0.$$ 

The space of all such $d \times n$ matrices is denoted by $SR_k$.

Based on Definition 4.2, if the given data are distributed in a union of $k$-subspaces, the data matrix $X$ is then a $k$-self-representative matrix. The intersection of the solution of $W$ and $BM_k$ are hence non-empty. We may then formulate the problem of finding $W$ as follows,

$$\min_W \|W\|_1 \text{ s.t. } X = XW, W \in BM_k. \quad (4.1)$$

In (4.1), it would be fundamentally difficult to constrain $W$ in $BM_k$ in the procedure of optimization. On the other hand, if we can get rid of this constraint without affecting the solution of $W_1(X)$, then the problem will degenerate to a classical $l_1$ minimization problem with linear constraint. In particular, we find that a sparse $W$ usually falls into $BM_k$, and therefore correctly reflects the UoS structure of $X$.

Intuitively, $BM_k$ and the space of sparse matrices overlap. In particular, let $n_i$ be the number of samples from $S^i$, and $b_i$ the dimension of block $W_I$ of $W$, then $n_i \geq b_i$. It follows that $b_i \leq \max_i\{n_i\}$. This condition constrains $W$ to be a sparse matrix, since $\rho(W) = \|W\|_0/n^2 \leq \max_i\{b_i\}/n \leq \max_i\{n_i\}/n$. To pursue a sparse matrix $W$, we therefore have the following problem,

$$\min_W \|W\|_1 \text{ s.t. } X = XW, \text{diag}(W) = 0. \quad (4.2)$$
With the relation among data samples recorded in $W$, we can then cluster the points from the same subspace together, and hence reconstruct the UoS structure in $X$.

### 4.3 Geometric interpretation on subspace detection property

In this section, we focus on the condition of the equivalence of (4.1) and (4.2). As discussed in Section 4.2, since the sparsity of $W$ is bounded below by $\max\{b_i\}/n$, where $b_i$ is the size of each block, we can see that the set of sparse matrices and $BM_k$ overlap. A natural question then would be under what condition can we simply use $l_1$ minimization to obtain an accurate $W$, i.e. reflecting the underlying subspace structure.

In a more formal way, if $W$ is the solution of the following problem,

$$\min_{W} \|W\|_1 \text{ s.t. } XW = X, \text{diag}(W) = 0,$$  \hspace{1cm} (4.3)

and $\text{supp}(W) \subseteq \text{supp}(A) \in BM_k$, then the solution of Eqn(4.3) is the same as that with a constraint $X \in BM_k$, where

$$A_{ij} = \begin{cases} 
1 & \text{if } x_i \text{ and } x_j \text{ are in the same subspace}, \\
0 & \text{otherwise}. 
\end{cases}$$ \hspace{1cm} (4.4)

In [93], Theorem 2.5 guarantees the correctness of the subspace segmentation, which they call $l_1$ subspace detection property. Intuitively, if the “subspace incoherence” for each subspace is high, and the distribution of points in each subspace is not skewed, then $w_{ij} \neq 0$ if and only if $x_i$ and $x_j$ are in the same subspace. In this section, we provide additional insight on this problem.

Specifically, we focus on each $x_i$ in $X$, and rewrite Eqn(4.3) as follows for each $x_i$,

$$\min_{w} \|w\|_1 \text{ s.t. } X_{-i}w = x_i$$ \hspace{1cm} (4.5)

We next give the $l_1$ subspace detection property as [93], and then provide a sufficient condition for the $l_1$ subspace detection property to hold.
**Definition** ($l_1$ subspace detection property) Let dataset $X$ lie in a union of subspaces $S = S^1 \cup S^2 \cup \ldots S^l$. For each $x_i \in X$, the optimal solution of Eqn(4.5) is $w_i$. Then we say the pair $(X, S)$ satisfies the $l_1$ subspace detection property if and only if $\text{supp}(w_i) \subseteq \{j | x_i, x_j \in S^j\}$.

Before presenting our main result, we would like to discuss the potential factors on this issue. On one hand, given the dataset $X$ in a union of subspaces, it would be easier to segment $X$ correctly if the "distance" between any two subspaces are sufficiently large. In the extreme case, if two subspaces overlap, then the identity of the points in the overlap region would not be well-defined. On the other hand, the density of samples in each subspace is important, in the sense that we need a subspace to be well-represented by the samples on it, so that we do not create "false outliers" by insufficient sampling. For example, in a two-dimensional subspace with a $x - y$ cartesian coordinate system, if we somehow only have one sample $p$ along $y$ coordinate, and all the rest along $x$ coordinate, then without knowing the underlying structure, it would be legitimate to assume that $p$ is an outlier, and is not able to be represented by other samples, and the rest of the data fall on a one-dimensional subspace. We therefore would expect a sufficient condition to include both of the above conditions: subspaces keeping a "safe distance" from each other, and each having enough samples on each of them.

In particular, the distance between two subspaces can be measured by the first principal angle between them as $\Theta(S_i, S_j)$. To provide some intuition here, if $\Theta(S_i, S_j) = 0$, then $S_i$ and $S_j$ overlap; and if $\Theta(S_i, S_j) = \pi/2$, we have $S_i \perp S_j$. On the other hand, to measure the sufficiency of samples, we need to first define the data density in an appropriate way. We hence next introduce concepts related to the measure of data sufficiency.

**Definition** (Conic Hull [21]) The conic hull of a set $C$ is

$$cone(C) = \{\alpha_1 x_1 + \cdots + \alpha_k x_k | x_i \in C, \alpha_i \geq 0, i = 1, \ldots, k\}$$

It is worth noting that $cone(C)$ is also the smallest convex cone that contains $C$ [21].

We then give the $\Delta$-density condition to measure the data sufficiency as follows,
**Definition** ($\Delta$-density condition) For all $x^l_i \in X^l$, if there exists an affine independent set $\{x^l_{k_1}, \ldots, x^l_{k_q}\}_{k_i \neq i} \subset \pm X^j$ such that $x^l_i \in C^l_i = \text{cone}(x^l_{k_1}, \ldots, x^l_{k_q})$, and the minimal circumscribed sphere in $S^l$ of $\{x^l_{k_1}, \ldots, x^l_{k_q}\}$ centered at $O_i$ obeys $\Theta(O_i, x^l_{k_j}) \leq \Delta$, $j = 1, \ldots, q$, then we say that $X^l$ in $S^l$ satisfies the $\Delta$-density condition.

Our main result now stated as the following theorem,

**Theorem 4.3.1** A dataset $X$ of unit-length points that lies in a union of subspaces $S = S^1 \cup S^2 \cup \ldots S^J$ satisfies the $l_1$ subspace detection property if $\forall x \in X$, $x$ satisfies the $\Delta$-density condition, and for any pair of $S^i$ and $S^j$, $\Theta(S^i, S^j) > \Delta$, where $\Theta(S^i, S^j)$ is the first principal angle between $S^i$ and $S^j$.

The proof is presented in Appendix A.2. The interpretation of Theorem 4.3.1 is straightforward: the angle between subspaces is bounded below by $\Delta$, which is exactly our measure for the data density, the maximum “size” of the smallest conic hull containing each sample. Specifically, if we have a higher density of samples, which means we have a clearer image of each subspace, then the segmentation of the union of subspaces can be accurately carried out with a more stringent condition, i.e. the angle between subspaces can be smaller. On the other hand, if the samples are sparse and far from each other, it would be more difficult to recover the underlying structure, and therefore we need the union of subspaces to be widely separated, i.e. a larger principal angle.

### 4.4 Algorithms for the SR model

Since the SR model (4.2) is convex, theoretically we can utilize linear programming to find its optimal solution. However, the high dimensionality of real-world data, e.g. images and videos, dramatically increases the computational cost of linear programming, and make it infeasible in practice. In this section, we introduce two efficient approaches to solve sparse models as (4.2), fast iterative shrinkage-thresholding algorithm (FISTA) [7] and Augmented Lagrangian method (ALM) [38] [65].
4.4.1 Augmented Lagrangian method

The story of ALM can be traced back in 1960’s. It was intensely discussed by Magnus Hestenes in 1969 [50], Michael Powell [80], Tyrrell Rockafellar [14] and Dimitri Bertsekas [38] as a general method for constrained optimization problems.

Specifically, consider the following optimization problem with linear equality constraint,

$$\min_x f(x) \text{ s.t. } h(x) = 0,$$  \hspace{1cm} (4.6)

its Lagrangian is then as follows,

$$L(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle.$$  \hspace{1cm} (4.7)

The optimal solution of (4.6) is then a saddle point of (4.7). We can therefore update the primal variable $x$ and the dual variable $\lambda$ alternatively to reach the optimal solution. Yet several obstacles may exist. On one hand, the convergence rate of the dual problem of (4.6) may be slow due to a large conditional number. On the other hand, the objective function $f(x)$ may not be locally convex in the neighborhood of the minimum point.

To fix these potential problems, an additional quadratic term is added into (4.6),

$$\min_x f(x) + \frac{1}{2}\mu \|h(x)\|^2 \text{ s.t. } h(x) = 0.$$  \hspace{1cm} (4.8)

Note that the optimal solution of (4.8) is same as (4.6). However, the quadratic term $\frac{1}{2}\mu \|h(x)\|^2$ decreases the condition number of the dual problem, and makes the objective function more “convex” in the neighborhood of the optimal point. The Lagrangian of (4.8), which is defined as the augmented Lagrangian of (4.6), is as follows,

$$L(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle + \frac{1}{2}\mu \|h(x)\|^2.$$  \hspace{1cm} (4.9)

We can see the quadratic term $\frac{1}{2}\mu \|h(x)\|^2$ as a penalty term. In contrast to the penalty method that $\mu$ has to be infinity to have the exact solution, Lagrangian term in (4.9) ensures that an exact solution can be reached with finite $\mu$. In practice, it implies
that a much smaller $\mu$ is needed to reach the same accuracy compared to the penalty method.

From an algorithmic perspective, we can solve (4.6) within the framework of ALM as Algorithm 3.

**Algorithm 3** Augmented Lagrangian method (ALM)

- **Input:** $\mathbf{x}$
- **Initialize:** $\lambda$, $\mu$, $\rho$
- **while** not converge **do**
  
  $x_{k+1} = \arg\min_{x} f(x) + \langle \lambda_{k}, h(x) \rangle + \frac{\mu_{k}}{2} \| h(x) \|^{2}$
  
  $\lambda_{k+1} = \lambda_{k} + \mu_{k} h(x_{k})$
  
  $\mu_{k+1} = \rho \mu_{k}$

- **end while**

- **return** $x_{k+1}$

The step in Algorithm 3 with highest computational cost is of solving the primal variable $\mathbf{x}$. It is hence critical to have an efficient way to update the primal variable at each iteration when using ALM. Fortunately, the SR model as (4.2) falls into this category. Specifically, the primal problem at each iteration for (4.2) is

$$
W_{k+1} = \arg\min_{W} \| W \|_1 + \langle Y_{k}, XW - X \rangle + \frac{\mu_{k}}{2} \| XW - X \|_F^2,
$$

(4.10)

where $Y$ is the Lagrangian multiplier. Utilizing the linearized ALM [103], we further have

$$
W_{k+1} = \mathcal{T}_{\frac{1}{\mu_{k} \eta}} \left( W_{k} - \frac{X^T(XW_{k} - X + Y_{k}/\mu)}{\eta} \right),
$$

(4.11)

where $\eta \geq \|X\|^2$.

Note that only soft-thresholding and matrix multiplication are involved in (4.11). We therefore have Algorithm 4 to solve the SR model.
Algorithm 4 ALM for the SR model

Input: $X$
Initialize: $Y_0, \mu, \rho, \eta$

while not converge do

$W_{k+1} = T_{\frac{1}{\mu k \eta}} \left( W_k - \frac{X^T (XW_k - X + Y_k / \mu)}{\eta} \right)$
$\text{diag}(W) = 0$
$Y_{k+1} = Y_k + \mu_k (XW_{k+1} - X)$
$\mu_{k+1} = \rho \mu_k$

end while

return $W_{k+1}$

4.4.2 Fast iterative shrinkage-thresholding algorithm

The fast iterative shrinkage-thresholding algorithm is proposed by Amir Beck and Marc Teboulle in [7]. It falls into the category of proximal-gradient methods, which solves a subproblem at each iteration to approximate the original primal problem. In particular, the proximal-point methods aim to minimize functions in the following formulation,

$$F(x) = f(x) + \lambda g(x), \quad (4.12)$$

where $f(x)$ is a smooth, convex function, and $g(x)$ is continuous, convex, yet non-smooth.

At each iteration, $x_{k+1}$ is generated to approximate $F(x)$ as follows,

$$x_{k+1} = \arg \min_x f(x_k) + (x - x_k)^T \nabla f(x_k) + \frac{\mu_k}{2} \|x - x_k\|^2_2 + \lambda g(x), \quad (4.13)$$

for some $\mu_k > 0$.

We can further simplify (4.13) as

$$x_{k+1} = \arg \min_x \lambda g(x) + \frac{\mu_k}{2} \|x - y\|^2_2, \quad (4.14)$$

where $y_k = x_k - \frac{\nabla f(x_k)}{\mu_k}$.

By following (4.14), we can achieve the optimal solution of (4.17) in a slow convergence rate $O(1/k)$. However, if we carefully choose a sequence $\{y_k\}$ using the information from previous $x_k$, the convergence rate can be improved further.
Specifically, as shown in [7], $Y_k$ can be generated as

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad (4.15)$$

$$y_{k+1} = x_k + (\frac{t_k - 1}{t_{k+1}})(x_k - x_{k-1}). \quad (4.16)$$

By using $\{y_k\}$, we can then enhance the convergence rate to $O(1/k^2)$. Note that although only first-order derivatives is evaluated at each iteration, $\{y_k\}$ is generated using $x_k$ and $x_{k-1}$ and hence can better approximate the original objective function $F(X)$.

We next consider to solve the SR model (4.2) within the FISTA framework. First, we can rewrite (4.2) as

$$\min_W \lambda \|W\|_1 + \frac{1}{2} \|XW - X\|^2_F \quad s.t. \quad \text{diag}(W) = 0. \quad (4.17)$$

We then have $g(\cdot) = \|\cdot\|_1$ and $\nabla f(W) = X^T(XW - X)$ in (4.13). It follows that

$$W_{k+1} = \arg \min_W \lambda \|W\|_1 + \frac{\mu_k}{2} \|W - W_k + \frac{X^T(XX_k - X)}{\mu_k}\|^2_F$$

$$= T_{\lambda/\mu_k} \left(W_k - \frac{X^T(XX_k - X)}{\mu_k}\right). \quad (4.18)$$

Using the same rules as (4.15) and (4.16), we can choose $Y_k$ to better approximate the original objective function, and therefore have Algorithm 5.

### 4.4.3 The comparison between ALM and FISTA

One of the major difference between ALM and FISTA is that the problems they solve are actually different. For ALM, the problem formulation is as (4.6), under a strict equality constraint. For FISTA, the problem formulation is as (4.12), which treats the equality constraint in the SR model as a penalty term and is essentially a non-constrained optimization problem. Generally, the equality only holds when the parameter $\lambda$ in (4.17) is infinitely close to zero.

Therefore, if the data is clean and a strict equality constraint is enforced, ALM is
Algorithm 5 Fast iterative shrinkage-thresholding algorithm (FISTA) for the SR model

Input: $\mathbf{X}$
Initialize: $\mathbf{Y}_0$, $L = 2\|\mathbf{X}\|_2^2$, $t_1 = 1$

while not converge do
    $\mathbf{W}_k = T_{\lambda/L}(\mathbf{Y}_k - \mathbf{X}^T(\mathbf{X}\mathbf{Y}_k - \mathbf{X}) / L)$
    $\text{diag}(\mathbf{W}_k) = 0$
    $t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$
    $\mathbf{Y}_{k+1} = \mathbf{W}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{W}_k - \mathbf{W}_{k-1})$
end while

return $\mathbf{W}$

preferable. However, if the data is contaminated with noises, and instead of an equality constraint, we have $\|\mathbf{X} - \mathbf{XW}\|_F^2 \leq \sigma$, then FISTA can address this problem better than ALM. Indeed, consider the following two convex optimization problems,

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } \|h(\mathbf{x})\|_2^2 \leq \sigma$$  \hspace{1cm} (P1)

and

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda\|h(\mathbf{x})\|_2^2,$$  \hspace{1cm} (P2)

where the set $\{\mathbf{x}|h(\mathbf{x}) = 0\}$ is non-empty, we can always find some $\lambda > 0$ in (4.17) such that the optimal solution of (P1) is the same as the one of (P2).

In particular, let $\mathbf{x}_1^*$ be the optimal solution of (P1) and $\mathbf{x}_2^*$ the one of (P2). If $\mathbf{x}^* = \arg\min f(\mathbf{x})$ is in $\{\mathbf{x}|h(\mathbf{x}) = 0\}$, the two problems are trivial with the same optimal solution $\mathbf{x}^*$. If $\mathbf{x}^*$ is out of $\{\mathbf{x}|h(\mathbf{x}) = 0\}$, then for (P1), the optimal solution is on the boundary of the constrained region. We can then always find $\lambda > 0$ in (P2) such that $\|h(\mathbf{x}_2^*)\|_2^2 = \|h(\mathbf{x}_1^*)\|_2^2 = \sigma$. It follows that $\mathbf{x}_2^*$ is a feasible solution for (P1), and

$$f(\mathbf{x}_2^*) + \lambda\|h(\mathbf{x}_2^*)\|_2^2 \leq f(\mathbf{x}_1^*) + \lambda\|h(\mathbf{x}_1^*)\|_2^2.$$  \hspace{1cm} (4.19)

We subsequently have $f(\mathbf{x}_2^*) \leq f(\mathbf{x}_1^*)$, and consider $\mathbf{x}_1^*$ the optimal solution of (P1), it follows that $f(\mathbf{x}_1^*) = f(\mathbf{x}_2^*)$. We therefore have $\mathbf{x}_2^*$ as the optimal solution of (P2) by
choosing $\lambda$ appropriately.

Practically, when there is no explicit information about the exact noise level, it would be easier to adjust $\lambda$ to implicitly address the noisy data rather than to constrain the data in a fixed region.

Another major difference between ALM and FISTA is that the speed of convergence is different. ALM converges linearly\textsuperscript{‡}, while FISTA converges sublinearly. Specifically, ALM converges at the speed of $O(1/\mu_k)$ \cite{38} \cite{65}, and $\mu_k = \rho \mu_{k-1}$ till $\mu_k$ hits its upper bound. On the other hand, FISTA converges at the speed of $O(1/k^2)$ \cite{7}. Hence when the noise level of the data is mild, ALM outperforms FISTA for its fast convergence.

\section{Conclusion}

We have introduced in this chapter the UoS model for high dimensional data analysis. The UoS structure can then be unveiled by pursuing a sparse self-representation of the given data. We have further proved a sufficient condition for the exact recovery of the underlying UoS structure by demanding the sample density and the principal angles between different subspaces. Moreover, we have introduced two efficient optimization methods and applied them to the SR model to pursue the UoS structure of a given dataset.

\textsuperscript{‡}Technically, the speed of convergence of ALM slow down to sublinear when $\mu_k$ hits its upper bound and keeps constant. However, we typically have a large upper bound for $\mu_k$ for a given practical problem, such that the optimal solution is achieved before $\mu_k$ reaching its upper bound.
Chapter 5

Robust Subspace Recovery via Bi-sparsity Pursuit

5.1 Introduction

Separating data from errors and noise has always been a critical and important problem in signal processing, computer vision and data mining [39]. Robust principal component pursuit is particularly successful in recovering low dimensional structures of high dimensional data from arbitrary sparse outliers [25]. Successful applications of sparse models in computer vision and machine learning [41] [87] have, however, increasingly hinted at a more general model, namely that the underlying structure of high dimensional data looks more like a union of subspaces (UoS) rather than one low dimensional subspace. Therefore, a natural and useful extension question is about the feasibility of such an approach in high dimensional data modeling where the union of subspaces is further impacted by outliers and errors. This problem is intrinsically difficult, since the underlying subspace structure is also corrupted by unknown errors, which may lead to unreliable measurement of distance among data samples, and make data deviate from the original subspaces.

Recent studies on subspace clustering [68] [44] [94] show a particular interesting and a promising potential of sparse models. In [68], a low-rank representation (LRR) recovers subspace structures from sample-specific corruptions by pursuing the lowest-rank repre-
sentation of all data jointly. The contaminated samples are sparse among all sampled data. The sum of column-wise norm is applied to identify the sparse columns in data matrices as outliers. In [44], data sampled from UoS is clustered using sparse representation. Input data can be recovered from noise and sparse errors under the assumption that the underlying subspaces are still well-represented by other data points. In [94], a stronger result is achieved such that data may be recovered even when the underlying subspaces overlap. Outliers that are sparsely distributed among data samples may be identified as well.

In this chapter, we consider a more stringent condition that all data samples may be corrupted by sparse errors. Therefore the UoS structure is generally damaged and no data sample is close to its original subspace under a measure of Euclidean metric. More precisely, the main problem can be stated as follows:

**Problem 5.1.1** Given a set of data samples $X = [x_1, x_2, \ldots, x_n]$, find a partition of $X$, such that each part $X_I$ can be decomposed into a low dimensional subspace (represented as low rank matrix $L_I$) and a sparse error (represented as a sparse matrix $E_I$), such that

$$X_I = L_I + E_I, I = 1, \ldots, J$$

Each $L_I$ then represents one low dimensional subspace of the original data space, and $L = [L_1|L_2|\ldots|L_J]$ the union of subspaces. Furthermore, the partition would recover the clustering structure of original data samples hidden from the errors $E = [E_1|E_2|\ldots|E_J]$.

Concretely, the goal of this problem is twofold: First, we wish to find out the correct partition of data so that data subset reside in a low dimensional subspace. Second, we wish to recover each underlying subspace from the corrupted data. It is worth noting that the corrupted data may highly affect the partition, and hence decoupling the two tasks would be problematic. In this paper, we propose an integral method to decompose the given corrupted data matrix into two parts, representing the clean data and sparse errors, respectively. The correct partition of data, as well as the individual subspaces, are also simultaneously recovered. Moreover, we prove a condition for the data to be exactly recovered as the global minimum of the proposed optimization problem, and provide an algorithm to approximate the global optimizer, which henceforth refer to as Robust
Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe).

The remainder of this chapter is organized as follows. In Section 5.2, we provide the fundamental concepts necessary for the development of our proper modeling. Building on this model, we reformulate in Section 5.3 Problem 5.1.1 as an optimization problem, and develop the rationale along with the condition for subspace recovery. In Section 5.4, we introduce the RoSuRe algorithm for robust subspace recovery. In Section 5.5, we finally present experimental results on synthetic data and real-world applications.

5.2 A union of subspaces with corrupted data

Consider the case that sample $l_i$ is corrupted by some sparse error $e_i$. Intuitively, we want to separate the sparse errors from the data matrix $X$ and present the remainder in $SR_k$ of Definition 4.2. Therefore Problem 5.1.1 can be formulated as

$$\min \|E\|_0$$

$$s.t. X = L + E, L \in SR_k.$$
We have some fundamental difficulties in solving this problem on account of the combinatorial nature of $\| \cdot \|_0$ and the complicated geometry of $SR_k$. For the former one, there are established results of using the $l_1$ norm to approximate the sparsity of $E$ [26][101]. We further elaborate the relations between the sparse model (4.2) and a UoS structure in $X$. The real difficulty, however, is that not only $SR_k$ is a non-convex space,* and even worse, $SR_k$ is not path-connected. Intuitively, it is helpful to consider $L_1, L_2 \in SR_k$, and let $\text{col}(L_1) \cap \text{col}(L_2) = 0$, then all possible paths connecting $L_1$ and $L_2$ must pass the origin, given that $L$ is a matrix with no zero columns, and $0 \notin SR_k$. $SR_k$ can hence be divided into at least two components $S_p$ and $SR_k/S_p$.

To avoid solving Eqn(5.1) with a disconnected feasible region, we opt to integrate this constraint into the objective function, and see the problem from a different angle. We hence have the following definition:

**Definition** ($W_0$-function on a matrix space). For any $d \times n$ matrix $X$, if there exists $W \in BM_k$, such that $X = XW$, then

$$W_0(X) = \min_W \|W\|_0, \quad \text{s.t. } X = XW, W_{ii} = 0, W \in BM_k.$$  

Otherwise, $W_0(X) = \infty$.

Then instead of Eqn(5.1), we consider the following optimization problem:

$$\min_{L,E} W_0(L) + \lambda \|E\|_0 \quad \text{(5.2)}$$

$$\quad \text{s.t. } X = L + E.$$  

The relation of Eqn(5.1) and Eqn(5.2) is established by the following lemma:

**Lemma 5.2.1** For certain $\lambda$, if $(\hat{L}, \hat{E})$ is a pair of global optimizer of Eqn(5.2), then $(\hat{L}, \hat{E})$ is also a global optimizer of Eqn(5.1).

---

*Consider $M_1, M_2 \in SR_1$, let $M_1 = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$ and $M_2 = \begin{pmatrix} 0 & 0 \\ 2 & 1 \end{pmatrix}$. It is easy to see that $M = (M_1 + M_2)/2 = \begin{pmatrix} 1/2 & 1 \\ 1 & 1/2 \end{pmatrix} \notin SR_1$. 

---

62
The proof of Lemma 5.2.1 is presented in Appendix A.3.

Next we will leverage the parsimonious property of \( l_1 \) norm to approximate \( \| \cdot \|_0 \).

First, the definition of \( W_0(\cdot) \) is extended to a \( l_1 \) norm-based function:

**Definition** (\( W_1 \)-function on a matrix space). For any \( d \times n \) matrix \( X \), if there exists \( W \in BM_k \), such that \( X = XW \), then

\[
W_1(X) = \min_W \|W\|_1, \quad s.t. \ X = XW, \ W_{ii} = 0, \ W \in BM_k. 
\]

Otherwise, \( W_1(X) = \infty \)

We then have the following problem,

\[
\min W_1(L) + \lambda \|E\|_1 \\
\text{s.t.} \ X = L + E
\] (5.3)

It is worth noting that formulation Eqn(5.3) bears a similar form to the problem of robust PCA in [25]. Intuitively, both problems attempt to decompose the data matrix into two parts: one with a parsimonious support, and the other also with a sparse support, however in a different domain. For robust PCA, the parsimonious support of the low rank matrix lies in the singular values. In our case, the sparse support of \( L \) lies in the matrix \( W \) in the \( W_0 \) function, meaning that columns of \( L \) can be sparsely self-represented.

### 5.3 Recovery of a union of subspaces: conditions and methodologies

#### 5.3.1 A sufficient condition for exact recovery

In this section, we discuss the important question of when the underlying structure can be exactly recovered by solving Eqn(5.3). This problem is essentially twofold: first, it is about when the solution of \((\hat{L}, \hat{E})\) is exact; and second, when \( \hat{W} \) correctly reflects the true clustering structure. For the former, we establish a sufficient condition of exact decomposition of \( L \) and \( E \) as follows:
Theorem 5.3.1 \((L_0, E_0)\) can be exactly recovered by solving Eqn(5.3) with \(\lambda > 0\), i.e. \((\hat{L}, \hat{E}) = (L_0, E_0)\), if \(\forall A\) the same dimension of \(L\), at least one of the following conditions is true:

1. for any partition of \(L_0 = [L_1|L_2|...|L_J]\), \(|col(L_I)| < k+1\), and \(A = [A_1|A_2|...|A_J]\) accordingly, \(\exists I\), such that \(L_I + A_I\) is of full rank.

2. \(\|P_{\Omega E}A\|_1 - \|P_{\Omega E}A\|_1 \geq \frac{\|W_0\|_1}{\lambda}\), where \(W_0 = W_1(L_0)\).

The proof of Theorem 5.3.1 is presented in Appendix A.4. Specifically, the first condition means that the perturbation \(A\) on \(L\) could lead to a non-feasible point, and the second condition states that \(E\) is sparse in a way that any feasible move will yield a larger component outside the support of \(E\). Intuitively, this theorem states that the space \(SR_k\) and \(E\) should be nearly ”incoherent” to each other, in the sense that any change of \(L_0\), i.e. \(A = L' - L_0\), will make \(E' = E_0 - A\) less sparse, and equivalently, any sparse solution \(E'\) will move the corresponding \(L'\) off of space \(SR_k\).

A more interpretive result can be derived from Theorem 5.3.1 with an additional condition. Specifically, we further constrain the column space of \(L\) to be invariant, i.e. \(col(L) = S\), and intend to recover the finer UoS structures with the presence of sparse errors. To that end we have the following problem from Eqn(5.3),

\[
\begin{align*}
\min & W_1(L) + \lambda\|E\|_1, \\
\text{s.t.} & X = L + E, col(L) = S.
\end{align*}
\]

With this additional constraint on the data space, we describe the “incoherence” between the space of \(L\) and that of the sparse errors \(E\) in a straightforward way, and subsequently have the condition of exact recovery by solving Eqn(5.4) as Theorem 5.3.2 below,

Theorem 5.3.2 Given any matrix \(A\) of the same dimension of \(L\) satisfying \(col(A) \subseteq col(L)\), if

\[
\|P_{\Omega E}A\|_1 - \|P_{\Omega E}A\|_1 \geq \frac{\|W_0\|_1}{\lambda},
\]

64
then \((L_0, E_0)\) can be exactly recovered by solving Eqn(5.4) with \(\lambda > 0\), i.e. \((\hat{L}, \hat{E}) = (L_0, E_0)\).

The proof of Theorem 5.3.2 is presented in Appendix A.4. Intuitively, condition Eqn(5.5) enforces the feasible region of \(L_0\) to be distant from \(E_0\). Consider any matrix \(A\) whose columns are in \(\text{col}(L)\), Theorem 5.3.2 ensures the exact recovery as long as the nonzero entries of \(A\) are not concentrated in the support of \(E_0\). From another perspective, If \(L_0\) is sparse and close to \(E_0\), then it would be intrinsically difficult to separate one sparse matrix from another sparse one. We therefore enforce the “incoherence” between \(L_0\) and \(E_0\) as described in Theorem 5.3.2.

After solving \(L\) and \(E\), the problem of finding sparse coefficients \(W\) is then equivalent to the subspace clustering problem without outliers.

### 5.3.2 An approximate solution via sparse modeling

Under the conditions stated in Theorem 4.3.1, we can subsequently modify \(W_1(L)\) into a convex function and define it in a connected domain by dropping the constraint \(W \in BM_k\). Specifically, we have

\[
\hat{W}_1(L) = \min_W \|W\|_1, \quad s.t. \ L = LW, W_{ii} = 0.
\] (5.6)

Substituting \(W_1(L)\) by \(\hat{W}_1(L)\) in Eqn(5.3) allows us to relax the constraints of Eqn(5.3) and directly work on the following problem,

\[
\min_{W,E} \|W\|_1 + \lambda \|E\|_1, \quad s.t. \ X = L + E, L = LW, W_{ii} = 0.
\] (5.7)

Other than posing this problem as a recovery and clustering problem, we may also view it from a dictionary learning angle. Note that the constraint \(X = L + E\) may be rewritten as \(X = LW + E\), to therefore reinterpret the problem of finding \(L\) and \(E\) as a dictionary learning problem. In addition to the sparse model, atoms in dictionary \(L\) are
brought from data samples with sparse variation. It may hence be seen as a generalization of [43] in the sense that we not only pick representative samples from the given data set using \( l_1 \) norm, but also adapt the representative samples so that they can “fix” themselves and hence be robust to sparse errors.

5.4 Algorithm: Subspace Recovery via Bi-Sparsity Pursuit

Obtaining an algorithmic solution to Eqn(5.7) is complicated by the bilinear term in constraints which lead to a non-convex optimization. In this section, we leverage the successes of alternating direction method (ADM) [65] and linearized ADM (LADM) [67] in large scale sparse representation problem, and focus on designing an appropriate algorithm to approximate the global minimum of Eqn(5.7).

Our method, what we refer to as robust subspace recovery via bi-sparsity pursuit (RoSuRe), is based on linearized ADMM [67]. Concretely, we pursue the sparsity of \( E \) and \( W \) alternatively until convergence. Besides the effectiveness of ADMM on \( l_1 \) minimization problems, a more profound rationale for this approach is that the augmented Lagrange multiplier (ALM) method can address the non-convexity of Eqn(5.7) [70] [84]. Although there is no guarantee on the convergence of general non-convex problems, Theorem 4 in [84] states that under the ALM setting, the duality gap may be zero when certain conditions are satisfied. We show the zero duality gap property of Problem Eqn(5.7) in Appendix A.5. It hence follows that with a sufficiently large augmented Lagrange multiplier \( \mu \), we can approximate the global optimizer by solving the dual problem.

Specifically, substituting \( L \) by \( X - E \), and using \( L = LW \), we can reduce Eqn(5.7) to a two-variable problem, and hence write the augmented Lagrange function of Eqn(5.7) as follows,

\[
L(E, W, Y, \mu) = \lambda \|E\|_1 + \|W\|_1 + \langle LW - L, Y \rangle + \frac{\mu}{2} \|(X - E)W - (X - E)\|^2_F, \tag{5.8}
\]
Algorithm 6 Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe)

Initialize: Data matrix $X \in \mathbb{R}^{m \times n}$, $\lambda$, $\rho$, $\eta_1$, $\eta_2$

while not converged do

Update $W$ by linearized soft-thresholding

$L_{k+1} = X - E_k$,

$W_{k+1} = \mathcal{T}_{\frac{1}{\eta_1}} \left( W_k + \frac{L_{k+1}^T(L_{k+1} W_k - Y_k)/\mu_k}{\eta_1} \right)$.

$W_{k+1}^{ii} = 0$.

Update $E$ by linearized soft-thresholding

$\hat{W}_{k+1} = I - W_k$,

$E_{k+1} = \mathcal{T}_{\frac{1}{\eta_2}} \left( E_k + \frac{(L_{k+1} W_{k+1} - Y_k)/\mu_k)\hat{W}_{k+1}^T}{\eta_2} \right)$.

Update the lagrange multiplier $Y$ and the augmented lagrange multiplier $\mu$

$Y_{k+1} = Y_k + \mu_k (L_{k+1} W_{k+1} - L_{k+1})$

$\mu_{k+1} = \rho \mu_k$

end while

where $Y$ is the Lagrange multiplier. Letting $\hat{W} = I - W$, we alternatively update $W$ and $E$,

$$W_{k+1} = \arg \min_W \|W\|_1 + \langle L_{k+1} W - L_{k+1}, Y_k \rangle$$

$$+ \frac{\mu}{2} \| L_{k+1} W - L_{k+1} \|_F^2, \quad (5.9)$$

$$E_{k+1} = \arg \min_E \|E\|_1 + \langle (E - X)\hat{W}_{k+1}, Y_k \rangle$$

$$+ \frac{\mu}{2} \| (E - X)\hat{W}_{k+1} \|_F^2. \quad (5.10)$$

The solution of Eqn(5.9) and Eqn(5.10) can be well approximated in each iteration by linearizing the augmented Lagrange term [67],

$$W_{k+1} = \mathcal{T}_{\frac{1}{\eta_1}} \left( W_k + \frac{L_{k+1}^T(L_{k+1} W_k - Y_k)/\mu_k}{\eta_1} \right), \quad (5.11)$$

$$E_{k+1} = \mathcal{T}_{\frac{1}{\eta_2}} \left( E_k + \frac{(L_{k+1} \hat{W}_{k+1} - Y_k)/\mu_k)\hat{W}_{k+1}^T}{\eta_2} \right), \quad (5.12)$$

where $\eta_1 \geq \|L\|_2^2$, $\eta_2 \geq \|\hat{W}\|_2^2$, and $\mathcal{T}_\alpha(\cdot)$ is a soft-thresholding operator.
In addition, the Lagrange multipliers are updated as follows,

\[ Y_{k+1} = Y_k + \mu_k (L_{k+1} W_{k+1} - L_{k+1}) \]  \hspace{1cm} (5.13)

\[ \mu_{k+1} = \rho \mu_k \]  \hspace{1cm} (5.14)

5.5 Experiments and Validation

5.5.1 Experiments on Synthetic Data

Section 5.3 discusses the sufficient condition to recover data structure by solving Eqn(5.1). In this section, we hence empirically investigate the viability extent of RoSuRe with various conditions. The recovery results are compared with Robust PCA [25] using the method presented in [65] and sparse subspace clustering using the algorithm in [45].

The data matrix is fixed to be a 200×200 matrix, and all data points are sampled from
a union of 5 subspaces. Fig.5.2 shows one example of the exact recovery and clustering. Note that \((L_{RoSuRe}, E_{RoSuRe})\) and \((L_0, E_0)\) are almost identical, and \(W_{RoSuRe}\) shows clear clustering properties such that \(w_{ij} \approx 0\) when \(l_i, l_j\) are not in the same subspace. In Fig.5.3 we compare with the result of Robust PCA, and demonstrate the big improvement of our method.

Fig.5.4 is the overall recovery results of RoSuRe, robust PCA and SSC. White shaded area means a lower error and hence amounts to exact recovery. The dimension of each subspace is varied from 1 to 15, and the sparsity of \(S\) from 0.5% to 15%. Each submatrix \(L_I = X_I Y_I^T\) with \(n \times d\) matrices \(X_I\) and \(Y_I\), are independently sampled from an i.i.d normal distribution. The recovery error is measured as \(err(L) = \|L_0 - \hat{L}\|_F / \|L_0\|_F\). We can see a significant larger range of RoSuRe compared to robust PCA and SSC. The reason to the result of RoSuRe and robust PCA is due the difference of data models. Concretely, when the sum of the dimension of each subspace is small, the UoS model degenerates to a "low-rank + sparse" model, which suits robust PCA very well. On the
other hand, when the dimension of each subspace increases, the overall rank of $L$ tend to be accordingly larger and hence the low rank model may not hold anymore. Since RoSuRe is designed to fit UoS model, it can recover the data structure in a wider range. For SSC, this method specifically fit the condition when only a small portion of data are outliers. Under the assumption that most of the data is corrupted, it is hence very difficult to reconstruct samples by other corrupted ones.

5.5.2 Experiments on Computer Vision Problems

Since UoS model has been intensively researched and successfully applied to many computer vision and machine learning problems [68] [45] [39], we expect that our model may also fit these problems. Here, we present experimental results of our method on video background subtraction and face clustering problem, as exemplars of the promising potential.

Video background subtraction

Surveillance videos can be naturally modeled as UoS model due to their relatively static background and sparse foreground. The power of our proposed UoS model lies in coping with both a static camera and a panning one with periodic motion. Here we test our method in both scenarios using surveillance videos from MIT traffic dataset [98]. In Fig.5.5, we show the segmentation results with a static background. For the scenario of a “panning camera”, we generate a sequence by cropping the previous video. The cropped region is swept from bottom right to top left and then backward periodically, at the speed of 5 pixels per frame. The results are shown in Fig.5.6. We can see that the results in the moving camera scenario are only slightly worse than the static case.

More interestingly, the sparse coefficient matrix $W$ provides important information about the relations among data points, which potentially may be used to cluster data into individual clusters. In Fig. 5.7(a), we can see that, for each column of the coefficient matrix $W$, the nonzero entries appear periodically. In considering the periodic motion of the camera, we essentially mean that every frame is mainly represented by the frames when the camera is in a similar position, i.e. a similar background, with the foreground
Figure 5.4: Overall recovery results of RoSuRe and Robust PCA. [0 0.2] is mapped to [1 0] of grayscale image.
moving objects as outliers. We hence permute the rows and columns of $\mathbf{W}$ according to the position of cameras, as shown in Fig. 5.7(b). A block-diagonal structure then emerges, where images with similar backgrounds are clustered as one subspace.

**Face clustering under various illumination conditions**

Recent research on sparse models implies that a parsimonious representation may be a key factor for classification [39] [62]. Indeed, the sparse coefficients pursued by our method shows clustering features in experiments of both synthetic and real-world data. To further explore the ability of our method, we evaluate the clustering performance on the Extended Yale face database B [63], and compare our results to those of state-of-the-art methods [102] [68] [45].

The database includes cropped face images of 38 different people under various illumination conditions. Images of each person may be seen as data points from one subspace, albeit heavily corrupted by entries due to different illumination conditions, as shown in Fig. 5.8. In our experiment, we adopt the same setting as [45], such that each image is downsampled to $48 \times 42$ and is vectorized to a 2016-dimensional vector. In addition, we use the sparse coefficient matrix $\mathbf{W}$ from RoSuRe to formulate an affinity matrix as
Figure 5.6: Background subtraction on traffic videos (panning camera)

\[ \mathbf{A} = \tilde{\mathbf{W}} + \tilde{\mathbf{W}}^T, \]  
where \( \tilde{\mathbf{W}} \) is a thresholded version of \( \mathbf{W} \). The spectral clustering method in [77] is utilized to determine the clusters of data, with affinity matrix \( \mathbf{A} \) as the input.

Table 5.1: Clustering error (%) on the Extended Yale Face Database B compared to state-of-the-art methods [45] [68] [102]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LSA</th>
<th>LRR</th>
<th>SSC</th>
<th>RoSuRe</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-subjects Mean</td>
<td>38.20</td>
<td>2.54</td>
<td>1.86</td>
<td>0.71</td>
</tr>
<tr>
<td>Median</td>
<td>47.66</td>
<td>0.78</td>
<td><strong>0.00</strong></td>
<td>0.39</td>
</tr>
<tr>
<td>5-subjects Mean</td>
<td>58.02</td>
<td>6.90</td>
<td>4.31</td>
<td>3.24</td>
</tr>
<tr>
<td>Median</td>
<td>56.87</td>
<td>5.63</td>
<td>2.50</td>
<td><strong>1.72</strong></td>
</tr>
<tr>
<td>10-subjects mean</td>
<td>60.42</td>
<td>22.92</td>
<td>10.94</td>
<td><strong>5.62</strong></td>
</tr>
<tr>
<td>Median</td>
<td>57.50</td>
<td>23.59</td>
<td>5.63</td>
<td><strong>5.47</strong></td>
</tr>
</tbody>
</table>

We compare the clustering performance of RoSuRe with the state-of-the-art methods such as local subspace analysis (LSA) [102], sparse subspace clustering (SSC) [45], and low rank representation (LRR) [68]. The best performance of each method is referenced in
Figure 5.7: Coefficient matrix $W$ (a) without rearrangement according to the position of the camera (b) with rearrangement according to the position of the camera

Figure 5.8: Sample face images in Extended Yale face database B

Table 5.1 for comparison. As shown in the table, RoSuRe has the lowest mean clustering error rate in all three settings, i.e. 2 subjects, 5 subjects and 10 subjects. In particular, in the most challenging case of 10 subjects, the mean clustering error rate is as low as 5.62% with the median 5.47%. In order to get a more intuitive understanding, we also present the affinity matrices in the scenario of 5 subjects and 10 subjects in Fig. 5.10. Similarly to the coefficient matrix $W$ in Section 5.5.2, we can see a clear block-diagonal structure from either matrix.

Additionally, we show the robustness of our method with respect to $\lambda$ in a 10-subject scenario. In Fig. 5.9, the correlation between the value of $\lambda$ and the cluster accuracy maintains above 98% with $\lambda$ varying from 500 to 15000.
In Fig. 5.11, we present the recovery results of some sample faces from the 10-subject clustering scenario. In most cases, the sparse term $E$ compensates the information missing caused by lightning condition. This is especially true when the shadow area is small, i.e. a sparser support of error term $E$, we can see a visually perfect recovery of the missing area. This result validates the effectiveness of our method to solve the problem of subspace clustering with sparsely corrupted data.

## 5.6 Conclusion

We have proposed in this chapter a novel approach to recover underlying subspaces of data samples from measured data corrupted by general sparse errors. We formulated the problem as a non-convex optimization problem, and a sufficient condition of exact recovery is proved. We also designed an effective algorithm named RoSuRe to well approximate the global solution of the optimization problem. Furthermore, experiments on both synthetic data and real-world vision data are presented to show a broad range of applications of our method.

Future work may include several aspects across computer vision and machine learning. It would first be interesting to understand and extend this work from a dictionary
learning angle, to learn a feature set for high dimensional data representation and recognition. Additionally, the sufficient condition proved in this paper is fairly strong, and a weaker condition is not only theoretically interesting, but would also be helpful for better understanding the problem.
Figure 5.11: Recovery results of human face images. The three rows from top to bottom are original images, the components $E$, and the recovered images, respectively.
Chapter 6

Toward Latent Variable Discovery: Analysis Dictionary Learning

6.1 Introduction

High dimensional data analysis has been a high focus of research in diverse areas, including machine learning, computer vision, and applied mathematics, on account of its theoretical complexity, and high relevance to big data problems. Dictionary learning has been one of the fundamental methodologies to handle high dimensional data, and has been successfully applied to feature extraction [41], denoising [41] [39] [87], recognition and classification [71], etc.

Specifically, a set of atoms learned from a given dataset are considered as a dictionary, and are expected to have the potential to analyze unknown incoming data. In order to construct an effective dictionary, signal models play a key role. One popular assumption is that high dimensional data is concentrated in a low-dimensional manifold embedded in a high dimensional ambient space. Dimension reduction was hence a natural way to characterize the data, and subsequently extended to a family of algorithms, so-called nonlinear dimension reduction [85][35]. On the other hand, recent research has shown that sparse models are crucial to learning a discriminating and robust dictionary [39]. In particular, data samples are assumed to lie on a union of subspaces (UoS), and a parsimonious constraint on the use of atoms to represent each data, helps identify the
underlying basis of each subspace [44] [17]. Both synthesis models and analysis models were proposed to reveal the sparsity property of data. In synthesis models, we have a synthesis dictionary $D = [d_1, \ldots, d_n]$ such that $x_i = \sum_{j \in S} d_j w_{ij}, \|S\|_0 \leq k$, where $x_i$ is our data, $d_j$ is the $j^{th}$ atom in the dictionary, and $w_{ij}$ the corresponding coefficient. On the other hand, an operator $H$ is adopted in an analysis model such that $H \circ x_i$ gives a sparse vector representing $x_i$ [16][87].

Another interesting problem invoking sparsity is the sparse null space problem (SNS), first proposed in 1986 by Coleman and Pothen [28], is yet to be related to the state-of-the-art methods in dictionary learning. The SNS problem may be stated as finding a sparse basis for the null space of a given matrix $A$. The derived elegant results of the sparse null space problem, turn out not only useful in helping us understand the dictionary learning problem, but as we further elaborate, also for solving practical problems.

In this chapter, we hence study the relation between the SNS problem and the dictionary learning problem, and as we shall show next, the SNS problem is equivalent to the analysis dictionary learning (ADL) problem. We then proceed to solve an ADL problem using methods for the SNS problem. Specifically, inspired by the existing results for the SNS problem and the state-of-the-art sparsity pursuit algorithms, we present a $l_1$ minimization-based greedy algorithm to solve the SNS problem. In contrast to current mainstream algorithms [39][87][71][57], the convergence of our method is assured by both the convergence of the greedy algorithm and the convex $l_1$ minimization. Moreover, we demonstrate its superior performance on both synthetic numerical data and real-world data.

### 6.2 From SNS to ADL

In this section, we reformulate the SNS problem and the ADL problem in a matrix form, and then proceed to establish the equivalence by way of the common solution they share.

Given any $m \times n$ matrix $A$ such that $\text{row}(A) \subset R^n$, the SNS problem may be defined as follows,

$$SNS(A) = \arg \min_{N} \|N\|_0, \ s.t. \ \text{col}(N) = \text{null}(A).$$  \hfill (6.1)
Let $X = [x_1, \ldots, x_n]$ be a generic data matrix, the ADL problem can then be rewritten as

$$ADL(X) = \arg \min_U \|U\|_0, \ s.t. \ DX = U, \ \text{row}(X) = \text{row}(U),$$

(6.2)

where $D$ is an analysis operator in matrix form, and $U$ is the corresponding sparse coefficient matrix. To avoid trivial solutions as $U = 0$, we further require $\text{row}(X) = \text{row}(U)$. Essentially, this is the maximum information we can preserve for $X$, since each row of $U$ is a linear combination of rows in $X$, and hence $\text{row}(U) \subseteq \text{row}(X)$. In practice, we may also consider the case that $\text{row}(U) \subset \text{row}(X)$ by further selecting a subset of $d_i$ in $D$. We are focusing on the generic formulation, i.e. $\text{row}(X) = \text{row}(U)$, in this section for the sake of theoretical analysis, and will elaborate on this issue later in the discussion of the detailed algorithm.

We note that finding a sparse representation of null space in Problem 6.1, is equivalent to sparsifying a given matrix $\hat{N}$ such that $\text{col}(\hat{N}) = \text{null}(A)$. This coincides with the goal of Problem 6.2 where the row space of data matrix $X$ is instead invoked. In particular, we have the following theorem,

**Theorem 6.2.1** Assume $\text{null}(A) = \text{row}(X)$, then a matrix $N$ is a minimizer of the SNS problem (as shown in (6.1)), if and only if $N^T$ is a minimizer of the ADL problem (as shown in (6.2)) .

**Proof** Assume $N$ is a minimizer of (6.1), then the constraints in (6.1) ensure that $\text{null}(A) = \text{col}(N)$. Since $\text{col}(X^T) = \text{row}(X) = \text{null}(A)$, we then have

$$\text{col}(X^T) = \text{col}(N).$$

(6.3)

We next consider any optimizer $U$ of problem (6.2). Since $\text{row}(X) = \text{row}(U)$, combined with the condition $\text{row}(X) = \text{null}(A)$, we have $\text{row}(U) = \text{row}(X) = \text{row}(N^T)$. $N^T$ is therefore a feasible solution of (6.2), and so is $U^T$ of (6.1). It follows that $\|U\|_0 = \|N^T\|_0$, and hence $N^T$ is also a minimizer of (6.2) and so is $U^T$ for (6.1).

This essentially tells us that we can solve a sparse dictionary learning problem, should we have access to an effective method to solve the corresponding SNS problem. Specifically,
given a data matrix $X = [x_1, \ldots, x_n]$, the analysis dictionary for $X$ may be constructed in the following three steps:

1. Build a matrix $A$ such that $\text{row}(A) = \text{null}(X)$, i.e. $XA^T = 0$ and $\text{rank}(A) + \text{rank}(X) = n$.

2. Find the sparse feature vectors $U^T$ by solving $N = SNS(A)$.

3. Construct the analysis operator $D$ from $DX = U$.

### 6.3 An iterative sparse null space pursuit

We have discussed the relation of SNS and ADL in Section 6.2, and have shown that they may be cast in one unified framework of sparse null space pursuit. Nevertheless, solving SNS is itself a difficult problem. Coleman and Pothen [28] have proved that SNS is essentially NP-hard, hence ruling out a polynomial time algorithm. We however show, it is still possible to approximate the sparse null space basis in polynomial time. In this section, we propose an iterative method based on $l_1$ minimization for sparse null space pursuit.

#### 6.3.1 A greedy algorithm for the SNS problem

Previous works on the SNS problem have shed light on finding a solution in polynomial time. In [28], the authors proposed a greedy algorithm for the SNS problem. For the convenience of further discussion, we present the greedy algorithm as Algorithm 7. Additionally, it has been proved in [28] that Algorithm 7 can be used to construct a sparse null space basis, as stated in Theorem 6.3.1 [28].

**Theorem 6.3.1** A matrix $N$ is a sparsest null basis of $A$ if and only if it can be constructed by the greedy algorithm.

It is worth noting that the maximum number of iterations $q$ in Algorithm 7 is constrained by the rank of $A$, i.e. $q = d - \text{rank}(A)$. Moreover, this greedy algorithm can find the global optimal solution for the SNS problem. This elegant result amounts to
Algorithm 7  A greedy algorithm for sparse null space problem

Initialize: matrix $A \in \mathbb{R}^{m \times d}$, $N = \emptyset$

for $i = 1, \ldots, q$ do
    Find a sparsest null vector $n_i$ such that $\text{rank}(N \oplus n_i) = i$.
    $N = N \oplus n_i$
end for

finding the sparsest null space basis of $A$ in exactly $q$ steps. The subproblem of finding a sparsest null vector itself, is however also a NP-hard problem [28]. We therefore next focus on finding a method to solve this subproblem in each iteration of Algorithm 7.

6.3.2 $l_1$-based search for sparse null space

We first reformulate the subproblem of finding a sparsest null vector in Algorithm 7 as follows,

$$
\min_{n_i} \|n_i\|_0,
\text{s.t. } \begin{cases} 
An_i = 0, \\
P_{N_i} n_i \neq 0
\end{cases}
$$

where $N_i$ is the subspace spanned by the constructed null space vectors in the previous $(i - 1)^{th}$ iteration. The condition $P_{N_i} n_i \neq 0$ implies that $n_i$ is not in the current span of $N$, and hence $\text{rank}(N_i \oplus n_i) = \text{rank}(N_i) + 1$.

There are two inherent difficulties in this formulation. First, $\| \cdot \|_0$ is of combinatorial nature, hence the reason of the NP-hardness of the problem. Second, the constraint in (6.4)

$$P_{N_i} n_i \neq 0,$$

defines a region that is neither compact nor convex. To address the first problem, we propose to take advantage of established results on sparsity pursuit via $l_1$ minimization [26][27]. While for the second one, and in order to have a convex and compact
feasible region, we further adjust the condition $P_{N_i^\perp} n_i \neq 0$ as follows,

$$\exists j \in \{1, \ldots, d\}, (P_{N_i^\perp} n_i)_j = c,$$  \hspace{1cm} (6.6)

where $c$ is a positive constant.

Additionally, we establish the following lemma to justify the change of the constraint from (6.5) to (6.6):

**Lemma 6.3.1** The solution of Problem 6.4 remains invariant if the constraint (6.5) is substituted by constraint (6.6).

The proof of Lemma 6.3.1 is presented in the Appendix A.6.

The meaning of Lemma 6.3.1 is that we may then separate the region defined by (6.6) into compact and convex regions based on $j$, i.e. the location of the forced nonzero element. Since the optimal solution must reside in one of these regions, we may search for the sparsest null vector in each region from $j = 1$ to $d$. We subsequently have a convex formulation of $l_1$ minimization for each $j$. Algorithm 7 may then be realized specifically as Algorithm 8.

**Algorithm 8** Sparse Null Space Basis Pursuit

```
Initialize: matrix $A$, $N = \emptyset$
for $i = 1, \ldots, p$ do
  for $j = 1, \ldots, d$ do
    Find $n_j^i = \arg \min \|n\|_1$, s.t. $An = 0$, $(P_{N^\perp} n)_j = c$
  end for
  $n_i = \arg \min \|n_j^i\|_0$
  $N = N \oplus n_i$
end for
```

This is tantamount to solving the following optimization problem for each $j$ in Algo-
It is worth noting that the exact recovery of each \( n \) via (6.7) is determined by the incoherence of the linear operator defined by the two constraints and the sparsity of each \( n \). To solve (6.7), we adopt the framework of augmented Lagrange method (ALM) on account of its superior performance on matrix-norm minimization problems [65] [67]. Specifically, we have the augmented Lagrange function of (6.7) as

\[
L(n, Y_1, Y_2, \mu) = \|n\|_1 + \langle Y_1, An \rangle + \langle Y_2, (P_{N^\perp}n)_j - c \rangle + \frac{\mu}{2} \|An\|^2 + \frac{\mu}{2} \|(P_{N^\perp}n)_j - c\|^2.
\] (6.8)

The primal variable \( n \) is first updated in each iteration with fixed dual variables \( Y_1, Y_2 \) and \( \mu \). By introducing an auxiliary variable \( \eta \), we have

\[
n_{k+1} = T_{\frac{1}{\mu_k \eta}} \left( n_k - \frac{n_1^k + n_2^k}{\eta} \right),
\] (6.9)

where \( T \) is the soft-thresholding operator, and \( \|\eta\|^2 \geq \|A\|^2 + \|P_{N^\perp}\|^2 \), and

\[
n_1^k = A^T \left( An_k + \frac{Y_1^k}{\mu_k} \right),
\] (6.10)

\[
n_2^k = P_{N^\perp} \left( (P_{N^\perp}n)_j - c + \frac{Y_2^k}{\mu_k} \right)_j.
\] (6.11)

\*The value of \( \eta \) is selected in this way to insure the convergence of the algorithm. The details are discussed in [67]
Next, the dual variables $Y_1, Y_2$ and $\mu$ are updated as

\[
Y_{1}^{k+1} = Y_{1}^{k} + \mu_{k} (A{n}_{k+1}), \tag{6.12}
\]
\[
Y_{1}^{k+1} = Y_{1}^{k} + \mu_{k} ((P_{N|n})_{j} - c), \tag{6.13}
\]
\[
\mu_{k+1} = \min\{\rho \mu_{k}, \mu_{max}\}. \tag{6.14}
\]

The strategy of linearized ALM method provides a fast convergence rate [67]. This effectively provides us a method (Algorithm 8), named Sparse Null Space Basis Pursuit (SNS-BP) in this paper, to efficiently solve the SNS problem.

### 6.3.3 Solving the ADL problem via sparse null space basis pursuit

In Section 6.2, we have discussed the equivalence of the ADL problem and the SNS problem. Here we further describe the details of solving the ADL problem (as in (6.2)) via SNS-BP.

For a typical ADL problem as in (6.2), the first step, as discussed in Section 6.2, is to construct a matrix $A$ whose transpose is the null space of $X$. Concretely, we have the following problem,

**Problem 6.3.1** Find $A$ such that $X A^T = 0$.

A simple way would be to consider a singular value decomposition of $X$, and keep the right singular vectors with zero singular values coinciding with the rows of $A$. For a common scenario where the data matrix $X$ is contaminated by Gaussian noise, we can set $A$ to the right singular vectors with small singular values, instead of exactly zero. This in fact offers an additional advantage of filtering out dense Gaussian noise from the data matrix $X$.

Upon constructing $A$, we have the sparse coefficient matrix $U^T = SNS(A)$. Here we obtain $U$ without calculating the analysis operator $D$. In the case that $D$ is needed to further process incoming data, since $DX = U$, we may have $D = UX^\dagger$, where $X^\dagger$ is the pseudo-inverse of $X$. In particular, if all entries of the dataset are all independent, i.e. $X$ is full row rank, then $X^\dagger = X^T (XX^T)^{-1}$.
Furthermore, in Section 6.2, we formulate the ADL problem with the constraint \( \text{row}(U) = \text{row}(X) \). When a more compact representation of \( X \) is however preferred, we may allow \( \text{row}(U) \subset \text{row}(X) \), by which the dimension of the original data space is further reduced. In particular, when \( X \) has been already separated as desired, we may stop SNS-BP before the pursuit null space reaches the maximum dimension.

6.4 Experiments and validations

In this section, we conduct a series of experiments, including both tests using numerical synthetic data and real-world data, to evaluate our algorithm. In the first part, we use synthetic numerical data that fit the formulation of SNS and ADL to validate the efficacy of our algorithm. In the second part, we show that our algorithm SNS-BP has the potential of solving real-world problems such as textural image classification.

6.4.1 Numerical experiments on SNS and ADL

For a better assessment of the capability of our algorithm, we synthesize data that is compatible with the model of SNS and ADL, and show that SNS-BP is able to reconstruct the sparse null space basis of the SNS problem/the sparse coefficients of the ADL problem.

First, we synthesize a \( d \times q \) sparse matrix \( N \) as the sparse null space basis of some matrix \( A \), where \( A \) can be constructed by considering the QR decomposition of \( N \) and keep the left singular vectors of nonzero singular values as the rows of \( A \), i.e. \( \text{row}(A) = \text{null}(N) \). All elements in \( N \) follow a binomial distribution as zero/nonzero entries. Moreover, the amplitude of each nonzero element is generated from a gaussian distribution.

The matrix \( A \) can then be seen as the input to SNS-BP, and we may therefore compare the recovered nullspace basis \( \hat{N} \) with the ground truth \( N \). In Fig. 6.1, we show one example of exact recovery of a sparse nullspace basis up to permutation and scale.

In Fig. 6.2, we present the sparsity level of \( \hat{N} \) with the sparsity of \( N \) varying from 0.01 to 0.2, i.e. 1% nonzero to 20% nonzero. If our method works well, we would expect it to find the sparsest basis, and therefore \( \rho(\hat{N}) \approx \rho(N) \), i.e. the relative sparsity
$\rho(\hat{N})/\rho(N) \approx 1$. In Fig. 6.2, 10 experiments have been carried out and the average sparsity is calculated. We can see that the sparse bases discovered by SNS-BP have similar sparsity with $N$, with $\rho(N)$ from 0.01 to 0.2. Additionally, define the relative error of $\hat{N}$ as

$$
err(\hat{N}) = \frac{\|\hat{N}P\Gamma - N\|_F}{\|N\|_F},
$$

(6.15)

where $P$ is an arbitrary permutation matrix, and $\Gamma$ is a diagonal matrix representing the scales of each sparse basis. The average relative error of all the experiments with the sparsity of $N$ from 0.01 to 0.2 is 1.69%.

We next test the analysis dictionary learning via SNS-BP by exploring data samples with hidden underlying sparse structures. In particular, data samples are randomly selected from a union of low-dimensional subspaces $S = S_1 \cup S_2 \cdots$, in which each subspace is also randomly constructed by using uniformly distributed vectors as its basis. Under this setting, each sample can be represented as a linear combination of other samples in the same subspace. The dataset, written as a matrix $X = [x_1, x_2, \cdots, x_n]$ as shown in Fig. 6.3(a), has a sparse intrinsic structure $W$ such that $X = XW$, where $W$ is a block-diagonal matrix as Fig. 6.3(b), and each block represents one subspace. In our experiment, we have data points distributed in five 3-dimensional subspaces within the ambient space $R^{100}$. It hence implies that the nullspace of $A$ constructed from $X$ is of dimension 15. An analysis dictionary $D$ is then trained using SNS-BP, and the associated sparse coefficient matrix $U$ is obtained as shown in Fig. 6.4(a). Specifically, we can see
that all nonzero entries of each sparse vector are clustered together, corresponding to
data samples from the same subspace. In other words, for an atom $d_i$ in the analysis
dictionary $D$, all the data points that have a significant response to it, are from one
subspace, and the rest of the data have zero response.

We next cluster the rows of $U$ and permute them accordingly, as is presented in
Fig. 6.4(b). It is interesting to see that the block-diagonal structure shows up again,
however in a more compact way. Note that in this example we find 15 atoms total, and
have three atoms each to support the data samples (have largely nonzero inner product)
in one subspace. This number corresponds to the intrinsic dimension of each subspace.
Having $D$ trained, it is then trivial to figure out which subspace a data point lies in:
we can simply separate $D$ into $D = [D_1, D_2, \ldots]$, where $D_i$ is the atom supporting $i^{th}$
subspace, and then find the maximal $\|D_i x\|$. It is therefore more efficient to recover
the underlying structure of a given dataset and represent this structure in a more compact
way.
6.4.2 Applications on real-world data

In this part, we further explore the potential of our method on images. The performance of our algorithm is evaluated on texture images from Brodatz database [83]. Each texture image is partitioned into a set of patches, and then the analysis operator learned from patches of different textures is applied to incoming data, which is also segmented into sets of patches. The properties of various textures may lead to different patterns of the corresponding sparse coefficients. For example, the texture with more randomness may lead to less sparse coefficients than more structured textures on account of a more spread distribution of its patches. We therefore apply the learned operator to incoming data, and compare the distribution of the associated coefficients with those from training sets. It is worth noting that we avoid the complexity of processing the order of patches by considering the distribution of the coefficients instead of matching the output vectors.

Specifically, we segment each texture image into $10 \times 10$ patches, and randomly pick a subset of 120 patches as the training set from each texture image, and the rest of the patches are used as a testing set. The texture images from Brodatz database [83], and the corresponding sample patches are shown in Fig. 6.5. In our experiment, we first train the analysis operator by using half of the data in the training set without knowing the label of each patch, and then calculate the distribution of the coefficients $P_i$ of the rest of the patches from the $i^{th}$ class of texture in the training set. In the next testing
(a) The sparse coefficient matrix $\mathbf{U}$ by ADL using SNS-BP

(b) Permute rows of $\mathbf{U}$ to show the structure of $\mathbf{X}$

Figure 6.4: Sparse coefficients using learned analysis dictionary

stage, texture images are used as a set of patches, and we compare the distribution of $\mathbf{U}_j = \mathbf{D}\mathbf{X}_j$ with all $\mathbf{P}_i$, and assign $\mathbf{X}_j$ to the class with the closest distribution, such as

$$\text{class}(\mathbf{X}_j) = \arg \min_i d(\mathbf{P}_i, \mathbf{P}_{\mathbf{U}_j}).$$

(6.16)

We use the total variation distance in ((6.16), as defined in [64]

$$d(p, q) = \|p - q\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |p(x) - q(x)|.$$  

(6.17)

In this experiment, the classification rate is 97.78% for the texture images shown in Fig. 6.5. The performance is higher than known state of the art methods based on predesigned features, such as [106] with a 86.63% classification rate, and comparable to the supervised dictionary learning algorithm [72]. It is worth noting that, the training set is only composed of around 1% of the dataset. A less stringent training set implies a lower computational cost. This also demonstrates the scalability of our method, in light of the competitive classification performance.
6.5 Conclusion

We have proposed in this chapter, a novel approach for sparse nullspace problem, and have proved the equivalence of the sparse nullspace problem and the analysis dictionary learning. We have presented SNS-BP, an iterative algorithm based on $l_1$ minimization, to pursue the solution of SNS problem. We have further applied this algorithm to analysis dictionary learning, and show the efficacy of our approach by experiments on both synthetic dataset and real-world data in texture classification.

Future work may include several aspects related to both SNS problem and ADL. The relation between SNS and nonlinear dimension reduction also need further investigation and may lead to results on graph embedding. Moreover, it would be interesting to explore the potential application of ADL on other high dimensional database, such as image/video classification.
Chapter 7

Conclusion and Future Research

In this thesis, we explored the low dimensional structures in high dimensional spaces, and utilized the proposed low dimensional data models to characterize real-world data. In order to overcome “the curse of dimensionality”, we need an effective high dimensional data model to find a low dimensional embedding, and to preserve the nonlinearity of data distribution simultaneously.

To mend the nonlinearity of image space, we map given images to a circulant operator subspace, with the Riemannian metric determined by the associated space. The optimality of the circulant operator subspace is further ensured by formulating the problem as a constrained rank minimization problem. We further model image sequences as a fibre bundle of the circulant operator space, and thereby propose an effective algorithm on video-based human activity recognition.

Considering the parsimonious degrees of freedom of high dimensional data compared to its dimensionality, we study the UoS model to characterize high dimensional data. The simplicity, as well as the ability to approximate the nonlinearity of data distribution, make this model very effective to reveal the underlying structures of data distribution. We further propose a bi-sparsity model with an effective algorithm RoSure, to recover the given data characterized by the UoS model from errors/corruptions. This framework shows superior performance for a wide range of problems, such as face clustering and video segmentation.

Additionally, toward the recovery of latent variables, we consider the analysis dic-
tionary learning problem within the UoS model, and show the equivalence of analysis
dictionary learning and the sparse null space problem. We thereby propose an efficient
algorithm, SNS-BP, to solve the analysis dictionary learning problem, and show the effi-
cacy by the experiments on both synthetic data and real-world texture data.

The future work may include several aspects: first, current low rank and sparse models
utilize very limited prior knowledge of a given dataset. The integration of prior knowledge
may lead to more powerful approaches for high dimensional analysis. Moreover, no
specific structure is enforced in the proposed sparse models. A possible direction would
be to make the sparse structure of the model adapt to the given data.


APPENDIX
Appendix A

Proofs

A.1 Proof of Convergence of Optimal Operator Subspace Pursuit

The convergence of the iterative method (3.25) for optimal operator subspace pursuit is confirmed by the following theorem:

**Theorem A.1.1** The sequence \( \{H_k\} \) converges to the unique solution of (3.24) when step size \( \delta_k \) satisfies \( 0 < \inf \delta_k < 2/\max L(A_i) \), where \( L(A_i) \) is the Lipschitz constant of function \( A_i(\cdot) \).

**Proof** Let \( w_i = \begin{pmatrix} y_i \\ s_i \end{pmatrix} \), \( F_i(H) = \begin{pmatrix} g - A_i(H) \\ -c \end{pmatrix} \), \( f_\tau(H) = \tau\|H\|_* + \frac{1}{2}\|H\|_F^2 \), and \( \{\hat{y}_i, \hat{s}_i, \hat{H}\} \) be the primal-dual optimal pair of (3.24). We intend to show that

\[
\lim_{k \to \infty} \|w_i^k - \hat{w}_i\|^2 = 0, \quad \lim_{k \to \infty} \|H_i^k - \hat{H}_i\|^2 = 0 \tag{A.1}
\]

First, consider the following equation,

\[
\|w_i^k - \hat{w}_i\|^2 = \|P(w_i^{k-1} + \delta_k F_i(H_i^k)) - \hat{w}_i\| \tag{A.2}
\]

\( P(\cdot) \) is the projection defined by (3.26).
Since \( \hat{w}_i \) is the optimal solution, which yields \( \hat{w}_i = P(\hat{w}_i + \delta_k F_i(\hat{H})) \), and also \( P(\cdot) \) is a contraction, we have

\[
\|w_i^k - \hat{w}_i\|^2 \leq \|w_i^{k-1} - \hat{w}_i - \delta_k F_i(\hat{H})\|^2 \\
= \|w_i^{k-1} - \hat{w}_i\|^2 + \delta_k^2 \|F_i(\hat{H}) - F_i(\hat{H})\|^2 \\
+ 2\delta_k \langle w_i^{k-1} - \hat{w}_i, F_i(\hat{H}) - F_i(\hat{H}) \rangle
\]

(A.3)

Consider the following relation and put it into (A.3),

\[
\|F_i(\hat{H}) - F_i(\hat{H})\|^2 = \| \begin{pmatrix} g - A_i(\hat{H}) \\ -c \end{pmatrix} - \begin{pmatrix} g - A_i(\hat{H}) \\ -c \end{pmatrix} \|^2 \\
= \|A_i(\hat{H}) - A_i(\hat{H})\|^2
\]

it follows that

\[
\sum_i \|w_i^k - \hat{w}_i\|^2 \leq \sum_i \|w_i^{k-1} - \hat{w}_i\|^2 + \delta_k^2 \sum_i \|A_i(\hat{H}) - A_i(\hat{H})\|^2 \\
+ 2\delta_k \sum_i \langle y_i^{k-1} - \hat{y}_i, A_i(\hat{H}) - A_i(\hat{H}) \rangle
\]

(A.4)

For the last part of (A.4), according to (3.25), for some \( J^k \in \partial f_\tau(\hat{H}), \hat{J} \in \partial f_\tau(\hat{H}) \), the following equations hold

\[
\sum_i A_i^*(y_i^{k-1}) + J^k = 0 \tag{A.5}
\]
\[
\sum_i A_i^*(\hat{y}_i) + \hat{J} = 0 \tag{A.6}
\]

Therefore, due to the strong convexity of \( f_\tau \), we have

\[
\sum_i \langle y_i^{k-1} - \hat{y}_i, A_i(\hat{H}) - A_i(\hat{H}) \rangle = \sum_i \langle A_i^*(y_i^{k-1}) - A_i^*(\hat{y}_i), H^k - \hat{H} \rangle \\
\leq -\|H^k - \hat{H}\|_F^2
\]

(A.7)
Now consider the second part of (A.4),

\[ \sum_i \| \mathbf{A}_i(\mathbf{H}^k) - \mathbf{A}_i(\hat{\mathbf{H}}) \|^2 \leq \sum_i L(\mathbf{A}_i)^2 \| \mathbf{H}^k - \hat{\mathbf{H}} \|^2 \]

\[ \leq (\max L(\mathbf{A}_i))^2 \sum_i \| \mathbf{H}^k - \hat{\mathbf{H}} \|^2 \] (A.8)

Substitute terms in (A.4) with (A.7) and (A.8), which yields

\[ \sum_i \| \mathbf{w}_i^k - \hat{\mathbf{w}}_i \|^2 \leq \sum_i \| \mathbf{w}^{k-1}_i - \hat{\mathbf{w}}_i \|^2 - 2\delta_k \sum_i \| \mathbf{H}^k - \hat{\mathbf{H}} \|^2 \]

\[ + \delta_k^2 (\max L(\mathbf{A}_i))^2 \sum_i \| \mathbf{H}^k - \hat{\mathbf{H}} \|^2 \] (A.9)

Under the assumptions that \( 0 < \inf \delta_k < 2/\| \max L(\mathbf{A}_i) \| \|^2 \), we have

\[ 2\delta_k - \delta_k^2 (\max L(\mathbf{A}_i))^2 \geq \alpha > 0, \forall k \] (A.10)

Then, it follows that

\[ \sum_i \| \mathbf{w}_i^k - \hat{\mathbf{w}}_i \|^2 + \alpha \sum_i \| \mathbf{H}^k - \hat{\mathbf{H}} \|^2 \leq \sum_i \| \mathbf{w}^{k-1}_i - \hat{\mathbf{w}}_i \|^2 \] (A.11)

which yields \( \lim_{k \to \infty} \| \mathbf{w}_i^k - \hat{\mathbf{w}}_i \|^2 = 0 \) and \( \lim_{k \to \infty} \| \mathbf{H}^k_i - \hat{\mathbf{H}}_i \|^2 = 0 \). Hence theorem (A.1.1) is proved.

### A.2 Proof of Theorem 4.3.1

Let \( \mathbf{X} \) represent the dataset with unit-length data and \( S = S^1 \cup S^2 \cup \cdots \cup S^q \) its underlying structure as a union of subspaces. Consider the partition of \( \mathbf{X} \) corresponding to \( S \) is \( \mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^q] \), then for any \( \mathbf{x}_i \in \mathbf{X}^j \), there is a linear combination of other samples in \( \mathbf{X}^j \) represent \( \mathbf{x}_i \) as \( \mathbf{x}_i = \sum_{\mathbf{x}_k \in \mathbf{X}^j, k \neq i} w_k \mathbf{x}_k \). We therefore have a feasible solution
for the following problem,

$$w^* = \arg \min_w \|w\|_1$$

$$s.t. \ X^j_{-i}w = x_i.$$  \hspace{1cm} (A.12)

Then the dual problem of Eqn(A.12) as follows also has at least one feasible point,

$$\max \langle x_i, \lambda \rangle \ s.t. \ \| (X^j_{-i})^T \lambda \|_{\infty} \leq 1.$$ \hspace{1cm} (A.13)

Let the support of $w^*$ be $Q_0$, and consider the dual vector $\lambda^*$ satisfying

$$\lambda^* = \arg \min_\lambda \| \lambda \|_2$$

$$s.t. \ (X^j_{Q_0})^T \lambda = sgn(w^*_{Q_0}), \ \| (X^j_{Q_0})^T \lambda \|_{\infty} \leq 1.$$ \hspace{1cm} (A.14)

It is worth noting that Eqn(A.12) and Eqn(A.14) imply that $x_i \in \text{cone}(X^j_{Q_0})$. Additionally, there are some properties of $\lambda^*$ which are crucial in the proof.

First, let $\lambda^* = \lambda^*_{S_j} + \lambda^*_{S_j^\perp}$. Since $\lambda^*$ is the feasible point with the least $l_2$ norm, and $(X^j_{Q_0})^T \lambda^*_{S_j} = 0, (X^j_{Q_0})^T \lambda^*_{S_j^\perp} = 0$, we have $\lambda^*_{S_j^\perp} = 0$, and therefore $\lambda^* \in S_j$.

Furthermore, the first constraint in Eqn(A.14) can be rewritten as

$$|\langle x^T, \lambda^* \rangle| = 1, \|x\|_2 = 1, \forall x \in X^j_{Q_0},$$ \hspace{1cm} (A.15)

which implies that $\lambda^*$ passes the origin of the circumscribed sphere of $\hat{X}^j_{Q_0}$, where $\hat{X}^j_{Q_0} \subset \pm X^j_{Q_0}$ and $\langle \hat{x}^j_q, \lambda^* \rangle = 1, \forall q \in Q_0$.

Now consider the $\Delta$-density condition for $x_i$, it follows that

$$\Theta(\lambda^*, x) \leq \Delta, \forall x \in \hat{X}^j_{Q_0}.$$ \hspace{1cm} (A.16)

Combined with $\|x\|_2 = 1$, we have

$$\|\lambda^*\|_2 \leq 1/ \cos(\Delta) \hspace{1cm} (A.17)$$
We then would like to utilize $\lambda^*$ and $w^*$ to further constrain the optimal solution of Eqn(4.5).

In particular, we have the following lemma from [93] using the dual certificate technique,

**Lemma A.2.1** Consider there exists $c \in \mathbb{R}^n$ which is feasible for the primal problem

$$\min_z \|z\|_1 \ s.t. \ Az = y, \quad (P)$$

and the support of $c$ is $R \subseteq Q$, then if there is dual vector $v$ satisfying

$$A_R^T v = \text{sgn}(c_R), \quad \|A_{Q \cup R^c}^T v\|_\infty \leq 1, \quad \|A_Q^T v\|_\infty < 1,$$

all optimal solutions $z^*$ to $(P)$ have $z^*_Q = 0$.

We next construct a primal feasible point for Eqn(4.5) by $w^*$. Consider the index set of $X^j$ in $X$ is $Q$, then $\bar{w}$ satisfying $\bar{w}_Q = w^*$, $\bar{w}_c^Q = 0$ is also feasible for Eqn(4.5). Additionally, since $X_{Q_0} = X^j_{Q_0}$, $X_{Q_0 \cup Q} = X^j_{Q_0}$, $\lambda^*$ have the following property from Eqn(A.14),

$$X_{Q_0}^T \lambda^* = \text{sgn}(\bar{w}_{Q_0}^c), \quad \|X_{Q_0 \cup Q}^T \lambda^*\|_\infty \leq 1 \quad (A.18)$$

Then according to Lemma A.2.1, if we further have $\|X_{Q^c}^T \lambda^*\|_\infty < 1$, then combined with the condition that $\bar{w}_{Q^c} = 0$, all optimal solutions $\hat{w}$ of Eqn(4.5) satisfy $\hat{w}_{Q^c} = 0$, which essentially implies the $l_1$ subspace detection property.

Consider that the principle angle between any pair of subspaces is larger than $\Delta$, we have

$$\|P_S x\|_2 < \|x\|_2 \cos(\Delta) = \cos(\Delta), \quad \forall x \in X_{Q^c} \quad (A.19)$$
Combined with Eqn(A.17), for all \( x \in X_Q \), it follows that

\[
|\langle x, \lambda^* \rangle| = |\langle P_{S^c}x, \lambda^* \rangle| \leq \| P_{S^c}x \|_2 \| \lambda^* \|_2
\]

\[
< \cos(\Delta) \cdot \frac{1}{\cos(\Delta)} = 1,
\]

and therefore Theorem 4.3.1 is proved.

### A.3 Proof of Lemma 5.2.1

At the beginning, we rewrite the objective function in Eqn(5.2) as

\[
f(L, E) = \frac{W_0(L)}{\lambda} + \| E \|_0.
\]

(A.21)

It is clear that this will not change the minimum value. In addition, we assume that there exists \( L \in SR_k \), otherwise the statement would be trivial, since Eqn(5.1) would be not feasible, and the value of the objective function in Eqn(5.2) would be infinite.

Let \((\hat{L}, \hat{E})\) be a global minimizer of Eqn(5.2), then \( \hat{L} \in SR_k \). If \( \exists E' \), such that \( \| E' \|_0 < \| \hat{E} \|_0 \) and \( L' = X - E' \in SR_k \), we have

\[
f(L', E') = \| E' \|_0 + 1 + \frac{W_0(L')}{\lambda} - 1
\]

\[
\leq \| \hat{E} \|_0 + \frac{W_0(L')}{\lambda} - 1.
\]

(A.22)

Since \((\hat{L}, \hat{E})\) is a global minimizer, \( f_2(\hat{L}, \hat{E}) < f_2(L', E') \). Combined with Eqn(A.22),

\[
0 < f(L', E') - f(\hat{L}, \hat{E}) \leq \frac{W_0(L') - W_0(\hat{L})}{\lambda} - 1.
\]

(A.23)

Then it follows that

\[
\lambda < W_0(L') - W_0(\hat{L}).
\]

(A.24)

Note that when \( L \in SR_k \), \( 0 < W_0(L) \leq n^2 \), where \( n \) is the number of columns of \( L \).
Therefore, letting $\lambda \geq n^2$ will violate Eqn(A.24) since

$$\lambda \geq n^2 > W_0(L') - W_0(\hat{L}).$$

(A.25)

Hence, with $\lambda \geq n^2$, $\hat{E}$ is also a solution of Eqn(5.1). Lemma 5.2.1 is proved.

### A.4 Proof of Theorem 5.3.1 and Theorem 5.3.2

We first focus on Theorem 5.3.1, and then prove Theorem 5.3.2 with the result from Theorem 5.3.1.

∀A such that $\text{dim}(A) = \text{dim}(L_0)$, we first prove by contradiction that for any partition of $L_0 = [L_1|L_2|\ldots|L_J]$, $\text{col}(L_I) \leq k + 1$, if there exists $I$, such that $L_I + A_I$ is full rank, then $(L', E') = (L_0 + A, E_0 - A)$ is not feasible.

Assume that $L' = L_0 + A$ is feasible, then there is a $W' \in BM_k$ such that $L' = L'W'$.

Partition $L'$ according to the block form of $W'$ as

$$L' = [L'_1|L'_2|\ldots|L'_J], \text{col}(L'_I) \leq k + 1,$$

then for each $I$, we have $L'_I = L'_I W'_I$ and $L'_I$ is not full-rank, which contradicts the assumption that $\exists I$ such that $L_I + A_I$ is full-rank.

Next we prove that if $\|P_{\Omega_{k}}A\|_1 - \|P_{\Omega_{k}}A\|_1 \geq \sigma$, then for $(L', E') = (L_0 + A, E_0 - A)$, $f(L, E) < f(L', E')$.

Consider

$$f(L', E') - f(L, E) = \|E_0 - A\|_1 - \|E_0\|_1$$

$$+ \frac{\|W'\|_1}{\lambda} - \frac{\|W_0\|_1}{\lambda},$$

(A.26)
by using the disjoint property of $\Omega_E$ and $\Omega^c_E$, we have

$$\|E_0 - A\|_1 - \|E\|_1 = \|E - P_{\Omega E}A - P^c_{\Omega E}A\|_1 - \|E\|_1$$

$$= \|E - P_{\Omega E}A\|_1 + \|P^c_{\Omega E}A\|_1 - \|E\|_1$$

$$\geq \|E\|_1 - \|P_{\Omega E}A\|_1 + \|P^c_{\Omega E}A\|_1 - \|E\|_1$$

$$= \|P^c_{\Omega E}A\|_1 - \|P_{\Omega E}A\|_1 \geq \sigma,$$  \hspace{1cm} (A.27)

then it follows that

$$f(L', E') - f(L, E) = \sigma + \frac{\|W'\|_1 - \|W_0\|_1}{\lambda}. \hspace{1cm} (A.28)$$

In addition, since $\lambda \geq \frac{\|W_0\|_1}{\sigma}$, we have

$$\frac{\|W'\|_1 - \|W_0\|_1}{\lambda} > -\frac{\|W_0\|_1}{\lambda} \geq -\sigma$$ \hspace{1cm} (A.29)

Plugging Eqn(A.29) into Eqn(A.28) yields

$$f(L', E') - f(L, E) > 0, \hspace{1cm} (A.30)$$

and therefore Theorem 5.3.1 is proved.

Now consider Theorem 5.3.2, in order to have $\text{col}(L_0)$ invariant, for any perturbation $A$, we have

$$\text{col}(A) \subseteq \text{col}(L_0). \hspace{1cm} (A.31)$$

Then according to the condition Eqn(5.5), $A$ also satisfies the second sufficient condition in Theorem 5.3.1. We therefore have Theorem 5.3.2 proved.

A.5 Zero Duality Gap of the Dual Problem

In Section 5.4, we elaborated our algorithm RoSuRe for Problem (5.7). Essentially, our algorithm can be seen as a dual method, which relies on solving the dual problem instead
of the primal one. However, as we mentioned in Section 5.4, a duality gap usually exists for general non-convex programming. We then use the framework of augmented Lagrange method to "convexify" the Lagrange function of (5.7). To substantiate our motives, in this section we would like to show the zero duality gap between the primal problem (5.7) and the associated "augmented" dual problem.

First of all, consider the nonlinear programming problem with equality constraints in the following general form,

$$\min f(x) \ s.t. \ h(x) = 0, \ x \in \Omega,$$  \hspace{1cm} (P)

then the primal function associated with (P) is defined as

$$p(z) = \inf \{ f(x) : h(x) \leq z, -h(x) \leq z, \ x \in \Omega \}. \hspace{1cm} (A.32)$$

In addition, the augmented Lagrange function is defined as

$$L(x, y, \mu) = f(x) + \langle y, h(x) \rangle + \frac{\mu}{2} \| h(x) \|_2^2, \ x \in \Omega, \hspace{1cm} (A.33)$$

which lead to the dual problem of (P) as follows,

$$\max g(y, \mu), \ \text{where} \ g(y, \mu) = \inf_{x \in \Omega} L(x, y) \hspace{1cm} (D)$$

Augmented Lagrange method for non-convex programming is intensively studied in [84], and a sufficient and necessary condition for a zero duality gap is further proved. In particular, two conditions, i.e. the quadratic growth condition and the stable of degree 0, are critical for a non-convex problem to be solved by a dual method. We therefore first give the definition of these two conditions, and then show that Problem (5.7) satisfies them.

**Definition** *(Quadratic Growth Condition)* We say that (P) satisfies the quadratic growth condition if for certain real number $q$,

$$L(x, 0, \mu) = f(x) + \frac{\mu}{2} \| H(x) \|_2^2 \geq q, \ \forall x \in \Omega. \hspace{1cm} (A.34)$$
**Definition** (Stable of degree $k$) If there is an open neighborhood $U$ of the origin of $R^n$, and a function $\omega: U \rightarrow R$ of class $C^k$, such that the primal function $p(z)$ of (P) satisfies the following condition:

$$p(z) \geq \omega(z), \ \forall z \in U, \text{with} \ p(0) = \omega(0),$$

then (P) is (lower) stable of degree $k$.

**Lemma A.5.1** The associate primal function of (5.7) satisfies the quadratic growth condition and is stable of degree 0.

**Proof.** We first show that the primal function $p(z)$ satisfies the quadratic growth condition. Note that the quadratic growth condition holds if $f(x)$ is bounded below on $\Omega$. In (5.7), $f(x) = \|W\|_1 + \lambda\|E\|_1 > 0$, and thus the associated $p(z)$ has a lower bound on $\Omega$.

We next show $p(z)$ is stable of degree 0. First of all, the stability of degree 0 is equivalent to the following condition [84]:

$$p(0) = \lim_{z \rightarrow 0} \inf p(z) > -\infty \quad (A.35)$$

Then constructing a compact set including $p(0)$ would suffice to (A.35). Specifically, a sufficient condition to (A.35) may be as follows: $\Omega$ is closed, $h(x)$ is continuous, and for some $z \in R_+^{d \times n}$ and $C > \inf p(z)$, the set

$$S = \{x \in \Omega \mid f(x) \leq C, -z \leq h(x) \leq z\}$$

is compact.

In problem (5.7), $\Omega = \{(W, E) \in R^{n \times n} \times R^{d \times n} \mid W_{ii} = 0\}$ is closed, and $h(x)$ is obviously continuous. To check the compactness of $S$, let $C > \lambda\|X\|_1$. It is easy to see that $(0, X)$ is a feasible point in the union of compact sets $S_1 = \{x \in \Omega \mid f(x) \leq C\}$ and $S_2 = \{x \mid -z \leq h(x) \leq z\}$. Then $S = S_1 \cap S_2$ is also a compact set. We therefore have the conclusion that $p(z)$ of (5.7) is stable of degree 0.

We finally have the sufficient condition, i.e. Lemma A.5.1 to show the zero duality gap of (P) and (D), given the theorem proved in [84]:

115
**Theorem A.5.1** The duality equation of \((P)\)

\[
\inf(P) = \sup(D)
\]

holds, if and only if \((P)\) satisfies the quadratic condition and is stable of degree 0.

**A.6 Proofs of Lemma 6.3.1**

Consider the subproblem in each iteration of Algorithm 7,

\[
\min_{n_i} \|n_i\|_0, \quad s.t. \, A_{n_i} = 0, P_{N_i} n_i \neq 0, \tag{A.36}
\]

Lemma 6.3.1 implies that we can substitute the non-convex and non-compact constraint

\[P_{N_i} n_i \neq 0, \tag{A.37}\]

to the following one,

\[\exists j \in \{1, \ldots, d\}, (P_{N_i} n_i)_j = c, \tag{A.38}\]

where \(c\) is a positive constant.

Lemma 6.3.1 essentially states that the following optimization problems, A and B, are equivalent.

\[
\min_{n} \|n\|_0, \quad s.t. \, A n = 0, P_{N} n \neq 0, \tag{A}
\]

\[
\min_{n} \|n\|_0, \quad s.t. \, A n = 0, \exists j \in \{1, \ldots, d\}, (P_{N} n)_j = c, \tag{B}
\]
We therefore prove the equivalence of the above two problems.

**Proof** First, we show that if $n$ is an optimal solution of (A), then for some real number $\alpha$, $\alpha n$ is also a minimizer of (B).

For any minimizer $n'$ of (B), it also satisfies the constraints of (A). It hence follows that

$$\|n'\|_0 \geq \|n\|_0.$$  \hfill (A.39)

Assume $|(P_{N^\perp} n)_k| = \|P_{N^\perp} n\|_\infty$, and note that $\|P_{N^\perp} n\|_\infty \neq 0$, we construct

$$\hat{n} = \frac{c}{(P_{N^\perp} n)_k} \cdot n = \alpha n.$$  \hfill (A.40)

Since $(P_{N^\perp} \hat{n})_k = c$, $\hat{n}$ is also feasible in (B). Consider the fact that $\|\hat{n}\|_0 = \|n\|_0 \leq \|n'\|_0$, we conclude that $\hat{n} = \alpha n$ is also a solution of (B), and therefore $\|\hat{n}\|_0 = \|n'\|_0$.

Then it is trivial to show that $n'$ is also a minimizer of (A), given the fact that $n'$ is a feasible solution of (A) and $\|n'\|_0 = \|\hat{n}\|_0 = \|n\|_0$. Lemma 6.3.1 is hence proved.
Appendix B

Linearized Soft-thresholding for Matrix-norm Minimization with Equality Constraints

Sparse representation and low-rank representation problems have drawn keen interest in machine learning and computer vision. The general underlying problem may be formulated as follows,

\[
\begin{align*}
\arg \min_{X} & \quad \|X\| \\
\text{s.t.} & \quad f_1(X) = A_1, f_2(X) = A_2, \ldots, f_n(X) = A_n,
\end{align*}
\]  

(B.1)

where \(\|\cdot\|\) is usually the nuclear norm (for low-rank representation) or 1-norm (for a sparse representation), and \(f_i(X)\) is a linear operator for matrix \(X\).

To solve this problem, the augmented Lagrange multiplier method is applied here. The augmented Lagrangian function of Eqn(B.1) is

\[
L(X, Y_1, \ldots, Y_n, \mu) = \|X\| + \sum_i \langle Y_i, f_i(X) - A_i \rangle + \sum_i \frac{\mu}{2} \|f_i(X) - A_i\|_F^2. 
\]  

(B.2)
Follow the procedure of the augmented Lagrange multiplier method, $X$ is updated as

$$X_{k+1} = \arg \min_X L(X, Y_{ik}, \mu_k).$$  \hfill (B.3)

For Eqn(B.3), it follows that

$$X_{k+1} = \arg \min_X L(X, Y_{ik}, \mu_k),$$  \hfill (B.4)
$$= \arg \min_X \|X\| + \sum_i \frac{\mu_k}{2} \|f_i(X) - f_i(X_k) + \frac{Y_{ik}}{\mu_k}\|^2_F.$$

If all linear operators on $X$ in $f_i$ are identity operators, then for the objective function as $\| \cdot \|_*$ or $\| \cdot \|_1$, there is a closed form solution for Eqn(B.4)\cite{65}\cite{25}. However, without such a condition, a closed form solution does not exist. Nevertheless, linearized alternating direction method is introduced in \cite{67}\cite{103} for solving general linear constraints in a low rank representation problem, and the convergence is proved. Here we elaborate a similar linearized thresholding method for Eqn(B.4) in details.

By introducing $X_k$ as an auxiliary variable, Eqn(B.4) can be rewritten as

$$X_{k+1} = \arg \min_X \|X\| + \sum_i \frac{\mu_k}{2} \|f_i(X) - f_i(X_k) + \frac{Y_{ik}}{\mu_k}\|^2_F,$$
$$= \arg \min_X \|X\| + \frac{\mu_k}{2} \sum_i \|f_i(X) - f_i(X_k)\|^2_F,$$
$$+ \sum_i \langle f_i(X) - f_i(X_k), \mu_k f_i^*(f_i(X_k) - A_i + Y_{ik}) \rangle.$$  \hfill (B.5)

Assume $\eta_i$ is an approximate parameter for $\|f_i\|^2$, we then approximate $X_{k+1}$ as

$$X_{k+1} = \arg \min_X \|X\| + \frac{\mu_k}{2} \sum_i \eta_i \|X - X_k\|^2_F + \sum_i \langle X - X_k, \mu_k f_i^*(f_i(X_k) - A_i + \frac{Y_{ik}}{\mu_k}) \rangle,$$
$$\hfill (B.6)$$

where $f_i^*$ is the adjoint of $f_i$. 

119
To proceed, let $\eta_0 = \sum_i \eta_i$. We have

$$X_{k+1} = \arg \min_X \|X\| + \frac{\mu_k \eta_0}{2} \|X - X_k\|_F^2 + \langle X - X_k, \sum_i \mu_k f_i^* (f_i(X_k) - A_i + \frac{Y_{ik}}{\mu_k}) \rangle.$$ (B.7)

$$= \arg \min_X \|X\| + \frac{\mu_k \eta_0}{2} \|X - X_k + \frac{1}{\eta_0} \sum_i f_i^* (f_i(X_k) - A_i + \frac{Y_{ik}}{\mu_k})\|_F^2$$

With the help of linear approximation of Eqn(B.4), we use the solution of Eqn(B.7) to update $X$ iteratively as

$$X_{k+1} = T_{\mu_k \eta_0} (X_k - \frac{1}{\eta_0} \sum_i f_i^* (f_i(X_k) - A_i + \frac{Y_{ik}}{\mu_k})).$$ (B.8)