# EFFICIENT LEARNING METHODS FOR HIGH DIMENSIONAL VISUAL DATA

A Dissertation
Submitted to the School of Computer Engineering
of Nanyang Technological University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

**Marcus Chen Caixing**

November 18, 2015

This thesis is dedicated to my beloved family.

# Abstract

High dimensional visual data, derived from images or videos, is ubiquitous as advance camera technologies enable more measurements per sample to be captured. Increasingly sophisticated visual data representations further contribute to an increase in data dimensionality. To facilitate high level visual analytic tasks, this thesis focuses on three areas of high dimensional data processing, namely direct graph embedding for sample class or cluster prediction, video tracking for temporal information extraction, and spatial segmentation for a compact representation of high resolution images.

To address the challenges of irrelevant, noisy, and highly correlational dimensions, a novel unified framework is proposed to simultaneously perform graph embedding and feature selection. This framework enables an efficient extraction of linear data intrinsic structures, which are low dimensional and robust to both noisiness in dimensions and outlier samples. This framework is computationally efficient and flexible to incorporate various data prior properties such as smoothness, sparsity, and locality.

In video analysis, efficient learning of high dimensional visual data often requires modeling of temporal evolution of object appearance and motion. Instead of analyzing all the visual data, object level temporal information can be extracted via visual tracking for more efficient learning. For a long video sequence, the object appearance will change due to variations in its poses and orientation, illumination, and occlusion. To track both the object appearance and position, it is necessary to have a robust tracker with an adaptive object appearance update. We propose a generative model to address the dual uncertainties in both the target positions and appearance simultaneously. A diffusion process on a Riemannian manifold allows a geodesic evolution of the target appearance.

Spatially, object segmentation can significantly simplify visual learning by grouping many pixels into a meaningful representation. However useful, object segmentation remains unsolved. We propose a novel object co-segmentation framework to learn object segmentation using a large image set. By leveraging on good segmentation results on the simplest images, we can propagate this to more and more complex images.

All the proposed methods enable efficient learning of high dimensional visual data. The proposed unified framework for simultaneous dimensionality reduction and feature selection provides an abstraction for many high dimensional learning methods in the literature. Upon this framework, more learning methods can be further developed. Both visual tracking and object segmentation are important steps for many complex visual applications such as visual recognition. The proposed methods enable robust and efficient exploitation of both temporal and spatial information in visual data.

# Acknowledgements

Over the past five years, I have received support and encouragement from a great number of individuals. Professor Cham Tat Jen has been a very encouraging mentor. At the beginning of my Ph.D. studies, Professor Cham made numerous trips to my company for discussions. His guidance and encouragement have made this a thoughtful and rewarding journey.

Professor Ivor Tsang has been a wise and helpful mentor. Occasionally, our technical discussion went past midnight on Skype. This is truly helpful and motivating, and has propelled me to work harder to achieve what I set out to do.

My company, DSO National Laboratories in Singapore, has generously sponsored my studies. This covers my undergraduate study at Carnegie Mellon University, Master's Degree studies at Stanford University, and Ph.D. studies at Nanyang Technological University. Dr. Pang Sze Kim, my company supervisor, has encouraged me throughout my Ph.D. studies, and his strong support has been very encouraging. Dr. Xavier Briottet, who supervised me during my stay in Toulouse, France, has been a friend and mentor.

My contributors of different projects have been very helpful. Dr. Karine Adeline, Dr. Santiago Velasco, and Dr. Alvina Goh have helped a lot in my learning journey. I am really grateful for this.

My fellowship group in Singapore has given me tremendous encouragement, supports and prayers. I feel truly blessed to be a member in this group.

Finally, my family has always encouraged me to pursue my interests. For this long journey, they have been very patient and understanding. Without them, I cannot imagine that this dream could ever be realized.

# Contents

# List of Abbreviations

| | |
|---|---|
| AD | Anomaly Detection |
| CCA | Canonical Correlation Analysis |
| EM | Expectation Maximization |
| IPCA | Incremental Principal Component Analysis |
| LDA | Linear Discriminant Analysis |
| LLE | Locally Linear Embedding |
| LPP | Locality Preserving Projections |
| LS | Laplacian Score |
| MDS | Multidimensional Scaling |
| ML | Maximum Likelihood |
| MLE | Maximum Likelihood Estimation |
| MMSE | Minimum Mean Square Error |
| PCA | Principal Component Analysis |
| PDF | Probability Density Function |
| RFE | Recursive Feature Elimination |
| SVM | Support Vector Machines |

# List of Notations

| | |
|---|---|
| $*$ | Convolution |
| $(\cdot)^\top$ | Matrix or vector transpose |
| $X \in \mathbb{R}^{d \times n}$ | A zero-mean data matrix with $d$ dimensions and $n$ samples |
| diag($\mathbf{x}$) | A diagonal matrix whose diagonal elements are the elements of vector $\mathbf{x}$ |
| $\|\|\mathbf{x} - \mathbf{y}\|\|^2$ | Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ |
| $\mathbb{R}$ | Real number |
| $\|X\|$ | Determinant of matrix $X$ |

# Chapter 1

# Introduction to High Dimensional Visual Data

In the era of big data [53], technological advances in visual data sensing, collection, and storage have motivated many industrial sectors to capture massive data at an accelerating pace. Massive data is often characterized by both a huge number of samples and a high dimensionality for representation. Besides advances in visual data technology, active research in visual data representation has also contributed significantly to the explosive growth in data dimensionality. With a higher data dimensionality, many visual analytic applications can achieve a much better performance in classification and recognition that cannot be easily obtained by lower dimensional data. However, these advances come with emerging challenges, significantly stressing classical statistical methods that are able to handle small dimensions [66].

This chapter introduces the concept of high dimensional visual data, its origin, characteristics, and the challenges for machine learning tasks.

## 1.1 High Dimensional Visual Data

The amount of data that exists in the world is enormous and continues to increase exponentially every day. About 90 percent of the data that exists in the world today was created in the last

two years alone, and a quintillion ($10^{18}$) bytes are created on a daily basis [172]. This enormous amount of data is often referred to as big data. Big data refers not only to a large number of data samples, but a high number of dimensions. **Data dimensionality also refers to the number of data attributes, explanatory variables, and features.** For example, in a business application, a customer's data could cover attributes such as age, gender, and salary. When a data set has many dimensions, it is referred to as high dimensional data.

Derived from images or videos, visual data contributes significantly to the data growth in the world. The social media websites provide a platform to undertake and record an explosive growth in both quantity and resolution of images and videos. For example, Instagram, an online mobile photo sharing website, has 300 million active users, 30 billion photos shared, and 70 million photos viewed and liked daily[1]. YouTube, a video sharing website, on average hosts additional 100 hours of video each minute, 6 billion hours of video watched each month[2].

High dimensional visual data can also be found throughout academic research. The phenomenon can be observed in datasets from two popular machine learning repositories, namely, the UC Irvine Machine Learning Repository and the LIBSVM databases (Table 1.1). Over the last two decades, data dimensionality has grown from less than 100 to more than 1 million.

| Data Name | Dimension | Year |
|---:|:---:|:---:|
| LETTER | 16 | 1991 |
| USPS | 256 | 1994 |
| MNIST | 784 | 1998 |
| GISETTE | 5,000 | 2003 |
| YOUTUBE MVG | 1,000,000 | 2013 |

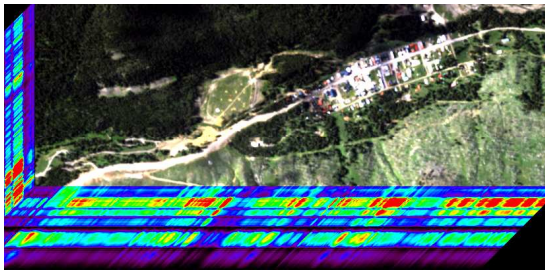
Table 1.1: Characteristics of some data sets collected from two popular data analytic repositories [172].

[1]Statistics were extracted from http://instagram.com/press/ on Dec 18, 2014.
[2]Statistics were extracted from https://www.youtube.com/yt/press/statistics.html on Dec 18, 2014.

## 1.2 The Origins of High Dimensional Visual Data

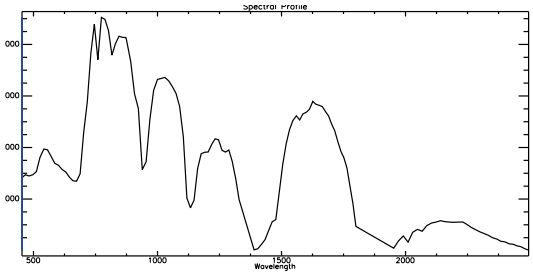
There are two main driving factors for the rapid growth in visual data dimensionality, namely, technological advancement and the development of new visual data representation techniques. We now examine these two motivating factors.

### 1.2.1 Technological Advancements



1.1.a: A photograph of a vegetation landscape with the output from hyperspectral imaging - this type of output is known as a data cube.

1.1.b: Photograph of the Colosseum, taken by a CMOS camera with a resolution of 10 megapixels. This number represents number of dimensions in the resulting image.




1.1.c: Plot that shows the vegetation spectra across the spectrum of wavelengths in the vegetation landscape image of Figure 1.1.a.

Figure 1.1: Examples of high dimensional data.

Advancements in sensor technology have contributed to the growth of high dimensional visual

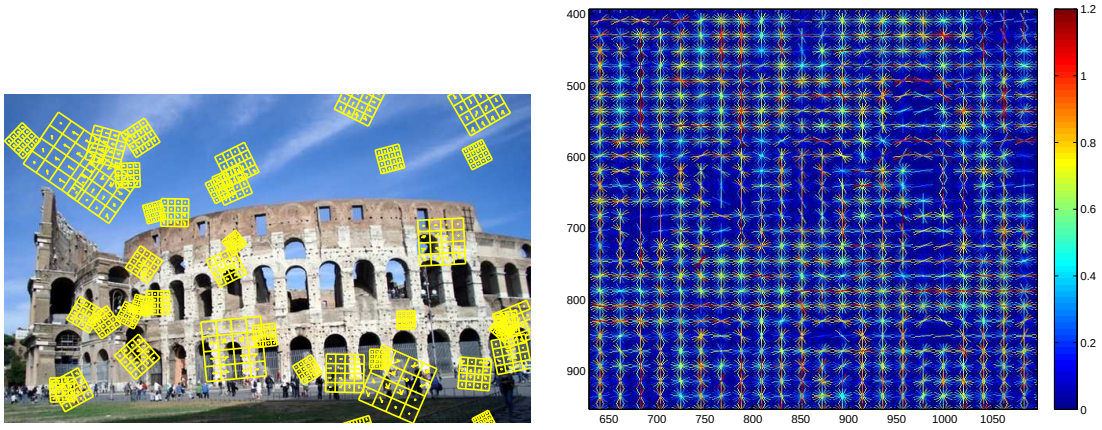
data. The increased spatial resolution allows fine details to be captured and results in higher quality photographs. For example, the latest camera phones such as the Nokia Lumia 1020 are able to capture photographs with resolutions up to 41 megapixels. Figure 1.1.b shows a photograph taken by a digital camera that results in an image with millions of pixels. High spatial resolution is also commonly referred to as high definition (HD).

Besides having high spatial resolution, consumer cameras also record HD image sequences at a fast frame rate (commonly 30 frames per second). This is the so-called HD video. An HD video will have about 1800 megapixels within one minute, assuming one megapixel per frame. Clearly, HD video data can easily have a very high data dimensionality.

Consumer cameras have evidenced a rapid increase in spatial resolution, allowing objects in images to resolve better with finer details. On the other hand, specialized cameras such as hyperspectral cameras can capture images across hundreds of bands in the electromagnetic spectrum (spectral measurements), including wavelengths that cannot be detected by the human eye. An example of an output from hyperspectral imaging is shown in Figure 1.1.c. This figure shows a vegetation landscape along with measurements at different wavelengths - this is called a "data cube". Figure 1.1.c shows vegetation spectra across the spectrum of wavelengths that exist in the image. In this example, over 200 dimensions exist in the output data.

### 1.2.2 Development of New Visual Data Representation Techniques

Besides technological advancement, continuing research in data representation has also brought a huge increase in data dimensionality. Data representation aims to mathematically represent data samples so that the corresponding distance measure can capture the desired properties of the original data. Taking image data for example, a simple way to represent it is as a vector, where each pixel of the image constitutes one dimension. Then, the distance between




1.2.a: SIFT feature extraction

1.2.b: HoG feature extraction

Figure 1.2: Examples of image features extraction techniques.

samples is simply a vector distance, capturing the difference between images when they are pixel-to-pixel aligned. Alternatively, a matrix can be used to represent images, then the distance between images is now a matrix distance, reflecting the similarities in both pixel values and spatial arrangement. Both representations can easily result in high dimensions.

The image processing community has developed much more advanced ways to represent image data. Advanced techniques have been invented to extract useful image features; these techniques are called feature extraction methods. Some popular feature extraction methods are included as follows [136, 155]:

(1) Encoding image local spatial information: Scale Invariant Feature Transformation features (SIFT, 128 dimensions), Local Binary Patterns features (LBP, 58 dimensions), and Gabor features. SIFT features (shown in Figure 1.2.a) encode gradient information on the scale of feature points. As shown in the image, each $4 \times 4$ rectangle represents a SIFT feature, in the scale, orientation, and magnitude inside each sub-rectangle. In Figure 1.2.b, the Histogram of Gradient (HoG, 512 dimensions) features capture the gradient

magnitudes in each direction (quantized to 8 bins).

(2) Encoding image global information: color histogram. Color histograms encode the empirical distribution of color intensities.

(3) Spatial-temporal information: spatio-temporal features for behavior recognition [52]. This is particularly important for video processing.

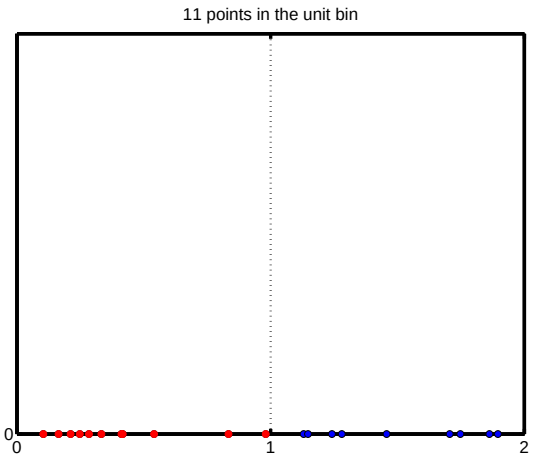
For a more comprehensive overview of current feature extraction methods, the reader is referred to [155]. Along with technological advancement in sensor technologies, advancement in feature extraction methods drives the significant growth of high dimensional data.

## 1.3 Characteristics of High Dimensional Visual Data

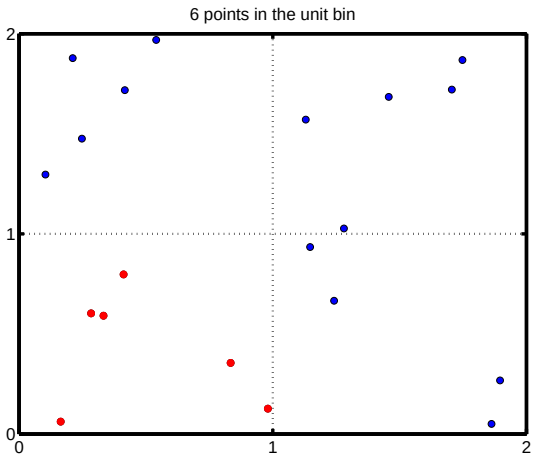
Inherently due to high dimensionality, visual data is sparsely distributed in the data space, tends to have a heavy tail distribution and have a high correlation among dimensions. These characteristics are elaborated in detail in the following subsections.

### 1.3.1 Sparsely Distributed in High Dimensional Space

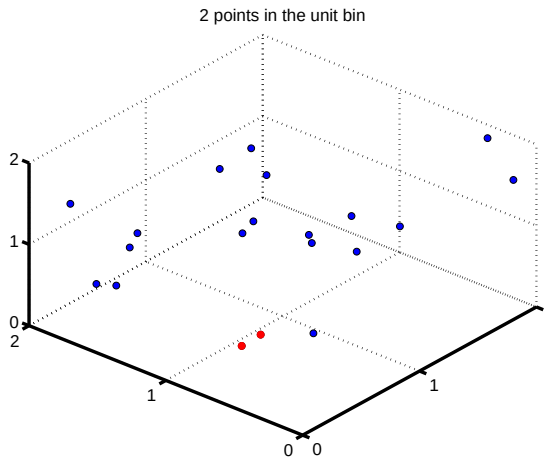
High dimensional data is sparsely distributed in data space. To illustrate this phenomenon, Figure 1.3 shows 20 points distributed randomly in 1-D, 2-D, 3-D spaces that have 2 units in each dimension. In 1-D, 10 out of the 20 points lie in the first unit of space. In 2-D, this number drops to 6, and in 3-D, this drops to 2. To achieve the same density of 10 data points per unit in a higher dimensional space as in 1-D space, it would require an exponential number of samples. Assuming every dimension is independent and identically distributed (*i.i.d.*) in a uniform distribution, then for $x \in \mathbb{R}^d$, it requires $10^{d-1}$ samples to ensure that on the average




1.3.a: Data distribution in 1-D



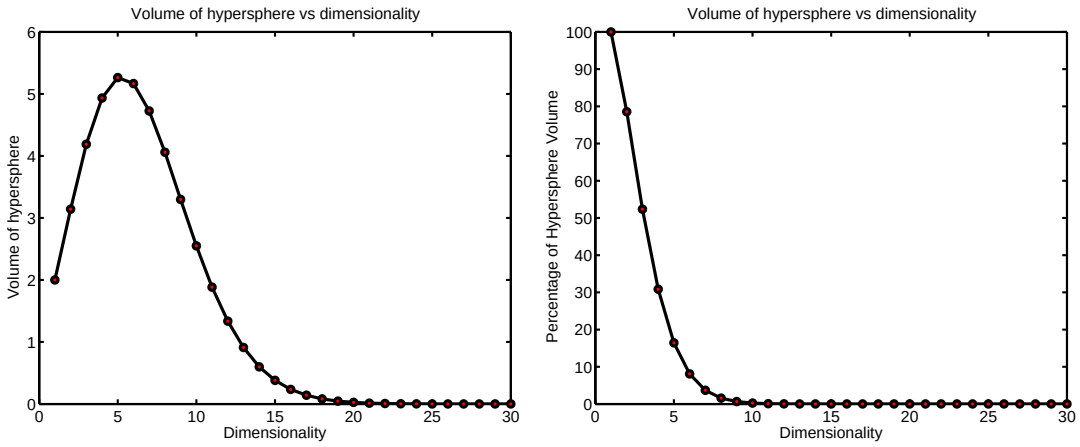

1.3.b: Data distribution in 2-D




1.3.c: Data distribution in 3-D

Figure 1.3: Data in only one dimensional space is relatively tightly packed. Adding a dimension stretches the points across that dimension, pushing them further apart. Additional dimensions spread the data even further, making high dimensional data extremely sparse.

every unit has 1 sample; this is hardly possible even for a small $d$, as the number of data samples required to reach a similar density is enormous. For example, to achieve the same density of 10 points per unit in a 5-D space, 1 million samples are required. In the light of this, it is difficult to estimate the data statistics from existing samples, and also computationally infeasible for generative model methods to generate sufficient samples to approximate the predefined distributions.




1.4.a: Volume vs dimensionality

1.4.b: Volume percentage vs dimensionality

Figure 1.4: Data is sparse in the shell of a hypersphere as dimensions increase.

Another interesting observation is that majority of high dimensional volume is at the shell of hypersphere. This is illustrated as follows. Consider the volume of a hyper-sphere in $\mathbb{R}^d$, $d$-dimensional space, given as follows:

$$V(d) = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2}+1)}, \tag{1.1}$$

where $d$ and $r$ are the dimensionality and radius, respectively, and $\Gamma(\frac{d}{2}+1)$ is a gamma function. Figure 1.4 shows how quickly a unit radius hypersphere drops in volume percentage as

dimensions increase. The hypersphere inscribes the hypercube with length of 2 units. Most of volume exists in the outer shell or the "corners" between the hypercube and the hypersphere. This is the so called *Empty Space Phenomenon* [138].

### 1.3.2 Heavy-Tailed Distribution

In 1-D space, a standard Gaussian distribution has 90% of samples in the interval $[-1.65, 1.65]$. This probability falls quickly as data dimensionality increases. The Empty Space Phenomenon indicates that the probability of samples falling within the "inscribed hyperball" converges to zero as the data dimensionality increases [20]. In other words, a large bulk of data samples exist in the tail ends. For data distribution modeling, it is necessary to take this into account.

### 1.3.3 High Correlation Among Dimensions

Besides the above two generic characteristics of high dimensional data, high dimensional visual data tends to have highly correlated dimensions. This is because visual data is derived from images and videos, which tend to be smooth and thus correlated in measurements. Visual data mainly captures three kinds of information, namely, spatial, spectral, and temporal information. An increase in the resolution contributes to a higher correlation among dimensions. Spatially, an image with more pixels within the same field of view has a higher resolution. It is very common for current consumer cameras to capture thousands of pixels in each image, resulting in a high spatial correlation. The spectral resolution refers to the measurements at different wavelengths, and is used in multispectral or hyperspectral cameras. Spectral cameras nowadays have a spectral resolution less than 10 $nm$. Equivalently, there are more than 100 continuous channels within 1 $\mu m$. It is also easy to imagine that those channels are highly correlated. Finally, temporal resolution refers to frame per second in videos. Current cameras can record

more than 100 frames per second; slow motion objects will have a high correlation in temporal dimensions.

## 1.4 The Curse of Dimensionality

High dimensionality poses new challenges in the processing of such data, including challenges in computational efficiency and data distribution modeling. The challenges that come from processing high dimensional data but do not occur as with low dimensional data is referred to as the curse of dimensionality.

### 1.4.1 Computational Complexity

The immediate challenge of high dimensional data lies in the computational complexity. The complexity of many optimization and learning algorithms is related to the data dimensionality by a polynomial distribution. For example, Principal Component Analysis (PCA), a popular algorithm, has a polynomial computational complexity of $O(d^2p)$, where $d$ is the data dimensionality and $p$ is the number of principal components. For each doubling of data dimensionality, these methods slow down by $4p$ times. Image applications involving high dimensional data cannot scale when its computational complexity is more expensive than linear complexity, i.e., $O(d)$.

### 1.4.2 Noisy and Correlated Dimensions

In many situations, high dimensional data is noisy and correlated among dimensions. Images can be corrupted by sensor artifacts such as dead pixels and non-calibrated pixels. Dust specks on the camera lens also cause noisy images. From these images, extracted features will be

noisy too. On the other hand, natural images are smooth spatially; many neighboring pixels are similar in readings. It is very likely that resultant image representations are highly correlated among different dimensions.

Noisiness and high correlation require us to construct robust statistical models for what is observed. Simply using all the dimensions for data analysis may cause the performance to be degraded due to unreliable data statistics and rank deficiency problems.

### 1.4.3 Data Distribution Modeling

High dimensional data also complicates data distribution modeling. In many statistical applications, covariance is an important statistic to calculate. However, the large number of dimensions makes it difficult to estimate covariance accurately in high dimensions, leading researchers to try find ways to reduce the number of parameters through sparsity [21] or with the structural model [134].

## 1.5 High Dimensional Data in A Visual Analytic Framework

High dimensional data, discussed so far, mainly exists in the low level processing layer of a typical visual analytic framework. Mid-level techniques are necessary for efficiently handling high dimensional data to extract semantically meaningful information for high level analysis, such as object recognition, event recognition, and activity inference. A possible framework is illustrated in Figure 1.5 (a similar framework is used in [99] to reduce semantic gap for image retrieval). The success of visual analysis depends on these three levels of processing, namely, low level feature extraction as discussed in Section 1.2.2, mid-level visual learning techniques such as machine learning, temporal tracking, and spatial object segmentation, and the high

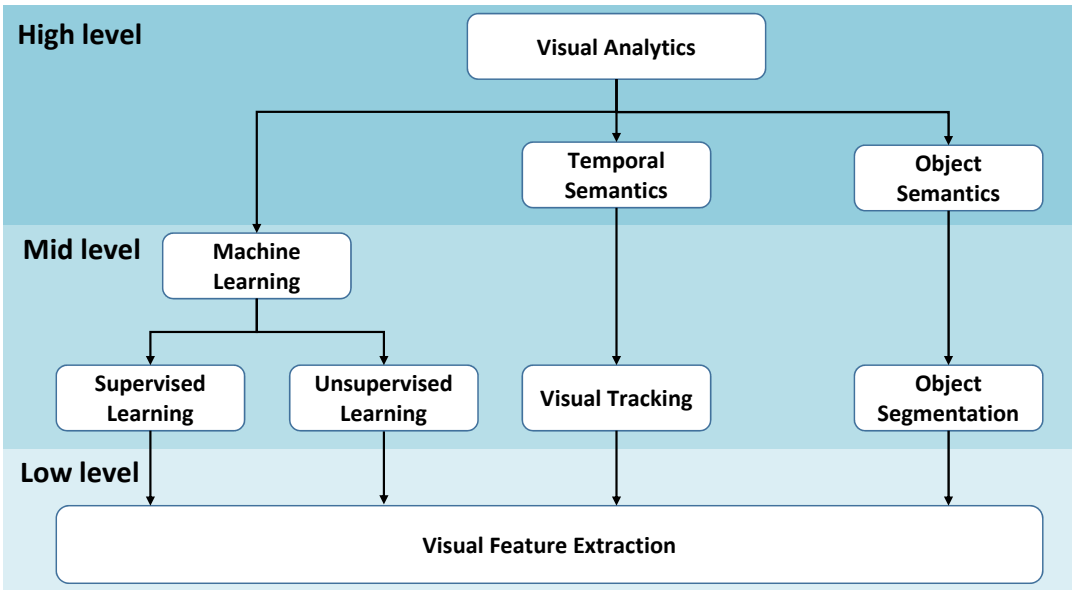



Figure 1.5: Visual analytic components and techniques.

level processing of the learning results from mid-level techniques.

While there are many other components in a visual analytic framework, this thesis chooses to focus on the three areas in the mid-level learning methods, namely, machine learning methods, visual tracking, and object segmentation. All these methods process low level features and provide meaningful compact information to high level visual analytic tasks.

- Machine learning. To extract insights from visual data, machine learning techniques can be used to predict visual data class labels (supervised learning methods) or appreciate data clustering (unsupervised learning methods). Typically, this approach predefines a learning task. For example, the learning task in [55] is to classify video data into five categories: tennis, basketball, volleyball, soccer, and table tennis.

- Temporal semantics. Image sequences have been used to extract high level temporal semantics, including event recognition [75, 168], target motion for surveillance [39], gesture recognition [141]. To support these high level semantic learning tasks, it is nec-

essary to associate objects throughout the video sequences, and visual tracking is the most common approach to achieve this.

- Object semantics. The ability to describe objects in an image is critical for visual understanding. For example, an image tagged with beach, man, and dogs can immediately help users visualize it as Figure 1.6. Object segmentation can efficiently help recognize objects in images.

Figure 1.6: An example of visual understanding based on object components. It is easy to visualize this image given the tags of man, dog and beach.

The mid-level processing layers aim to efficiently process the low level high dimensional data, so that more relevant and compact visual representations may be achieved for high level processing. While machine learning methods directly achieve the pre-defined learning tasks, both visual tracking and object segmentation will compress the high dimensional visual data into compact information like object attributes.

## 1.6 Organization of This Thesis

In this chapter, we have presented an overview of high dimensional data and its characteristics, how advances in technology and continuing research have led to the increase in data dimensionality, and the challenges posed by higher dimensionality. We also discuss different levels of high dimensional data processing layers in visual analytic applications.

As shown in Figure 1.7, this thesis is organized as follows. Chapter 2 briefly surveys existing

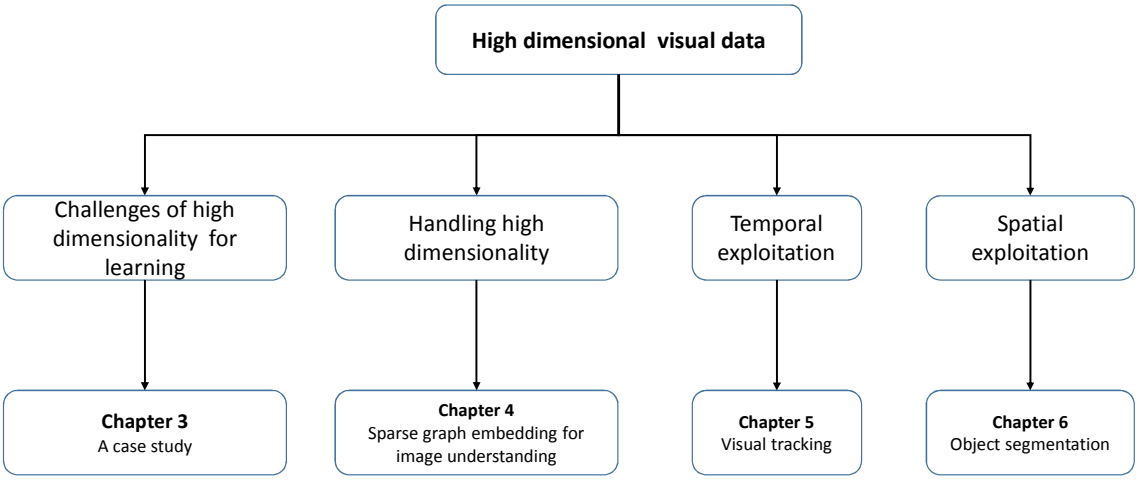



Figure 1.7: Organization of this thesis.

methods of handling high dimensional data in the literature. We conduct a case study on challenges of high dimensionality in Chapters 3. In order to deal with challenges of high dimensionality, Chapter 4 proposes a efficient feature selection framework for linear graph embedding. Chapter 5 and 6 propose methods for visual tracking and object segmentation. Finally, Chapter 7 concludes this thesis and briefly describes some future work.

# Chapter 2

# A Brief Review of Dimensionality Reduction and Feature Selection Methods for High Dimensional Data

Natural data is intrinsically low dimensional. One intuitive approach for handling high dimensional data is to extract the low dimensional intrinsic data structures. There are two main approaches, namely, feature selection and dimensionality reduction. This chapter briefly reviews some existing methods on feature selection and dimensionality reduction. This will serve as the foundation for the subsequent chapters.

## 2.1 Feature Selection

Recognizing the importance of feature selection, many methods have been proposed in the literature. Subsequently, a few survey papers have been published trying to organize and summarize various methods according to different machine learning areas [48]. This may be because feature evaluation criteria could differ from one application to another. Therefore, feature selection methods are often proposed for specific learning contexts. For example, the Recursive Feature Elimination (RFE) scheme [67] defines the least relevant feature as the one contributing the least to the classification margin through a recursive elimination method using Support

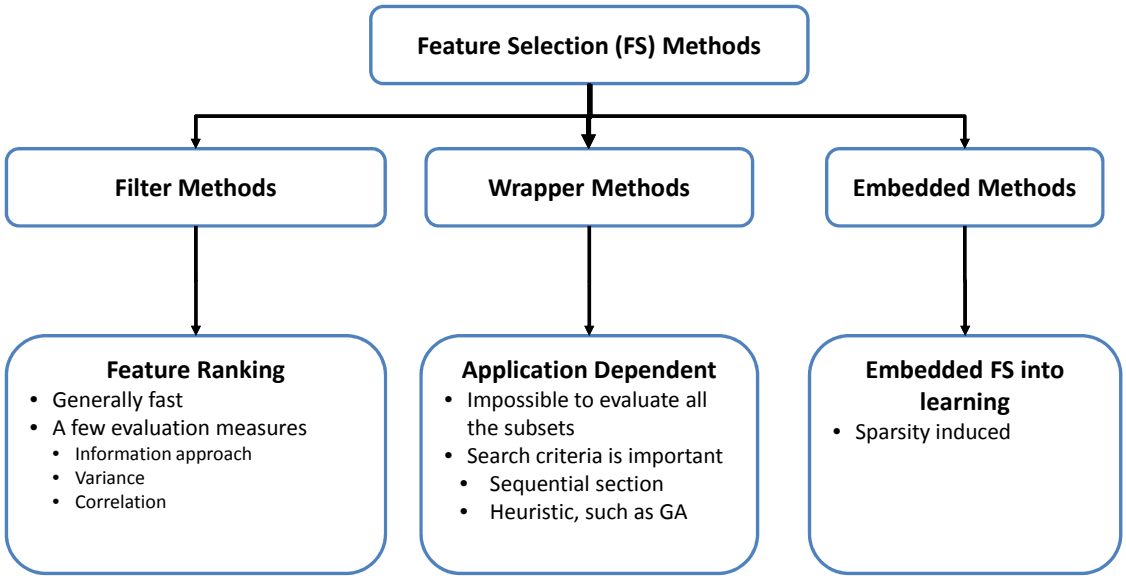



Figure 2.1: Some feature selection methods.

Vector Machines (SVM).

In the literature, there are a few popular feature importance measures adopted in existing methods [97]. These measures are designed based on the following characteristics. For supervised methods, there are two main feature importance measures, namely, distance based measures and the correlation based measures. Specifically, in the distance measure, it defines the important features as those that separate classes better and cluster the samples of the same classes, such as Linear Discriminant Analysis (LDA) based feature selection methods [67]. In the correlation based measure, the important features are those that correlate well with class labels and give better prediction results, such as the method in [68]. In the unsupervised methods, due to the absence of class labels, several criteria have been proposed to evaluate the feature importance based on different learning contexts, such as information measure [125], variance measure [101], and locality measure [71], summarized as follows:

- information measure: good features gain more information when included. Information

gain can be computed via mutual information methods.

- variance measure: good features capture more variance in the data.
- locality measure: good features can preserve data locality better.

The existing feature selection methods are also often classified as filter, wrapper and embedded approaches [66]. Overall, feature selection methods may be categorized as shown in Figure 2.1. First, filter methods aim to pre-define some desired intrinsic data properties and rank feature subsets accordingly. It is independent of its subsequent applications. Filter methods employ either forward selection or backward elimination search strategies, often resulting in local optimality. Second, wrapper methods are application dependent; the subset of features giving the best performance of a predetermined learning task will be chosen. The evaluation of features subsets involves performing the application processes, and is thus very computationally expensive. Generally, the wrapper methods give a better performance than the filter methods [96]. Finally, the embedded methods formulate feature selection process in their optimization objectives such as sparsity regularization [76].

Recent works in literature focus on graph based feature ranking and sparsity induced embedded methods [96]. They are explained in more detail in the next two sections.

### 2.1.1 Graph-based Methods for Features Ranking

Graph, a powerful tool to encode sample neighborhood relationships, is often used to model intrinsic data structures. For example, ISOMAP [149] constructs shortest paths between any pairs of samples on a graph. By utilizing a graph model, good features preserve the intrinsic data structures well. He et al. [69] proposed to encode data locality via a Laplacian Score (LS) for feature ranking. Similarly, Zhao and Liu [175] proposed to evaluate features based on SPECtrum decomposition (SPEC) on the Laplacian matrix. These two methods employ

feature level ranking, and redundancy between features is not considered. To address this, Nie *et al.* [120] proposed a feature-subset trace ratio method to rank features that best discriminate between-class and preserve within-class relationships. It reformulates the subset level selection into a sum of individual features based on the monotonicity property. In other words, if feature subset A is better than B, then $\{$A,c$\}$ is better than $\{$B,c$\}$.

### 2.1.2 Sparsity-induced Embedded Methods

Another popular approach to select features is to formulate the problem as a sparse regression problem, with the sparsity regularization term controlling features cardinality. This formulation benefits from many efficient and optimal solvers available in the literature.

The sparse regression method LASSO [154] is popular in the literature due to its simplicity and efficiency in computation. Its formulation of the feature selection is as follows:

$$\min_{w} \ \|X^\top w - y\|^2 + \gamma\|w\|_1, \tag{2.1}$$

where $X \in \mathbb{R}^{d\times n}$ is the data matrix, $y$ is the response variable vector, and $\gamma$ controls the sparsity of the solution (a larger $\gamma$ generally results in a more sparse solution, and vice versa). LASSO can only choose at most $n$ variables when $d > n$ and is not well defined unless $\|w\|_1$ is bounded by a certain value [180]. The LASSO shrinkage may produce biased estimates for large coefficients of $w$ [179]. Zhou and Hastie [180] then proposed the Elastic net method by adding an $L_2$ norm term, i.e. $\gamma\|w\|_1 + \beta\|w\|^2$. However, both LASSO and the Elastic net method do not guarantee the choice of the same set of features when regressing over different principal components. To address this issue, Cai *et al.* [30] proposed a simple heuristic method to rank the features based on the maximum absolute weight of different subspaces. This method is termed as MultiClusters Feature Selection (MCFS) with preserving data locality as its main

learning objective.

Recently, the Convex Relaxations for Subset Selection (CRSS) method [10] was shown to achieve better results than the Least Angle Regression method (LAR) method [60]. It proposes two search methods. One is a branch and bound search method and is combinatorial in computational complexity. Another is based on random generation of Gaussian vectors. Both LASSO and CRSS select features based on sequential iteration of principal components. In this manner, tuning parameters for different components is complicated [163]. Therefore, they do not optimize for the collective objectives.

A recent work [163] optimizes the collective objectives for sparse PCA via a Fantope projection approach as follows:

$$\begin{aligned} \text{Maximize} \quad & < S, X > -\lambda \|X\|_{1,1} \\ \text{s.t.} \quad & X \in \mathbb{F}^d, \end{aligned}$$

where $S$ is the covariance matrix, $X$ is the estimator with sparsity constraint in a Fantope $\mathbb{F}^d$, $< S, X >$ is the inner product of $S$ and $V$, and $\|X\|_{1,1}$ is the sum of all the absolute entries of $X$. This approach is computationally efficient and optimal [163].

## 2.2 Dimensionality Reduction

Different from feature selection methods, dimensionality reduction approaches may employ all the features to find a lower dimensional representation of the original data. There are two major types of dimensionality reduction methods: linear and non-linear dimensionality reduction as shown in Figure 2.2. In this section, we first briefly review the recent literature.

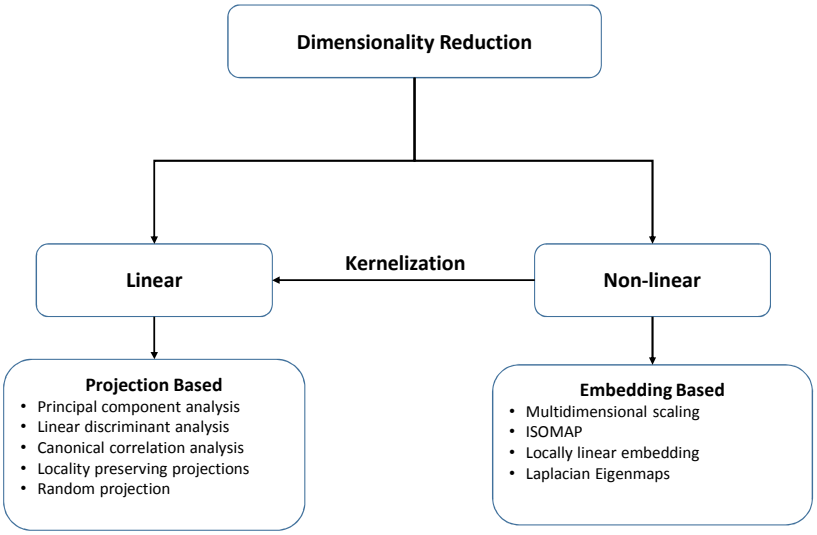



Figure 2.2: Some dimensionality reduction methods.

### 2.2.1 Linear Dimensionality Reduction

Linear dimensionality reduction methods result in new projected features (dimensions) being linear combinations of original dimensions. Let $X \in \mathbb{R}^{d \times n}$ be a dataset with zero-mean $d$-dimensional $n$ samples, $W = [\omega_1, \omega_2, ...]$ be the projection matrix, and $Y$ be the projected low dimensional representation. Note that for non-zero-mean data, a linear projection can be applied first. Linear techniques can then be summarized in the following formulation:

$$Y = W^\top X. \tag{2.2}$$

#### 2.2.1.1 Principal Component Analysis

Principal Component Analysis (PCA) finds the subspace that maximizes the explained variance,

$$\underset{w}{\operatorname{argmax}} \quad \frac{w^\top X X^\top w}{w^\top w}. \tag{2.3}$$

Often, the first few leading components explain the most of variance, and the rest can be discarded with minimal loss of information. The projection matrix $W$ consists of orthogonal principal components; as a result, the new features are uncorrelated.

PCA is probably the earliest algorithm that has inspired many modern algorithms. It has a few limitations.

- Ignoring high order statistics: although PCA decorrelates data, it ignores the correlation in the third and higher orders.
- It does not use sample label information, and thus does not optimize for classification tasks [56].

#### 2.2.1.2 Locality Preserving Projections

While PCA preserves the overall data variance globally, Locality Preserving Projections (LPP) preserve data locality so that the sample neighborhood relationship is retained as much as possible in the lower-dimensional space. Its formulation is as follows:

$$\begin{aligned} &\underset{w}{\operatorname{argmin}} \quad \frac{w^\top X L X^\top w}{w^\top X D X w}, \\ &\underset{w}{\operatorname{argmax}} \quad \frac{w^\top X A X^\top w}{w^\top X D X w}, \end{aligned} \tag{2.4}$$

where $L$ is the Laplacian matrix, $D$ is a diagonal matrix, $L = D - A$, and $A$ is the affinity matrix.

#### 2.2.1.3 Fisher's Linear Discriminant Analysis

Both PCA and LPP are unsupervised and do not use any sample label information; they are often used for general pre-processing, clustering, or visualization. For classification, supervised dimensionality reduction methods generally achieve a better accuracy as they explore the label information. For example, LDA optimizes for between-class separation ($S_b$) and within-class closeness ($S_w$) as follows:

$$\underset{w}{\operatorname{argmax}} \ \frac{w^\top S_b w}{w^\top S_w w}. \tag{2.5}$$

LDA considers class-level global data structures, and assumes a Gaussian distribution for the samples in each class. To include local data properties, Marginal Fisher Analysis [169] considers both local sample neighborhood relationships and between-class relationships.

#### 2.2.1.4 Random Projection

Random projection is a simple yet powerful method to project high dimensional data to a lower dimensional space. In the general formulation, $Y = W^\top X$, $W$ is simply a random matrix. It is computationally efficient and generally produces decent results. Kaski [86] shows that given sufficient projected subspaces, random projection preserves the sample angular distance.

#### 2.2.1.5 Kernelization

Linear projection methods represent the resultant low dimensional features by linearly combining the original features. They could have some severe limitation when data is not linear. There are a few ways to extend the linear methods to handle non-linear data. One of the ways is to apply the *kernel trick*. The kernelized versions of projective methods are also popular, for example Kernel PCA, Kernel LDA, and Kernel LPP.

Assume that some algorithms only depend on the inner product of samples, and the dot product is well defined in kernel space of the data such that for all samples $x_i, x_j$,

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle,$$

where $k$ is the kernel function, $\Phi(x)$ is the transformation of data from the original space to the kernel space.

The **kernel trick** refers to that instead of computing the dot product of transformed features $\langle \Phi(x_i), \Phi(x_j) \rangle$, the algorithm computes the dot product in the original space before applying the kernel function like $k(x_i, x_j)$. Taking kernel PCA (KPCA) as an example. The derivation of Kernel PCA can be found in [27], and the final form of KPCA is as follows:

$$\overline{K}\omega = n\lambda\omega, \tag{2.6}$$

where $\overline{K} = PKP$, $\overline{K} \in \mathbb{R}^{n \times n}$ is the centered kernel matrix, $P = \mathbf{1} - I$, $\mathbf{1} \in \mathbb{R}^{n \times n}$ is the matrix whose entries are all 1, and the entries of the kernel matrix $K_{i,j} = k(x_i, x_j)$ are the dot products of pairs of data samples. Some typical kernel functions include:

- linear: $k(x_i, x_j) = \langle x_i, x_j \rangle$,

- polynomial of degree $m$: $k(x_i, x_j) = \langle x_i, x_j \rangle^m$,
- exponential: $k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2d^2}\right)$.

Based on the formulation, it is clear that KPCA has a similar algorithm as PCA with an additional step on the construction of the centered kernel matrix.

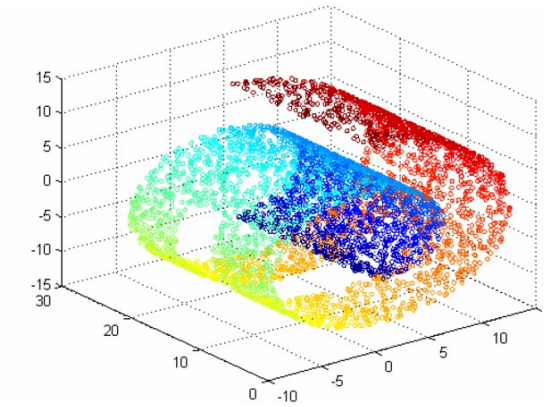
### 2.2.2 Non-linear Dimensionality Reduction

Linear dimensionality reduction techniques find low dimensional representations by a linear projection of the original data. Most of them are reasonably fast and are often used to obtain the first impression of data. Generally, except LPP, they do not consider sample neighborhood relationships. This may result in a loss of sample neighborhood relationships in some cases and degradation in the subsequent application performance. For example, the famous toy data, Swiss Roll, shown in Figure 2.3.a, cannot be "unwrapped" by linear dimensionality reduction methods. Data neighbors generally share the same class labels in the original space, but these relationships would be destroyed if projected using linear techniques such as PCA.

For such cases, non-linear dimensionality reduction methods are required to model the sample neighborhood relations via manifolds. Some common examples of manifolds include variations in images such as apple images, and golf swings images [1], as shown in Figure 2.3.

One of the earliest non-linear learning methods is Multidimensional Scaling (MDS) [25], which attempts to preserve all the pair-wise distances. MDS has been commonly used for visualizing high dimensional data. More recent developments in manifold learning methods such as ISOMAP [149], Locally Linear Embedding (LLE) [135], Hessian Eigenmaps [54], and Laplacian Eigenmaps [14] have created a huge impact in the non-linear modeling of high dimensional data.

[1] Images were extracted from `http://www.golfswingphotos.com/` on Sept 11, 2014.




2.3.a: Swiss Roll data


2.3.b: Variations in appearance

2.3.c: Golf swing

Figure 2.3: Some examples of non-linear data.

The ISOMAP, LLE, Hessian Eigenmaps, and Laplacian Eigenmaps methods based on K-nearest neighbors (KNN) graphs typically follow the three steps below:

(1) Find the K nearest neighbors,

(2) Estimate the local properties of the manifold by examining the KNN graphs,

(3) Find a global embedding by preserving the properties in Step (2).

ISOMAP preserves the property of shortest path between any two points, using Dijkstra's algorithm. This step is called geodesic approximation. Subsequently, MDS is applied to the distance matrix. It preserves the global structure, and generalizes to high dimensional data if the connectivity and metric information of the manifold are correctly formulated [12]. As shown in [12], ISOMAP is sensitive to noise. The choice of $K$ for KNN is crucial as a large

$K$ may introduce "short-circuit" edges in the graph, and a small $K$ may cause the graph to become too sparse to approximate geodesic properties well.

LLE on the other hand aims to minimize the reconstruction error of each data sample using its neighbors. The local properties of manifold is based on local linear relationships between neighbors. Instead of using geodesic distance as in ISOMAP, LLE does not need to resort to a distance measure, which may not be appropriate for some data. In addition, LLE is globally optimal and has one free parameter, and similar to ISOMAP, LLE is prone to noise.

## 2.3 Conclusion

In this chapter, we have seen that high dimensional data tends to have low dimensional intrinsic structures, exploiting which enables us to benefit from high dimensionality and avoid the associated problems. This chapter briefly discusses linear and non-linear embedding methods that are used to extract and model the data in the low dimensional space, interested readers may refer to a more detailed survey in [1]. These methods are classified into two main types of methods - dimensionality reduction techniques and feature selection methods. Dimensionality reduction techniques use all the dimensions to embed the original high dimensional data in the desired low dimensional representation, while feature selection methods aim to choose a subset of important features only. Both approaches aim to mitigate the challenges of high dimensionality, and they share so many similarities that they can be unified into a single framework. In Chapter 3, we conduct a case study on the challenges of high dimensional data before proposing a unified feature selection framework for a generalized class of dimensionality reduction techniques.

# Chapter 3

# High Dimensional Covariance Matrix Estimation for Anomaly Detection in Hyperspectral Images

The previous chapters describe some challenges in high dimensional data and the existing methods in the literature. Before introducing our methods, we examine some of these challenges in hyperspectral images as a case study. This is because hyperspectral data is naturally high dimensional, noisy, and highly correlated among dimensions. Furthermore, in high dimensional learning, covariance matrix estimation is important for many dimensionality reduction methods such as PCA, LDA, and CCA. Therefore, this chapter focuses on covariance matrix estimation methods in the literature on Hyperspectral imaging (HSI).

## 3.1 Introduction

Hyperspectral (HS) imagery provides rich information both spatially and spectrally. In contrast to from the conventional RGB camera, HS images measure scientifically the radiance received at fine divided bands across a continuous range of wavelengths. These images enable grain-fine classification of materials otherwise undistinguishable in spectrally reduced sensors. Anomaly detection (AD) using HS images is particularly promising in discovering the subtlest difference

among a set of materials. AD is a target detection problem, in which there is no prior knowledge about the spectra of the target of interest [143]. In other words, it aims to detect spectrally anomalous targets. However, the definition of anomalies varies. Practically, HS anomalies are referred to as materials semantically different from the background, such as a target in the homogeneous background [37]. Unfortunately, often the backgrounds are a lot more complex due to presence of multiple materials, which could be spectrally mixed at pixel levels.

This chapter comparatively surveys the existing AD methods via background modeling by covariance matrix estimation techniques. We analyze the AD in the context of optimal statistical detection, where the covariance matrix of the background is required to be estimated. The aim of covariance matrix estimation is to compute a matrix $\widehat{\mathbf{\Sigma}}$ that is "close" to the actual, but unknown, covariance $\mathbf{\Sigma}$. We use "close" because that $\widehat{\mathbf{\Sigma}}$ should be an approximation that is useful for he given task at hand. The *sample covariance matrix* (SCM) is the maximum likelihood estimator, but it tends to overfit the data when $n$ does not greatly exceed $d$. Additionally, in the presence of multiple clusters, this estimation fails to characterize the background well. For these reasons, a variety of regularization schemes have been proposed [40, 41], as well as several robust estimation approaches [2, 19, 63, 74, 108, 146, 152, 159, 166]. In order to comparatively evaluate different methods, a series of experiments have been conducted using synthetic data from distribution with covariance matrix $\mathbf{\Sigma}$ from real HS images. The rest of this manuscript is organized as follows:

## 3.2 Hyperspectral Anomaly Detector

This section briefly describes the *RX anomaly detector* (Reed and Xiaoli Yu [131]) before reviewing some covariance matrix estimation methods in the literature.

### 3.2.1 The RX-detector

AD may be considered as a binary hypothesis testing problem at every pixel as follows:

$$\begin{aligned} \mathcal{H}_0 : \quad & \mathbf{X} \sim f_{\mathbf{X}|\mathcal{H}_0}(\mathbf{x}), \\ \mathcal{H}_1 : \quad & \mathbf{X} \sim f_{\mathbf{X}|\mathcal{H}_1}(\mathbf{x}), \end{aligned} \tag{3.1}$$

where $f_{\mathbf{X}|\mathcal{H}_i}(\cdot)$ denotes the probability density function (PDF) conditioned on the hypothesis $\mathcal{H}_0$ when the target is absent (*background*), and $\mathcal{H}_1$ when the target is present. Usually, the background distribution $f_{\mathbf{X}|\mathcal{H}_0}(\mathbf{x})$ is assumed to follow a multivariate Gaussian model due to theoretical simplicity. Targets have a uniform distribution [151]. Based on these two assumptions, the hypothesis testing in the well-known *RX anomaly detector* has the following test statistics:

$$\begin{aligned} \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &\underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau_0, \\ \Rightarrow \mathtt{AD}_{\mathtt{RX}}(\mathbf{x},\tau_1) = (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1, \end{aligned} \tag{3.2}$$

where $|\mathbf{\Sigma}|$ is the determinant of matrix $\mathbf{\Sigma}$, $x \in \mathbb{R}^d$ is data sample, $\tau_0$ and $\tau_1$ are thresholds, above which $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. In other words, the RX-detector is a threshold test on the *Mahalanobis distance* [102]. In most of the cases, $\mathbf{\Sigma}$ is unknown and needs to be estimated. SCM is the maximum likelihood estimator (MLE) of $\mathbf{\Sigma}$ [8], and is computed as follows:

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top, \tag{3.3}$$

where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are samples from a $d$-variate Gaussian distribution with known mean $\boldsymbol{\mu} \in \mathbb{R}^d$.

### 3.2.2 The RX-detector in High Dimensional Space

To help better understand the implication of high dimensionality in the RX-detector, we develop an alternative expression for (3.2) based on the *Singular Value Decomposition* (SVD) of the covariance matrix $\boldsymbol{\Sigma}$, as follows:

$$\mathtt{AD}_{\mathtt{RX}}(\mathbf{x}, \tau_1) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{U}^{-1} \boldsymbol{\Lambda}^{-1} \mathbf{U} (\mathbf{x} - \boldsymbol{\mu}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1,$$

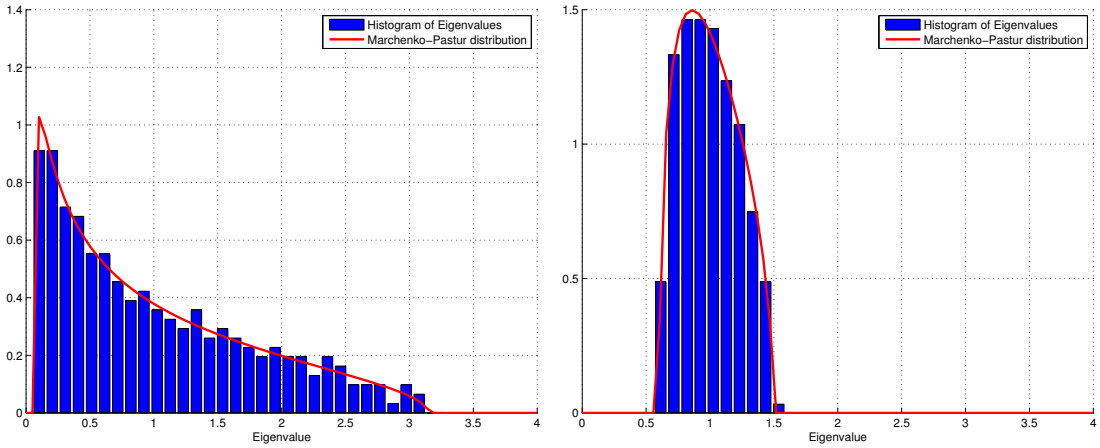
where $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ with $\boldsymbol{\Lambda}$ a diagonal matrix and $\mathbf{U}$ an orthogonal matrix. The eigenvalues $\{\lambda_i\}_{i=1}^d$ in $\boldsymbol{\Lambda}$ correspond to the variances along the individual eigenvectors and sum up to the total variance of the original data. Define the diagonal matrix $\boldsymbol{\Omega} = \{\omega_{ii}\}_{i=1}^d = \{1/\sqrt{\lambda_i}\}_{i=1}^d$, then $\boldsymbol{\Omega}^2 = \boldsymbol{\Lambda}^{-1}$. Since $\mathbf{U}^{-1} = \mathbf{U}^\top$, we can rewrite the RX-detector as follows:

$$\begin{aligned} \mathtt{AD}_{\mathtt{RX}} =& (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{U} \boldsymbol{\Omega} \boldsymbol{\Omega} \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1 \\ =& ||\boldsymbol{\Omega}\mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu})||_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_1. \end{aligned} \tag{3.4}$$

As we can see from this decomposition, the RX-detector in (3.2) is equivalent to the weighted Euclidean norm by the eigenvalues along the principal components. Note that as $\lambda_i \to 0$, the detector $\mathtt{AD}_{\mathtt{RX}}(\mathbf{x}, \tau_1) \to \infty, \forall \mathbf{x}$, resulting in an unreasonable bias towards $\mathcal{H}_1$ to $\mathcal{H}_0$. This fact is well-known in the literature as bad conditioning, i.e., the condition number [1] of $cond(\boldsymbol{\Sigma}) \to \infty$.

Before looking at the possible solutions to the ill-conditioning issue, we analyze the eigenvalue

[1] The condition number of a real matrix $\boldsymbol{\Sigma}$ is the ratio of the largest singular value to the smallest singular value. A well-conditioned matrix means its inverse can be computed accurately.




3.1.a: $n = 300, c = 1$ 3.1.b: $n = 6000, c = 0.01$

Figure 3.1: Empirical distribution of eigenvalues of SCM $\widehat{\mathbf{\Sigma}}$ and the corresponding Marchenko-Pastur Distribution for two different values of sample size $n$ with $d = 300$.

distribution of covariance matrices in the theory of random matrices [7, 57, 106]. Denoting the eigenvalues of $\widehat{\mathbf{\Sigma}}$ by $\lambda_1, \lambda_2, \ldots, \lambda_n$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. The *Marchenko-Pastur (M-P) law* states that the distribution of the eigenvalues of empirical covariance matrix, i.e., the empirical spectral density,

$$f(\lambda) = \frac{1}{n}\delta_{\lambda_i}(\lambda), \tag{3.5}$$

converges to the deterministic M-P distribution, when $d,\ n \to \infty$ and $\frac{d}{n} \to c$ [106] [7]. In the case of $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the M-P law describes the asymptotic behavior of $f(\lambda)$

$$f(\lambda) = \frac{\sqrt{(\lambda - a)(b - \lambda)}}{2\pi c\lambda}, \lambda \geq 0, \tag{3.6}$$

where $a = (1 - \sqrt{c})^2,\ b = (1 + \sqrt{c})^2$. A simple analysis of previous equation illustrates that, when $n$ does not greatly exceed $d$, the SCM will have eigenvalues in the vicinity of zero. This is illustrated in Fig. 3.1 with two different $\frac{d}{n}$ (dimensions to sample size ratio) values. Additionally, one can compute the integral of (3.6) between 0 and a small $k$ as function of $c$ to understand the effect of the ratio $\frac{d}{n}$ in (3.4). This is exactly what Figure 3.2 shows for

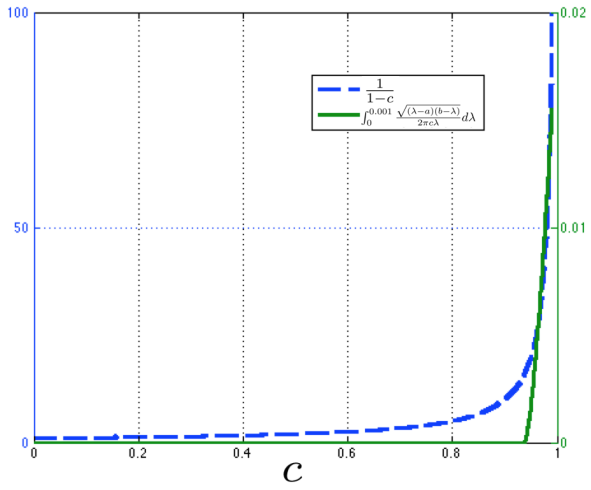



Figure 3.2: Approximation of estimation accuracy of $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}}$ from [50] (in blue) and probability of eigenvalues less than $k = 0.001$ both as function of $c = \frac{d}{n}$ (in green).

$k = 0.001$. It gives the intuition that as $c \to 1$, i.e. $\frac{d}{n}$, there is a significant increase in the probability of having very small eigenvalues and consequently ill-conditioning issues in (3.4). Similarly, the analysis of the estimation accuracy of $\boldsymbol{\Sigma}$ elaborated in [130] and [50], with the same distribution assumption, provides more clues about the relationship between $c$ and the performance of the RX-detector. Furthermore, the precision in the $\boldsymbol{\Sigma}$ estimation by $\widehat{\boldsymbol{\Sigma}}$ can be approximated by $\frac{1}{1-c}$ for large $d$ [50]. This simple expression shows that if $c = \frac{d}{n} = 0.1$, there are more $11\%$ overestimation on average (depending on $d$). Thus, a value less than $c = 0.01$ is needed to achieve $1\%$ estimation error on average.

### 3.2.3 Robust Estimation in Non-Gaussian Cases

Presence of outliers can distort both mean and covariance estimates in computing Mahalanobis distance. This section describes two types of robust estimators for covariance matrix.

#### 3.2.3.1 M-estimators

In a Gaussian distribution, the SCM $\widehat{\mathbf{\Sigma}}$ in (3.3) is the MLE of $\mathbf{\Sigma}$. This can be extended to a larger family of distributions. *Elliptical distributions* is a broad family of probability distributions that generalize the multivariate Gaussian distribution and inherit some of its properties [8, 62]. The $d$-dimension random vector $\mathbf{X}$ has a multivariate elliptical distribution, written as $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \psi)$, if its characteristic function can be expressed as, $\psi_{\mathbf{X}} = \exp(i\mathbf{t}^T\boldsymbol{\mu})\psi\left(\frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\right)$ for some vector $\boldsymbol{\mu}$, positive-definite matrix $\mathbf{\Sigma}$, and for some function $\psi$, which is called the characteristic generator. From $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \psi)$, it does not generally follow that $\mathbf{X}$ has a density $f_{\mathbf{X}}(\mathbf{x})$, but, if it exists, it has the following form:

$$f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}, g_d) = \frac{c_d}{\sqrt{|\mathbf{\Sigma}|}} g_d \left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{3.7}$$

where $c_d$ is the normalization constant and $g_d$ is some non-negative function with $(\frac{d}{2}-1)$-moment finite. Many applications including AD requires a robust estimator for data sets sampled from distributions with heavy tails or outliers. A commonly used robust estimator of covariance is the Maronna's M estimator [107] as follows:

$$\widehat{\mathbf{\Sigma}}_M = \frac{1}{n}\sum_{i=1}^{n} u((\mathbf{x}_i-\boldsymbol{\mu})^\top\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))((\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top, \tag{3.8}$$

where the function $u : (0,\infty) \to [0,\infty)$ determines a whole family of different estimators. In particular, a special case $u(x) = \frac{d}{x}$ is shown to be the most robust estimator of the covariance matrix of an elliptical distribution with form (3.7), in the sense of minimizing the maximum asymptotic variance. This is the *Tyler's method* [159] which is given by

$$\widehat{\mathbf{\Sigma}}_{\text{Tyler}} = \frac{d}{n}\sum_{i=1}^{n}\frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top}{(\mathbf{x}_i-\boldsymbol{\mu})^\top\widehat{\mathbf{\Sigma}}_{\text{Tyler}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}. \tag{3.9}$$

Tyler [159] establishes the conditions for the existence of a solution to (3.9), and shows that the estimator is unique up to a positive scaling factor, i.e., that $\mathbf{\Sigma}$ solves (3.9) if and only if $c\mathbf{\Sigma}$ solves (3.9) for some positive scalar $c > 0$. Another interpretation to (3.9) can be found by considering normalized samples defined as $\{\mathbf{s}_i = \frac{\mathbf{x}_i - \boldsymbol{\mu}}{||\mathbf{x}_i - \boldsymbol{\mu}||}\}_{i=1}^n$. Then, the PDF of $\mathbf{s}$ takes the form [62]:

$$f_{\mathbf{S}}(\mathbf{s}) = \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} \det(\mathbf{\Sigma})^{-\frac{1}{2}} (\mathbf{s}^\top \mathbf{\Sigma}^{-1} \mathbf{s})^{-\frac{d}{2}},$$

and the MLE of $\mathbf{\Sigma}$ can by obtained by minimizing the negative log-likelihood function:

$$\mathcal{L}(\mathbf{\Sigma}) = \frac{d}{2} \sum_{i=1}^{n} \log(\mathbf{s}_i^\top \mathbf{\Sigma}^{-1} \mathbf{s}_i) + \frac{n}{2} \log\det(\mathbf{\Sigma}). \tag{3.10}$$

If the optimal estimator $\widehat{\mathbf{\Sigma}} > 0$ of (3.10) exist, it needs to satisfy the equation (3.9) [62]. When $n > d$, Tyler proposed the following iterative algorithm based on $\{\mathbf{s}_i\}$:

$$\widetilde{\mathbf{\Sigma}}_{k+1} = \frac{d}{n} \sum_{i=1}^{n} \frac{\mathbf{s}_i \mathbf{s}_i^\top}{\mathbf{s}_i^\top \widehat{\mathbf{\Sigma}}_k^{-1} \mathbf{s}_i}, \widehat{\mathbf{\Sigma}}_{k+1} = \frac{\widetilde{\mathbf{\Sigma}}_k}{\mathrm{tr}(\widetilde{\mathbf{\Sigma}}_k)}. \tag{3.11}$$

It can be shown [159] that the iteration process in (3.11) converges and does not depend on the initial setting of $\widehat{\mathbf{\Sigma}}_0$. Accordingly, the initial $\widehat{\mathbf{\Sigma}}_0$ is usually set to be the identity matrix of size $d$. We have denoted the iteration limit $\widehat{\mathbf{\Sigma}}_\infty = \widehat{\mathbf{\Sigma}}_{\text{Tyler}}$. Note that the normalization by the trace in the right side of (3.11) is not mandatory but it is often used in Tyler based estimation to make easier the comparison and analysis of its spectral properties without any decrement in the detection performance. Recently, a similar M-P law to (3.6) for the empirical eigenvalues of (3.11) has been shown in [46, 174].

#### 3.2.3.2 Multivariate $t$-distribution Model

Firstly, we evoke a practical advice to perform AD in real-life HS images from [37]. They have indicated that the quality of the AD can be improved by means of considering the correlation matrix $\mathbf{R}$ instead of the covariance matrix $\mathbf{\Sigma}$, also known as the *R-RX-detector* [51]. However, notice that writing the $j$-th coordinate of the vector $\mathbf{z}$ as $z_{(j)} = \frac{\mathbf{x}_{(j)} - \mu_{(j)}}{\sqrt{\sigma_{(jj)}}}$, we have $\mathbf{z} = (z_1, \ldots, z_d) = \boldsymbol{\sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\sigma} = \text{diag}(\sqrt{\sigma_1, \ldots, \sigma_d})$. Now, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ is zero-mean, and $cov(\mathbf{Z}) = \boldsymbol{\sigma}^{-1/2}\mathbf{\Sigma}\boldsymbol{\sigma}^{-1/2} = \mathbf{R}$, the correlation matrix of $\mathbf{X}$. Thus, the correlation matrix of $\mathbf{X}$ is the covariance matrix of $\mathbf{Z}$, i.e., the standardization ensuring that all the variable in $\mathbf{Z}$ are on the same scale. Additionally, note that [51] gives a characterization of the performance of the R-RX-detection. They conclude that the performance of R-RX depends not only on the dimensionality $d$ and the deviation from the anomaly to the background mean but also on the squared magnitude of the background mean. That is an unfavorable point in the case that $\boldsymbol{\mu}$ needs to be estimated. At this point, we are interested in characterizing the MLE solution of the correlation matrix $\mathbf{R}$ by means of $t$*-distribution*. A $d$-dimensional random vector $\mathbf{x}$ is said to have the $d$-variate $t-$distribution with degrees of freedom $v$, mean vector $\boldsymbol{\mu}$, and correlation matrix $\mathbf{R}$ (and with $\mathbf{\Sigma}$ denoting the corresponding covariance matrix) if its joint PDF is given by:

$$f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}, v) = \frac{\Gamma(\frac{v+d}{2})|\mathbf{R}|^{-1/2}}{(\pi v)^{\frac{d}{2}}\Gamma(\frac{v}{2})\left[1 + \frac{1}{v}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{R}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{\frac{v+d}{2}}},$$

where the degree of freedom parameter $v$ is also referred to as the shape parameter, because the peakedness of (3.12) may be diminished or increased by varying $v$. Note that if $d = 1, \boldsymbol{\mu} = 0$, and $\mathbf{R} = 1$, then (3.12) is the PDF of the *univariate Student's t distribution* with degrees of freedom $v$. The limiting form of (3.12) as $v \to \infty$ is the joint PDF on the $d$-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Hence, (3.12) can be viewed as a

generalization of the multivariate normal distribution. The particular case of (3.12) for $\boldsymbol{\mu} = 0$ and $\mathbf{R} = \mathbf{I}_d$ is a normal density with zero means and covariance matrix $v\mathbf{I}_d$ in the scale parameter $v$. However, the MLE does not have closed form and it should be found through *Expectation Maximization (EM) algorithm* [113] [94]. The EM algorithm takes the following form of iterative updates, using the current estimates of $\boldsymbol{\mu}$ and $\mathbf{R}$ to generate the weights:

$$\widehat{\boldsymbol{\mu}}_{k+1} = \frac{\sum_{i=1}^{n} w_k^i \mathbf{x}_i}{\sum_{i=1}^{n} w_k^i}, \text{ and} \tag{3.12}$$

$$\widehat{\mathbf{R}}_{k+1} = \frac{1}{n}\sum_{i=1}^{n}(w_k^i(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{k+1})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{(k+1)})^\top), \tag{3.13}$$

where $w_{k+1}^i = \frac{v+d}{v+(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k)^\top \mathbf{R}_k^{-1}(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k)}$. For more details, interested readers may refer to [94,116]. In our case with zero mean data, this approach becomes:

$$\widehat{\mathbf{R}}_{k+1} = \frac{v+d}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{v + \mathbf{x}_i^\top \widehat{\mathbf{R}}_k^{-1}\mathbf{x}_i} \tag{3.14}$$

For the case of unknown $v$, [91] showed how to use EM to find the joint MLEs of all parameters $(\boldsymbol{\mu}, \mathbf{R}, v)$. However, our preliminary work [161] shows that the estimation of $v$ does not give any improvement in AD task. Therefore, we limit ourselves to the case of $t$-distribution with known degrees of freedom $v$.

### 3.2.4 Estimators in High Dimensional Space

The SCM $\widehat{\boldsymbol{\Sigma}}$ in (3.3) has the advantages of being simple and unbiased, i.e., its expected value is equal to the covariance matrix. However, as illustrated in Section 3.2.2, in high dimensions the eigenvalues of the SCM are poor estimates for the true eigenvalues. The sample eigenvalues spread over the positive real numbers with the smallest eigenvalues approaching zero the largest to infinity [58, 92]. Consequently, SCM tends to perform poorly for large covariance matrix

estimation problems.

#### 3.2.4.1 Shrinkage Estimator

To overcome this drawback, it is common to regularize the estimator $\widehat{\mathbf{\Sigma}}$ with a highly structured estimator $\mathbf{T}$ via a linear combination $\alpha\widehat{\mathbf{\Sigma}} + (1-\alpha)\mathbf{T}$, where $\alpha \in [0,1]$. This technique is called regularization or *shrinkage*, since $\widehat{\mathbf{\Sigma}}$ is "shrunk" towards the structured estimator. The shrinkage helps to condition the estimator and avoid the problems of ill-conditioning in (3.4). The notion of shrinkage is based on the intuition that a linear combination of an *over-fit* sample covariance with some *under-fit* approximation will lead to an intermediate approximation that is "just-right" [152]. A desired property of shrinkage is to maintain eigenvectors of the original estimator while conditioning on the eigenvalues. This is called *rotationally-invariant estimators* [142]. Typically, $\mathbf{T}$ is set to $\rho\mathbf{I}$, where $\mathbf{I}$ is the identity matrix for some $\rho > 0$ and $\rho$ is set by $\rho = \frac{1}{d}\sum_{i=1}^{d}\sigma_{ii}$. In this case, the same shrinkage intensity is applied to all sample eigenvalue, regardless of their position. To illustrate the eigenvalues behavior after shrinkage, let us consider the case of linear shrinkage intensity equal to 1/4, 1/2 and 3/4. Figure 3.3 illustrates this case. As it was shown in [93], in the case of $\alpha = 1/2$, every sample eigenvalue is moved half-way towards the grand mean of all sample eigenvalues. Similarly, for $\alpha = 1/4$ eigenvalues are moved a quarter towards the mean of all sample eigenvalues. An alternative is the non-rotationally invariant shrinkage method, proposed by Hoffbeck and Landgrebe [74], uses the diagonal matrix $\mathbf{D} = \text{diag}(\widehat{\mathbf{\Sigma}})$ which agrees with the SCM the diagonal entries, but shrinks the off-diagonal entries toward zero:

$$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{diag}} = (1-\alpha)\widehat{\mathbf{\Sigma}} + \alpha\text{diag}(\widehat{\mathbf{\Sigma}}) \tag{3.15}$$

However, in the experiments, we use a normalized version of (3.15), considering the dimension of the data, i.e.,

$$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}} = (1-\alpha)\widehat{\mathbf{\Sigma}} + \alpha \text{Id}(\widehat{\mathbf{\Sigma}}) \tag{3.16}$$

where $\text{Id}(\mathbf{\Sigma}) = \frac{\text{tr}(\widehat{\mathbf{\Sigma}})\mathbf{I}}{d}$. This is sometimes called *ridge* regularization.

#### 3.2.4.2 Regularized Tyler-estimator

Similarly, shrinkage can be applied to other estimators such as the robust estimator in (3.11). The idea was proposed in [3, 40, 165]. Wiesel [165] proposes to compute a robust and well-conditioned estimator of $\mathbf{\Sigma}$ by:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{k+1} &= \frac{d}{n(1+\alpha)} \sum_{i=1}^{n} \frac{(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\top}}{(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \tilde{\mathbf{\Sigma}}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu})} + \frac{\alpha}{1+\alpha} \frac{d\mathbf{T}}{\text{tr}(\widehat{\mathbf{\Sigma}}_k^{-1}\mathbf{T})} \\ \widehat{\mathbf{\Sigma}}_{k+1} &:= \frac{\tilde{\mathbf{\Sigma}}_{k+1}}{\text{tr}(\tilde{\mathbf{\Sigma}}_{k+1})}.\end{aligned} \tag{3.17}$$

This estimator is a trade-off between the intrinsic robustness from M-estimators in (3.11) and the well-conditioning of shrinkage based estimators in section 3.2.4.1. The existence and uniqueness of this approach has been shown in [146].

#### 3.2.4.3 Geodesic Interpolation in Riemannian Manifold

The shrinkage methods discussed so far involve a linear interpolation between a covariance matrix estimator and a structured target matrix. The interpolation can be extended to the Riemannian manifold space since covariance matrix is positive semi-definite [126]. In general, Riemannian manifold are analytical manifolds endowed with a distance measure, which allows the measurement of similarity or dissimilarity (closeness or distance) of points. In this representation, the distance, called *geodesic distance*, is the minimum length of the curvature

path that connects two points [39], and it can be computed by

$$\text{Dist}_{\text{Geo}}(\mathbf{A}, \mathbf{B}) := \sqrt{\text{tr}(\log^2(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}))}. \tag{3.18}$$

This nonlinear interpolation, here called a *geodesic path* from $\mathbf{A}$ to $\mathbf{B}$. A complete analysis of (3.18) and the geodesic path via its representation as ellipsoids have been presented in [18]. We have included a *Geodesic Stein estimation* with the same intuition behind equation (3.16) as follows,

$$\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}} = \text{Geo}_{\alpha}(\widehat{\mathbf{\Sigma}}, \text{Id}(\widehat{\mathbf{\Sigma}})), \tag{3.19}$$

where $\alpha \in [0, 1]$ determines the trade-off between the original estimation $\widehat{\mathbf{\Sigma}}$ and the well-conditioning $\text{Id}(\widehat{\mathbf{\Sigma}})$.

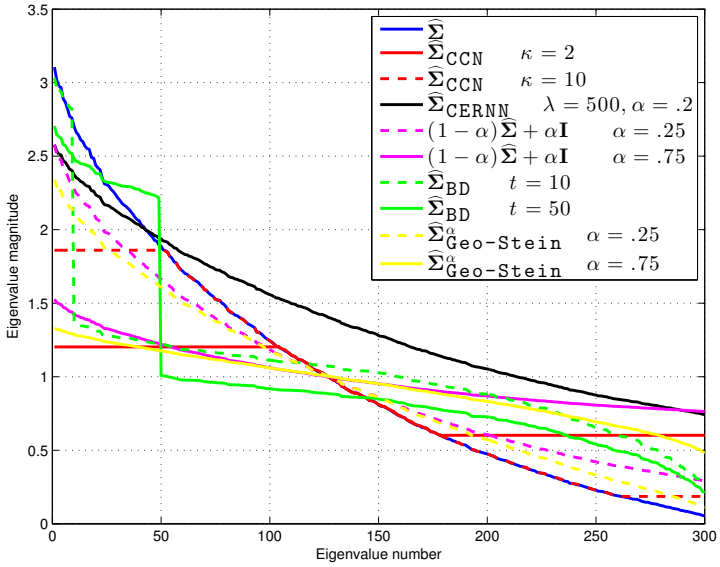



Figure 3.3: CCN truncates extreme sample eigenvalues and leaves the moderate ones unchanged and CERNN gives the contrary effect. Linear and geodesic shrinkages moves eigenvalues towards the grand mean of all sample eigenvalues. However, the effect of geodesic shrinkage is more attenuated for extreme eigenvectors than in linear case. The effect of BD-correction depends on the eigenvalues sets defined by $t$.

#### 3.2.4.4 Constrained MLE

As we have shown in Section 3.2.2, even when $n > d$, the eigenstructure tends to be systematically distorted unless $d/n$ is extremely small, resulting in ill-conditioned estimators for $\mathbf{\Sigma}$. Recently, several works have proposed regularizing the SCM by explicitly imposing a constraint on the condition number. [166] proposes to solve the following constrained MLE problem:

$$\text{maximize } \mathcal{L}(\mathbf{\Sigma}) \text{ subject to } cond(\mathbf{\Sigma}) \leq \kappa \tag{3.20}$$

where $\mathcal{L}(\mathbf{\Sigma})$ stands for the log-likelihood function in the Gaussian distributions. This problem is hard to solve in general. However, Won *et al.* [166] proves that in the case of rotationally-invariant estimators, (3.20) reduces to an unconstrained univariate optimization problem. Furthermore, the solution of (3.20) is a nonlinear function of the sample eigenvalues given by:

$$\widehat{\lambda}_i = \begin{cases} \eta, & \lambda_i(\widehat{\mathbf{\Sigma}}) \leq \eta \\ \lambda_i(\widehat{\mathbf{\Sigma}}), & \eta < \lambda_i(\widehat{\mathbf{\Sigma}}) < \eta\kappa \\ \kappa\eta, & \lambda_i(\widehat{\mathbf{\Sigma}}) \geq \eta\kappa \end{cases} \tag{3.21}$$

for some $\eta$ depending on $\kappa$ and $\lambda(\widehat{\mathbf{\Sigma}})$. We refer this method to as *Condition Number-Constrained* (CCN) estimation.

#### 3.2.4.5 Covariance Estimate Regularized by Nuclear Norms

Instead of constrain the MLE problem in (3.20), Chi and Lange [43] proposes to penalize the MLE as follows,

$$\text{maximize } \mathcal{L}(\mathbf{\Sigma}) + \frac{\lambda}{2}\left[\alpha||\mathbf{\Sigma}||_* + (1-\alpha)||\mathbf{\Sigma}^{-1}||_*\right] \tag{3.22}$$

where $||\mathbf{\Sigma}||_*$ is the nuclear norm of a matrix $\mathbf{\Sigma}$ and is the sum of the eigenvalues of $\mathbf{\Sigma}$, $\lambda$ is a positive strength constant, and $\alpha \in (0, 1)$ is a mixture constant. We refer this approach by the acronym CERNN (Covariance Estimate Regularized by Nuclear Norms).

#### 3.2.4.6 Ben-David and Davidson correction

Given zero-mean[2] data with normal probability density $\mathbf{x} \sim N(0, \mathbf{\Sigma})$, its sampled covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ follows a central Wishart distribution with $n$ degrees of freedom. The study of covariance estimators in Wishart distribution where the sample size $(n)$ is small in comparison to the dimension $(d)$ is also an active research topic [17,112,115]. Firstly, Efron and Morris proposed a rotationally-invariant estimator of $\mathbf{\Sigma}$ by replacing the sampled eigenvalues with an improved estimation [59]. Their approach is supported by the observation that for any Wishart matrix, the sampled eigenvalues tend to be more spread out than population eigenvalues, in consequence, smaller sampled eigenvalues are underestimated and large sampled eigenvalues are overestimated [17]. Accordingly, they find the best estimator of inverse of the covariance matrix of the form $a\widehat{\mathbf{\Sigma}}^{-1} + b\mathbf{I}/\text{tr}(\widehat{\mathbf{\Sigma}})$ by:

$$\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}} = \left((n-d-1)\widehat{\mathbf{\Sigma}}^{-1} + \frac{d(d+1)-2}{\text{tr}(\widehat{\mathbf{\Sigma}})}\mathbf{I}\right)^{-1}. \tag{3.23}$$

It is worth mentioning that other estimations have been developed following the idea behind Wishart modeling and assuming a simple model for the eigenvalue structure in the covariance matrix (usually two phases model). Recently, Ben-David and Davidson [17] have introduced a new approach for covariance estimation in HSI, termed as *BD-correction*. From the SVD of $\widehat{\mathbf{\Sigma}} = \mathbf{U}\Lambda_{\widehat{\mathbf{\Sigma}}}\mathbf{U}^\top$, they proposed a rotationally-invariant estimator by correcting the eigenvalues

[2]Or $\boldsymbol{\mu}$ known, in which case, one might subtract $\boldsymbol{\mu}$ from the data.

by means of two diagonal matrices,

$$\widehat{\mathbf{\Sigma}}_{\text{BD}} = \mathbf{U}\Lambda_{\text{BD}}\mathbf{U}^{\top}, \quad \text{with} \quad \Lambda_{\text{BD}} = \Lambda_{\widehat{\mathbf{\Sigma}}}\Lambda_{\text{Mode}}\Lambda_{\text{Energy}}. \tag{3.24}$$

They first estimate the apparent multiplicity $p_i$ of the $i$-th sample eigenvalue as

$$p_i = \sum_{j=1}^{d} card[a(j) \leq b(i) \leq b(j)],$$

where $a(i) = \Lambda_{\widehat{\mathbf{\Sigma}}}(i)(1-\sqrt{c})^2$ and $b(i) = \Lambda_{\widehat{\mathbf{\Sigma}}}(i)(1+\sqrt{c})^2$. One can interpret the concept of "apparent multiplicity" as the number of distinct eigenvalues that are "close" together [17]. Second, BD-correction conditions the $i$-th sample eigenvalue via its apparent multiplicity $p_i$ as $\Lambda_{\textit{Mode}}(i) = \frac{(1+p_i/n)}{(1-p_i/n)^2}$ and as

$$\Lambda_{\textit{Energy}}(i) = \begin{cases} \sum_{i=1}^{t} \Lambda_{\widehat{\mathbf{\Sigma}}}(i) / \sum_{i=1}^{t} (\Lambda_{\widehat{\mathbf{\Sigma}}}(i)\Lambda_{\text{Mode}}(i)) \\ \sum_{i=t+1}^{d} \Lambda_{\widehat{\mathbf{\Sigma}}}(i) / \sum_{i=t+1}^{d} (\Lambda_{\widehat{\mathbf{\Sigma}}}(i)\Lambda_{\text{Mode}}(i)) \end{cases} \tag{3.25}$$

for a value $t \in [1, \min(n, d)]$ indicating the transition between large and small eigenvalues. Interested readers can refer to [17] for an optimal selection of $t$. A comparison of correction in the eigenvalues by CCN, CERNN, the linear shrinkage in (3.16), the geodesic Stein in (3.19) and the BD-correction is illustrated in Figure 3.3 for three values of regulation parameter. We can see that CCN truncates extreme sample eigenvalues and leaves the moderate ones unchanged, while CCN (3.20) and CERNN (3.22) condition more the larger eigenvalues more and less the smaller eigenvalues.

#### 3.2.4.7 Sparse Matrix Transform

Recently, [31, 152] introduced the *sparse matrix transform* (SMT). The idea behind is the estimation of the SVD from a series of *Givens rotations*, i.e., $\widehat{\mathbf{\Sigma}}_{\text{SMT}} = \mathbf{V}_k \mathbf{\Lambda} \mathbf{V}_k^\top$, where $\mathbf{V}_k = \mathbf{G}_1 \mathbf{G}_2 \cdots \mathbf{G}_k$ is a product of $k$ *Givens rotation* defined by $\mathbf{G} = \mathbf{I} + \mathbf{\Theta}(i, j, \theta)$ where

$$\mathbf{\Theta}(a, b, \theta) = \begin{cases} \cos(\theta) - 1, & \text{if } r = s = a \text{ or } r = s = b \\ \sin(\theta), & \text{if } r = a \text{ and } s = b \\ -\sin(\theta), & \text{if } r = b \text{ and } s = a \\ 0, & \text{otherwise,} \end{cases}$$

where each step $i \in \{1, \ldots, k\}$ of the SMT is designed to find the single Givens rotation that minimize $\text{diag}(\mathbf{V}_i^\top \widehat{\mathbf{\Sigma}} \mathbf{V}_i)$ the most. The details of this transformation are given in [31,32]. The number of rotations $k$ is a parameter and it can be estimated from heuristic Wishart estimator as in [152]. However, in the numerical experiments, this method of estimating $k$ tended to over-estimate. As such, SMT is compared with $k$ as function of $d$ in our experiments. Table 3.2.4.7 summarizes the different covariance matrix estimators considered in the experiments.

| Methods | Notation | Formula |
|---|---|---|
| SCM | $\widehat{\mathbf{\Sigma}}$ | $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$ |
| Stein Shrinkage [92] | $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$ | $(1-\alpha)\widehat{\mathbf{\Sigma}} + \alpha \text{Id}(\widehat{\mathbf{\Sigma}})$ |
| Tyler [159] | $\widehat{\mathbf{\Sigma}}_{\text{Tyler}}$ | $\widehat{\mathbf{\Sigma}}_{j+1} = \frac{d}{n}\sum_{i=1}^{n} \frac{(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top}{(\mathbf{x}_i - \boldsymbol{\mu})^\top \widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}$ |
| Tyler Shrinkage [41] | $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ | $\tilde{\mathbf{\Sigma}}_{k+1} = \frac{1}{1+\alpha}\frac{d}{n}\sum_{i=1}^{n} \frac{\mathbf{x}\mathbf{x}^\top}{\mathbf{x}_i^\top \widehat{\mathbf{\Sigma}}_k^{-1}\mathbf{x}_i} + \frac{\alpha}{1+\alpha}\frac{d\mathbf{T}}{\text{tr}(\widehat{\mathbf{\Sigma}}_k^{-1}\mathbf{T})}$ |
| Sparse Matrix Transform (SMT) [152] | $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ | $\mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_k\mathbf{\Lambda}(\mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_k)^\top$ |
| $t$ distribution [91] | $\widehat{\mathbf{\Sigma}}_{\text{t}}$ | $\widehat{\mathbf{\Sigma}}_{j+1} = \frac{1}{n}\sum_{i=1}^{n} \frac{(v+d)(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top}{v + (\mathbf{x}_i - \boldsymbol{\mu})^\top \widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}$ |
| Geodesic Stein | $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$ | $\text{Geo}_\alpha(\widehat{\mathbf{\Sigma}}, \text{Id}(\widehat{\mathbf{\Sigma}}))$ |
| Constrained condition number [166] | $\widehat{\mathbf{\Sigma}}_{\text{CCN}}$ | (3.21) |
| Covariance Estimate Regularized by Nuclear Norms [43] | $\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$ | (3.22) |
| Efron-Morris Correction [59] | $\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}}$ | (3.23) |
| Ben-Davidson Correction [17] | $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ | (3.24) |

Table 3.1: Covariance matrix estimators considered in this chapter.

## 3.3 Experiments

A series of experiments were conducted to compare the performance of the different methods of covariance matrix estimation, using simulation by considering $\mathbf{\Sigma}$ from some well-known HS images. All the covariance matrix estimators are normalized (trace equal to $d$) to have comparable "size". Note that we are not interested in the joint estimation of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ in this work, then mean vector $\boldsymbol{\mu}$ is assumed known throughout the experiments, i.e. the data matrix is centered by $\boldsymbol{\mu}$.

The experiments were designed to address the following three issues in the covariance estimation methods:

(1) The effect of covariance ill-conditioning due to limited data samples and high dimension ($c$ close to one).

(2) The effect of contamination due to anomalous data being included in the covariance computation.

(3) The effect of non-Gaussian distribution.

### 3.3.1 Performance Measure

There are a few performance measures for anomaly detectors. First, we consider the probability of detection (PD) and the false alarm (FA) rate. This view yields a quantitative evaluation in terms of *Receiver Operating Characteristics* (ROC) curves [61]. A detector is good if it has a PD and a low FA rate, i.e., if the curve is closer to the upper left corner. One may reduce the ROC curve to a single value using the *Area under the ROC curve* (AUC). A detector with a greater AUC is said to be "better" than a detector with a smaller AUC. The AUC value depicts

the general behavior of the detector and characterizes how near it is to perfect detection (AUC equal to one) or to the worst case (AUC equal to $1/2$) [61].

Besides AUC, another measure is to find the one with better data fitting. That is the intuition behind the approach proposed by [151]. It is a proxy that measures the *volume* inside an anomaly detection surface for a large set of false alarm rates. Since in practical applications, the AD is usually calibrated to a given false alarm rate, one can construct a coverage log-volume versus log-false alarm rate to compare detector performances [150, 153]. Accordingly, for a given threshold radius $\eta$, the volume of the ellipsoid contained within $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \leq \eta^2$ is given by

$$\texttt{Volume}(\boldsymbol{\Sigma}, \eta) = \frac{\pi^{d/2}}{\Gamma(1+d/2)} |\boldsymbol{\Sigma}|^{1/2} \eta^d. \tag{3.26}$$

Given an FA rate, a smaller $\texttt{Volume}(\boldsymbol{\Sigma}, \eta)$ indicates a better fit of real structure of the background and thus is preferred. In this chapter, we compute the logarithm of (3.26) to reduce the effect of numerical issues in the computation.

### 3.3.2 Simulations on Elliptical Distribution

We start with experiments in the case of multivariate $t$ distribution in (3.12) with $v$ degrees of freedom. It can be interpreted as generalization of the Gaussian distribution (or conversely, the Gaussian as a special case of the $t$-distribution when the degree of freedom tends to infinity). As $\boldsymbol{\Sigma}$, we have used the covariance matrix of one homogeneous zone in Pavia University HSI (pixels in rows 555 to 590 and columns 195 to 240) [160] in 90 bands (from 11 to 100). $\boldsymbol{\Sigma}$ is normalized to have trace equal to one. Its condition number is large, $2.7616 \times 10^5$. Anomalies in this example are generated by the same distribution but using an identity matrix of trace equal to one as parameter of the distribution. We perform estimations varying three components:

(1) The degrees of freedom of the distribution from where the multivariate sample is generated.

(2) The size of the sample to calculate covariance matrix estimators in Table 3.2.4.7.

(3) The number of anomalies included in the sample to compute covariance matrix estimators in Table 3.2.4.7.

With that in mind, we have generated 4000 random vectors (half of them anomalies) and we have set the parameters in each estimator by minimizing the volume calculated in the threshold corresponding to a false alarm rate of $0.001$. Different volumes by varying parameters in the estimation can be compared in (b,d,f,h,j,l) of Figure 3.4, 3.5 and 3.6 in all the explored cases. In the experiments, the number of rotations in $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ is fixed to $i$ times the dimension $d$, for $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$, the regularization parameter $\alpha$ is $i$, for $\widehat{\mathbf{\Sigma}}_{\text{CCN}}$ the regularization parameter is $2^{i+1}$, in $\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$ the value $\alpha = i/20$, and for $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ the value $t$ is equal to $i + 1$. Different values of $i$ from $1$ to $20$ are shown in x-axis. We highlight that the estimators yield detectors with AUC close to one. Additionally, to compare the general performance from the "best estimation" in each approach, we have plot the coverage log-volume versus log-false alarm rate in (a,c,e,g,i,k) of Figure 3.4, 3.5 and 3.6. The interpretation of these figures can be done in three directions:

- *From left to right*, we provide the evolution of the performance by varying the degrees of freedom $v$. Note, that the limiting form of (3.12) as $v \to \infty$ is the joint pdf of the $d$-variate normal distribution with covariance matrix $\mathbf{\Sigma}$. Hence, we use a large value of degrees of freedom, $v = 1000$, to generate the Gaussian case. In $v = 1$, is the case of multivariate Cauchy distributions. Note that Cauchy distributions look similar to Gaussian distributions. However, they have much heavier tails. Thus, it is a good indicator of how sensitive the estimators are to heavy-tail deviation from normality.

- *From up to down*, we illustrate the effect of the relative value $c = d/n$ in the performance

of the estimation. We have used, in the first row, five times the number of sample than the dimension, i.e. $c = 0.2$, and in the second row, only 100 samples which correspond to a difficult scenario where $c = 0.9$.

- *From Figure 3.4 to Figure 3.6*, we show the consequence of including anomalies in the sample where the estimation is performed. Three cases are considered: Figure 3.4 is a free noise case, Figure 3.5 includes a low level of contamination ($1\%$), and Figure 3.6 shows a level of noisy samples equal to $10\%$.

At this stage, we can have some conclusion about the performance of studied estimators:

- In the more "classical" scenario, i.e., Gaussian distribution, no contamination and much more samples than dimensions ($c = 0.2$ in Figure 3.4), the approaches $\widehat{\mathbf{\Sigma}}_{\text{BD}}$ and $\widehat{\mathbf{\Sigma}}_{\text{Efron-Morris}}$ based on correction of eigenvalues (section 3.2.4.6) performed slightly better than the other alternatives. However, as soon as the sample size was reduced, the data was contaminated or the distribution of data was "less" Gaussian, their performances seemed to be drastically affected.

- In Gaussian cases with contaminated samples and $c = 0.2$, the robust approaches, $\widehat{\mathbf{\Sigma}}_{\text{t}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ performed better than other approaches. However, $\widehat{\mathbf{\Sigma}}_{\text{t}}$ was unquestionably affected by the nocuous decreasing of the sample size in the case of $c = 0.9$, producing detector with huge volumes.

- In the scenario of Gaussian data and $c = 0.9$, $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ did the best job followed by the shrinkage approaches, i.e., $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$, $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ and $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$. Another important point to note is that $\widehat{\mathbf{\Sigma}}_{\text{SMT}}$ was more affected by the contamination in the data than shrinkage-based methods.

- In the case of Cauchy distributions, $\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Tyler}}$ was in general less affected by heavy-tails than other approaches. Additionally, geodesic interpolation ($\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Geo-Stein}}$) clearly outper-

formed linear interpolations ($\widehat{\mathbf{\Sigma}}^{\alpha}_{\text{Stein}}$) in these heavy-tails scenarios. $\widehat{\mathbf{\Sigma}}_{\text{CERNN}}$ and $\widehat{\mathbf{\Sigma}}_{\text{CCN}}$ were robust in this difficult case of heavy tails with contaminated data.

Finally, to summarize, the best three performances according to the coverage log-volume versus log-false alarm curve in each scenario are included in Table 3.3.3. Now, we move forward along the difficulty of the studied problems by including simulations on a more complex scenario.

### 3.3.3 Simulations on Dirichlet Distributions

In HS images, spectral diversity can be considered in a set of proportions, called abundances [88]. In this type of data, called *compositional data* [5], the Dirichlet family of distributions is usually the first candidate employed for modeling the data. The rationale behind this choice is the following [118]:

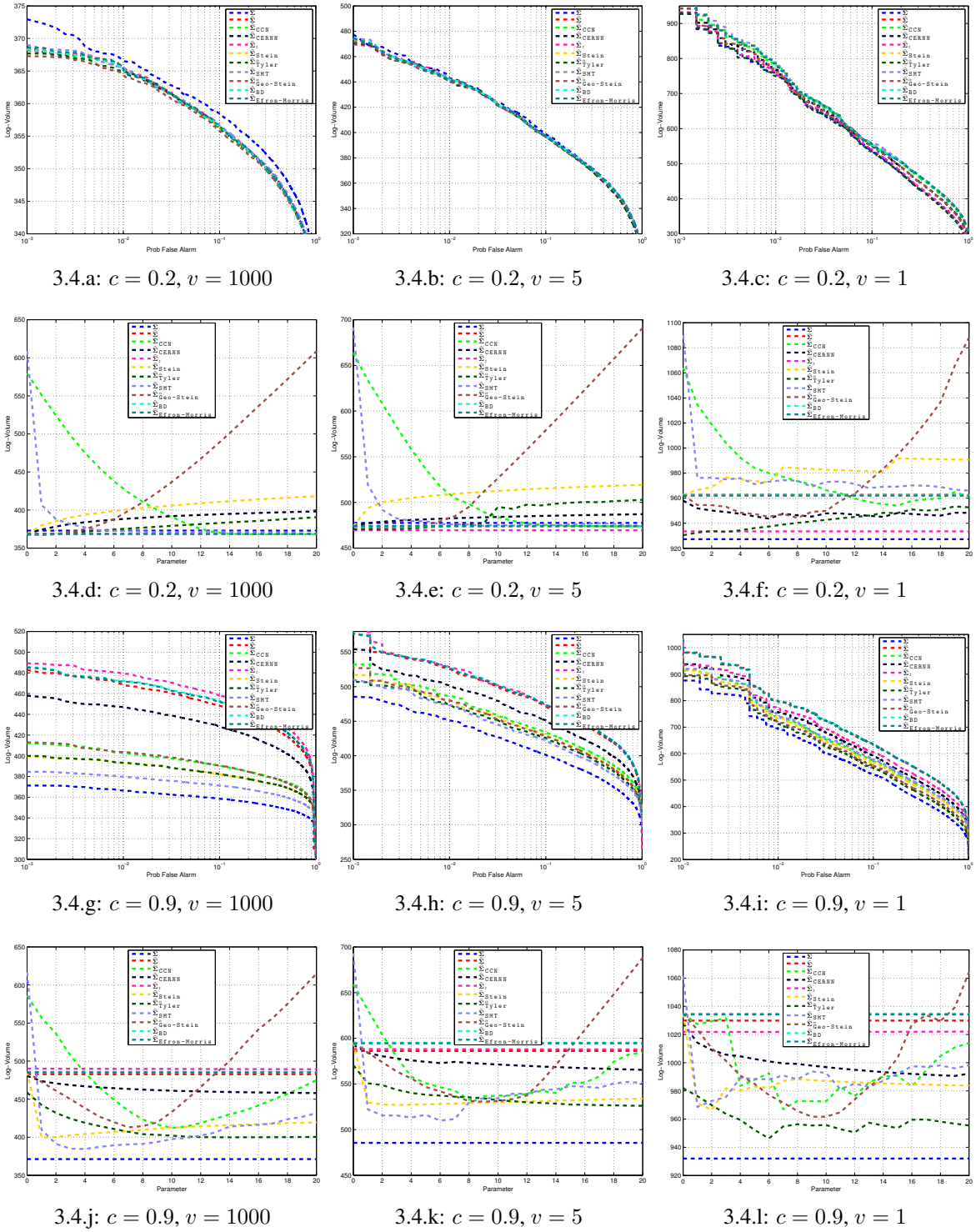
(1) The Dirichlet density automatically enforces the non-negativity and sum-to-one constraint, which is natural in the linear mixture model.

(2) Mixtures of density allow one to model complex distributions in which the mass probability is scattered over a number of clusters inside the simplex [118].

A $d$-dimensional vector $\mathbf{p} = (p_1, p_2, \ldots, p_d)$ is said to have the *Dirichlet distribution* with parameter vector $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_d)$, $\rho_i > 0$, if it has the joint density

$$f(\mathbf{p}) = B(\rho) \prod_{i=1}^{d} p_i^{\rho_i - 1}, \tag{3.27}$$

where $B(\rho) = \frac{\Gamma(\sum_{i=1}^{d} \rho_i)}{\prod_{i=1}^{d} \Gamma(\rho_i)}$, $p_i \geq 0, \sum_{i=1}^{d} p_i = 1$. We write $\mathbf{p} \sim \mathcal{D}(\rho)$.



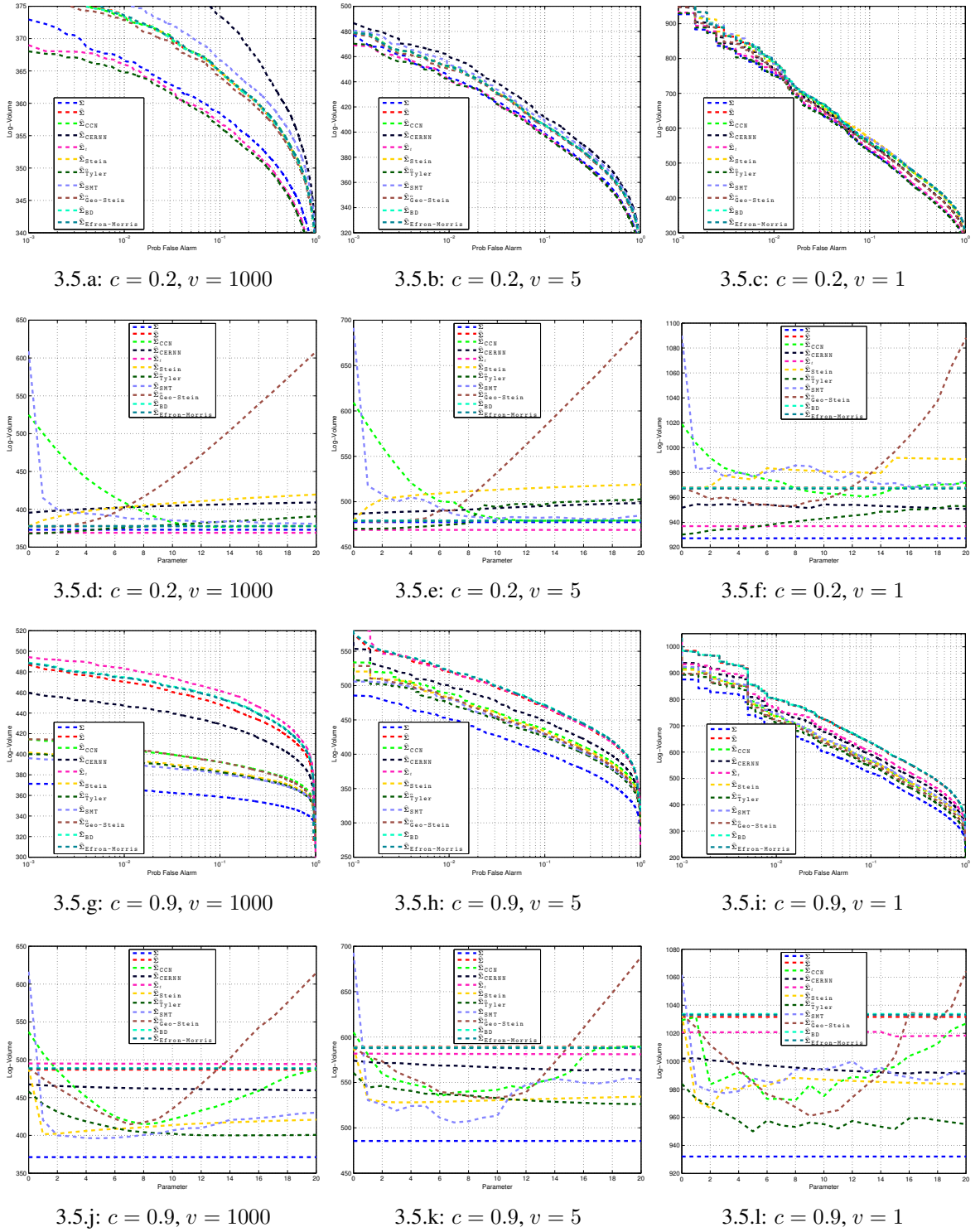

3.4.a: $c = 0.2$, $v = 1000$ 3.4.b: $c = 0.2$, $v = 5$ 3.4.c: $c = 0.2$, $v = 1$

3.4.d: $c = 0.2$, $v = 1000$ 3.4.e: $c = 0.2$, $v = 5$ 3.4.f: $c = 0.2$, $v = 1$

3.4.g: $c = 0.9$, $v = 1000$ 3.4.h: $c = 0.9$, $v = 5$ 3.4.i: $c = 0.9$, $v = 1$

3.4.j: $c = 0.9$, $v = 1000$ 3.4.k: $c = 0.9$, $v = 5$ 3.4.l: $c = 0.9$, $v = 1$

Figure 3.4: Non-contamination case. Performance of covariance matrices with $d = 90$ in different settings from Multivariate Gaussian (large $v$) to Multivariate Cauchy distribution ($v$=1). First row: $n = 450, c = 0.2$. Second row: $n = 100, c = 0.9$. In (d,e,f,j,k,l) volumes are calculated in the threshold corresponding to a false alarm rate of $0.001$.



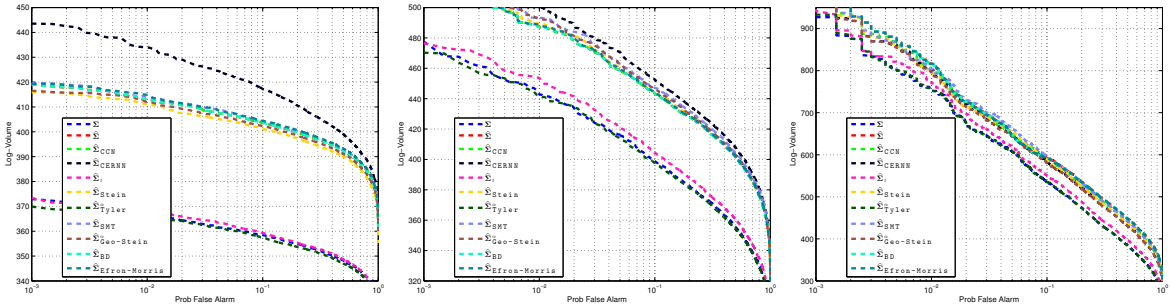

3.5.a: $c = 0.2$, $v = 1000$ 3.5.b: $c = 0.2$, $v = 5$ 3.5.c: $c = 0.2$, $v = 1$

3.5.d: $c = 0.2$, $v = 1000$ 3.5.e: $c = 0.2$, $v = 5$ 3.5.f: $c = 0.2$, $v = 1$

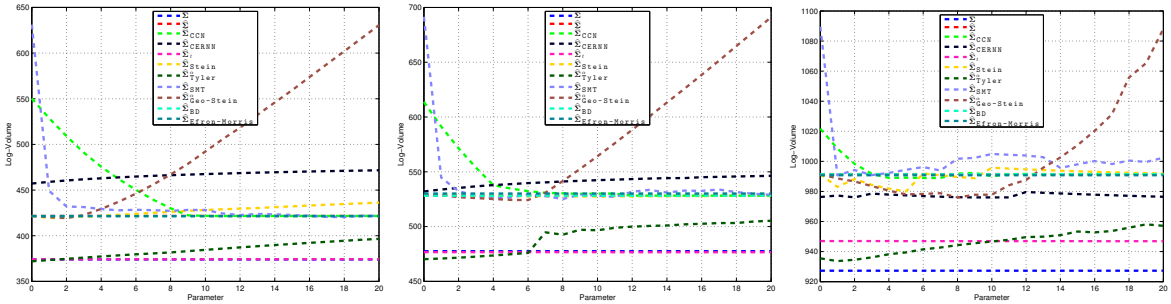
3.5.g: $c = 0.9$, $v = 1000$ 3.5.h: $c = 0.9$, $v = 5$ 3.5.i: $c = 0.9$, $v = 1$

3.5.j: $c = 0.9$, $v = 1000$ 3.5.k: $c = 0.9$, $v = 5$ 3.5.l: $c = 0.9$, $v = 1$

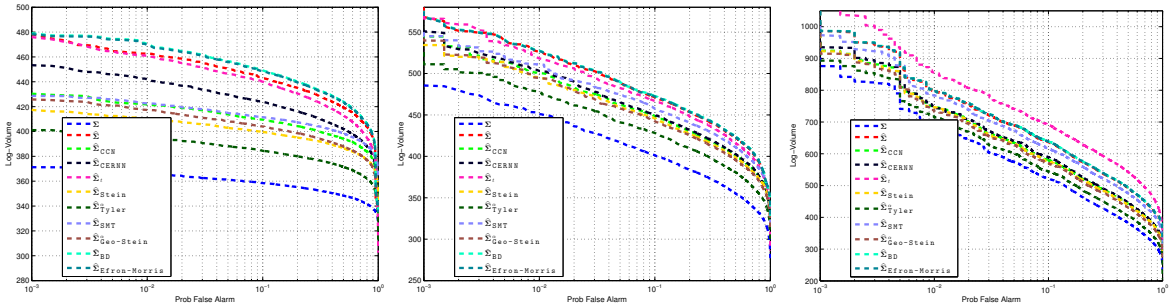
Figure 3.5: Contamination case with $1\%$ outliers. Performance of covariance matrices are estimated considering background vectors in $d = 90$ and $1\%$ of anomalies with different degrees of freedom $v$. First row: $n = 450, c = 0.2$. Second row: $n = 100, c = 0.9$. In (d,e,f,j,k,l) volumes are calculated in the threshold corresponding to a false alarm rate of $0.001$.



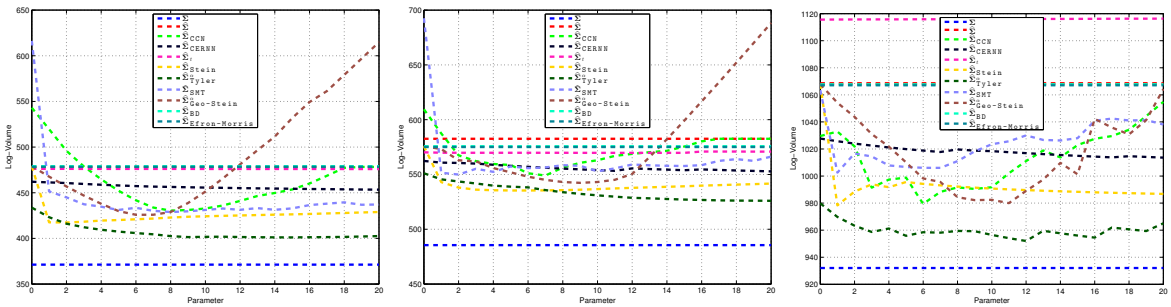

3.6.a: $c = 0.2$, $v = 1000$ 3.6.b: $c = 0.2$, $v = 5$ 3.6.c: $c = 0.2$, $v = 1$




3.6.d: $c = 0.2$, $v = 1000$ 3.6.e: $c = 0.2$, $v = 5$ 3.6.f: $c = 0.2$, $v = 1$




3.6.g: $c = 0.9$, $v = 1000$ 3.6.h: $c = 0.9$, $v = 5$ 3.6.i: $c = 0.9$, $v = 1$



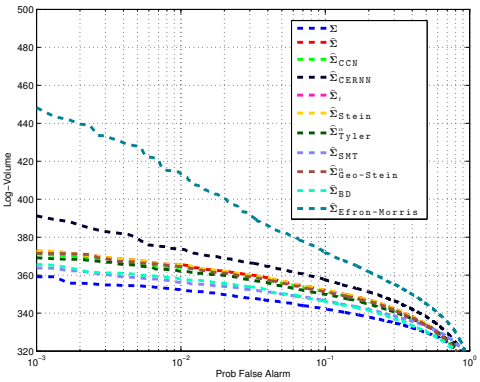

3.6.j: $c = 0.9$, $v = 1000$ 3.6.k: $c = 0.9$, $v = 5$ 3.6.l: $c = 0.9$, $v = 1$

Figure 3.6: Contamination case with $10\%$ outliers. Performance of covariance matrices are estimated considering background vectors in $d = 90$ and $10\%$ of anomalies vs different degrees of freedom $v$. First row: $n = 450, c = 0.2$. Second row: $n = 100, c = 0.9$. In (d,e,f,j,k,l) volumes are calculated in the threshold corresponding to a false alarm rate of $0.001$.

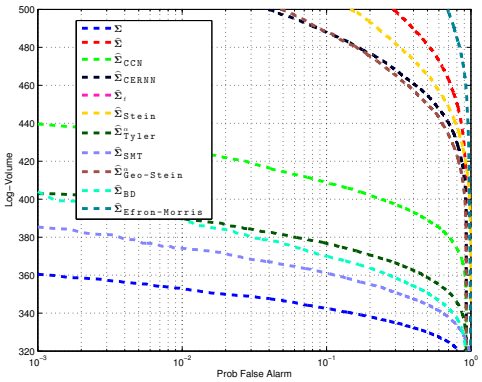
| S/N | Settings | Computation of $\widehat{\Sigma}$ |
|---|---|---|
| 1 | No contamination | $\widehat{\Sigma}$ is estimated based on 970 samples from D1 |
| 2 | No contamination | $\widehat{\Sigma}$ is estimated based on 215 samples from D1 |
| 3 | 10% contamination | $\widehat{\Sigma}$ is estimated based on 970 samples, 90% from D1 and 10% from D2,D3 |
| 4 | 10% contamination | $\widehat{\Sigma}$ is estimated based on 215 samples, 90% from D1 and 10% from D2,D3 |

Table 3.2: Experiment settings based on Dirichlet distributions.

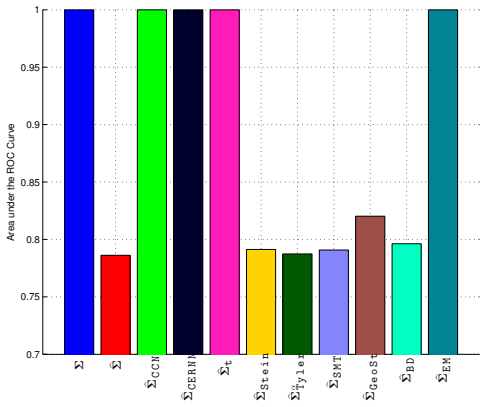
An experiment is carried out by selecting 15 endmembers from a real HS image (World Trade Center) by the popular endmember extraction method *Vertex Component Analysis* (VCA) [117]. subsequently, the samples are generated by generating the abundance matrix (how much endmembers are mixed to form a sample) and adding a random Gaussian noise. Our motivation is to have a realistic low-rank covariance matrix, which appears often in many HS images [105]. In the experiments, we generate 4000 samples from three Dirichlet distributions $\mathcal{D}_1([9,\ldots,9])$ (3000 samples), $\mathcal{D}_2([3,9,1,\ldots,1])$ (500 samples) and $\mathcal{D}_3([1,1,3,9,9,1,\ldots,1,1])$ (500 samples). The background covariance is estimated based on the settings in Table 3.2. Detection results are shown in Figure 3.7. We can see that some techniques fail to correctly detect the anomalies. Among the estimators with AUC close to one, the best performance according to volume is clearly given by the $\widehat{\Sigma}_{\text{SMT}}$. After that, $\widehat{\Sigma}_{\text{BD}}$ and $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ performed better than other approaches. From this point, we would like to analyze the behavior of the estimator in the presence of contaminated samples. Accordingly, we substitute $10\%$ of the sample with vectors from $\mathcal{D}_2$. Thus, the AUC and the volume vs false alarm rate for the studied estimators are illustrated in Figure 3.7 (c) and (e) with $10\%$ and two sample sizes (215 and 970). We can see, that the idea of constrain the estimation of covariance matrix by the condition number provides detectors ($\widehat{\Sigma}_{\text{CCN}}$), which outperformed all the other methods in the particular task of AD for Dirichlet distributions even if we reduce the sample size from 970 to 215. Finally, to



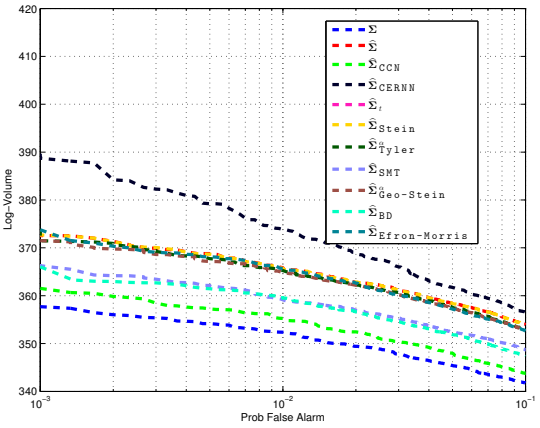

3.7.a: Log-volume versus log-false alarm rate in $n = 970, c = 0.2$ without contamination.



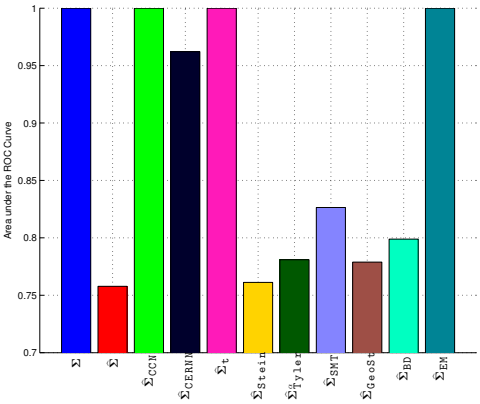

3.7.b: Log-volume versus log-false alarm rate in $n = 215, c = 0.9$ without contamination.



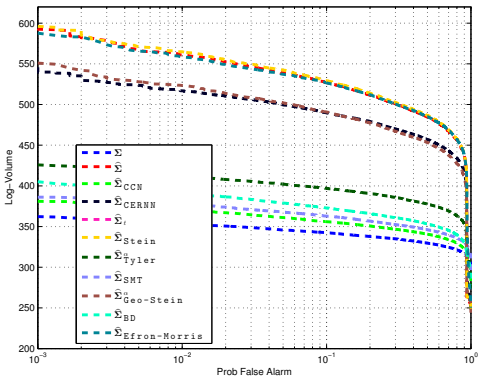

3.7.c: AUC for $n = 970, c = 0.2$ with $10\%$ of contamination.




3.7.d: Log-volume versus log-false alarm rate in experiment of (c)




3.7.e: AUC for $n = 215, c = 0.9$ with $10\%$ of contamination.




3.7.f: Log-volume versus log-false alarm rate in experiment of (e)

Figure 3.7: Dirichlet case. Covariance matrices are estimated considering background vectors in $d = 194$. Parameters in each covariance matrix estimator are set to minimize the volume in the threshold corresponding to a false alarm rate of $0.001$.

| Distribution | Anomalies | $c = \frac{d}{n}$ | Top three performances | | |
|---|---|---|---|---|---|
| Gaussian | No | 0.2 | $\widehat{\Sigma}_{\text{BD}}$ | $\widehat{\Sigma}_{\text{Efron-Morris}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ |
| Gaussian | 1% | 0.2 | $\widehat{\Sigma}_{\text{t}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | — |
| Gaussian | 10% | 0.2 | $\widehat{\Sigma}_{\text{t}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | — |
| Gaussian | No | 0.9 | $\widehat{\Sigma}_{\text{SMT}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Stein}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ |
| Gaussian | 1% | 0.9 | $\widehat{\Sigma}_{\text{SMT}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Stein}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ |
| Gaussian | 10% | 0.9 | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Stein}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ |
| Cauchy | No | 0.2 | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ | $\widehat{\Sigma}_{\text{t}}$ |
| Cauchy | 1% | 0.2 | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ | $\widehat{\Sigma}_{\text{CERNN}}$ |
| Cauchy | 10% | 0.2 | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}_{\text{t}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ |
| Cauchy | No | 0.9 | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}_{\text{CERNN}}$ |
| Cauchy | 1% | 0.9 | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ | $\widehat{\Sigma}_{\text{CERNN}}$ |
| Cauchy | 10% | 0.9 | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$ | $\widehat{\Sigma}_{\text{CCN}}$ |
| Dirichlet | No | 0.2 | $\widehat{\Sigma}_{\text{SMT}}$ | $\widehat{\Sigma}_{\text{BD}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ |
| Dirichlet | 10% | 0.2 | $\widehat{\Sigma}_{\text{CCN}}$ | — | — |
| Dirichlet | No | 0.9 | $\widehat{\Sigma}_{\text{SMT}}$ | $\widehat{\Sigma}_{\text{BD}}$ | $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ |
| Dirichlet | 10% | 0.9 | $\widehat{\Sigma}_{\text{CCN}}$ | — | — |

Table 3.3: Top-3 performances in different analyzed scenarios

summarize, the best performances according to the coverage log-volume versus log-false alarm curve in each scenario have been included in Table 3.3.3 to make easier the comparison with the result of previous sections.

## 3.4 Conclusion

This chapter presents a case study on high dimensional data learning based on anomaly detection in HS image processing. We investigated the performance of covariance matrix estimators in the AD problem in three scenarios: different ratios between sample size and dimension, outlier contamination in the estimation, and non-Gaussian distributions of data samples (Cauchy and linear mixing model from Dirichlet distribution).

In the simple Gaussian case without outlier contamination and a large sample size, the Ben-

Davidson Correction [17] method outperformed the other alternatives. However, its performance decreased when the samples contained a few outliers or there are insufficient the samples. With a few outlier samples, the $t$ distribution [91] method $\widehat{\Sigma}_{\text{t}}$ could obtain satisfactory performance, but its performance decreased with a small sample size. Additionally, Geodesic interpolations ($\widehat{\Sigma}^{\alpha}_{\text{Geo-Stein}}$) performed better than linear interpolations ($\widehat{\Sigma}^{\alpha}_{\text{Stein}}$) in most of the cases, especially in heavy-tails distributions. Overall, $\widehat{\Sigma}^{\alpha}_{\text{Tyler}}$ and $\widehat{\Sigma}_{\text{SMT}}$ showed the best performance in most of the explored cases. However, note that $\widehat{\Sigma}_{\text{SMT}}$ was more affected by the contamination than shrinkage-based methods. In contrast, $\widehat{\Sigma}_{\text{SMT}}$ could adapt better to the data samples generated from linear mixture models. Finally, the recent approach by constraining the condition number ($\widehat{\Sigma}_{\text{CCN}}$) performed exceptionally well in the difficult case of heavy tails distributions with contaminated data, in addition to all the explored cases in Dirichlet simulations. Without any mitigation, the sample covariance estimation performed poorly in the experiments.

This extensive comparative review of covariance estimation methods also reinforces the need to deal with the challenges of high dimensionality efficiently. In the next few chapters, we propose some efficient methods in dealing with these challenges.

# Chapter 4

# A Unified Feature Selection Framework for Graph Embedding on High Dimensional Data

In the previous chapters, we discuss challenges of high dimensional visual data, and some current state-of-the-art methods in handling high dimensionality. A case study on hyperspectral data demonstrates the challenges of high dimensionality and benefits of the handling of high dimensionality. This chapter proposes a unified framework for graph embedding on high dimensional data, providing an abstraction of a generalized class of machine learning methods for image understanding.

## 4.1 Introduction to Graph Embedding on High Dimensional Data

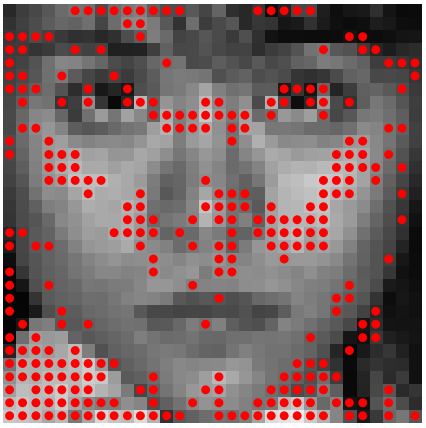
High dimensional data is ubiquitous in many real world applications. However, directly learning a classifier on high dimensional data may significantly degrade the performance of many applications, especially when data features are highly correlated and the sample size is relatively small. This is commonly referred to as the *curse of dimensionality* [16]. To alleviate

this, one possible approach is to transform high dimensional data into a lower dimensional representation while preserving the intrinsic data structure. This is known as dimensionality reduction.

Graph embedding has been shown to be a powerful tool for dimensionality reduction [70, 169]. In particular, some popular dimensionality reduction methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) [110], ISOMAP [149], Locally Linear Embedding (LLE) [135], and Locality Preserving Projections (LPP) [70] can be formulated into graph embedding methods [169]. By employing full dimensional features for learning tasks, the graph embedding methods try to learn a low dimensional projection, preserving some intrinsic data structures. Intrinsic data structures can have both local and global properties, depending on the applications. Local properties often refer to the local neighborhood relationship such as in LPP, while examples of global properties include class separation in LDA, the global variance in PCA, and the global shortest path between any pairs of data samples in the ISOMAP method. Graph embedding of high dimensional data suffers from two weaknesses. First, it is hard to interpret the resultant features when using all dimensions for projection. Second, original data inevitably contain noisy feature measurements. Simply incorporating these noisy features could make graph embedding not reliable and noisy [81, 163]; therefore, it is important to employ only significant features for graph embedding.

Both the graph embedding and feature selection methods define their own paradigms to preserve data intrinsic structures. In existing work in the literature, these two tasks are mostly done independently or mutually exclusively. This work instead proposes a novel paradigm to unify these two schemes by performing feature selection and graph embedding simultaneously. As an application illustrated in Figure 4.1.b, the proposed paradigm is applied to the LPP graph embedding of some face images. The proposed method can automatically choose some important pixels that capture the variations in illumination conditions, poses, and facial expressions.

4.1.a: Sample face images with 1024 ($32 \times 32$) pixels.



4.1.b: Overlaying a face sample with 300 features using the proposed method.

Figure 4.1: Important pixels that capture variations in the face images under different illumination conditions, poses, and facial expressions.

The main contributions of this work are summarized as follows.

- By exploiting the least squares formulation of graph embedding, we introduce a binary feature selector to directly constrain the desired number of features. We further reformulate the resultant problem as a convex semi-infinite programming (SIP) problem. This novel feature selection scheme can be applied to unsupervised, supervised, and semi-supervised learning tasks in preserving the corresponding intrinsic data structures via low dimensional embedding.

- By exploiting the observation that only a few constraints are active in the resultant SIP problem, we proposed an efficient cutting plane method, which essentially conducts a sequence of accelerated proximal gradients only on a set of features. Therefore, a major advantage of the proposed method is its ability to handle ultrahigh dimensions efficiently due to its low computation cost and memory requirements. Moreover, the proposed method is guaranteed to converge globally.

- The proposed method addresses the learning in a holistic way, resulting in both gener-

alized graph embedding and the desired cardinality of the features. A wide range of data sets have been tested in the experiments to verify the effectiveness of the proposed framework for unsupervised, supervised, and semi-supervised learning tasks.

The organization of the rest of the chapter is as follows. In Section 4.2, we briefly review some graph embedding methods for dimensionality reduction and introduce the least squares formulation of graph embedding. Section 4.3 details the proposed approach. In Section 4.4, we conduct some experiments to compare our results with the current state-of-the-art algorithms. Section 4.5 concludes this work.

## 4.2 Related Work and Preliminaries

In this section, we first briefly review the recent literature on graph embedding and feature selection. An overview of the generalized subspace problem is also provided since this is a foundation of our method.

### 4.2.1 Graph Embedding for Dimensionality Reduction

Coherent structures in high dimensional data, such as neighboring pixels in images, naturally induce a high correlation among dimensions. To alleviate the curse of dimensionality, scientists have proposed to transform data into a low dimensional manifold via graph embedding [69, 169]. The number of data samples, data dimensionality, and the number of classes are denoted by $n, d$, and $k$, respectively. $X \in \mathbb{R}^{d \times n}$ is a zero-mean data matrix, and $Y = \{y_1, y_2, ...y_n\} \in \mathbb{R}^{k \times n}$ if available represents class label information. Furthermore, the symmetric positive semi-definite matrix, $S \in \mathbb{R}^{n \times n}$, encodes the desired data properties. Graph embedding for a class of dimensionality reduction methods aims to find the projection vector $\omega$ for the following

| S/N | Methods | $S$ |
|---|---|---|
| 1 | **PCA** | $X^\top X$ |
| 2 | **CCA** | $Y^\top (YY^\top)^{-1} Y$ |
| 3 | **LDA** | $Y^\top (YY^\top)^{-1} Y$ |
| 4 | **LPP** | $D^{-1/2} A D^{-1/2}$ |
| 5 | **HSL** | $I - L$ |
| 6 | **SDA** | $(I + \beta D)^{-1/2}(W_b + \beta A)(I + \beta D)^{-1/2}$ |

Table 4.1: $S$ computations for common generalized eigen problems. $X$ is a data matrix, $Y$ is a regression response matrix in CCA and class label matrix in LDA and Partial Least Squares (PLS), $A$ is an affinity matrix, $D$ is a diagonal matrix whose diagonal entries are the row sum of $A$, $L$ is a Laplacian matrix, and $\beta$ is a regularization constant.

generalized eigenvalue problem [145]:

$$XSX^\top \omega = \lambda XX^\top \omega. \tag{4.1}$$

Many dimensionality reduction methods such as PCA, LDA, CCA, LPP, and Hypergraph Spectral Learning (HSL) can be formulated into the above graph embedding framework [169]. The definitions of $S$ for the above mentioned methods are tabulated in Table 4.1. More details can be found in [145] and references therein.

Both PCA and LPP are unsupervised as they do not consider class labels. They are often used for general pre-processing, clustering, or visualization. For classification, supervised graph embedding, such as LDA, generally can achieve better performance since class label information is considered. The graph embedding can be readily extended to the semi-supervised setting, which utilizes large unlabelled data sets and small labelled data sets to model intrinsic data structures [15, 28, 100, 177]. To achieve this, a possible model is a weighted graph whose vertices are both labelled and unlabelled samples and edges reflect samples similarity. For ex-

ample, the semi-supervised discriminative analysis (SDA) method [29] builds upon the LDA and LPP graph embedding.

### 4.2.2 The Least Squares Formulation for Graph Embedding

Solving the generalized eigenvalue problem in (4.1) is very expensive for large-scale and high dimensional problems. To reduce the computational burden for large-scale problems, Sun *et al.* [145] formulates it into a least squares problem, as in (4.3). Specifically, since $S \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix, it can be decomposed as $S = HH^\top$, where $H \in \mathbb{R}^{n\times r}$ and $r \leq n$ is the number of significant singular values of $S$. Furthermore, let $HP = QR$ be the *QR-decomposition* of $H$ with the permutation matrix $P$, where $Q \in \mathbb{R}^{n\times r}$ is a matrix with $r$ orthonormal columns. Moreover, let $R = U_R\Sigma_R V_R^\top$ be the compact singular value decomposition (SVD) of $R$, and the regression response $T \in \mathbb{R}^{n\times r}$ is computed as follows:

$$T = QU_R. \tag{4.2}$$

Then the generalized eigenvalue problem can be cast as the following least squares regression problem,

$$\min_W \ \|X^\top W - T\|_F^2, \tag{4.3}$$

where $X \in \mathbb{R}^{d\times n}$ and $W \in \mathbb{R}^{d\times r}$ are data matrix and weight matrix, respectively. In practice, to improve the robustness to noise and avoid overfitting, a regularization term could be added as follows:

$$\min_W \ \|X^\top W - T\|_F^2 + \gamma\|W\|_F^2. \tag{4.4}$$

## 4.3 General Framework for Feature Selection

After transforming the generalized eigenvalue problem into a regression problem, the formulation (4.4) can benefit from many existing efficient least squares solvers. However, the regularizer $\gamma\|W\|_F^2$ may not produce sparse solutions. In other words, it cannot achieve the feature selection task. To induce sparsity, we introduce a binary vector $p \in \{0,1\}^d$, whose entries are $1$ for the selected features and $0$ otherwise. To select $m$ desired features, exactly $m$ entries in $p$ will be set to 1, where $m \ll d$. Let $\mathbb{P} = \{p : p \in \{0,1\}^d,\ \ p^\top \mathbf{1} = m\}$ be the domain of $p$, and $\mathbf{1} \in \mathbb{R}^d$ denotes a vector with all entries equal to 1. The proposed least squares formulation for graph embedding based feature selection is cast as the following optimization problem,

$$\begin{aligned} \min_{p\in\mathbb{P}} \min_{W} \ & \frac{1}{2}\|\Xi\|_F^2 + \frac{\gamma}{2}\|W\|_F^2 \\ s.t. \ & \Xi = X^\top \text{diag}(p) W - T, \end{aligned} \tag{4.5}$$

where $T \in \mathbb{R}^{n\times r}$ is the response matrix, $\text{diag}(p)$ is the matrix whose diagonal is the feature selector vector $p$, and $\Xi \in \mathbb{R}^{n\times r}$ denotes the residual matrix.

The proposed formulation has several advantages over the conventional approaches, which impose sparsity directly on $W$ like the sparsity objective $\sum_i \|W_i\|_1$ in the sparse PCA (SPCA) method [178]. First, it selects features naturally with the desired cardinality. This is much more efficient than the sparsity induced methods, in which a regularizer constant controls cardinality. Second, the proposed model can be transformed to a convex programming problem [148], based on which an efficient solver can be developed. The similar schemes used in [148] and [65] are designed for the Fisher score method and classification method, respectively. These two methods can be seen as special cases of the proposed framework in (4.5).

In general, the problem in (4.5) is NP-hard to solve due to the combinatorial integral constraints

on $p$. To address it, it is necessary to make some transformations and relaxations. It is not difficult to verify that the inner minimization problem with a fixed $p$ can be solved equivalently in its dual. By introducing $V \in \mathbb{R}^{n\times r}$, the dual variable, to the constraint $\Xi = X^\top \text{diag}(p)W - T$, we can solve the inner regression problem in its dual. Specifically, the Lagrangian function of the inner regression problem is

$$\mathcal{L}(W, \Xi, V) = \frac{1}{2}\|\Xi\|_F^2 + \frac{\gamma}{2}\|W\|_F^2 + \langle V,\ \Xi - X^\top \text{diag}(p)W + T\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. By setting the first derivatives of $\mathcal{L}(W, \Xi, V)$ w.r.t. $W$ and $\Xi$ to zero, we can obtain the Karush-Kuhn-Tucker (KKT) conditions, namely, $\gamma W = \text{diag}(p)XV$ and $V = -\Xi$. By substituting these results into the Lagrangian function, the problem in (4.5) can be transformed into the following dual formulation:

$$\min_{p\in\mathbb{P}} \max_{V\in\mathbb{R}^{n\times k}} f(V, p), \tag{4.6}$$

where

$$f(V, p) = tr(V^\top T) - \frac{1}{2} tr\left(V^\top (\frac{1}{\gamma} X^\top \text{diag}(p) X + I)V\right).$$

However, this problem is still a non-convex problem since the main optimization variable $p$ is in discrete values. Following the convex relaxation in [148], we have

$$\min_{p\in\mathbb{P}} \max_{V\in\mathbb{R}^{n\times k}} f(V, p) \geq \max_{V\in\mathbb{R}^{n\times k}} \min_{p\in\mathbb{P}} f(V, p).$$

Moreover, this convex relaxed problem can be further transformed into a convex QCQP prob-

lem by introducing an additional variable $\theta \in \mathbb{R}$,

$$\begin{aligned} \max_{V \in \mathbb{R}^{n \times k}, \theta \in \mathbb{R}} \quad & \theta \\ s.t. \;\; & \theta \leq f(V, p), \;\; p \in \mathbb{P}. \end{aligned} \tag{4.7}$$

Note that the constraint domain $\mathbb{P}$ contains a combinatorial number of $p$'s, making the optimization problem intractable even for small-sized $p$ and $m$.

### 4.3.1 Sparse Graph Embedding for Feature Selection

The optimization problem in (4.7) has a combinatorial number of constraints. However, only a few of them are active. Exploiting this observation, we adopt the cutting plane algorithm to solve the QCQP problem (4.7). The cutting plane algorithm [87, 114] iteratively finds the most active constraint and adds it to the active constraint set $\Pi$, which is initialized to an empty set $\varnothing$. The active constraint set $\Pi$ is always a subset of $\mathbb{P}$, i.e., $\Pi \subseteq \mathbb{P}$. Given the updated set $\Pi$, we solve the following subproblem,

$$\begin{aligned} \max_{V \in \mathbb{R}^{n \times k}, \theta \in \mathbb{R}} \quad & \theta \\ s.t. \;\; & \theta \leq f(V, p^t), \;\; \forall p^t \in \Pi. \end{aligned} \tag{4.8}$$

We term our proposed procedure Sparse Graph Embedding (or **SparGE** for short), described in Algorithm 1.

**Algorithm 1** Sparse Graph Embedding for Feature Selection

Input: data $X \in \mathbb{R}^{d \times n}$, a positive semi-definite matrix $S$, the desired feature cardinality $m$.

(1) Initialize $\Pi = \varnothing$, and compute $T$ according to (4.2). Assign $t := 1$.

(2) Iterate the following two steps until convergence.

(a) Update $V$ by solving the subproblem in (4.8).

(b) Find the most active constraint, which is indicated by $p^t$, by solving $p^t = \operatorname{argmax}_p f(V, p)$, based on $V$. Update $\Pi$ by $\Pi := \Pi \cup \{p^t\}$ and $t$ by $t := t + 1$.

Output: $\Pi = \{p^1, p^2, ..., p^k\}$, with each $p^i$ indexing the selected features.

Given $V$, Step 2b) of Algorithm 1 requires us solving

$$p^t = \operatorname*{argmax}_p f(V, p) = \operatorname*{argmax}_p \|\text{diag}(p)XV\|_F^2$$

in order to find the most active constraint of problem (4.7). Let $A = XV \in \mathbb{R}^{d \times r}$, and define $s_i = \sum_{j=1}^{r} (A_{i,j})^2$. The optimization problem becomes:

$$\operatorname*{argmax}_p \|\text{diag}(p)XV\|_F^2 = \operatorname*{argmax}_p \sum_{i=1}^{d} s_i p_i. \tag{4.9}$$

Apparently, problem (4.9) can be solved readily by sorting $s$ and then setting its top $m$ values in $s$ to 1 and the rest to 0. In other words, the most active constraint can be identified by choosing the features with the $m$ highest values in $s$. The algorithm for the most active constraint analysis is summarized in Algorithm 2. The most active constraint $p^t$ obtained is then added to the active constraint set $\Pi := \Pi \cup \{p^t\}$.

**Algorithm 2** The Most Active Constraint Selection

Input: data $X \in \mathbb{R}^{d\times n}$, dual variable $V$, the desired number of features $m$, and the selection vector $p$.

(1) Set all the entries of $p$ to 0.

(2) Compute $s_i = \sum_{j=1}^{k}(A_{i,j})^2$, $\forall i = 1, ..., d$.

(3) Sort $s$ in descending order.

(4) Set $m$ entries of $p$ w.r.t. the top $m$ values of $s$.

Output: $p$ which defines the most active constraint.

### 4.3.2 The Subproblem Optimization

After updating the active constraint set $\Pi$, we then solve the subproblem in (4.8) with reduced constraints as defined by $\Pi$. Since the number of constraints in $\Pi$ is no longer large, this problem is readily solved by a sub-gradient method, such as simpleMKL [65, 148]. However, solving this problem w.r.t. the dual variables $V$ can be very expensive, in particular when $n$ is very large.

Assume there are $\kappa$ active constraints in $\Pi$, i.e., $\kappa = |\Pi|$. Even though there are a large number of features in $X$, at most $m\kappa$ features are chosen by $\Pi \subseteq \mathbb{P}$, where $m\kappa \ll d$. Based on this observation, the subproblem in (4.8) might be solved more efficiently w.r.t. the primal variables $W$. To be more specific, following [11], we have the following proposition.

**Proposition 1.** *The subproblem in (4.8) can be equivalently addressed in the following primal form:*

$$\min_{\Omega} \ \frac{\gamma}{2}(\sum_{t=1}^{C}\|\Omega^t\|_F)^2 + \frac{1}{2}\|\Xi\|_F^2, \tag{4.10}$$

*where* $\Xi = T - \sum_{t=1}^{\kappa} X^\top diag(p^t)\Omega^t$ *denotes the regression residual matrix and* $\Omega^t$ *denotes the*

*weight matrix defined on the features indicated by $p^t$. Moreover, the dual variable $V$ in (4.8) can be recovered by $V = \Xi$, which is required for the most active constraint selection.*

The proof of this proposition is included in Appendix A. Problem (4.10) is a non-smooth problem due to the regularization term $\frac{\gamma}{2}(\sum_{t=1}^{C} \|\Omega^t\|_F)^2$. However, there are at most $m\kappa$ (where $m\kappa \ll d$) features involved in this problem, making it easier to be solved.[1] For convenience, we define $\Omega = [\Omega^1, \Omega^2, ..., \Omega^\kappa] \in \mathbb{R}^{d\times\kappa r}$ by stacking $\Omega^i \in \mathbb{R}^{d\times r}$. Let $P(\Omega) = \frac{\gamma}{2}(\sum_{t=1}^{\kappa} \|\Omega^t\|_F)^2$ and $f(\Omega) = \frac{1}{2}\|\Xi\|_F^2$. Following [147], we propose to solve the primal problem using the accelerated proximal gradient method (APG), which iteratively minimizes the following quadratic approximation of (4.10):

$$\begin{aligned} Q(\Omega, \Omega_t) &= f(\Omega_t) + < \nabla f, \Omega - \Omega_t > + \frac{\tau}{2}\|\Omega - \Omega_t\|_F^2 + P(\Omega) \\ &= \frac{\tau}{2}\|\Omega - G\|_F^2 + P(\Omega) + f(\Omega_t) - \frac{1}{2\tau}\|\nabla f\|_F^2, \end{aligned} \tag{4.11}$$

where $\nabla f$ denotes the gradient of $f$ at point $\Omega_t$, $\tau > 0$ denotes the Lipschitz constant of $f(\Omega)$, and $G = \Omega_t - \frac{1}{\tau}\nabla f = [G^1, G^2, ..., G^\kappa] \in \mathbb{R}^{d\times\kappa r}$ w.r.t. $\Omega = [\Omega^1, \Omega^2, ..., \Omega^\kappa]$. Note that $f(\Omega_t) - \frac{1}{2\tau}\|\nabla f\|_F^2$ is constant w.r.t. $\Omega$, and thus we just need to solve the following projection problem:

$$\min_{\Omega} \frac{\tau}{2}\|\Omega - G\|_F^2 + P(\Omega). \tag{4.12}$$

This problem has a unique global closed-form solution, which can be calculated as follows via Moreau Projection [124].

**Proposition 2.** *Suppose the optimal solution to problem (4.12) is*

$$S_\tau(G) = [S_\tau(G^1), S_\tau(G^2), ..., S_\tau(G^\kappa)] \in \mathbb{R}^{d\times\kappa r}$$

[1] In practice, the optimization is conducted on those selected features only.

*and* $o = [o_1, o_2, ..., o_\kappa]' \in \mathbb{R}^\kappa$ *is an intermediate variable. Then* $S_\tau(G)$ *is unique and its* $t$*-th component,* $S_\tau(G^t)$*, can be calculated as follows:*

$$S_\tau(G^t) = \begin{cases} \frac{o_t}{||G^t||_F} G^t, & \textit{if } o_t > 0. \\ \\ \mathbf{0}, & \textit{otherwise}. \end{cases}, \tag{4.13}$$

*where* $t \in \{1, 2, ..., \kappa\}$*. The intermediate vector* $o_t$ *can be calculated via a soft-threshold operator soft*$(u, \varsigma)$ *[49, 124]:*

$$o_t = [\textit{soft}(u, \varsigma)]_t = \begin{cases} u_t - \varsigma, & \textit{if} \quad u_t > \varsigma, \\ 0, & \textit{Otherwise}, \end{cases} \tag{4.14}$$

*where the threshold value* $\varsigma$ *can be calculated as in Step 4 of Algorithm 3.*

**Algorithm 3** Moreau Projection $S_\tau(G)$

Given input $G = [G^1, G^2, ..., G^\kappa]$ and $s = \frac{1}{\tau}$.

1: Calculate $\widehat{u}_t = \|G^t\|_F$ for all $t = 1, ..., \kappa$.

2: Sort $\widehat{u}$ to obtain $u$ such that $u_{(1)} \geq ... \geq u_{(\kappa)}$.

3: Find $\rho = \max \left\{ t | u_t - \frac{s}{1+ts} \sum_{i=1}^{t} u_i > 0, t = 1, ..., \kappa \right\}$.

4: Calculate the threshold value $\varsigma = \frac{s}{1+\rho s} \sum_{i=1}^{\rho} u_i$.

5: Compute $o = soft(\widehat{u}, \varsigma)$.

6: Compute and output $S_\tau(G)$.

The overall APG algorithm for solving problem (4.10) is summarized in Algorithm 4, where $F(\Omega) = \frac{\gamma}{2}(\sum_{t=1}^{C} \|\Omega^t\|_F)^2 + \frac{1}{2}\|\Xi\|_F^2$. Interested readers can find more details and the convergence derivation of Algorithm 1 and Algorithm 4 in [147].

**Algorithm 4** Accelerated Proximal Gradient for Solving Problem (4.10)

Initialization: Initialize the Lipschitz constant $L_t = L_{t-1}$ and set $\Omega^{-1} = \Omega^0$ by warm start, $\tau_0 = L_t$, $\eta \in (0, 1)$, parameter $\varrho^{-1} = \varrho^0 = 1$, and $k = 0$.

1: Set $V^k = \Omega^k + \frac{\varrho^{k-1}-1}{\varrho^k}(\Omega^k - \Omega^{k-1})$.

2: Set $\tau = \eta\tau_k$.

Repeat

Set $G = V^k - \frac{1}{\tau}\nabla f(V^k)$, compute $S_\tau(G)$.

If $F(S_\tau(G)) \leq Q(S_\tau(G), V^k)$,

set $\tau_k = \tau$, stop, break;

else

$\tau = \min\{\eta^{-1}\tau, L_t\}$.

End

Until convergence $F(S_\tau(G)) \leq Q(S_\tau(G), V^k)$

3: Set $\Omega^{k+1} = S_{\tau_k}(G)$.

4: Let $\varrho^{k+1} = \frac{1+\sqrt{(1+4(\varrho^k)^2)}}{2}$. Let $k = k + 1$.

5: Quit if the stopping condition is achieved. Otherwise, go to step 1.

6: Let $L_t = \eta^2\tau_k$ and return.

### 4.3.3 Handling High Dimensional Sparse Problems

Given an ultrahigh dimensional sparse data matrix, removing the data mean (zero-centering) could make the matrix very dense. The data matrix $\begin{pmatrix} X \\ \mathbf{1}_{1\times n} \end{pmatrix}$ can be used instead for regression to remove the data offset. As for the proposed framework, zero-centering can be performed in each sub-problem. Zero-centering could also affect the computation of some regression responses $T$, such as PCA as in Table 4.1, which assumes zero mean. In this case, $XX^\top$ can

be first computed and centering can then be applied to both rows and columns as follows:

$$X\left(I-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right), \tag{4.15}$$

$$S=\left(I-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)XX^\top\left(I-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right). \tag{4.16}$$

### 4.3.4 Computational Complexity

There are a few components in the proposed framework. Given $S \in \mathbb{R}^{n\times n}$ with $r$ significant singular values, the proposed framework requires a Cholesky decomposition $S = HH^\top$ and a QR-decomposition of $HP = QR$, SVD of $R$, finding $p^t$, and solving the subproblem. The decomposed matrices are compact, i.e., $Q \in \mathbb{R}^{n\times r}$, $R \in \mathbb{R}^{r\times r}$. Table 4.2 shows the computational complexity of different components. For $r \leq n$ and the desired feature cardinality $m\kappa \ll d$, the overall computational complexity is $O(ndr)$. This is much more efficient than the Fantope Projection and Selection method [163], whose complexity is $O(d^3 + nd^2)$ especially for high dimensional data applications where $d \gg m\kappa$, $d \gg n$, and $d \gg r$.

## 4.4 Experiments

In the experiments, we evaluated the proposed framework for unsupervised, supervised, and semi-supervised graph embedding, including PCA, LPP, LDA, and semi-supervised discriminant analysis (SDA). We termed them as SparGE-PCA, SparGE-LPP, SparGE-LDA, and SparGE-SDA, respectively, where SparGE stands for the proposed sparse graph embedding. We compared them with some current state-of-the-art algorithms surveyed in Section 4.1. A series of experiments on a wide range of data sets were conducted to compare the proposed methods with the current state-of-the-art algorithms.

| Modules | Cholesky & QR decompositions | SVD | Finding $p^t$ | Subproblem |
|---|---|---|---|---|
| Details | $S = HH^\top$, $H = QR$ | $U_R \Sigma_R V_R^\top$ | Compute $XV$, $s$ in (4.9), and sort $s$ | regression |
| Complexity | $O(nr^2)$ | $O(r^3)$ | $O(ndr + dr + m \log d)$ | $O(m\kappa nr)$ |

Table 4.2: Computational complexity of the proposed framework.

### 4.4.1 Experiments on Unsupervised Sparse Embedding

| S/N | **Data** | **# dimensions** | **# instances** |
|---|---|---|---|
| 1 | Toy data | 550 | 5,000 |
| 2 | Ramaswamy data | 16,063 | 144 |

Table 4.3: Data sets used to compare the SPCA, FPS, and SparGE-PCA methods.

The proposed SparGE-PCA method chooses feature subsets that maximize the variance in the data in an unsupervised manner. It optimizes over all principal components simultaneously. On the other hand, most of the sparse PCA methods [72, 84, 171, 178] select features sequentially over the principal components (PC). It is assumed that the feature subset derived from the first PC should be more important, but this may not be true. A simple counterexample is to find one feature explaining the most variance of the covariance matrix $[3.2\ 0\ 0;\ 0\ 3\ 3;\ 0\ 3\ 3]$. In this case, SPCA [178] will select either the second or the third feature based on the first PC, but the correct selection should actually be the first feature. Only the recent concurrent work on Fantope Projection and Selection (FPS) by Vu *et al.* [163] shares a similar optimization objective as our proposed method.

To compare the proposed SparGE-PCA method with FPS and SPCA, we conducted experiments on both simulated data and a real gene data data set shown in Table 4.3. The percentage

of explained variance $r_\Sigma$ was used to measure the quality of the selected feature set as follows:

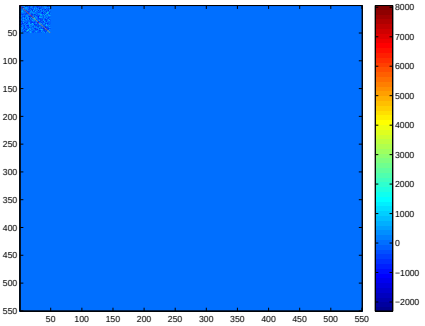
$$r_\Sigma = \frac{\text{trace}(\Sigma_{fs})}{\text{trace}(\Sigma)} \times 100\%, \tag{4.17}$$

where $\Sigma_{fs}$ and $\Sigma$ are the covariances of the selected features and of all features, respectively. A larger $r_\Sigma$ indicates a better feature subset. The results of 50 independent runs are reported. For SPCA, we chose the features with the highest absolute magnitudes of the weight matrix, similar to [30].

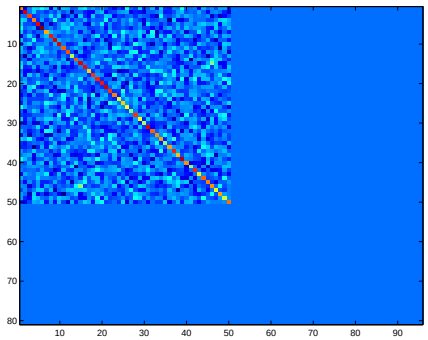
#### 4.4.1.1 The Results on Simulated Data

Optimizing for global variance, a good sparse embedding method should be able to identify a feature subset explaining the most variance and also remove noisy features. To test the optimality of the selected features of the proposed SparGE-PCA method, a simulation experiment is carried out. In this experiment, a toy data is generated with 50 significant features and 500 noisy features. Its covariance is shown in Figure 4.2.a. The explained variance should converge at 50 features. Non-zero values in the off-diagonal entries indicate a correlation among the first 50 significant features.

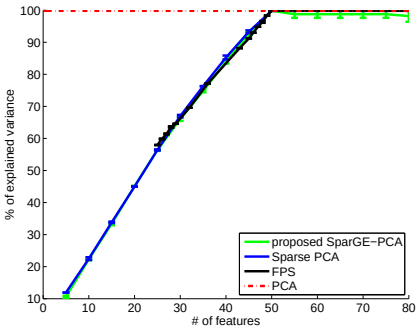
As shown in Figure 4.2.c, FPS, SPCA, and the proposed SparGE-PCA method all converged to the optimal variance at 50 features in accordance to the groundtruth. Computationally, SPCA was an order of magnitude slower than the proposed method, as shown in Figure 4.2.d. FPS shared a similar computational time with SPCA.



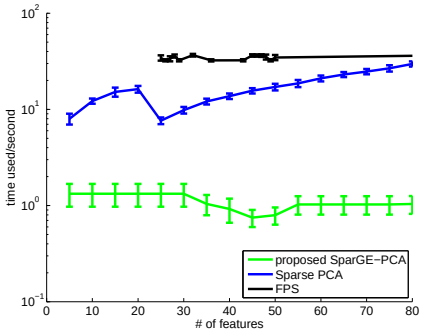
4.2.a: Covariance of the toy data



4.2.b: Enlarged view of the covariance in (a), there are 25 important and correlated features and 500 small and noisy features.

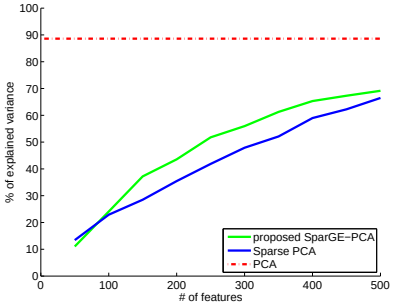



4.2.c: Variance explained

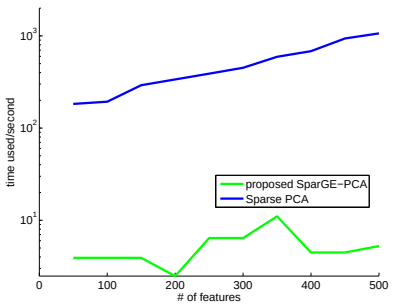



4.2.d: Computational time

Figure 4.2: Toy experiments to show the explained variance and run time vs features cardinality. The proposed SparGE-PCA method performed similarly to SPCA and FPS in the explained variance on a toy data set with the first 80 important features, but it was at least an order of magnitude faster than SPCA and FPS. PCA with 25 subspaces using all of the dimensions explains about $99.8\%$.




4.3.a: Variance explained




4.3.b: Computational time

Figure 4.3: Comparing to the SPCA method on the Ramaswamy data for variance and computational time, the proposed SparGE-PCA method outperformed the SPCA method in the explained variance and was much more efficient. PCA with 25 significant subspaces explains about 89% of the total variance.

#### 4.4.1.2 The Results on Real Data Sets

PCA is often used for analyzing high dimensional data with small samples, especially biological data. In this experiment, the microarray data, Ramaswamy data [178], is used to find the meaningful genes from very high dimensional data. The data has a very high dimensionality of 16,063 (genes) and 144 samples only. Only the first 25 principal components were used to select features.

In this experiment, the FPS method was unable to handle such a high dimensionality since the its computational complexity is $O(d^3+nd^2)$ when performing a Fantope Projection. Therefore, only the SparGE-PCA and SPCA methods were chosen for comparison.

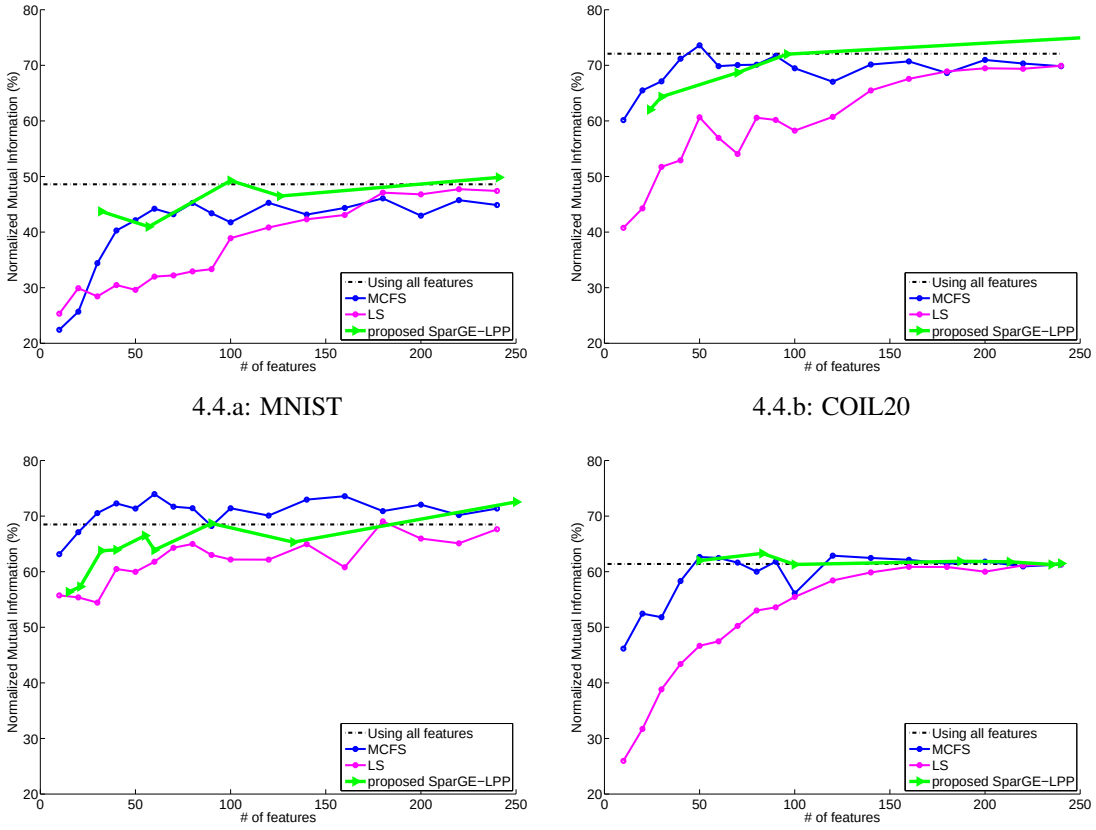
The proposed SparGE-PCA method outperformed SPCA significantly in the explained variance, by about 10% between 200 and 350 features as shown in Figure 4.3.a. Both SPCA and the proposed SparGE-PCA method converged to 70% in the explained variance with 500 features (only about 3.1% of the total features). Computationally, SPCA was at least two orders of magnitude slower than the proposed method as shown in Figure 4.3.b.

### 4.4.2 Experiments on Feature Selection for Clustering

| S/N | **Data** | **# dimensions** | **# instances** | **# classes** |
|---|---|---|---|---|
| 1 | MNIST | 784 | 4,000 | 10 |
| 2 | COIL20 | 1,024 | 1,440 | 20 |
| 3 | ORL | 1,024 | 400 | 40 |
| 4 | USPS | 256 | 9,298 | 10 |

Table 4.4: Image data sets used for clustering.

Besides sparse graph embedding for PCA, our proposed framework can be used to identify important features for clustering tasks. As discussed in Section 4.1, unsupervised LPP can model




4.4.a: MNIST

4.4.b: COIL20

4.4.c: ORL 32x32

4.4.d: USPS

Figure 4.4: Comparison of different feature selection methods on clustering tasks.

the local data structure well, and thus its embedding could improve clustering performance. Since image data sets, such as digits and faces images, usually lie on a low dimensional manifold, four popular image data sets, namely MNIST, COIL20, ORL, and USPS as shown in Table 4.4 were chosen.

To evaluate the quality of selected features, we apply k-means clustering on the data with the chosen features, where $k$ is set to the number of classes. The normalized mutual information defined in [69] is used as the performance measure. The clustering baseline employed all of the features. Besides the baseline, the feature ranking methods such as Laplacian Scores (LS) and MultiClusters Feature Selection (MCFS) were also chosen for a comparative evaluation.

The result was shown in Figure 4.4 with up to 250 features only. Interestingly, the baseline results could be achieved with as few as 50 to 70 features. Both MCFS and the proposed SparGE-LPP method performed much better than the LS feature selection method. The proposed SparGE-LPP method also outperformed MCFS on MNIST, COIL20, and USPS. Note that MCFS employed a simple ranking method to choose features from a sequential learning framework. Theoretically, it is unclear whether this approach could converge to optimality. Furthermore, it does not learn the collective weights based on the common feature subsets. In contrast, the proposed framework could achieve both weight learning and feature subset selection simultaneously.

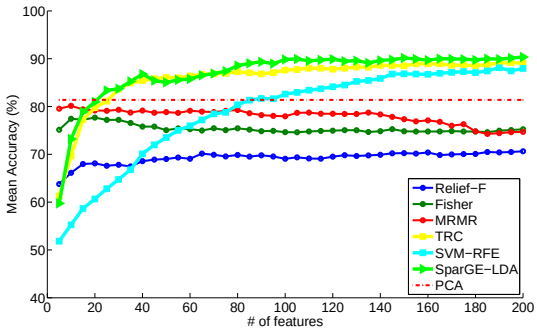
### 4.4.3 Experiments on Supervised Feature Selection

| S/N | **Data** | **# dimensions** | **# instances** | **# classes** | **Type** |
|---|---|---|---|---|---|
| 1 | PCMAC | 3,289 | 1,943 | 2 | Text |
| 2 | ORL10P | 10,304 | 100 | 10 | Face image |
| 3 | GLI-85 | 22,283 | 85 | 2 | Microarray data |
| 4 | Real-Sim | 20,958 | 72,309 | 2 | Text |
| 5 | RCV1 | 47,236 | training:20,242, testing: 67,739 | 2 | Text |
| 6 | News20 | 62,060 | training:15,935, testing: 3,993 | 20 | Text |

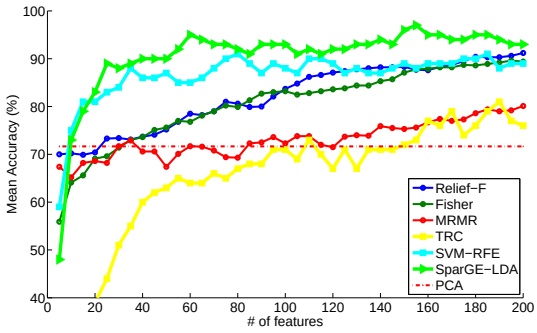
Table 4.5: A set of data used for classification experiments. Unless indicated otherwise, data sets were split for 10-fold cross validation.

| Methods | Real-Sim | RCV1 | News20 |
|---|---|---|---|
| SparGE-LDA | 228 | 302 | 460 |
| Fisher | 53 | 91 | 329 |
| SVM-RFE | 155 | 344 | 24,160 |

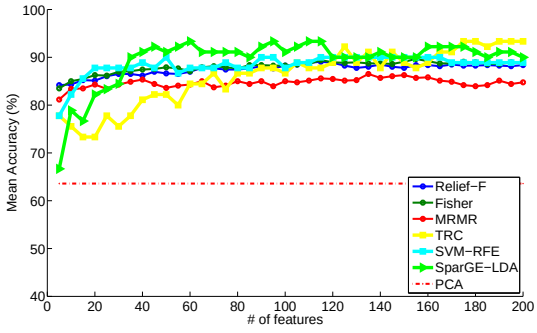
Table 4.6: Computational time in seconds of SparGE-LDA, the Fisher Score method, and SVM-RFE on ultrahigh dimensional data.



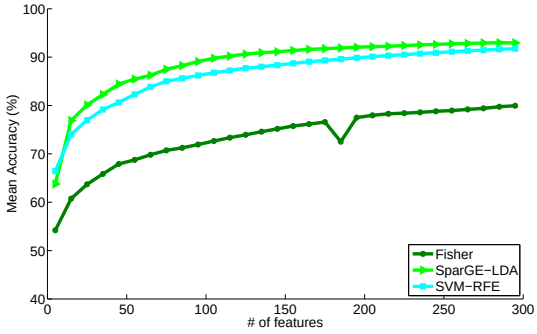

4.5.a: PCMAC



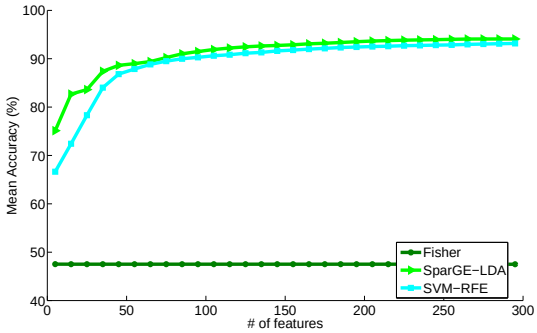

4.5.b: ORL10P



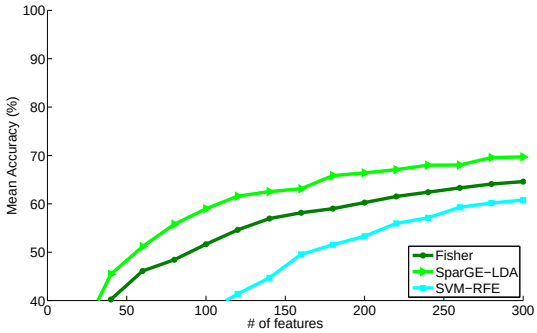

4.5.c: GLI-85




4.5.d: Real-Sim




4.5.e: RCV1




4.5.f: News20

Figure 4.5: Comparison of different feature selection methods on classification tasks. Classification results of PCA as pre-processing is also included as a benchmark for the first three data sets.

In this section, we compare the proposed SparGE-LDA method with some chosen feature selection methods on classification tasks. Six data sets as shown in Table 4.5 are used for comparison. The first three data sets ranging from text, images, and microarray data, have medium dimensionality and small sample sizes, and are collected from the Arizona State University (ASU) feature selection repository [176]. The other three higher dimensional and large-scale text data sets are from the LIBSVM [38] repository.

Several popular feature selection methods, such as Relief-F [95], the method based on max-dependency, Max-Relevance, and Min-Redundancy (MRMR) [125], Fisher score, and SVM Recursive Feature Elimination (SVM-RFE) are chosen for comparison. Beside these, we also compared the most recent method by Nie *et al.* [120], the Trace Ratio Criterion (TRC) method. TRC was shown to outperform many methods in the literature. PCA is also included as a benchmark for classification using SVM. However, PCA cannot handle data sets such as Real-Sim, RCV1, and News20, which have both ultrahigh dimensions and large data sample sizes. Therefore, no comparison with PCA for these three data sets is included. For the RCV1 and News20 data sets, the accuracies on the test sets are reported. For the rest, the mean accuracies of 10-fold cross-validation are reported.

The results are shown in Figure 4.5. We do not report the results of Relief-F, TRC, and MRMR on the three high dimensional and large-scale data sets because these methods are computationally inefficient on these data sets. Generally, both TRC and the proposed SparGE-LDA method can handle global feature subset directly, they outperformed the feature-level selection processes such as MRMR, Relief-F, and Fisher score methods. Our SparGE-LDA outperformed the TRC method in the ORL10P and GLI-85 significantly, was marginally better than TRC for the PCMAC data set. In the TRC method, the optimization enforces only one feature in each column of the weight matrix; it is thus more constrained and may lead to sub-performance compared to the proposed method.

| | | 1 labelled samples | | | 2 labelled samples | | | 3 labelled samples | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Datasets** | **# feats** | SparGE-SDA | SVM-RFE | Fisher | SparGE-SDA | SVM-RFE | Fisher | SparGE-SDA | SVM-RFE | Fisher |
| **COIL** | 100 | **59.7** | 56.8 | 31.9 | **64.3** | 61.8 | 39.7 | **71.3** | 70.3 | 50.4 |
| | 200 | **61.2** | 57.6 | 37.0 | **67.1** | 62.6 | 54.9 | **74.5** | 70.8 | 61.1 |
| | 300 | **64.2** | 57.9 | 50.0 | **69.1** | 63.4 | 58.5 | **76.7** | 71.1 | 65.5 |
| **ORL** | 100 | **59.0** | 48.5 | 25.3 | **65.3** | 63.3 | 56.8 | 70.1 | **73.9** | 34.8 |
| | 200 | **64.7** | 52.2 | 40.0 | **71.5** | 66.5 | 64.9 | **77.5** | 75.9 | 68.9 |
| | 300 | **65.0** | 53.9 | 48.5 | **71.9** | 66.6 | 69.2 | 76.8 | **76.9** | 74.2 |

Table 4.7: Feature selection results for the SparGE-SDA, SVM-RFE and Fisher score methods.

On the other hand, the SparGE-LDA method outperformed the SVM-RFE for the ORL10P, PCMAC, Real-Sim, RCV1, and News20 data sets. On the first three small data sets, SVM-RFE performed well too as it directly optimized for the classification methods. However, SVM-RFE is a greedy method, and its performance dropped significantly in the higher dimensional and large-scale data shown in Figure 4.5.d, 4.5.e, and 4.5.f. Compared to SVM-RFE, SparGE-LDA achieved 20% better in performance on the News20 data set, 5% on the Real-Sim data set, and 10% on the RCV1 data set. The Fisher score method significantly underperformed the proposed SparGE-LDA method in accuracy on the last three high dimensional data sets.

Computationally, the Fisher score method was the fastest, but SparGE-LDA was also efficient and completed the tasks within minutes as shown in Table 4.6. SVM-RFE was generally fast, but it was very slow on the News20 data set, which had the highest dimension and number of classes among the three ultrahigh dimensional data sets.

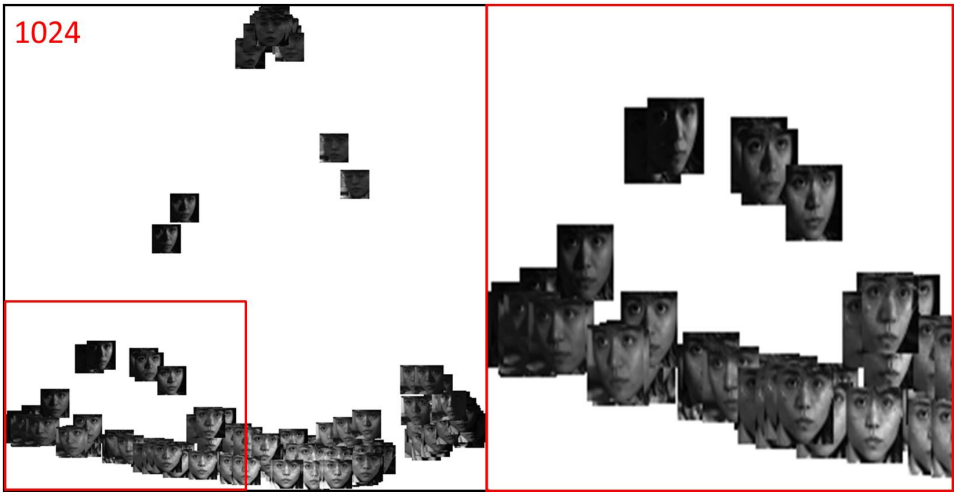
### 4.4.4 Experiments on Semi-supervised Feature Selection

So far, the experiments have demonstrated the effectiveness of our proposed framework for both unsupervised and supervised learning settings. By incorporating small labelled samples, the semi-supervised discriminant analysis based on the proposed framework (SparGE-SDA) can achieve a good classification rate.

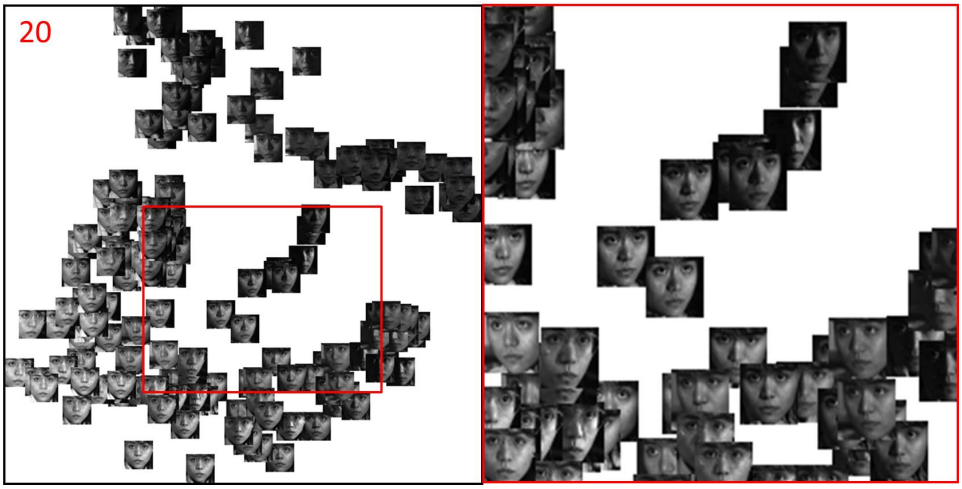
Two data sets (COIL and ORL) from Table 4.4 are used for evaluation. The proposed SparGE-SDA method was compared with the SVM-RFE and Fisher score feature selection methods. The classification accuracies are tabulated in Table 4.7. From the table, we can see that SparGE-SDA performed much better than the Fisher score method and better than SVM-RFE in the most of cases. SVM-RFE performed well with more labelled data as it optimizes for SVM classifiers. Ours optimizes between-class separation. Therefore, it is possible that SVM-RFE performed better in some cases.

### 4.4.5 Experiments on Data Visualization

In this section, we intend to identify the important face features that can preserve the data locality, i.e., the neighborhood relationships of these image samples. The data locality is visualized using a 2-D linear embedding of face images by LPP. From the CMU-PIE database [140], 170 face images of the person shown in Figure 4.1.a were chosen for this experiment. The linear embedding of these images is shown in Figure 4.6. In Figure 4.6.a, it can be observed that the illumination increases from left to right, and her face turns from left to right. The faces with expressions are far apart from the rest showing on the top, indicating a large difference. A similar observation can be drawn using the proposed SparGE-LPP with only 20 features. Variations in illumination, poses, and expressions are much more gradual on the embedding shown in Figure 4.6.b, indicating a better fit of the underlying manifold.




4.6.a: Linear embedding of face images using LPP on all of the pixels




4.6.b: Linear embedding of face images using the proposed SparGE-LPP method on 20 pixels

Figure 4.6: Linear embedding of face images using all of the pixels and the 20 selected pixels by the proposed SparGE-LPP method. The proposed SparGE-LPP method has a better embedding as indicated by a gradual change in different poses, illumination, and expressions. The respective enlarged portions are shown in the red boxes.

## 4.5 Conclusion

This chapter proposes a novel unified framework to select features for generalized graph embedding. It utilizes a feature selector to directly optimize feature subsets for graph embedding in modeling the intrinsic data structures, enabling a more robust embedding, especially for high dimensional data with a small sample size. Its efficiency and effectiveness have been demonstrated with a series of experiments for clustering, classification, and visualization. In the experiments, the proposed methods outperformed the current state-of-the-art algorithms for unsupervised, supervised, and semi-supervised learning tasks. The proposed framework demonstrated its computational and memory efficiency in handling ultrahigh dimensional data for classification.

## Appendix A: Proof of Proposition 1

In Proposition 1, the dual form of problem (4.10) has the same form as in problem (4.8).

Let $\Omega = [\Omega^1, \Omega^2, ..., \Omega^\kappa]$ be the stack of $\Omega^t$. Define the cone $\mathcal{Q} = \{(\Omega, v)\} \in (\mathbb{R}^{d\times r}, \mathbb{R}) \mid \|\Omega\|_F \leq v\}$. Let $z_t = \|\Omega^t\|_F$, $z = \sum_{t=1}^{\kappa} z_t$. The optimization problem in (4.10) is equivalent to the following problem:

$$\begin{aligned} \min_{z,\Omega} \ & \frac{\lambda}{2} z^2 + \frac{1}{2} ||\Xi||_F^2, \\ \text{s.t. } & \Xi = T - \sum_{t=1}^{\kappa} X^\top \text{diag}(p^t)\Omega^t, \\ & \sum_{t=1}^{\kappa} z_t \leq z, \quad (\Omega^t, z_t) \in \mathcal{Q}_r. \end{aligned} \tag{4.18}$$

The Lagrangian function of (4.18) can be written as:

$$\begin{aligned}\mathcal{L} =& \frac{\gamma}{2}z^2 + \frac{1}{2}\|\Xi\|_F^2 - tr\left(V^\top(\sum_{t=1}^{\kappa} X^\top \text{diag}(p^t)\Omega^t - T + \Xi)\right) \\ &+ \eta(\sum_{t=1}^{\kappa} z_t - z) - \sum_{t=1}^{\kappa}\left(tr((\xi^t)^\top\Omega^t) + \mu_t z_t\right),\end{aligned}$$

where $V \in \mathbb{R}^{n\times r}$, $\eta \in \mathbb{R}$, $\xi^t \in \mathbb{R}^{n\times r}$, and $\mu_t \in \mathbb{R}$ are the Lagrangian dual variables for the corresponding constraints. By setting the derivatives of $\mathcal{L}$ w.r.t. $z, z_t, \Omega^t$, and $\Xi$ to zero, we obtain the KKT conditions as follows:

$$\gamma z = \eta = \mu_t, \xi^t = -X\text{diag}(p^t)V, \Xi = -V, \|\xi^t\|_F \leq \eta.$$

Substitute these results into the Lagrangian function, and we obtain the dual problem as follows:

$$\max_{V,\eta} \; tr(V^\top T) - \frac{1}{2}tr(V^\top V) - \frac{1}{2\gamma}\eta^2 \tag{4.19}$$

$$\text{s.t. } \|X\text{diag}(p^t)V\|_F \leq \eta, t = 1, ..., k. \tag{4.20}$$

Setting $\theta = tr(V^\top T) - \frac{1}{2}tr(V^\top V) - \frac{1}{2\gamma}\eta^2$ and $f(V, p^t) = tr(V^\top T) - \frac{1}{2\gamma}\|X\text{diag}(p^t)V\|_F^2 - \frac{1}{2}tr(V^\top V)$, then the problem becomes as follows:

$$\begin{aligned}\max_{V\in\mathbb{R}^{n\times k},\eta\in\mathbb{R}} \quad & \theta \\ \text{s.t. } & \theta \leq f(V, p^t), \;\; \forall p^t \in \Pi.\end{aligned} \tag{4.21}$$

# Chapter 5

# Visual Tracking Via A Diffusion Process on the Riemannian Manifold of Covariances

The proposed unified framework in the previous chapter can handle a generalized class of linear graph embedding, including PCA, FDA, CCA, pLS, LPP, and SDA. Among these, PCA and LPP are unsupervised. The other methods are supervised learning methods as they require label information. Both supervised and unsupervised learning methods require a predefinition of the learning task. In visual analytic applications, task definition may be defined in a higher level after the mid-level processing. For example, high level temporal semantics may first require visual tracking to continuously associate the target across a sequence of images. This chapter proposes a novel tracking method for long sequences.

## 5.1 Introduction to Visual Tracking

Tracking essentially means to follow a target temporally, and visual tracking refers to achieving this using image sequences. Visual tracking is an important vision research topic that has many applications, ranging from motion-based recognition [34] and surveillance [78], to human-computer interaction [47]. Visual tracking covers many aspects of computer vision problems, such as target feature representation [170], feature selection [44], and feature learning [64]. Generally, long term tracking first requires a good representation of the target (termed

as features), then in each subsequent frame, searching for the locations which have similar features through a feature comparison function, and updating the features accordingly (feature learning).

Even though visual tracking has been actively researched for decades, many challenges remain especially with changes in target poses and appearance, and illumination in a long video sequence. Unfortunately, these challenges are common in many real life applications. The target pose consists of its position and orientations in 3D space, and changes as the target moves. 3D movements of a target can easily induce appearance changes in 2D images to be captured by the camera. Appearance changes can also be caused by changes in illumination, especially when the target enters a shadowed region. Figure 5.1 shows two simple examples of how a target can vary over a short time interval. The first target is a man's head, which changes appearances significantly when he changes his expression and takes off his spectacles. The second target is a toy dog, which undergoes a series of 3D movements. Consequently, for long stable tracking in many real life tasks, it is necessary to deal with these challenges. Note that in this work, the target feature, i.e., target representation, is not limited to the image patches, and we interchange it with the target template due to its wider usage. There are several choices for a target template found in the literature. For example, [144] uses the histogram of oriented gradients, while [22] uses the color histogram, [167] uses L1 sparse representation, [109] uses active appearance model, [133] uses principal subspace of image patches, and [129] uses features' covariance.

The template update problem can be expressed mathematically as follows:

$$\overline{T_t} = f\left(T_t, \overline{T_{t-1}}\right), \tag{5.1}$$

where $T_t, \overline{T_t}, t \in [1, 2, ...]$ are the estimated template and the updated template, respectively at time $t$. However, as shown in [109], target template updating is a very challenging task. There

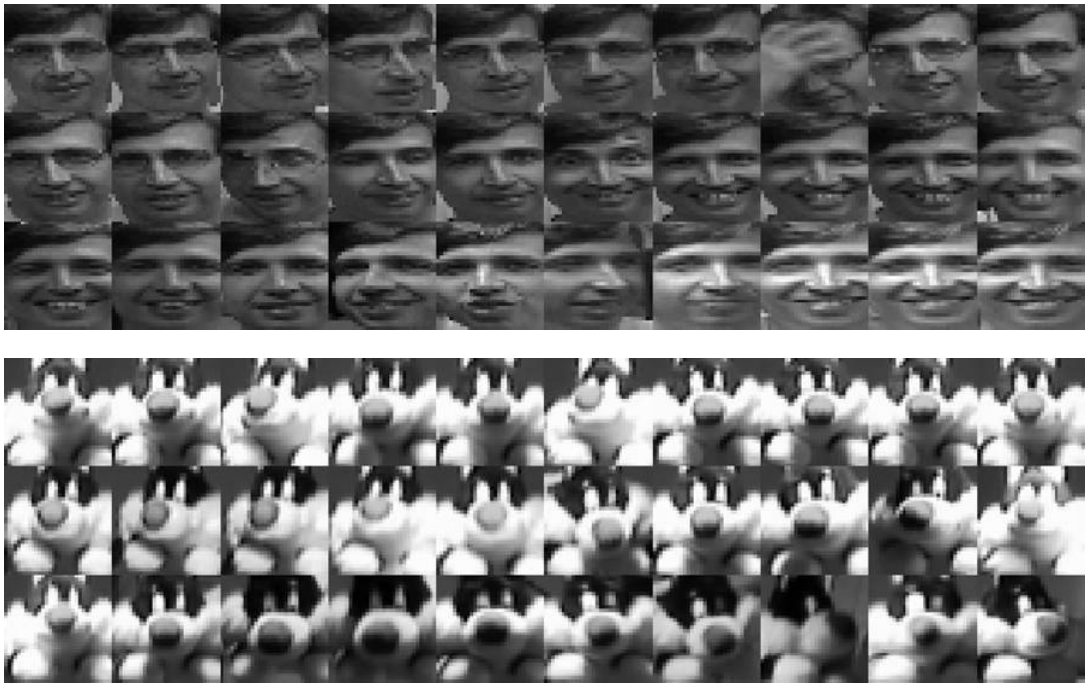


Figure 5.1: Target patches for successive 871 frames, from #1, 31, ...871 from 2 video sequences. Target changes in both illumination, poses, appearances even after being warped to a standard size.

are two intuitive and common ways to approach the template updating problem: update every frame or never update and just use the first frame template. According to [109] and as shown in

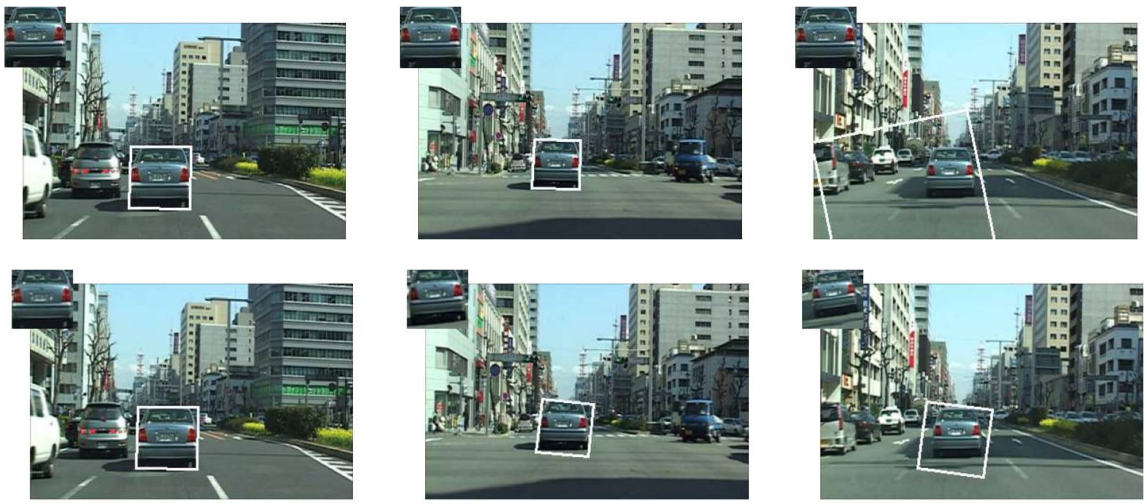


Figure 5.2: An example video sequence that is used in [109]. Frames 10, 250, and 350 are shown here. Top row: no update at all, tracking fails quickly; bottom row: update every frame, tracking drifts.

Figure 5.2, if the template was not updated at all, the template would become outdated shortly and cannot be used for matching as the target appearance would have undergone changes temporally. On the other hand, updating at every frame would result in the accumulation of small errors, and eventually a template drift and the loss of target information. Template drift occurs when the template includes more and more background information and mistakes it as target information.

The challenges of the template update problem include what to use for an update and how to update given a hypothesis of the target region. During tracking, as the hypothesized target location contains uncertainty, questions arise including how uncertainties should be incorporated into the update and how an unstable frame should be handled.

Theoretically, the problem of template updating is about incorporating uncertain information into the existing pool of information. This is often encountered in many machine learning applications. The ability to make full use of the uncertain information would allow the machine to learn how to deal with real life applications in a principled way.

## 5.2 Literature Review of Methods Handling Template Variation

There are generally three common approaches to dealing with target appearance variations. The first is to use robust or invariant target features such as the scale invariant feature transformation (SIFT) and color histograms [22]. The SIFT feature has been empirically shown to be superior to any other features for recognition of affine-transformed images. It aims to capture the histogram of gradients (descriptors) in different scales, and it has a deterministic way of locating scaled descriptors. This allows SIFT to be robust to scaling and rotation. On the other

hand, a color histogram describes the local color probability distribution of a target. In this way, the color histogram is also robust to scaling and rotation. However, in reality, targets of interest often undergo a lot more complicated 3D motions, rather than 2D affine transformations. A simple example is shown in Figure 5.1; as illustrated, the target appearance can change significantly over time, and ends up being totally different from the one in the starting frame due to variations in target poses and image illumination.

The second approach is to employ a complete set of possible target models [23], aiming to model all the possible target variations rather than aiming to make one model as invariant and robust as possible. This requires learning of the target models in advance and can hardly be scalable. This could be applied to some simple constrained environments, in which both target 3D motions and appearance variations can be easily predicted.

Finally, the last approach is to update the template gradually as it evolves. Recognizing the importance of a template update, many methods have been proposed. One common and intuitive approach is to use a linear updating function in the respective feature spaces, such as [129] on the covariance manifold. This will smoothen the changes between the estimated template and the updated template. Similarly, a Kalman filter has also been used in [119] to track template features, but not the target trajectory. On the other hand, there are three well-known template updating algorithms in the literature, namely, the Template Alignment [109] method, Online Expectation and Maximization (EM) [79], and the Incremental Subspaces method [133]. Here, we briefly survey these three algorithms.

In the template alignment algorithm [109], the authors proposed a heuristic but robust algorithm by using the first template to correct any template drift. The problem formulation is as follows. The deformation parameters $P$ at the $t^{th}$ frame are found via the following minimiza-

tion problem:

$$P_t = \arg\min_{P} \sum_{\mathbf{x} \in T_t} \left( I_t(W(\mathbf{x}; P) - T_t(\mathbf{x}) \right)^2, \tag{5.2}$$

where $\mathbf{x} = (x, y)^\top$ are the coordinates of a pixel, $W(\mathbf{x}; P)$ is the warping function, $I_t(\cdot)$ is the intensity value of the image, and $T_t(\cdot)$ is the template at the $t^{th}$ frame. Matthews *et al.* [109] shows how a gradient descent algorithm starting with $P = P_{t-1}$ is used to solve Problem (5.2) and that this equation can be rewritten as:

$$P_t = \operatorname*{gd\,min}_{P=P_{t-1}} \sum_{\mathbf{x} \in T_t} \left[ I_t(W(\mathbf{x}; P) - T_t(\mathbf{x}) \right]^2, \tag{5.3}$$

where $\left( \operatorname*{gd\,min}_{P=P_{t-1}} \right)$ means "to perform a gradient descent minimization starting with $P = P_{t-1}$". In order to correct for the template drift, an alternative update of the parameters is proposed as follows:

$$P_t^* = \operatorname*{gd\,min}_{P=P_t} \sum_{\mathbf{x} \in T_1} \left[ I_t(W(\mathbf{x}; P) - T_1(\mathbf{x}) \right]^2. \tag{5.4}$$

Problem (5.4) shows the correction of the drift is done against the first template $T_1$. The template is only updated if the solved optimal deformation parameters in Problem 5.4 do not differ much from that in Problem (5.2) by enforcing the following condition:

$$T_{t+1}(x) = \begin{cases} I_t(W(\mathbf{x}; P_t^*)) & \text{if } \|P_t^* - P_t\| \leq \varepsilon, \\ T_t(\mathbf{x}) & \text{otherwise.} \end{cases} \tag{5.5}$$

Figure 5.3 shows some of the frames used in an example video sequence used in [109]. The

Figure 5.3: An example video sequence that is used in [109]. Frames 10, 250, and 350 are shown here.

constraint imposed in (5.5) restricts any large changes of the current template from both the previously updated template and the first template. However, such a constraint, which is imposed on every pixel, is not valid in many scenarios. For example, it is unlikely that a single set of optimal deformation parameters can be found for a video sequence with a soccer player running on the field as there is a wide variation in poses as well as an issue with pixel-pixel misalignment.

The second algorithm, online EM, is proposed in [79]. More precisely, three templates are used to capture three types of variations in the wavelet-based appearance of the target. The first template is known as the *long term stable template*. This template is learned over a long sequence, varies slowly temporally, and aims to capture the stable appearance of the target. The second template is the *interframe variational template*. This template changes every frame, and its aim is to capture the sudden changes in the target due to changes in the illuminations or poses. The last template is the *outlier template*, which models the occlusion of the target, outliers or missing tracking.

Specifically, the target is modeled using Gaussian mixtures. Each of the three templates consists of $N$ pixels and each pixel is modeled independently by a Gaussian distribution with a mean and variance. Therefore, there are a total of $3N$ Gaussian models, and one would need to estimate $6N$ Gaussian parameters and another $2N$ mixture ratios. The target model is updated via the incremental update of the $8N$ Gaussian parameters using the online Expectation

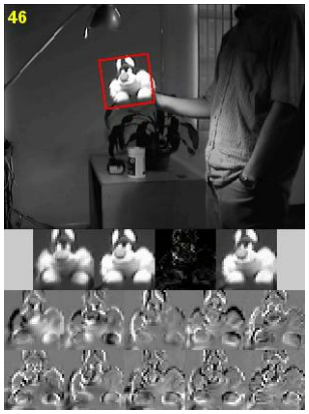
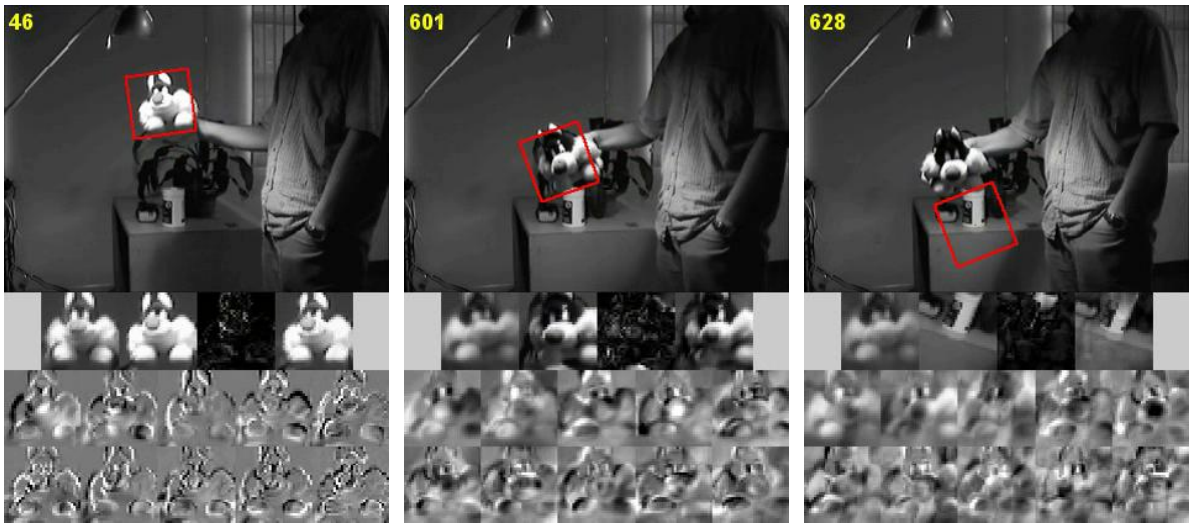



Figure 5.4: The $1^{st}$ row shows an example frame. The $2^{nd}$ row images are the current sample mean, tracked region, reconstructed image, and the reconstruction error, respectively. The $3^{rd}$ and $4^{th}$ rows are the top $10$ principal eigenvectors.

and Maximization method. Since each pixel is modeled independently of each other, pixels with small variations contribute more to the similarity measure. Thus, this would result in a tendency to include more and more background pixels and the templates would drift quickly to the more stable background.

The last algorithm is known as the incremental subspace update method, proposed in [133]. A subspace is used to represent the target and the subspace is updated via incremental Principal Component Analysis (PCA). PCA is performed on the most likely estimator of the target template together with the previous stored templates and the mean is estimated. The authors tested this algorithm on many datasets and showed that their algorithm is very robust. Figure 5.4 shows an example of the results.

There are a few issues with the incremental PCA approach. First of all, it is well-known that PCA seeks to maximize the variance, and hence, it is reasonable to assume that the first eigenbasis, which corresponds to the largest eigenvalue, is the most stable template, whereas the subsequent eigenbases corresponding to smaller eigenvalues to be the less stable templates. However, PCA assumes that the target templates are distributed in a Gaussian manner. While this assumption generally holds true for the slow interframe variations, it is not true otherwise.




5.5.a: Representative eigenbasis

5.5.b: Eigenbasis of mis-aligned target regions

5.5.c: Non-representative eigenbasis

Figure 5.5: Results of the incremental subspace method on the Sylvester sequence. Pixel-wise misalignment could render the eigenbasis non-representative.

This can be seen in Figure 5.5, taken from [133]. One can see that from frames 600 to 636, the eigenbases are not representative anymore and the tracker loses track of the target. During these frames, the difference between the target template and the target region is more than the difference between the target template and the background region due to pixel-wise misalignment. Finally, PCA can be computationally expensive when using a large target template.

So far, most of the current state-of-the-art algorithms update templates in an out-of-chain manner, by assuming the posterior estimate is "good enough" for template updating with pixel-wise alignment. If the posterior estimate is inaccurate or there is a mis-alignment between the estimated and last updated template, the update methods will gradually drift. On the other hand, if the template update is not good, then the posterior estimate of target poses is unlikely to be accurate. These coupled dual problems often render these methods unable to track well when the targets undergo fast changes in poses or non-rigid transformations. However, robustness to fast target poses has many real life applications such as human tracking and maritime target

tracking.

## 5.3 Modeling Target Using Riemannian Manifold of Covariance Matrices

To solve these dual problems faced by the existing state-of-the-art algorithms, we introduce a novel approach to simultaneously quantify these two uncertainties by including both of them into the state space of a Bayesian framework, instead of just target poses in the existing methods. In this manner, the updating is not performed on one single estimated temperate; instead, better matched multiple hypothesized templates are propagated automatically.

We give a detailed analysis of the framework of a simultaneous propagation model of kinetics and template state space. Using a similar target feature set as [39, 157], we show that on the manifold, the target template dynamics tend to be small, random, and gradual, indicating the feasibility of a random walk model. We propose a novel superior template propagation mechanism in the log-transformed space of the manifold to free the constraints imposed inherently by positive semidefinite matrices, leading to a greater ability in dealing with template variations. Our resultant method outperforms the state-of-the-art Incremental PCA algorithm (IPCA) [133] in dealing with fast moving and changing targets, as will be clearly shown in the experiments.

### 5.3.1 Covariance Descriptor

A covariance descriptor is defined as follows:

$$C = \frac{1}{N-1} \sum_{i=1}^{N} \left(f(i) - \bar{f}\right) \left(f(i) - \bar{f}\right)^{\top}, \tag{5.6}$$

where $f$ is a feature vector, and $\bar{f} = 1/N \sum_{i=1}^{N} (f(i))$ is the mean of the feature vector over $N$ pixels in the target region. In this research, we use the following 9-dimensional feature vector:

$$f(i) = \left[ x_w, y_w, I(x_w, y_w), |I_{x_w}|, |I_{y_w}|, \sqrt{I_{x_w}^2 + I_{y_w}^2}, \arctan \frac{|I_{x_w}|}{|I_{y_w}|}, |I_{xx_w}|, |I_{yy_w}| \right]. \tag{5.7}$$

The feature vector includes $x, y$ coordinates, pixel intensity, $x, y$ directional intensity gradients, gradient magnitude and angle, and second order gradients, respectively. $w$ denotes that these features are extracted after warping image patches to a standard size.

Since its proposed use in human detection [156], the covariance descriptor quickly gains popularity for many applications, such as face recognition [123], license plate detection [128], and tracking [129, 167]. Some main advantages of choosing the covariance descriptor [157] to model the template include its lower dimensionality of $(d^2 + d)/2$ ($45$ in this research as $d = 9$), compared to its number of target pixels ($32 \times 32 = 1024$ in this research), its ability to fuse multiple possibly correlated features, and its robustness to match targets in different views and poses. Using covariance descriptors has the following advantages [158]:

(1) The distance between two covariance matrices is invariant to the scaling/offset of any features $f_i$. As a result, the first two features $(x, y)$ do not require any normalization.

(2) A single covariance matrix extracted from a region is usually sufficient to match the region in different views and poses. Large rotations and illumination changes are taken

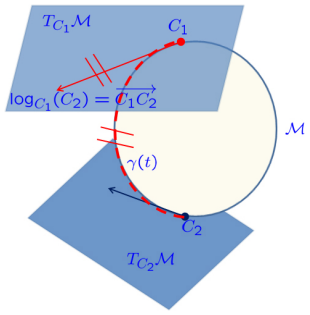



Figure 5.6: The geodesic distance is the norm of a vector on the tangent space $T_{C_1}\mathcal{M}$ at point $C_1$ on the manifold $\mathcal{M}$

care of in the covariance matrices.

(3) The covariance matrix of features has a much smaller dimensionality of $(d^2 + d)/2$, compared to using the raw values of the region.

(4) The covariance matrix is a natural way to fuse multiple possibly correlated features.

By its definition, a covariance matrix is clearly a positive semi-definite matrix, which lies on a Riemannian manifold. We will now briefly explain some basic operations on the Riemannian manifold.

### 5.3.2 Riemannian Manifold

Figure 5.6 shows an example of a manifold $\mathcal{M}$. The tangent space of $\mathcal{M}$ at a point $C_i \in \mathcal{M}$, denoted as $T_{C_i}\mathcal{M}$, is defined as the span of the tangent vectors for all the possible smooth curves $\gamma$, where $\gamma(t) : \mathbb{R} \to \mathcal{M}$, passing through $C_i$. A curve between two points $C_i$ and $C_j$ with minimum length is called a geodesic. We define the *exponential map* $\exp_{C_i} : T_{C_i}\mathcal{M} \to \mathcal{M}$, which maps each tangent vector $ty \in T_{C_i}\mathcal{M}$ to the point $\gamma(t) \in \mathcal{M}$ obtained by following the geodesic $\gamma(t)$ (parameterized with arc-length) passing through $C_i$ with direction $y$ for a distance $t$. The *logarithm map* is defined as $\log_{C_i} = \exp_{C_i}^{-1}$. Table 5.3.2 shows the operations in Euclidean and Riemannian spaces.

| Euclidean space | Riemannian manifold |
|---|---|
| $C_j = C_i + \overrightarrow{C_i C_j}$ | $C_j = \exp_{C_i}(\overrightarrow{C_i C_j})$ |
| $\overrightarrow{C_i C_j} = C_j - C_i$ | $\overrightarrow{C_i C_j} = \log_{C_i}(C_j)$ |
| $\text{dist}(C_i, C-j) = \|C_j - C_i\|$ | $\text{dist}(C_i, C_j) = \|\overrightarrow{C_i C_j}\|_{C_i}$ |

Table 5.1: Operations in Euclidean and Riemannian spaces

The Riemannian space of covariance matrices has been extensively studied [127] and the Riemannian operations can be found in closed form. For two tangent vectors $y_k, y_l \in T_{C_i}\mathcal{M}$ at a point $C_i \in \mathcal{M}$, the Riemannian metric is given as

$$\langle y_k, y_l \rangle_{C_i} = \text{tr}\left( C_i^{\frac{1}{2}} y_k C_i^{-1} y_l C_i^{-\frac{1}{2}} \right). \tag{5.8}$$

The exponential map $\exp_{C_i} : T_{C_i}\mathcal{M} \to \mathcal{M}$, takes a tangent vector $y$ at point $C_i$ and maps it to another point $C_j$:

$$C_j = \exp_{C_i}(y) = {C_i}^{\frac{1}{2}} \exp\left( {C_i}^{-\frac{1}{2}} y {C_i}^{-\frac{1}{2}} \right) {C_i}^{\frac{1}{2}}, \tag{5.9}$$

The inverse of the exponential map is the logarithm map, which takes a starting point $C_i$ and destination $C_j$ and maps to the tangent vector $y$ at point $C_i$:

$$y = \log_{C_i} C_j = {C_i}^{\frac{1}{2}} \log\left( {C_i}^{-\frac{1}{2}} C_j {C_i}^{-\frac{1}{2}} \right) {C_i}^{\frac{1}{2}}. \tag{5.10}$$

Note that $\exp_{C_i}(y)$ and $\log_{C_i} C_j$ are both $d \times d$ matrices. In addition, $\exp_{C_i}(\cdot)$ and $\log_{C_i}(\cdot)$ are maps on the Riemannian manifold, whereas $\exp(\cdot)$ and $\log(\cdot)$ denote the normal matrix exponential and logarithmic operations, which are defined as follows:

$$\exp A = B = \sum_{k=0}^{\infty} \frac{1}{k!} A^k, \tag{5.11}$$

$$\log(B) = A. \tag{5.12}$$

The distance between two covariance matrices $C_i$ and $C_j$ is given as

$$d(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k (C_i, C_j)}, \tag{5.13}$$

where $\lambda_k (C_i, C_j)$ are the generalized eigenvalues of $C_i$ and $C_j$. That is, $\lambda_k C_i v_k - C_j v_k = 0$, and $d$ is the dimension of the covariance matrices.

Note that $\exp_{C_i}(\cdot)$ and $\log_{C_i}(\cdot)$ are maps on the Riemannian manifold, whereas $\exp(\cdot)$ and $\log(\cdot)$ denote the normal matrix exponential and logarithmic operations. Both $\exp_{C_i}(y)$ and the tangent vector $y$ are $d \times d$ matrices here.

## 5.4 Bayesian Framework

In this chapter, we use a standard Bayesian framework [132] to formulate the tracking of both the template and the kinetics as follows: $z_t$ is the measurement, $\mathbf{s}_t$ is the kinetic state variable, $C_t$ is the covariance descriptor, $P(C_t, \mathbf{s}_t | z_{1:t})$ is the posterior probability of the target template and pose given the measurement, $P(z_t | C_t, \mathbf{s}_t)$ is the observational model, and $P(\mathbf{s}_t, C_t | \mathbf{s}_{t-1}, C_{t-1})$ is the dynamical model. They are further elaborated in the following subsections.

### 5.4.1 Dynamical Model

The state space in our research includes both target kinetic variables $\mathbf{s}_t$ and template covariance descriptor $C_t$. The state variables are defined in (5.14) and (5.15), and we would like to estimate them through the Bayesian framework in (5.16). These state variables are propagated

from time $t-1$ to $t$ through a dynamical model $P(\mathbf{s}_t, C_t|\mathbf{s}_{t-1}, C_{t-1})$ .

$$\mathbf{s}_t = [x_t, y_t, \dot{x}_t, \dot{y}_t, h_t, \theta_t], \tag{5.14}$$

$$C_t = \text{cov}\left(x_w, y_w, I(x_w, y_w), |I_{x_w}|, |I_{y_w}|, \arctan\frac{|I_{y_w}|}{|I_{x_w}|}, \sqrt{I_{x_w}^2 + I_{y_w}^2}, |I_{xx_w}|, |I_{yy_w}|\right) \tag{5.15}$$

where $t$ is the frame number, $x_t, y_t$ are the spatial coordinates of the target, whose velocities are $\dot{x}_t$, $\dot{y}_t$, its scaling factor is $h_t$, its orientation is $\theta_t$, $x_w, y_w$ are the coordinates of a pixel on the standard target patch warped from $x_t, y_t$, $I(x_w, y_w)$ is the pixel intensity, $\{I_{x_w}, I_{y_w}\}$ are the patch intensity gradients, $\{I_{xx_w}, I_{yy_w}\}$ are the second order gradients. Assuming independence between kinetic variables and covariance, we model the joint dynamics as follows:

$$P(\mathbf{s}_t, C_t|\mathbf{s}_{t-1}, C_{t-1}) = P(\mathbf{s}_t|\mathbf{s}_{t-1})P(C_t|C_{t-1}), \tag{5.16}$$

$$\mathbf{s}_t = k(\mathbf{s}_{t-1}) + u_t, \tag{5.17}$$

$$C_t = \exp_{C_{t-1}}(n_t). \tag{5.18}$$

$k$ is the kinetic model and we use a near constant velocity linear model $k(\mathbf{s}_{t-1}) = A\mathbf{s}_{t-1}$. $u_t$ is generated with an interacting Gaussian model with a jumping probability of $[0.9, 0.1]$ to model sudden changes in target poses. As for the template dynamical model, $n_t \in T_{C_i}\mathcal{M}$ is a random process on the tangent plane of manifold $\mathcal{M}$. An example of this could be the Brownian motion process as described by [77]. In our previous work [39], we model the random process by using a random distribution on each eigenvalue of $C_{t-1}$ and keeping the same eigenvectors. Clearly, the distribution of generated covariance matrices $C_t$ will cluster around $C_{t-1}$ on the manifold mainly due to the small random differences between their eigenvalues. However, the generated random samples may be influenced by the eigenvalues of $C_{t-1}$ as shown by the blue points in Figure 5.7. Since the template dynamics are random and small, ideally, it requires a random process whose diffusion spread is independent of previous samples. To fulfill this, we choose to

model the template dynamical model in the log-transformed space of the manifold as follows:

$$C_t = \exp(\log(C_{t-1}) + w_t), \tag{5.19}$$

$$P(\log(C_t)) \propto \exp\left(-\frac{1}{2}\sum_{i\leq j, i,j\in[1,d]}\frac{w_{i,j}^2}{\sigma_{i,j}^2}\right), \tag{5.20}$$

$$w_t = \begin{pmatrix} N(0,\sigma_{1,1}^2) & N(0,\sigma_{1,2}^2) & \cdots & N(0,\sigma_{1,d}^2) \\ N(0,\sigma_{1,12}^2) & N(0,\sigma_{2,2}^2) & \cdots & N(0,\sigma_{2,d}^2) \\ \vdots & \vdots & \ddots & \vdots \\ N(0,\sigma_{1,d}^2) & N(0,\sigma_{2,d}^2) & \cdots & N(0,\sigma_{d,d}^2) \end{pmatrix},$$

where $w_t$ is simply a random symmetric matrices and $N(0,\sigma_{i,j}^2), i,j \in [1,d]$ are normal distributions. According to [9], the matrix exponential function maps a symmetric matrix to its corresponding positive semidefinite matrix as follows:

$$exp : Sym(d) \rightarrow Sym^+(d).$$

This mapping is one-to-one. Therefore, based on the proposed random generation model, the generated samples of $C_t$ are always positive semidefinite (PSD) matrices. This frees the inherent constraints of positive eigenvalues in a PSD matrix. The proposed distribution may be considered as a log-normal distribution of the PSD matrices as defined in [137]. In the following derivation, we show that the distance, $d(C_t, C_{t-1})$ is constrained by the generated symmetric noise matrix $w_t$ independent of $C_{t-1}$. Consequently, the extremal bounds of the eigenvalues of $w_t$ control the diffusion spread of this random process. The derivation is as

follows:

$$
\begin{aligned}
C_t^{-1}C_{t-1} &= \exp(-\log(C_{t-1}) - w_t)\exp(\log(C_{t-1})), \\
&= exp(-w_t), \\
\lambda_k C_t v - C_{t-1}v &= 0, \\
C_t^{-1}C_{t-1}v &= \lambda_k v, \\
exp(-w_t)v &= \lambda_k v, \\
d(C_{t-1}, C_t) &= \sqrt{\sum_{k=1}^{d}\left[\ln^2 \lambda_k \left(exp(-w_t)\right)\right]}, \\
&= \sqrt{\sum_{k=1}^{d}\lambda_k^2(w_t)}, \quad if\ \exists w_t^{-1}.
\end{aligned}
$$

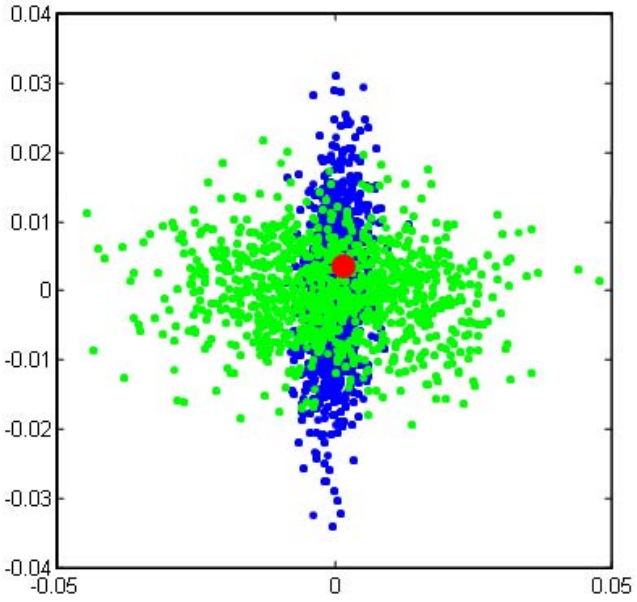
In Figure 5.7, the green samples generated by our new template dynamical model. In this research, for $d = 9$, $w_t$'s eigenvalues $\lambda_1(w_t) \geq \lambda_2(w_t) \geq ... \geq \lambda_9(w_t)$ can be bounded according to [173], assuming the entries of the noise matrix are bounded by $[a, b]$, i.e., $a \leq w_t(i, j) \leq b$:

$$
\lambda_9(w_t) \geq \begin{cases} \frac{1}{2}\left(9a - \sqrt{a^2 + 80b^2}\right) & |a| < b \\ 9a & \text{otherwise.} \end{cases} \tag{5.21}
$$

$$
\lambda_1(w_t) \leq \begin{cases} \frac{1}{2}\left(9b + \sqrt{a^2 + 80b^2}\right) & |a| < b \\ 9b & \text{otherwise.} \end{cases} \tag{5.22}
$$

In other words, the eigenvalues are roughly within an order of magnitude of $\max(\sigma_{i,j})$ for this random process. In this way, the template diffusion spread on the manifold can be easily managed by choosing an appropriate $\max(\sigma_{i,j})$ in $w_t$.

5.7.a: A target sample




5.7.b: Red: target, Blue: the previously proposed method in [39], Green: proposed method

Figure 5.7: Comparison between two methods of generating random samples on covariance Riemannian manifold.

### 5.4.2 Observation Model

The observation model $P(z_t|C_t, \mathbf{s}_t)$ measures the likelihood of a target given target poses and template values. It is modeled as follows:

$$P(z_t|C_t, \mathbf{s}_t) \sim N(0, \sigma^2), \tag{5.23}$$

$$z_t = d(C_t, C_t^*),$$

$$C_t^* = g(\mathbf{s}_t, \text{Image}),$$

$$P(z_t|C_t, \mathbf{s}_t) \propto \exp(-\frac{1}{2\sigma^2}d^2). \tag{5.24}$$

Here, $d(C_t, C_t^*)$ is given by (5.13), $g$ is the covariance computation operator, it takes the kinetic value $\mathbf{s}_t$ of each particle at time $t$ and warps the region to a standard size (in this research, $32 \times 32$) before computing the covariance.

## 5.5 Overall Framework

We employ a new set of features for covariance, modified models of kinetic and template dynamics. A particle filter is used to solve the sequential inference problem. The overall framework is as follows:

(1) **Initialization**. The particle filter is initialized with a known realization of target state variables, including the target initial state values. The covariance of the target $C_0$, i.e., initial template is extracted for comparison later. The parameters of the covariance generative process, i.e., the template dynamical model, are also determined.

(2) **Propagation**. Each particle is propagated according to the propagation model in (5.17) and (5.19). Both the kinetic variables and the template are generated through these ran-

dom processes.

(3) **Measure the likelihood**. At each particle $i$, the covariance descriptor $C_t^*(i)$ extracted is compared to its corresponding template $C_t(i)$. The likelihood of the particle is then estimated as given in (5.24).

(4) **Posterior estimation**. The posterior estimate gives the estimate of the current target state, given all its previous information and measurements. This could be the maximum *a posteriori* probability estimate or the minimum mean square error estimate (MMSE). In this research, we use MMSE.

(5) **Resampling**. To avoid any degeneracies, resampling is conducted to redistribute the weights of the particles.

(6) **Loop**. Repeat the process from step 2 to 5 as time progresses.

## 5.6 Analysis of the Template Generation Process

In this section, we show that the covariance descriptor is a good representation of the target as well as the motivation behind performing a random walk as given in Section 5.4.1. Two reasonable criteria for a good target representation are as follows:

- the representation evolves gradually as the target undergoes changes in poses, appearance,
- there is clear separation of the target and background.

To help visualize the distribution of target covariance matrices on the manifold, we use multidimensional scaling [89] to construct a visualization of the distribution of the covariance matrices. The distance matrix is constructed using the Riemannian distance as given in (5.13).

The visualization shows the relative positions of targets (red) and backgrounds (blue). Visualizing the PETS 2003 Soccer sequence and the Dudek Face sequence in Figures 5.8 and 5.9, we noticed that our representation of the targets tended to cluster together as they evolved gradually. This evolution is smoother and easier to model on the manifold as compared to the evolution of its original feature values at each pixel. This observation motivated us to model the template variations by using a random walk on the Riemannian manifold. Based on (5.17) and (5.19), Figure 5.9 illustrates a realization of the random walk. This shows that our template dynamical model can model the actual target appearance variations. Changes in facial expression and face poses cause the covariance template (shown as red points) to evolve slowly on the manifold, and they are well modeled by the generated covariances on the manifold (shown as green points).

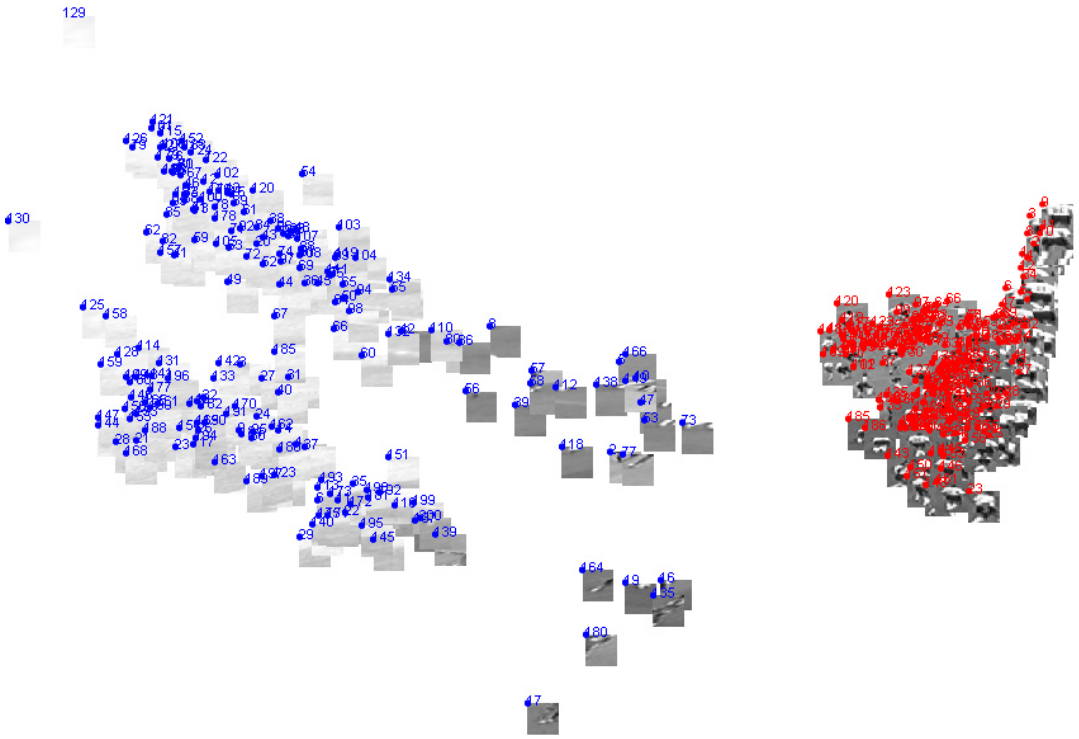
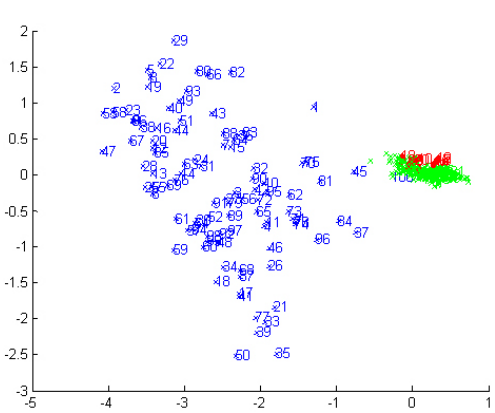
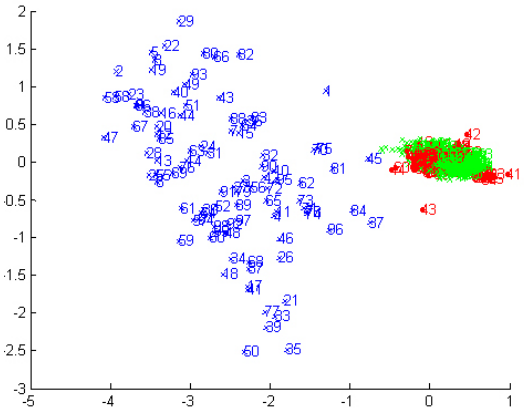


Figure 5.8: Visualization of the target in the soccer sequences on the covariance manifold. Red: target patches, blue: background patches.



5.9.a: Propagation for frames 1, 6, 11, ...,100



5.9.b: Propagation for frames 1, 6, ...,500

5.9.c: Corresponding target patches for frames 1, 51, ...,500

Figure 5.9: Visualizing template propagation on Riemannian manifold. Red: target, blue: background, green: random walk.

## 5.7 Experiments and Results

## 5.8 Experiments and Results

### 5.8.1 Experimental Data

We tested our algorithm on some of popular tracking datasets, David Ross's sequences including plush toy (toy Sylv), toy dog, David, car 4 sequences from his website, Dudek Face sequences, and vehicle tracking sequences from PETS2001, soccer sequence from PETS2003. The test data information is tabulated in Table 5.2.

| Test sequences | Source | No. of frames | Characteristics |
|---|---|---|---|
| Plush Toy (Toy Sylv) | David Ross | 1344 | fast changing, 3D Rotation, Scaling, Clutter, large movement |
| Toy dog | David Ross | 1390 | fast changing, 3D Rotation, Scaling, Clutter, large movement |
| Soccer player 1 | PETS 2003 | 1000 | Fast changing, white team, good contrast with background, occlusion |
| Soccer player 2 | | | |
| Soccer player 3 | | | |
| Soccer player 4 | | | Fast changing, gray(red) team,poor contrast with background, occlusion |
| Dudek Face Sequence | A.D. Jepson | 1145 | Relatively stable, occlusion, 3D rotation |
| Truck | PETS 2001 | 200 | relatively stable, scaling |
| David | David Ross | 503 | relatively stable 2D rotation |
| Car 4 | David Ross | 640 | Relatively stable, scaling, shadow, specular effects |

Table 5.2: Test Sequences

### 5.8.2 Performance Measure

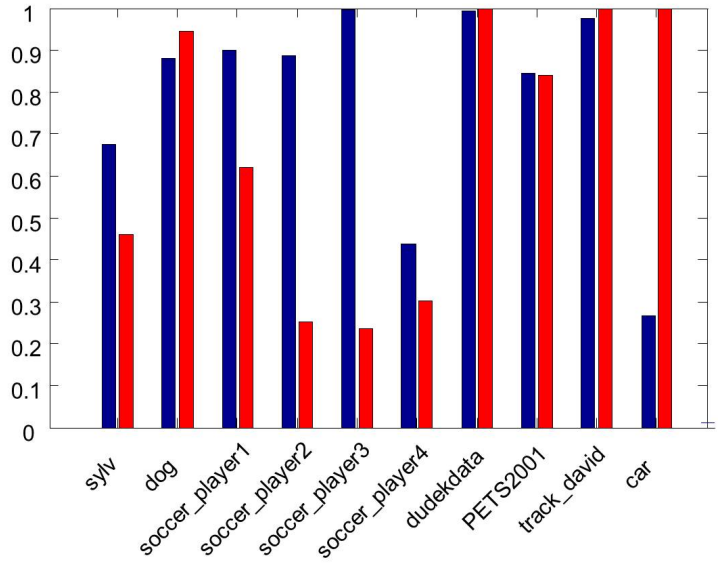
As discussed in [104], a good measure should include both overall tracking and goodness of track. This work uses the ratio between on-track length and sequence length to capture the performance of overall tracking, and on-track accuracy for goodness of track. Define tracking errors as: $e_x(t) = \|g_x(t) - x(t)\|, e_y(t) = \|g_y(t) - y(t)\|$, where $e_x(t), e_y(t), g_x(t), g_y(t)$ are the errors in $x, y$ and ground truth in $x, y$ at time $t$ respectively.

$$\gamma_{ontrack} = \frac{1}{2}\left(\frac{e_x(t)}{H_x(t)} + \frac{e_y(t)}{H_y(t)}\right) \leq 1 \tag{5.25}$$

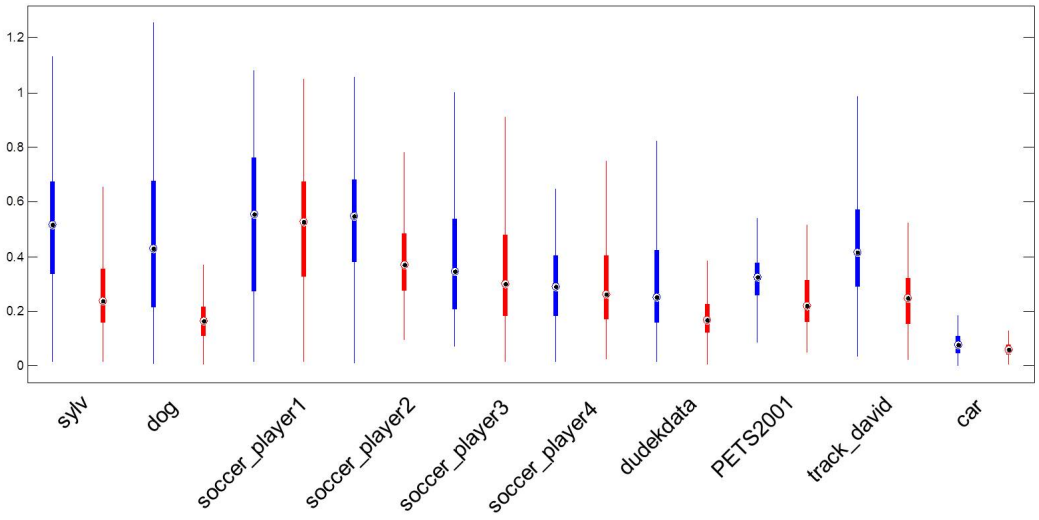
$$r_{ontrack} = \frac{\gamma_{ontrack}}{l} \tag{5.26}$$

$$rms_{ontrack} = \sqrt{\left(\frac{e_x(t)}{H_x(t)}\right)^{1/2} + \left(\frac{e_y(t)}{H_y(t)}\right)^{1/2}} \tag{5.27}$$

$H_x(t), H_y(t)$ are the ground truth target size at time $t$. In this work, ground truth on the target center is manually annotated, the target size is assumed to as those of the first frame (this may not be applicable to frames with a large change in target size).




5.10.a: Track duration rate $r_{ontrack}$




5.10.b: Track accuracy $rms_{ontrack}$

Figure 5.10: The results statistics, our results in blue, IPCA in red.

### 5.8.3 Results and Discussion

In the literature, there are many ways to make tracking better, such as using better features, additional cues like motion indicators. However, handling template drift using a generative

model is the main contribution of the chapter. As such, we only compared our method with the current state-of-the-art generative algorithm, the incremental PCA (IPCA) method by David *et al.* [133]. Our results are shown in red and the IPCA in green from Figures 5.11 to 5.17. In PLUSH TOY SYLV sequences shown in Figure 5.11, the IPCA failed to recover tracking from frame #609 when it locked onto the background, which looks more similar to the upright SYLV. Fast poses changes around frame #609 caused the IPCA eigenbases non-representative as shown in Figure 4.

Similarly, in Figure 5.12, the IPCA failed to follow through when target underwent a fast motion towards the frame #1351. This shortcoming of the IPCA is better reflected in Soccer Sequences of PETS2003. the IPCA started to drift off from frame #628 shown in Figure 5.13 when the player moved his legs fast, and lost track shortly. In the same sequence in Figure 5.14, the IPCA found it hard to track the opposite team players who wore dark clothes after a short occlusion at frame #285.

In Figure 5.15, Dudek Face sequences, both methods perform well despite of his rich facial expressions, which have more effects on our covariance descriptor. In the more stable vehicle sequence from PETS2001 in Figure 5.16, again both methods could track well. Figure 5.17 shows an example of a car sequence, in which our method did not perform satisfactorily. Our method locked onto the background whereas the IPCA showed robustness to the illumination changes. The possible explanation is that our template dynamics was unable to account for this dramatic and non-smooth transition of the template when the car went into a shadowed region. Also, a closer look showed that the IPCA eigenbasis looked similar to the target template in shadows.

The overall tracking performance on the test cases is summarized in Figure 5.10. Note that images sequences of Sylv, PETS2001 and soccer player 4 have targets out of the images, this explained the small track duration performance. Nevertheless, our method shown in red generally

had longer track length. On the hand, given frames that were on track for both trackers, IPCA showed better track accuracy shown in Figure 5.10b. For the sequences with frequent changes in target appearance such as soccer sequences, the track goodness was comparable. The video sequences may be found on the website, http://www.youtube.com/watch?v=KaSrVbGyvq4.

**Discussion**. In stable tracking cases, good pixel-wise alignment enabled the IPCA to track very well. The IPCA was generally very robust to blurring, even illumination changes, as eigenbasis tended to encompass these changes. In other words, some eigenbasis looked similar to blurred or illumination-changed templates. The distance measure in the IPCA uses a norm of all corresponding pixels difference; as such, it tends to be very stable and well aligned in the stable target cases. On the other hand, it is likely to favor the relatively stable regions in the target. When such regions are too similar to the background and target poses changes at the same time, then the IPCA may lose track very quickly in the Soccer Sequence in Figure 5.14. On the other hand, our method uses covariance of gradients and intensity; the template feature descriptor is much smaller in dimension. This may cause our method slightly less precise than the IPCA shown in Figure 5.12, which our method did not match to pixel accuracy. Figure5.17, our method lost track when the vehicle entered the shadowed region, because the both gradients and intensity changed significantly and for an interval.

Although our method was slightly not as precise in the stable cases, it gain much more flexibility in the non-stable tracking scenarios. In the cases of non-rigid or fast motion of targets, mis-alignment in the posterior estimate (the new template sample to add to the eigen space in the IPCA) and eigenbases may accumulate over a short interval and consequently render eigenbases non-representative at all. This inevitably leads to loss in tracking. Our method could deal with these scenarios a lot better for two reasons. Firstly, the template descriptor did not require pixel-wise alignment and is robust to mis-alignment. Secondly, the generative process could accommodate multiple hypothesis of the template on the covariance Riemannian manifold, and

it automatically selects the better hypothesis as the target template evolves as shown in Figure 5.9.

However, there are some limitations in our algorithm. One of them is to the need to careful choose a suitable region for tracking. Since we used the published features such as intensity and gradients, and second order gradients for covariance, these features are sensitive to specular effects, dark shadows as shown in Figure 5.17. It is also important to choose a target region with fairly good gradients variations, otherwise the covariance descriptor may be ill-conditioned consequently affecting both eigenvalues estimation and distance measurements in Equation 5.13.

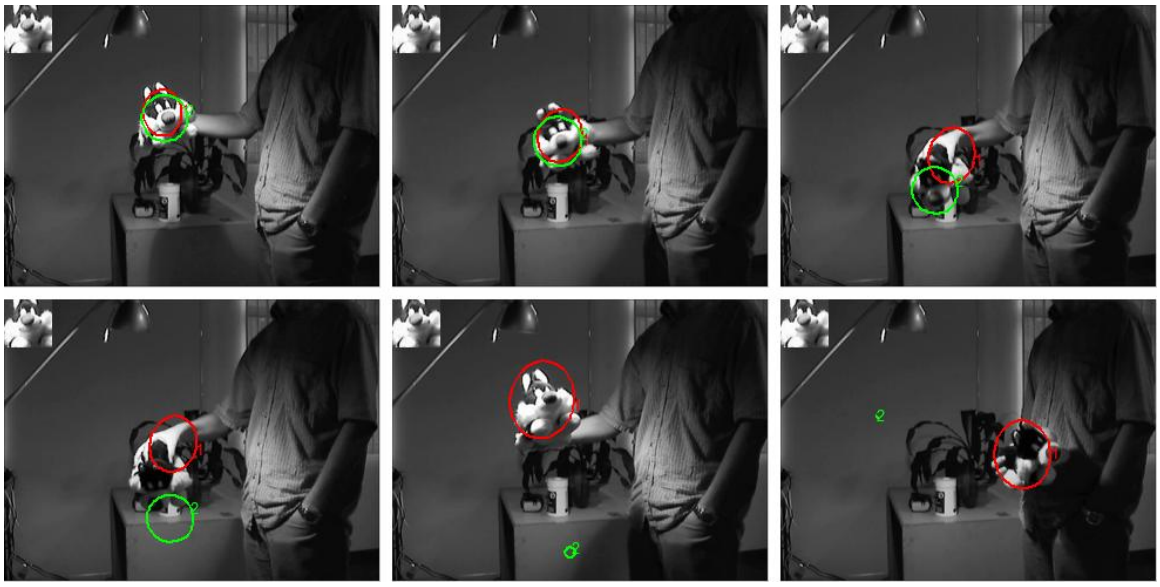


Figure 5.11: Tracking results on the PLUSH TOY SYLV sequences, frame #133, 594, 609, 613, 957, and 1338, Green: IPCA, red: our results. The IPCA failed to recover the track from frame # 609.

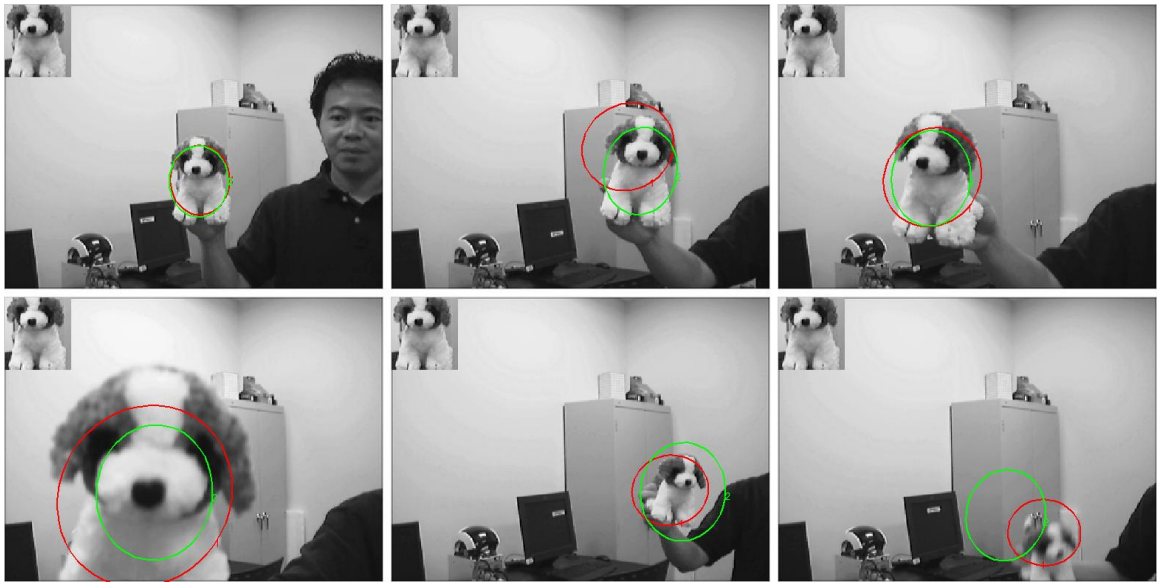


Figure 5.12: Tracking results on toy dog sequences, frame #1, 450, 715, 1014, 1271, and 1351. Green: IPCA, red: our results. The IPCA was slightly more localized in the stable case, but failed to follow through when the target underwent a fast motion towards frame #1351.

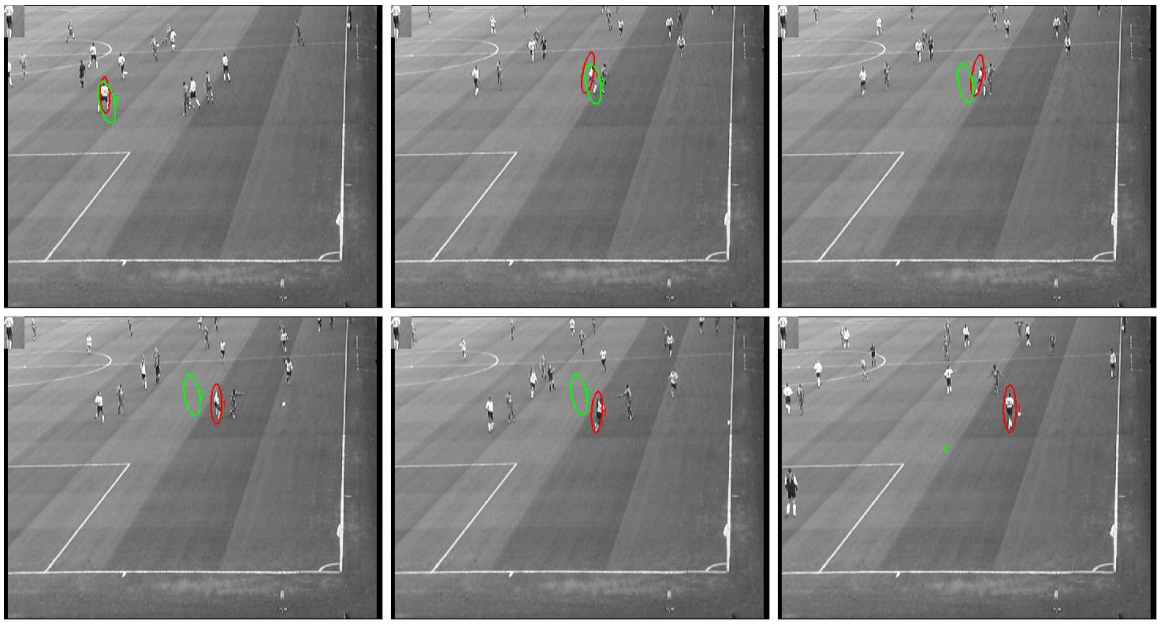


Figure 5.13: Tracking results on soccer sequences, frame #246, 628, 630, 661, 686, and 996. Green: IPCA, red: our results. The IPCA started to drift off from frame #628 when the player's legs moved fast, and lost track shortly.

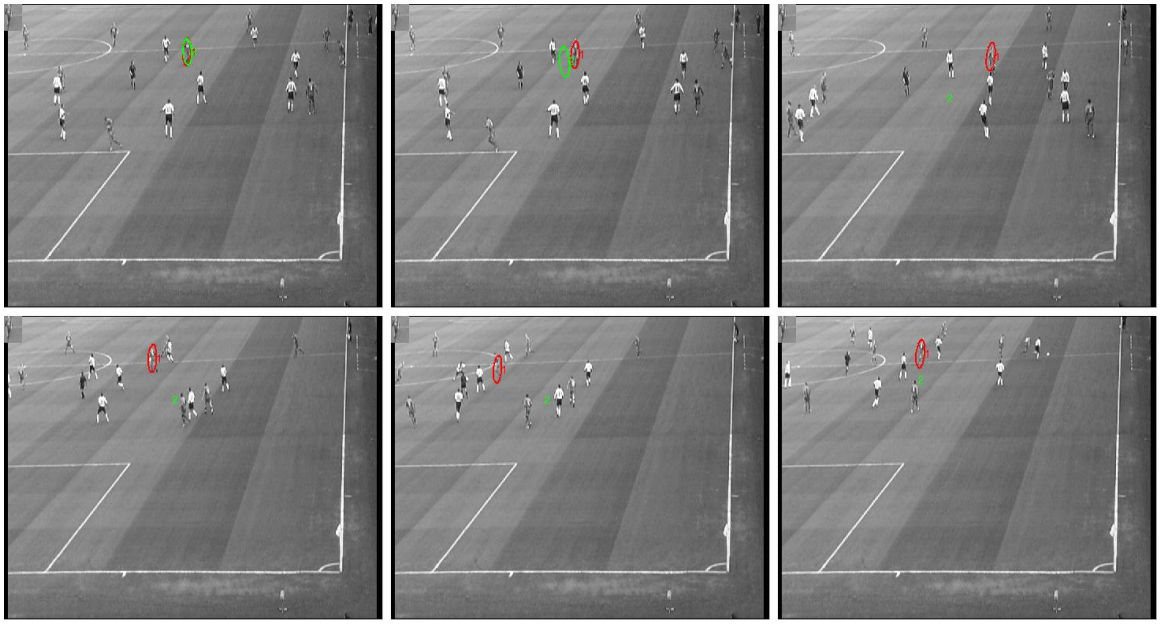


Figure 5.14: Tracking results on soccer sequences, frame #10, 15, 122, 248, 285, and 360. Green: IPCA, red: our results. The IPCA started to drift off from frame #15 due to low contrast between the target and the background.

Figure 5.15: Tracking results on DUDEK FACE sequences, frame #1, 361, 459, 605, 795, and 1095. Green: IPCA, red: our results. Both results were comparable despite his rich facial expressions, which had more effect on our covariance descriptor.

Figure 5.16: Tracking results on PETS2001 vehicle sequences, frame #1, 25, 50, 75, 100, and 125. Green: IPCA, Red: our results. Both results were comparable.

Figure 5.17: Tracking results on car sequences, frame #1, 132, 150, 168, 184, 227, Green: IPCA, Red: our results. The IPCA method performed better and was robust to illumination changes, but our method mainly used template gradients, which changed dramatically due to shadows and lack of reflection of the car plate. At frame #227, the arrow sign might look too similar to the target in gradients.

## 5.9 Conclusion

In this chapter, we have proposed a novel superior template propagation mechanism in the log-transformed space of the covariance manifold to free the constraints inherently imposed by positive definite matrices. We have shown that the simple generative process can allow the template to evolve naturally with target appearance variations. The simultaneous modeling of poses and template uncertainties enable us to better tackle these dual problems together. By employing a particle filter, our method can deal with multi-modal distributions in both poses and template variations. In the experiments, our algorithm outperformed the current state-of-the-art algorithm, IPCA, particularly when the target underwent a fast and non-rigid pose changes, and our method also maintained a comparable performance when the target was more stable.

The proposed method focuses on robust tracking of objects in a long video sequence, enabling efficient extraction of temporal semantic for high level visual analytic applications. For example, object trajectory information was used in [80] to infer any anomalous events. This is particularly the case for high resolution images, where high dimensionality in data could result in significantly expensive computation in extracting temporal information. So far, we only consider using the image rectangular patch to present targets, and may thus include background information when the object is not rectangular. For this, object segmentation information may also be included for better tracking. In the next chapter, we will propose an efficient way for object segmentation.

# Chapter 6

# Object Co-Segmentation: Propagated from Simpler Images

As discussed in Chapter 1, robust estimation of high dimensional statistics is challenging, especially when the data is highly correlated and noisy. High spatial resolution images are common in our daily life, enabling us to resolve objects in the images better with clear boundaries. However, this also brings in the challenges of highly correlated high dimensional data to the automatic processing of images. This is particularly true when we have to model spatial relationships between neighboring pixels for object segmentation. With good object segmentation, object semantics can then be derived for high level visual analytic applications. This chapter proposes an unsupervised method for robust object co-segmentation using a large image set.

## 6.1 Introduction

Image segmentation is used in many image applications for classification and recognition. Segmentation results often serve as spatial priors for object-based analysis [103] such as in remote sensing [24]. Without a clear definition of subsequent applications, segmentation by itself is not well defined; i.e., the definitions of complete objects vary according to their utilization. For example, if an image contains a pedestrian wearing a hat, a good segmentation for hat recognition would be just the hat, but pedestrian recognition may require both the person and the

hat as one segment. This problem is alleviated when a common object exists in many images. Given a large image set, the common object becomes *a priori* information for segmentation. Furthermore, accurate segmentation could significantly improve automatic recognition performance [103]. Therefore, it is important to develop an efficient and accurate objects segmentation algorithm for large image sets. Image co-segmentation is typically defined as the task of jointly segmenting something similar in an image set.

Images co-segmentation has been actively researched recently [13,33,35,36,73,82,162]. Most co-segmentation methods are unsupervised except [13], which requires interaction with users. They usually leverage on similarities in foregrounds and backgrounds among different images, and integrate pixel classification into the segmentation. In modeling, these methods aim to extract what is common in all images in terms of visual features such as the Scale Invariant Feature Transform (SIFT) [33]. A recent work [82] utilized the well-known discriminative clustering method for segmentation and reformulates the problem into an optimization problem. However, it cannot handle object variations well. This is extended [83] to a multi-class segmentation using both spectral and discriminative clustering for a probabilistic estimation.

The existing co-segmentation methods face a few challenges. First, the common objects may vary substantially in appearance, color, and orientations across images. Some images are easier to segment, but some are much more difficult due to cluttered backgrounds. It is hard to model pixel spatial relationships for segmentation in feature spaces holistically, especially when image features are high dimensional and samples are few. Second, simultaneous modeling of simple and cluttered images may result in a non-discriminative model and poor segmentation in simple images as illustrated in Figure 6.1. Third, a large image set may contain multiple foreground objects. Without knowing their labels, it is not easy for the existing co-segmentation techniques to work well. Lastly, the images may also contain undesired common backgrounds such as leaves, as shown in Figure 6.1.

Figure 6.1: Examples of co-segmentation results. First row: original images, second row: results from [82].

To meet these challenges, this chapter proposes a co-segmentation paradigm to segment images sequentially, from easy to increasingly difficult images. We first propose a novel image ranking measure to rank image segmentation easiness based on a saliency measure. With a saliency prior, single image segmentation is applied to the simplest images to extract complete objects. The complete object masks are then propagated to more complex images, on which the common objects are less salient. This propagation gives a probabilistic estimation of foreground objects in the images, which are then segmented using a graph cut. This process is efficient in computation and memory, and can segment multiple classes object simultaneously without class label information. In experiments, it achieved better object segmentation than the current state-of-the-art algorithms, especially on more complex images.

## 6.2 Recent Works

### 6.2.1 Recent Works on Segmentation

There is a recent trend of performing image segmentation in a superpixel representation, which aims to group pixels with spatial and intensity homogeneity. This kind of representation allows us to perform pixel group level processing and much faster. Two recent popular superpixels methods are Simple Linear Iterative Clustering (SLIC) [4] and Entropy Rate Superpixel (ERS) [98]. Both methods can handle object boundaries well, and are computationally efficient. SLIC employs k-means clustering in a local manner with a weighted distance measure combining both color and spatial proximity. On the other hand, ERS formulates the superpixel segmentation problem as an optimization problem on graph topology. The ERS method has a parameter on the expected number of superpixels in an image. Usually, superpixels are the intermediate steps for segmentation, as in [98].

Although using superpixels has many benefits, clustering superpixels into a complete object is not an easy task. Some well-known segmentation methods such as mean-shift [45], graph cut [26], and normalized cut methods [139] have been used in an attempt to segment out objects of interest. The mean-shift method recursively shifts the means of regions as they expand to include neighboring pixels. A cluster of pixels converges to a local distribution forming a segment, and small statistically close segments are merged into bigger segments. Both graph cut and normalized cut methods build a graph to represent pixels and their neighborhood relationships. The graph nodes are image pixels and the graph edges model the affinity between pixels. Graph-based methods have proven to be flexible to include multiple desired properties of segmentation, such as known pixels relationships and symmetry.

Unsupervised segmentation methods do not utilize any *a priori* information on the foreground

and background. These methods mainly focus on spatial grouping, rather than on foreground segmentation of objects . The resultant images may not give complete foreground objects.

The concurrent work by [111] formulates the co-segmentation problem using an energy minimization approach, which takes into account of both intra-group information within each image and inter-group information between images. This approach may be used in the propagation step of our proposed method.

### 6.2.2 Segmentation Propagation

There is also a trend towards transferring knowledge to learn a new class from a few training examples by leveraging examples from related classes. Most of these works are intended for object recognition or detection, but not segmentation. [90] proposes a segmentation propagation approach. Initialized with some known segmentation masks, it propagates the masks to the most similar unsegmented images, serving as segmentation prediction. Segmentation is then performed using a graph cut to refine the prediction. These segmented images will form sources for segmentation transfer in the subsequent step. This novel approach maps the segmentation results to the test images based on patch similarity. The underlying assumption is that similar patches share a similar foreground and background segmentation. This approach is supervised as it requires initial labelled images for training.

## 6.3 Unsupervised Image Set Segmentation

Building upon the success and challenges of the existing unsupervised segmentation methods, this chapter proposes a paradigm to leverage on their success together with high level prior information. In simple images without cluttered backgrounds, salient objects can quickly capture

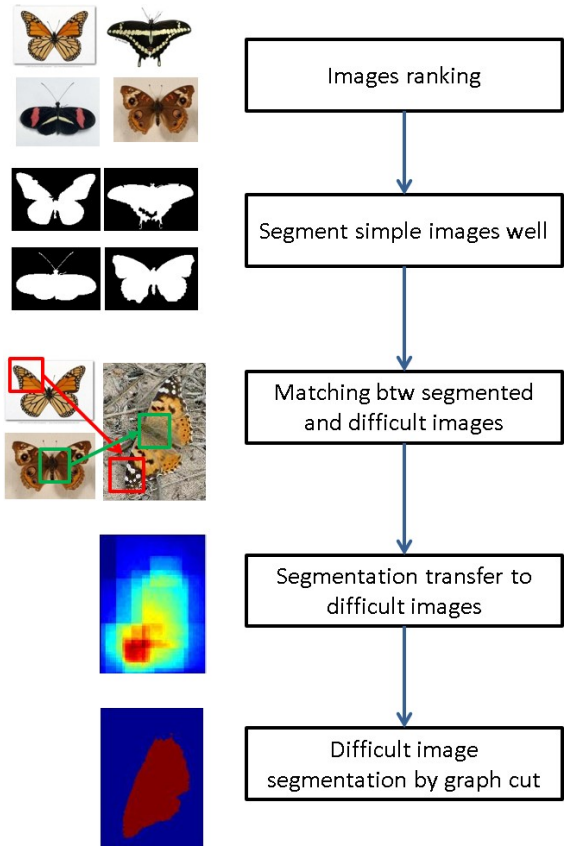



Figure 6.2: Overall algorithm framework.

the user's attention. The common foreground objects in simple images can be segmented out readily and completely. Given a sufficient number of well-segmented images, the foreground object masks can be propagated to increasingly difficult images in the manner similar to [90]. The overall algorithm framework is illustrated in Figure 6.2 and is explained in more detail as follows.

### 6.3.1 Ranking of Segmentation Easiness

An image is easy to segment if the foreground stands out readily from the homogeneous background. For such images, there should be a clean separation between foreground and background with clear boundaries, and the resultant segments should contain complete foreground

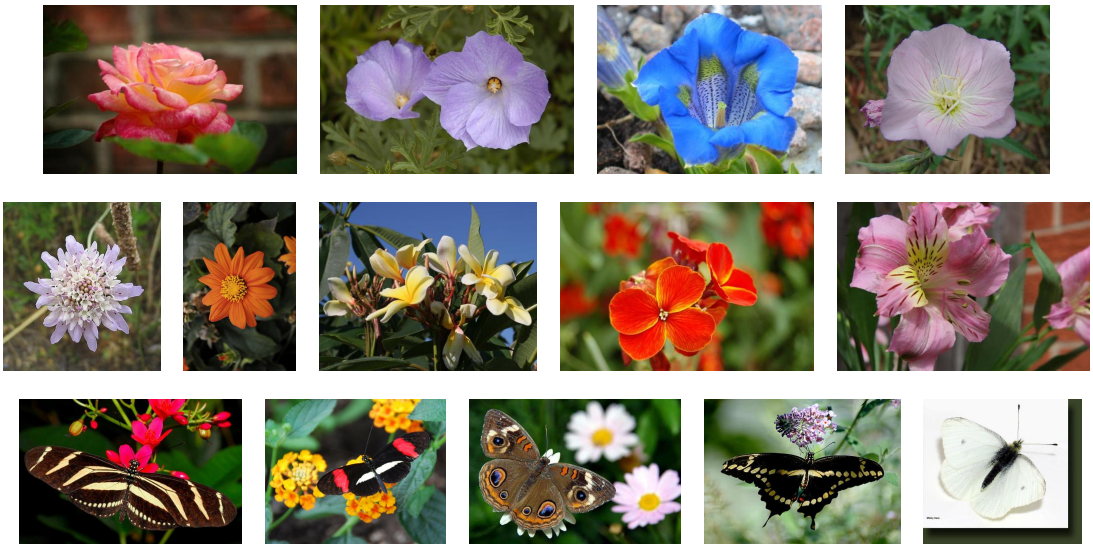


Figure 6.3: Ranking of segmentation easiness based on saliency. Images in row 1 have higher $R_{\text{sal}}$ than in row 2, and are easier to segment. Row 3 images are difficult images in the butterfly data set.

objects. This chapter proposes a saliency-based continuous measure for segmentation easiness $R_{\text{sal}}$ as follows:

$$R_{\text{sal}} = \frac{\sum_{i \in \text{fg}} S(i)}{\sum_i S(i)}, \tag{6.1}$$

where $\sum_{i \in \text{fg}} S(i)$ is the sum of saliency scores over a foreground region, and $\sum_i S(i)$ is the sum over the whole image. The foreground and background regions in an image are determined by a binary segmentation, the more salient regions are assumed to be the foreground. The saliency score of every image region $S(i)$ is estimated via a global contrast saliency score as in [42]. This score is based on the region's color contrast with respect to the whole image, with weighted sum contributions from the neighboring regions. Subsequently, more salient regions are segmented out using a graph cut. Upon segmentation, the saliency ranking $R_{\text{sal}}$ is computed. An image with a high $R_{\text{sal}}$ ranking should be easy to segment. Examples of ranking by (6.1) are presented in Figure 6.3 and the corresponding saliency maps in Figure 6.4.

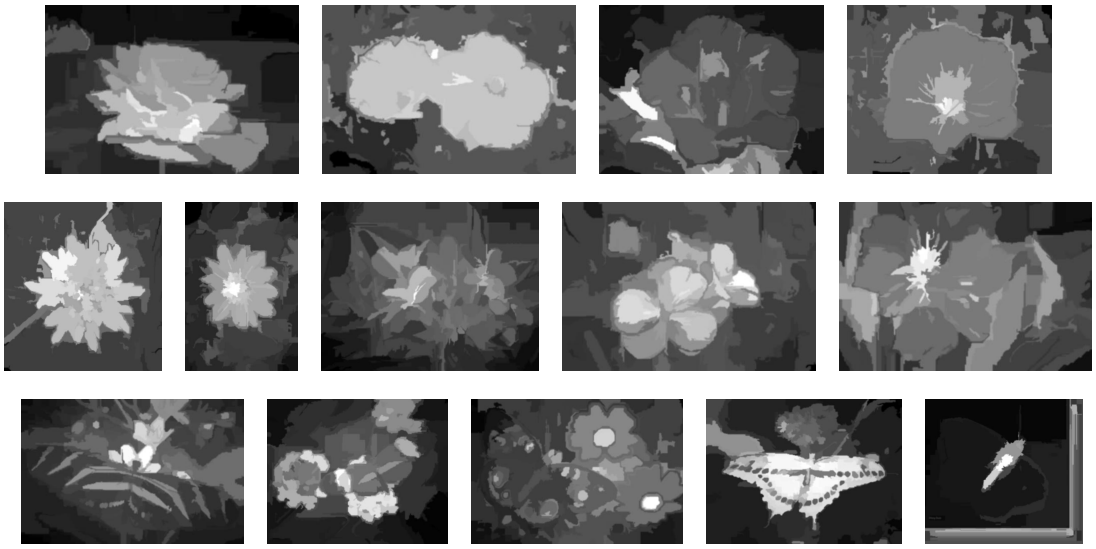


Figure 6.4: The corresponding saliency maps for images in Figure 6.3.

### 6.3.2 Segmentation Propagation

Simple images are segmented out using a graph cut on the saliency map. The resultant well defined object masks are then propagated to more difficult images as segmentation prior. The segmentation results from simple images are generally quite good, due to a clear separation between foreground and background in these images. Even in some images that may not be well segmented, the results can be further improved by passing them to the propagation step.

Let the image set be $\{I_1, I_2, ..., I_t, I_{t+1}, ...\}$, where $I_k$ is an image according to our ranking $R_{\text{sal}}$. $\{I_1, ..., I_t\}$ have been segmented, and $I_{t+1}$ is to be segmented. In more difficult images, objects may not be as salient and there could be more background clutter as shown in Figure 6.3. This is where well segmented images can help, and similar object regions should have similar segmentation boundaries. Since the similarity between images may be affected by varying backgrounds, we adopt image patches for comparison. Possible object patches are extracted from the image $I_{t+1}$, and matched to the closest $K$ patches $l$ in the segmented set $\{I_1, ..., I_t\}$.

The resultant segmentation prior of patch $x$ in image $I_{t+1}$ is defined as follows:

$$P(x) = \frac{1}{K} \sum_{l=1}^{K} exp\left(-d^2(x,l)/2\sigma^2\right), \tag{6.2}$$

where $P(x)$ is the prior probability of patch $x$ being in the foreground, $d(x,l)$ is the Euclidean distance between patches $x$ and $l$, and $\sigma$ is a parameter to set. To compute $d(x,l)$, features are extracted from these two patches. Here, we use GIST features as in [122]. Note that image patches are extracted based on *objectness* as defined in [6], the patches may overlap. In this manner, every pixel on the test patch will have a probability of being in the foreground and being in the background.

### 6.3.3 Segmentation with Prior Information

To segment an image, a graph cut is used to solve the energy minimization problem as follows:

$$E(L) = \sum_{i} U(L_i) + \sum_{i,j} V(L_i, L_j), \tag{6.3}$$

where $E(L)$ is the energy to minimize, $U(L_i)$ is the unary potential of pixel $i$ being labelled as $L_i$, and $V(L_i, L_j)$ is the potentials term modeling the spatial coherence between two neighboring pixels $i, j$. The unary potentials term is defined as $U(L_i) = -\sum_k \log(P(L_i|C_k)P(C_k))$, where $P(L_i|C_i)$ is the probability of pixel $i$ belonging to class $k,\ k \in \{0,1\}$, and $P(C_k)$ is the prior probability of class $k$ computed from (6.2). $P(L_i|C_i)$ is computed based on a Gaussian mixture model. The pair-wise potentials are defined as:

$$V(L_i, L_j) \propto d(i,j)^{-1} \exp\left(-\gamma \sum_{k=R,G,B} \left|I_i(k) - I_j(k)\right|_1\right),$$

where $d(i,j)$ is the pixel spatial distance and $\left|I_i(k) - I_j(k)\right|_1$ is the intensity difference across RGB channels. The minimization and pixels labelling are carried out iteratively until there

is no change in pixel labels. It took about 8 hours to finish segmenting 6000 images using a 16GB, 2.13 GHz Xeon machine.

## 6.4 Experiments

### 6.4.1 Data Sets

In the experiments, two data sets are chosen: the Leeds butterfly data set [164] and the flower database from Oxford University [121]. The butterfly data set has 10 categories and 832 images as shown in Table 6.1. The butterflies in these images appear in different positions, sizes, orientations, and poses. The flower database consists of 102 flower categories, commonly seen in the United Kingdom. Each class consists of between 40 and 258 images. There are a total of 5198 images chosen. All images are scaled to maximum size of 500 pixels and no category information is used in the experiments. Both have a wide variety of appearances, poses, and cluttered backgrounds. They have multiple categories of foreground objects, and their segmentation ground truths are also provided. There are both simple and difficult images in these two data sets. In this work, we used the average accuracy as the performance measure. The accuracy on one image is defined as follows:

$$\text{acc} = \frac{TP}{TP + FP + FN} \times 100\%, \tag{6.4}$$

where TP, FP correspond respectively to the number of the foreground pixels classified correctly and incorrectly, and FN corresponds to the number background pixels classified as foreground.

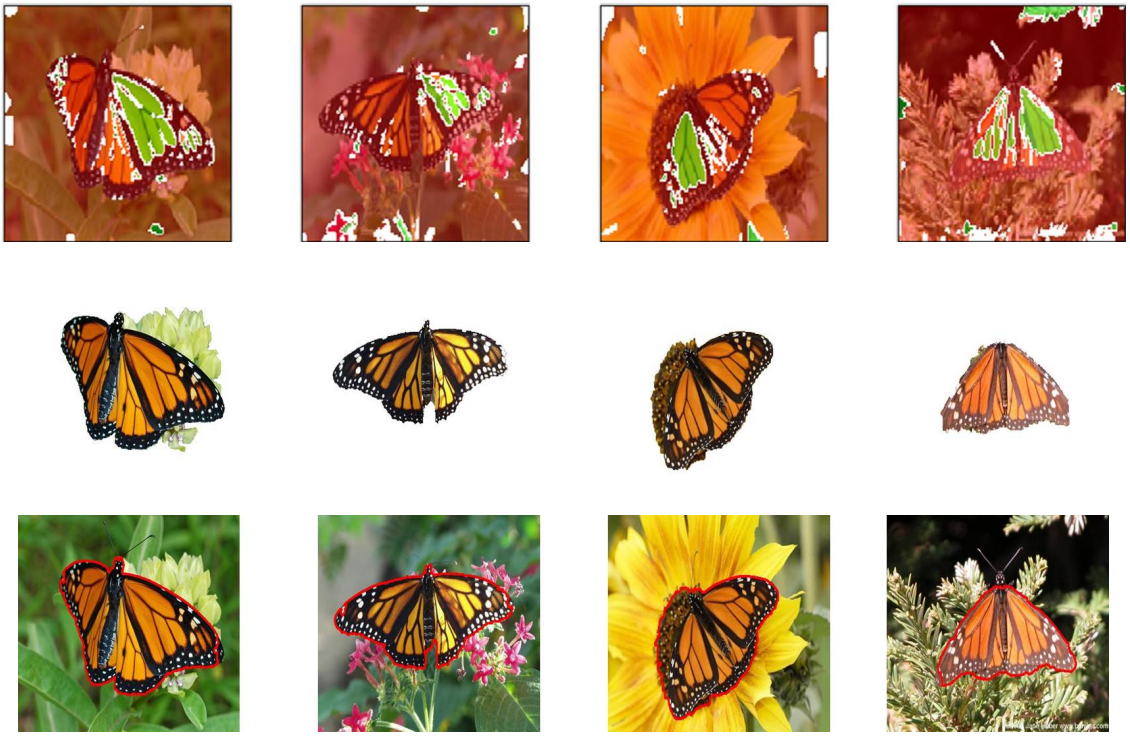


Figure 6.5: Examples of co-segmentation results. 12 images were used. The results are in the sequence of DCM [82], SC [42], and our results.

### 6.4.2 Results

The discriminative clustering method (DCM) [82], the Saliency Cut method (SC) [42], and our proposed method are compared using these two data sets. In a more complex example shown in Figure 6.3, the butterflies are of different orientations, slightly different illuminations, and different sizes. In this case, the DCM method performed poorly in segmenting out the butterflies. Both SC and our proposed methods could segment out the butterflies well as shown in Figure 6.5. Since the DCM method could not handle variations in images and multi-class objects well and it was computationally infeasible to process these two data sets, only SC was chosen for qualitative evaluation.

Table 6.1 shows the segmentation results on the butterfly data set. Our method significantly

| S/N | Names | SC [42] | Our Method |
|---|---|---|---|
| 1 | Danaus plexippus | 85.8 | **87.7** |
| 2 | Heliconius charitonius | 56.4 | **72.0** |
| 3 | Heliconius erato | 73.7 | **79.7** |
| 4 | Junonia coenia | 61.4 | **72.7** |
| 5 | Lycaena phlaeas | **79.2** | 79.1 |
| 6 | Nymphalis antiopa | **87.5** | 83.5 |
| 7 | Papilio cresphontes | 76.7 | **83.6** |
| 8 | Pieris rapae | 54.7 | **75.2** |
| 9 | Vanessa atalanta | **84.1** | 83.0 |
| 10 | Vanessa cardui | 79.2 | **82.1** |
| | Mean | 73.9 | **79.9** |

Table 6.1: Segmentation results on butterfly data set, accuracy in %.

outperformed SC, especially on the following categories: Heliconius charitonius (56.4% vs 72.0%), Junonia coenia (61.4% vs 72.7%), Papilio cresphontes (76.7% vs 83.6%), and Pieris rapae (54.7%, 75.2%), as shown in the third row of Figure 6.3. It is clear that these categories of butterflies do not have salient features, the distinction between foreground and background objects is minimal especially for Pieris rapae. Our method could propagate the segmentation results to handle these difficult cases.

For the flower data set, the results are shown in Figure 6.6. Our method outperformed the SC method significantly on most of the flower categories with the average score 78.8% as compared to 72.1% by SC.

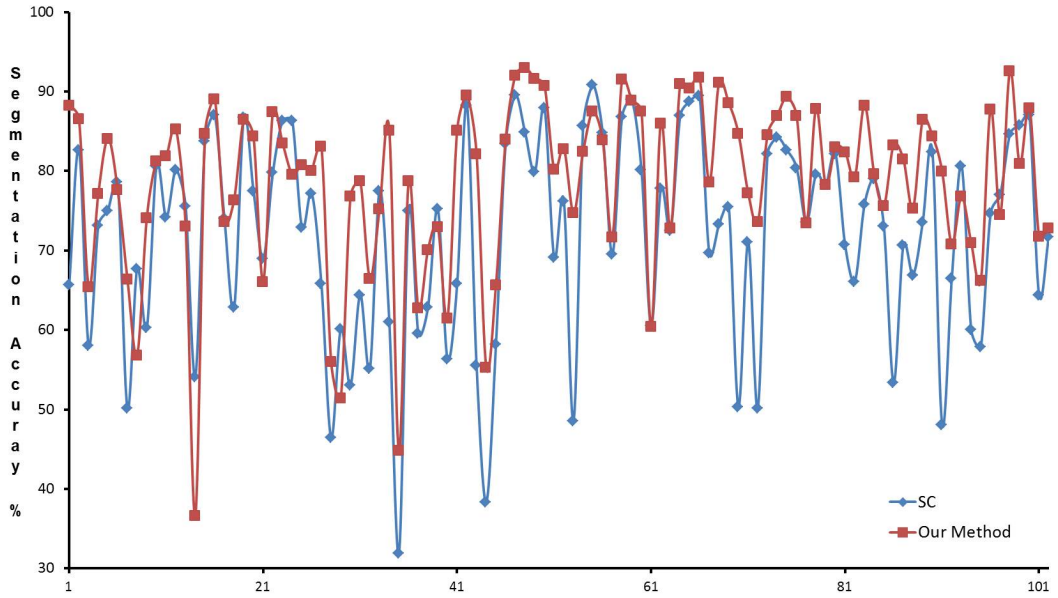



Figure 6.6: Segmentation results on the flowers data set with 102 categories.

## 6.5 Conclusion

This chapter proposes a novel unsupervised co-segmentation method for large image sets. The proposed method first ranks the image set according to segmentation difficulty. Using top level information such as object saliency, it can perform segmentation on simpler images very accurately. It is unsupervised, and no class label information is required. Equipped with the knowledge of both foreground objects and their accurate masks, the proposed method then transfers the segmentation knowledge to more difficult images. This sequential, simple-to-complex manner allows the proposed method to be very robust. It can handle complex images, in which objects are not as salient. In the experiments, more than 6,000 images were tested for robustness. The results were compared to current state-of-the-art algorithms, unsupervised saliency cut, and images co-segmentation using discriminative clustering. Our results outperformed them significantly, especially in complex images.

The proposed object segmentation method is efficient and automatic, can be used for handling

a large image set, particularly useful for high resolution images. In this case, visual analysis could then look at the object level for semantic interpretation instead of the pixel level. Together with the visual tracking method proposed in the previous chapter, the proposed methods form the mid-level processing for the high level visual analytic applications.

# Chapter 7

# Conclusion

High dimensional visual data, ubiquitous in modern data analysis, gives rise to potentially better data modeling and class separation for visual analytic applications. Along with these new developments emerge new challenges such as sample insufficiency problems, data sparsity in high dimensional space, and non-scalable computational costs. This thesis focuses on efficient learning methods for high dimensional visual data. First, we propose a sparse graph embedding methods for a generalized class of machine learning methods. Subsequently, two efficient methods are proposed for learning of temporal and spatial semantics, respectively. They form the important mid-level processing layer for high level visual analytic applications. This chapter summarizes some contributions, and discusses some possible areas to further extend this research.

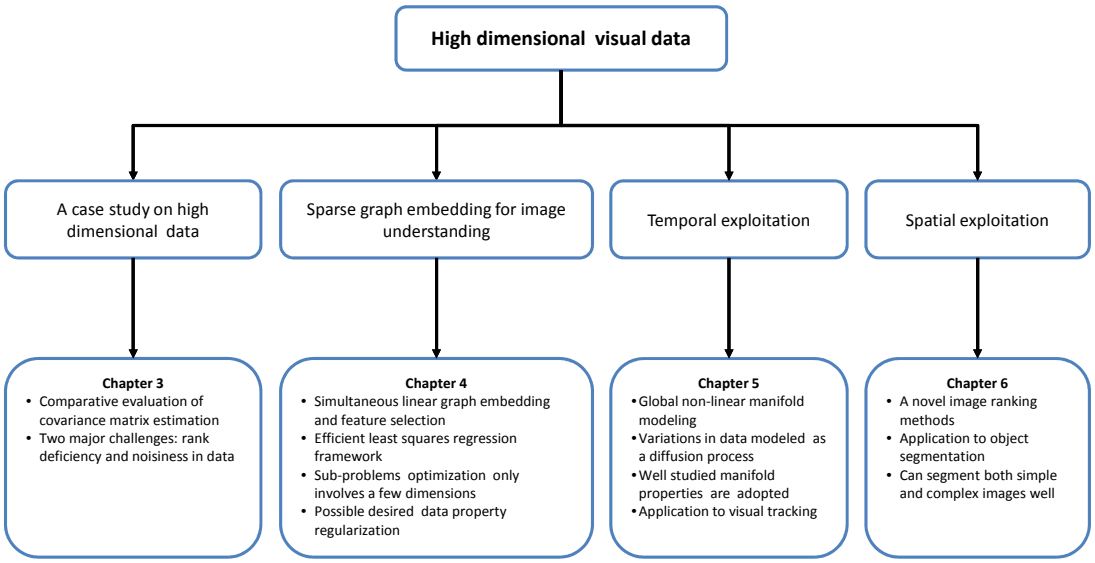



Figure 7.1: Contributions of this thesis for efficient learning of high dimensional visual data.

## 7.1 Summary of Contributions

This thesis has made a few contributions to efficiently learning of high dimensional data for visual analytic applications. Specifically, they include high dimensional covariance matrix estimation for hyperspectral data, an efficient framework to perform linear graph embedding and feature selection simultaneously, manifold modeling of covariance matrices and its application to visual tracking, and object co-segmentation on a large image set.

1 **A case study on high dimensional covariance matrix estimation for hyperspectral data**

To cast light on the adverse effects of high dimensionality, we survey the literature for covariance matrix estimation on hyperspectral data. High correlation and noisiness among the dimensions of hyperspectral data make covariance matrix estimation particularly relevant for understanding the impacts of high dimensionality. The findings show that

the sample covariance of high dimensional data severely degraded the performance of anomaly detection as dimensionality increases. The performance was made worse by the data outliers. A high correlation among data dimensions tends to cause rank deficiency in the covariance matrix and presence of outliers yields unreliable estimates of data statistics. On the other hand, conditioning covariance matrix with a sparse structure and regularization can significantly mitigate rank deficiency and the outlier issue. This demonstrates the necessity of efficient and robust methods for high dimensional data learning.

2 **An efficient framework to perform graph embedding and feature selection simultaneously**

Dimensionality reduction and feature selection are two main successful approaches in handling high dimensional data. Although these two approaches have their own pros and cons, they share a common objective in modeling data intrinsic structures. To unify them, we propose a novel paradigm to perform feature selection and graph embedding simultaneously. In this paradigm, a novel feature selection scheme is presented using a regularized least squares formulation, which is then solved efficiently using an accelerated proximal gradient method.

The proposed framework is flexible enough to cater for a generalized class of linear dimension reduction techniques, such as Principal Component Analysis, Linear Discriminant Analysis, Locality Preserving Projections, Canonical Correlation Analysis, and Hypergraph Spectral Learning. It can be also used for unsupervised, supervised, and semi-supervised methods in preserving the corresponding intrinsic data structures via low dimensional embedding.

Furthermore, the proposed framework is very computationally efficient and can readily handle very high dimensional data, more than 50,000 dimensions, for which the common

dimensionality techniques cannot achieve easily.

Computational efficiency and optimality in this framework enables an abstraction of generalized linear graph embedding, facilitates future developments of both linear dimensionality reduction techniques and feature selection methods.

3 **A diffusion process on Riemannian manifold for covariance evolution**

While the proposed unified framework for linear graph embedding can efficiently handle high dimensional data in linear modeling, non-linear modeling using manifolds has many applications, especially on image analysis. A robust distance measure for image analysis has to handle variations in image illuminations, object poses and appearances - these variations are all within-class. It is therefore necessary to have a compact representation in the low dimensional manifold. Many manifold learning methods have been proposed for image modeling, many involve a pair-wise distance computation, which is neither scalable nor easy to extend to incremental new samples.

We propose an efficient generative model on the Riemannian manifold of covariances to handle a multiple-model distribution of covariances. The proposed model can effectively incorporate the uncertainty in the labelled data for an incremental learning. This is particularly applicable to long-term tracking, where there is inevitably a drift in target positions, orientations, appearance, and possible occlusion. It is necessary to deal with the dual uncertainties in the target kinetics (positions and orientations) and representations (appearance). The proposed generative target representation model, together with a target motion model, enables us to simultaneously address the dual problems in tracking via a Bayesian framework.

The proposed method exploits a compact representation on a well-defined manifold, effectively handles the dual uncertainties in visual tracking. This method can be extended

to incremental learning problems, where there is an uncertainty in data labeling.

4 **Object co-segmentation: propagated from simpler images**

Object segmentation allows simplification of images for efficient visual analysis, it remains challenging to perform this task on a single image. We propose a novel framework to tackle this task by co-segmenting a large image set. First, we note that the modern image segmentation techniques work well for simple images, but not for complex images. This is due to lack of a good spatial relationship model to group object components together. We then propose to propagate object segmentations results from the simpler images to more difficult ones. A novel image ranking mechanism is proposed so that simpler images can be filtered for object segmentation first. Segmentation results from simple images serve as training samples for unsegmented more complex images via a local KNN-graph. In this manner, the object co-segmentation model is much more smooth and compact, and is capable of handling both simple and difficult images very well.

This application illustrates a neat way to address a complex visual analytic problem. By first accomplishing simpler tasks well, a more efficient and compact learning model can be constructed for more complex tasks in a sequential manner.

5 **Spectral, spatial and temporal processing**

This thesis covers all three domains of visual data, namely, spectrally across different color channels up to hyperspectral channels, spatially for object segmentation, and temporally for object tracking. The comprehensive coverage will enable potential fusion of the information from all three domains for complete visual data exploitation.

## 7.2 Future Work

Some possible future work is to extend our methods to high level visual analytic applications or to better integrate the vertical processing chain in the visual analytic framework shown in Figure 1.5. The proposed methods in this thesis can be used to individually to facilitate extraction of the respective high level semantics. Also they can be used together to achieve even better performance. They are further elaborated as follows:

- **Simultaneous detection, segmentation, learning and tracking**

  The recent work on tracking-learning-detection [85] integrates the three areas of work

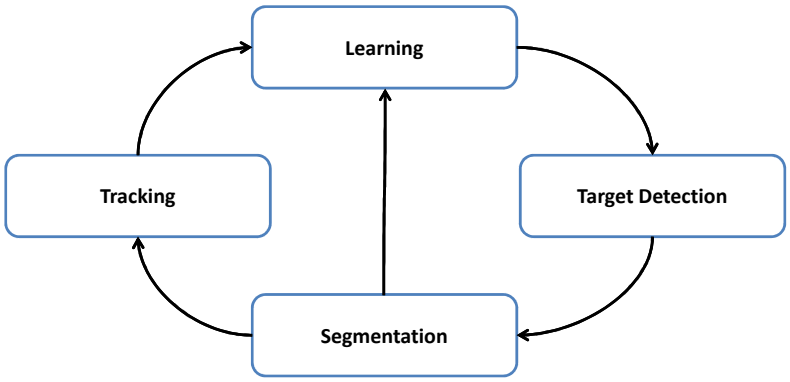



Figure 7.2: Visual analytic components and techniques.

for robust tracking, namely, visual tracking, machine learning, and target detection. It was shown to outperform most of current state-of-the-art algorithm. Most of the current tracking algorithms have a suitable underlying kinetic model to predict target potential regions, and often fail to predict for fast motion changes. Instead of searching for a sophisticated model, a target detection model can be integrated to better estimate the potential target locations. Upon detection, the target size and poses can be further refined by a segmentation algorithm. The proposed object co-segmentation in Chapter 5

is particularly suitable for this case, as it is sequential and instance-based learning. The idea is that detection is mostly based on the pure object patch information without considering the surrounding information, and segmentation complements it by including the surrounding information for a more complete object. Subsequently, tracking is to associate temporally the detected and segmented objects and enables extraction of temporal semantics such as object trajectories and target appearance changes. Finally, the learning module can take in the tracked target segmentation and update its classifier, which will help in detection. By this, a cycle is completed as shown in Figure 7.2.

- **Extension to high level visual analytic applications**

  In this thesis, we focus on mid-level learning methods for visual analytic applications. With a more complete framework as shown in Figure 7.2, we can then extract more robustly high level semantic information, such as event recognition, anomaly activities, and spatial semantic extraction for object recognition. Trajectory information was used in [80] to infer any anomalous events. More sophisticated models such as Hidden Markov Model was also used in [75] to infer activities such as "converse", "approach then stop", "approach then leave". Given more constrained environments, such as pedestrian path in the park, the number of common activities are limited, high level visual analytic tools can be used to identify potential anomalous activities such as "snatching".

  Besides the temporal event recognition, our work can be extended for visual recognition. Our segmentation and tracking work can be used to extract better training samples for classification, therefore, enabling better discrimination among similar images. One of possible examples is face recognition in closed-circuit television videos.

- **Applying to a deep representation**

  Throughout the work of this thesis, we have focused on efficient methods for high dimensional learning, assuming the data representation is known. The recent development

of unsupervised feature representation through deep learning has been shown to significantly improve many machine learning tasks. The basic idea is to build multi-level representations. In each hierarchical level, unsupervised learning methods are applied to learn a new transformation method on the feature transformation derived from the previous level. To leverage the success of unsupervised feature representation, we can apply the proposed efficient learning methods in each level of the multi-level representations so to benefit from high dimensional representations without suffering from the adverse effects discussed so far.

# Publications

**The following publications are included in the thesis:**

(1) Objects Co-segmentation: Propagated from Simpler Images. Marcus Chen, Santiago Velasco-Forero, Ivor Tsang, Tat Jen Cham, *in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

(2) Comparative Analysis of Covariance Matrix Estimation for Anomaly Detection in Hyperspectral Images. Santiago Velasco-Forero, Marcus Chen, Alvina Goh, Pang Sze Kim. *Journal of Selected Topics in Signal Processing*, Vol.9, no.6, pp. 1061-1073, June 2015.

(3) A Unified Feature Selection Framework for Graph Embedding on High Dimensional Data. Marcus Chen, Ivor Tsang, Mingkui Tan, Tat Jen Cham. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, Vol.27, no.6, pp.1465-1477, December 2014.

(4) Visual Tracking with Generative Template Model Based on Riemannian Manifold of Covariances, Marcus Chen, Sze Kim Pang, Tat Jen Cham, and Alvina. Goh, *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp.1-8, IEEE, 2011.

(5) A Comparative Analysis of Covariance Matrix Estimation in Anomaly Detection, Santiago Velasco-Forero, Marcus Chen, Pang Sze Kim, Alvina Goh. $6^{th}$ *IEEE-Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2014.

**Author's other publications include the following:**

(1) Shadow Detection in Very High Spatial Resolution Aerial Images: a Comparative Study,

Karine Adeline, Marcus Chen, Xavier Briottet, Sze Kim Pang, Nicolas Paparoditisf, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.80, pp.21-38, June 2013.

(2) Efficient Empirical Reflectance Retrieval in Urban Environments, Marcus Chen, Kailing Cheryl Seow, Xavier Briottet, Sze Kim Pang. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE*,Vol.6, no.3, pp.1596-1601, May 2013.

(3) Anomaly Detection and Important Bands Selection for Hyperspectral Images via Sparse PCA, Santiago Velasco-Forero, Marcus Chen, Pang Sze Kim, Alvina Goh. $6^{th}$ *IEEE-Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2014.

(4) Detection and Recognition of Alert Traffic Signs, Chia-Hsiung Chen, Marcus Chen, Tianshi Gao, *Publications of Stanford University*, 2008

# Bibliography


[1] "Special section - dimensionality reduction methods," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 1–1, March 2011. [↑26]

[2] Y. I. Abramovich and O. Besson, "Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach-Part 1: The Over-Sampled Case," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5807–5818, 2013. [↑28]

[3] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (lnsmi) for outlier-resistant adaptive filtering," in *ICASSP 2007*, vol. 3. IEEE, 2007, pp. 1111–1105. [↑38]

[4] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012. [↑118]

[5] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society. Series B*, pp. 139–177, 1982. [↑48]

[6] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," 2012. [↑123]

[7] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*. Cambridge University Press, 2010, no. 118. [↑31]

[8] T. W. Anderson, *An introduction to multivariate statistical analysis*. Wiley, 1958. [↑29], [↑33]

[9] V. Arsigny, P. Fillard, X. Pennec, N. Ayache *et al.*, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, p. 328, 2008. [↑99]

[10] F. Bach, S. D. Ahipasaoglu, and A. d'Aspremont, "Convex relaxations for subset selection," *arXiv preprint arXiv:1006.3601*, 2010. [↑19]

[11] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proceedings of the Twenty-first International Conference on Machine learning*. ACM Press, 2004, p. 6. [↑66]

[12] M. Balasubramanian, E. L. Shwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford, "The isomap algorithm and topological stability," *Science*, 2002. [↑25]

[13] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3169–3176. [↑116]

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." in *NIPS*, vol. 14, 2001, pp. 585–591. [↑24]

[15] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006. [↑60]

[16] R. Bellman and R. Bellman, *Dynamic Programming*, ser. P (Rand Corporation). Princeton University Press, 1957. [↑56]

[17] A. Ben-David and C. E. Davidson, "Eigenvalue estimation of hyperspectral wishart covariance matrices from limited number of samples," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 50, no. 11, pp. 4384–4396, 2012. [↑41], [↑42], [↑43], [↑55]

[18] A. Ben-David and J. Marks, "Geodesic paths for time-dependent covariance matrices in a Riemannian manifold," *IEEE Transactions on Geosc. and Rem. Sens. Letters*, vol. 11, pp. 1499–1503, 2014. [↑39]

[19] O. Besson and Y. I. Abramovich, "Regularized Covariance Matrix Estimation in Complex Elliptically Symmetric Distributions Using the Expected Likelihood Approach–Part 2: The Under-Sampled Case," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5819–5829, 2013. [↑28]

[20] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Database Theory - ICDT99*. Springer, 1999, pp. 217–235. [↑9]

[21] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008. [↑11]

[22] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *CVPR*, 1998, pp. 232–. [↑85], [↑87]

[23] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *IJCV*, vol. 26, no. 1, pp. 63–84, 1998. [↑88]

[24] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 65, no. 1, pp. 2–16, 2010. [↑115]

[25] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer, 2005. [↑24]

[26] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006. [↑118]

[27] C. J. C. Burges, "Dimension Reduction: A Guided Tour," *Foundations and Trends in Machine Learning*, vol. 2, no. 4, pp. 275–365, 2009. [↑23]

[28] D. Cai, X. He, and J. Han, "Semi-supervised regression using spectral techniques," *Dept. Comput. Sci., Univ. Illinois at Urbana-Champaign, Urbana, Tech. Rep. UIUCDCS*, 2006. [↑60]

[29] ——, "Semi-supervised discriminant analysis," in *Proc. Int. Conf. Computer Vision (ICCV'07)*, 2007. [↑61]

[30] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, pp. 333–342. [↑18], [↑72]

[31] G. Cao, L. R. Bachega, and C. A. Bouman, "The Sparse Matrix Transform for Covariance Estimation and Analysis of High Dimensional Signals," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 625–640, mar 2011. [↑43]

[32] G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," *Advances in Neural Information Processing Systems*, vol. 21, pp. 225–232, 2009. [↑43]

[33] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8. [↑116]

[34] C. Cedras and M. Shah, "Motion-based recognition a survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, 1995. [↑84]

[35] Y. Chai, V. Lempitsky, and A. Zisserman, "Bicos: A bi-level co-segmentation method for image classification," 2011. [↑116]

[36] Y. Chai, E. Rahtu, V. Lempitsky, L. Van Gool, and A. Zisserman, "Tricos: A tri-level class-discriminative co-segmentation method for image classification," in *Com-*

*puter Vision–ECCV 2012*. Springer, 2012, pp. 794–807. [↑116]

[37] C.-I. Chang and S.-S. Chiang, "Anomaly detection and classification for hyperspectral imagery," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 40, no. 6, pp. 1314–1325, 2002. [↑28], [↑35]

[38] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011. [↑78]

[39] M. Chen, S. Pang, T. Cham, and A. Goh, "Visual tracking with generative template model based on riemannian manifold of covariances," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8. [↑12], [↑39], [↑93], [↑98], [↑101]

[40] Y. Chen, A. Wiesel, and A. O. Hero, "Shrinkage estimation of high dimensional covariance matrices," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 2937–2940. [↑28], [↑38]

[41] ——, "Robust shrinkage estimation of high-dimensional covariance matrices," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4097–4107, 2011. [↑28], [↑43]

[42] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 409–416. [↑121], [↑125], [↑126]

[43] E. C. Chi and K. Lange, "Stable estimation of a covariance matrix guided by nuclear norm penalties," *Computational statistics & data analysis*, vol. 80, pp. 117–128, 2014. [↑40], [↑43]

[44] R. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *PAMI*, vol. 27, no. 10, pp. 1631–1643, 2005. [↑84]

[45] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002. [↑118]

[46] R. Couillet, F. Pascal, and J. W. Silverstein, "The random matrix regime of Maronna's $M$-estimator with elliptically distributed samples," *arXiv:/1311.7034*, 2013. [↑34]

[47] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001. [↑84]

[48] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997. [↑15]

[49] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004. [↑68]

[50] C. E. Davidson and A. Ben-David, "Performance loss of multivariate detection algorithms due to covariance estimation," in *SPIE Europe Remote Sensing*, 2009, pp. 74 770–74 770. [↑32]

[51] ——, "On the use of covariance and correlation matrices in hyperspectral detection," in *Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2011, pp. 1–6. [↑35]

[52] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72. [↑6]

[53] D. L. Donoho *et al.*, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1–32, 2000. [↑1]

[54] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003. [↑24]

[55] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *Multimedia, IEEE Transactions on*, vol. 7, no. 6, pp. 1066–1083, 2005. [↑12]

[56] P. E. Duda and O. Richard, "Hart, pattern classification and scene analysis," 1973. [↑21]

[57] A. Edelman and N. R. Rao, "Random matrix theory," *Acta Numerica*, vol. 14, no. 1, pp. 233–297, 2005. [↑31]

[58] B. Efron and C. Morris, "Data analysis using Steinś estimator and its generalizations," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 311–319, 1975. [↑36]

[59] ——, "Multivariate empirical bayes and estimation of covariance matrices," *The Annals of Statistics*, pp. 22–32, 1976. [↑41], [↑43]

[60] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004. [↑19]

[61] J. P. Egan, *Signal Detection and ROC Analysis*. Academic Press, 1975. [↑44], [↑45]

[62] G. Frahm, "Generalized elliptical distributions: theory and applications," Ph.D. dissertation, Universität zu Köln, 2004. [↑33], [↑34]

[63] J. Frontera-Pons, M. Mahot, J. Ovarlez, and F. Pascal, "Robust Detection using M-estimators for Hyperspectral Imaging," in *4th IEEE GRSS Workshop on Hyperspectral*

*Image and Signal Processing: evolution in remote sensing (WHISPERS 2012)*. IEEE, 2012. [↑28]

[64] M. Grabner, H. Grabner, and H. Bischof, "Learning features for tracking," in *CVPR*. IEEE, 2007, pp. 1–8. [↑84]

[65] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012. [↑62], [↑66]

[66] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003. [↑1], [↑17]

[67] I. Guyon, J. Weston, S. Barnhill, and V. N. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, Jan. 2002. [↑15], [↑16]

[68] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999. [↑16]

[69] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems*, vol. 18, p. 507, 2006. [↑17], [↑59], [↑75]

[70] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing systems*, vol. 16, 2004, p. 153. [↑57]

[71] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507–514. [↑16]

[72] M. Hein and T. Bühler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA," *arXiv preprint arXiv:1012.0774*, 2010. [↑71]

[73] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 269–276. [↑116]

[74] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 763–767, 1996. [↑28], [↑37]

[75] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 129–162, 2004. [↑12], [↑135]

[76] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *Transaction on Cybernetic*, 2013. [↑17]

[77] K. Ito, "The brownian motion and tensor fields on riemannian manifold," *Kiyosi Itō selected papers*, p. 298, 1987. [↑98]

[78] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," *ECCV 2002*, pp. 439–443, 2006. [↑84]

[79] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *PAMI*, vol. 25, no. 10, pp. 1296–1311, Oct. 2003. [↑88], [↑90]

[80] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image and vision computing*, vol. 14, no. 8, pp. 609–615, 1996. [↑114], [↑135]

[81] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104,

no. 486, 2009. [↑57]

[82] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1943–1950. [↑116], [↑117], [↑125]

[83] ——, "Multi-class cosegmentation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 542–549. [↑116]

[84] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, 2010. [↑71]

[85] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1409–1422, 2012. [↑134]

[86] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, vol. 1. IEEE, 1998, pp. 413–418. [↑22]

[87] J. Kelley, "The cutting-plane method for solving convex programs," *Journal of the Society for Industrial and Applied Mathematics*, pp. 703–712, 1960. [↑64]

[88] N. Keshava and J. F. Mustard, "Spectral unmixing," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 44–57, 2002. [↑48]

[89] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964. [↑103]

[90] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 459–473. [↑119], [↑120]

[91] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the $t$ distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989. [↑36], [↑43], [↑55]

[92] O. Ledoit and M. Wolf, "Honey, i shrunk the sample covariance matrix," *UPF Economics and Business Working Paper*, no. 691, 2003. [↑36], [↑43]

[93] ——, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004. [↑37]

[94] C. Liu and D. B. Rubin, "ML estimation of the $t$ distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995. [↑36]

[95] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. Chapman and Hall/CRC, 2007. [↑78]

[96] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining." *J. Mach. Learn. Res.-Proceedings Track*, vol. 10, pp. 4–13, 2010. [↑17]

[97] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 4, pp. 491–502, Apr. 2005. [↑16]

[98] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2097–2104. [↑118]

[99] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007. [↑11]

[100] Y. Liu, F. Nie, J. Wu, and L. Chen, "Efficient semi-supervised feature selection with noise insensitive trace ratio criterion," *Neurocomputing*, vol. 105, pp. 12–18, 2013. [↑60]

[101] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proceedings of the 15$^{th}$ international conference on Multimedia*, ACM. ACM Press, 2007, pp. 301–304. [↑16]

[102] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936. [↑29]

[103] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," 2007. [↑115], [↑116]

[104] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo, "Performance evaluation of object detection and tracking in video," *Computer Vision–ACCV 2006*, pp. 151–161, 2006. [↑106]

[105] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 79–116, 2003. [↑52]

[106] V. A. Marv̈cenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Sbornik: Mathematics*, vol. 1, no. 4, pp. 457–483, 1967. [↑31]

[107] R. Maronna, "Robust $M$-estimators of multivariate location and scatter," *The Annals of Statistics*, pp. 51–67, 1976. [↑33]

[108] S. Matteoli, M. Diani, and G. Corsini, "Improved estimation of local background covariance matrix for anomaly detection in hyperspectral images," *Optical Engineering*,

vol. 49, no. 4, pp. 1–16, 2010. [↑28]

[109] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *PAMI*, vol. 26, pp. 810–815, 2004. [↑85], [↑86], [↑88], [↑89], [↑90]

[110] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2004, vol. 544. [↑57]

[111] F. Meng, J. Cai, and H. Li, "On multiple image group cosegmentation," in *ACCV*, vol. 1, no. 1, 2014, pp. 1–15. [↑119]

[112] R. Menon, P. Gerstoft, and W. Hodgkiss, "Asymptotic eigenvalue density of noise covariance matrices," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3415–3424, 2012. [↑41]

[113] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996. [↑36]

[114] A. Mutapcic and S. Boyd, "Cutting-set methods for robust convex optimization with pessimizing oracles," *Optimization Methods & Software*, vol. 24, no. 3, pp. 381–406, 2009. [↑64]

[115] R. Nadakuditi and J. W. Silverstein, "Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples," *IEEE Journal of Sel. Topics in Signal Processing*, vol. 4, no. 3, pp. 468–480, 2010. [↑41]

[116] S. Nadarajah and S. Kotz, "Estimation methods for the multivariate $t$ distribution," *Acta Applicandae Mathematicae*, vol. 102, no. 1, pp. 99–118, 2008. [↑36]

[117] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geosc. and Rem. Sens.*,

vol. 43, no. 4, pp. 898–910, 2005. [↑52]

[118] ——, "Hyperspectral unmixing based on mixtures of Dirichlet components," *IEEE Transactions on Geosc. and Rem. Sens.*, vol. 50, no. 3, pp. 863–878, 2012. [↑48]

[119] H. Nguyen, M. Worring, and R. Van Den Boomgaard, "Occlusion robust adaptive template tracking," in *ICCV*, vol. 1. IEEE, 2001, pp. 678–683. [↑88]

[120] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proceedings of the 23rd national conference on Artificial intelligence*, vol. 2, 2008, pp. 671–676. [↑18], [↑78]

[121] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec. 2008. [↑124]

[122] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001. [↑123]

[123] Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 7, pp. 989–993, 2008. [↑94]

[124] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013. [↑67], [↑68]

[125] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005. [↑16], [↑78]

[126] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 1, pp. 127–154, 2006. [↑38]

[127] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *IJCV*, vol. 66, pp. 41–66, 2006. [↑96]

[128] F. Porikli and T. Kocak, "Robust license plate detection using covariance descriptor in a neural network framework," in *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*. IEEE, 2006, pp. 107–107. [↑94]

[129] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *CVPR*, vol. 1, Jun. 2006, pp. 728–735. [↑85], [↑88], [↑94]

[130] I. S. Reed, J. D. Mallett, and L. E. Brennan, "Rapid convergence rate in adaptive arrays," *Aerospace and Electronic Systems, IEEE Transactions on*, no. 6, pp. 853–863, 1974. [↑32]

[131] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct 1990. [↑28]

[132] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004. [↑97]

[133] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental Learning for Robust Visual Tracking," *IJCV*, vol. 77, no. 1, pp. 125–141, 2008. [↑85], [↑88], [↑91], [↑92], [↑93], [↑108]

[134] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 177–

186, 2009. [↑11]

[135] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000. [↑24], [↑57]

[136] H. Samet, *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006. [↑5]

[137] A. Schwartzman, "Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data," Ph.D. dissertation, Stanford University, 2006. [↑99]

[138] D. W. Scott and J. R. Thompson, "Probability density estimation in higher dimensions," in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, vol. 528. North-Holland, Amsterdam, 1983, pp. 173–179. [↑9]

[139] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000. [↑118]

[140] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (pie) database," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 46–51. [↑80]

[141] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 12, pp. 1371–1375, 1998. [↑12]

[142] C. Stein, "Estimation of a covariance matrix," *Rietz Lecture*, 1975. [↑37]

[143] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 58–69, 2002. [↑28]

[144] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *Intelligent Vehicles Symposium, 2006 IEEE*. Ieee, 2006, pp. 206–212. [↑85]

[145] L. Sun, S. Ji, and J. Ye, "A least squares formulation for a class of generalized eigenvalue problems in machine learning," in *Proceedings of the 26$^{th}$ Annual International Conference on Machine Learning*. ACM, 2009, pp. 977–984. [↑60], [↑61]

[146] Y. Sun, P. Babu, and D. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143–5156, 2014. [↑28], [↑38]

[147] M. Tan, I. W. Tsang, and L. Wang, "Towards ultrahigh dimensional feature selection for big data," *Journal of Machine Learning Research*, vol. 15, pp. 1371–1429, 2014. [↑67], [↑68]

[148] M. Tan, L. Wang, and I. W. Tsang, "Learning sparse svm for feature selection on very high dimensional datasets," in *Proceedings of the 27$^{th}$ International Conference on Machine Learning*, 2010, pp. 1047–1054. [↑62], [↑63], [↑66]

[149] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. [↑17], [↑24], [↑57]

[150] J. Theiler, "Ellipsoid-simplex hybrid for hyperspectral anomaly detection," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2011. [↑45]

[151] ——, "By definition undefined: Adventures in anomaly (and anomalous change) detection," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2014. [↑29], [↑45]

[152] J. Theiler, G. Cao, L. Bachega, and C. Bouman, "Sparse matrix transform for hyperspectral image processing," *IEEE Journal of Sel. Topics in Signal Processing*, vol. 5, no. 3, pp. 424–437, 2011. [↑28], [↑37], [↑43]

[153] J. Theiler and D. Hush, "Statistics for characterizing data on the periphery," in *International Geoscience and Remote Sensing Symposium*. IEEE, 2010, pp. 4764–4767. [↑45]

[154] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996. [↑18]

[155] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008. [↑5], [↑6]

[156] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *ECCV*, pp. 589–600, 2006. [↑94]

[157] ——, "Human detection via classification on riemannian manifolds," in *CVPR*, Jun. 2007, pp. 1–8. [↑93], [↑94]

[158] ——, "Region Covariance: A Fast Descriptor for Detection and Classification," in *ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin / Heidelberg, 2006, vol. 3952, pp. 589–600. [↑94]

[159] D. E. Tyler, "A distribution-free $M$-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987. [↑28], [↑33], [↑34], [↑43]

[160] S. Velasco-Forero and J. Angulo, "Classification of hyperspectral images by tensor modeling and additive morphological decomposition," *Pattern Recognition*, vol. 46, no. 2, pp. 566–577, 2013. [↑45]

[161] S. Velasco-Forero, M. Chen, A. Goh, and S. K. Pang, "A comparative analysis of covariance matrix estimation in anomaly detection," in *IEEE Workshop on Hyperspectral Image and Signal Processing*, 2014. [↑36]

[162] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 465–479. [↑116]

[163] V. Q. Vu, J. Cho, J. Lei, and K. Rohe, "Fantope projection and selection: A near-optimal convex relaxation of sparse PCA," in *Advances in Neural Information Processing Systems*, 2013, pp. 2670–2678. [↑19], [↑57], [↑70], [↑71]

[164] J. Wang, K. Markert, and M. Everingham, "Learning models for object recognition from natural language descriptions," in *Proceedings of the British Machine Vision Conference*, 2009. [↑124]

[165] A. Wiesel, "Unified framework to regularized covariance estimation in scaled gaussian models," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 29–38, 2012. [↑38]

[166] J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam, "Condition-number-regularized covariance estimation," *Journal of the Royal Statistical Society: Series B*, vol. 75, no. 3, pp. 427–450, 2013. [↑28], [↑40], [↑43]

[167] Y. Wu, H. Ling, E. Blasch, L. Bai, and G. Chen, "Visual tracking based on log-Euclidean riemannian sparse representation," *Advances in Visual Computing*, pp. 738–747, 2011. [↑85], [↑94]

[168] D. Xu and S.-F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1985–1997, 2008. [↑12]

[169] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 1, pp. 40–51, Jan. 2007. [↑22], [↑57], [↑59], [↑60]

[170] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 2006. [↑84]

[171] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *arXiv preprint arXiv:1112.2679*, 2011. [↑71]

[172] Y. Zhai, Y. Ong, and I. Tsang, "The emerging "big dimensionality"," *Computational Intelligence Magazine, IEEE*, vol. 9, no. 3, pp. 14–26, 2014. [↑2]

[173] X. Zhan, "Extremal eigenvalues of real symmetric matrices with entries in an interval," *SIAM journal on matrix analysis and applications*, vol. 27, no. 3, pp. 851–860, 2006. [↑100]

[174] T. Zhang, X. Cheng, and A. Singer, "Marchenko-pastur law for Tyler's and Maronna's $M$-estimators," *arXiv:1401.3424*, 2014. [↑34]

[175] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24$^{th}$ International Conference on Machine learning*. ACM, 2007, pp. 1151–1157. [↑17]

[176] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," *ASU Feature Selection Repository*, 2010. [↑78]

[177] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, University of Wisconsin-Madison*, vol. 2, p. 3, 2006. [↑60]

[178] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006. [↑62], [↑71], [↑74]

[179] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006. [↑18]

[180] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. [↑18]