Sparse Support Matrix Machines for the Classification of Corrupted Data

Muhammad Imran Razzak

School of Computer Science
Faculty of Engg. & IT
University of Technology Sydney
NSW - 2019, Australia

Sparse Support Matrix Machines for the Classification of Corrupted Data

A thesis submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Muhammad Imran Razzak

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

November 2019

AUTHOR'S DECLARATION

I, *Muhammad Imran Razzak* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

[Muhammad Imran Razzak]

DATE: 22nd November, 2019

PLACE: Sydney, Australia

Acknowledgments

I would like to express my sincere gratitude to Guandong Xu, for providing continuous support and motivation through out PhD.

Dedication

To my family ...

List of Publications

Published/Accepted

1. Imran Razzak, Raghib Abu Saris, Michael Blumenstein, and Guandong Xu, "Integrating Joint Feature Selection into Subspace Learning: A Formulation of 2DPCA for Outliers Robust Feature Selection." Neural Networks (Elsevier), doi.org/10.1016/j .neunet.2019.08.030, **[Core rank: A/JCR IF. 7.197/Q1]** [80] .

2. Imran Razzak, Michael Blumenstein, Guandong Xu, Multi-Class Support Matrix Machines by Maximizing the Inter-class Margin for Single Trial EEG Classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering, doi: 10.1109/TNSRE.2019.2913142 **[Core rank: A*/JCR IF. 3.478/Q1]**[73].

3. Imran Razzak, Raghib Abu Saris, Michael Blumenstein, and Guandong Xu. "Robust 2D Joint Sparse Principal Component Analysis With F-Norm Minimization For Sparse Modelling: 2D-RJSPCA." International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-7 **[Conference core rank: A]** [85].

4. Imran Razzak, Ibrahim A Hameed, Michael Blumenstein and Guandong Xu, "Sparse Representation and Support Matrix Machines for Epileptic EEG Signal Classification" Journal of Translational Engineering in Health and Medicine, doi:10.1109/JTEHM.2019.2942017 **[JCR IF. 2.05/Q2]** [74].

5. Imran Razzak, Muhammad Imran, Guandong Xu, Efficient Brain Tumor Segmentation with Multiscale Two-Pathway-Group Conventional Neural Networks, IEEE Journal of Biomedical and Health Informatics, 2018 Oct 4. doi: 10.1109/JBHI.2018.287403 **[Core rank: A*/JCR IF. 4.217/Q1]** [76].

6. Imran Razzak, Muhammad Imran, Guandong Xu, "Big Data Analytics for Preventive Medicine", Neural Computing and Application, DOI https://doi.org/10.1007/s00521-019-04095-y **[JCR IF. 4.664/Q1]** [82].

7. Zafar Saeed, Rabeeh Abbasi, Imran Razzak, Guandong Xu, "Event Detection in Twitter Stream using Weighted Dynamic Heartbeat Graph Approach", IEEE Computational Intelligence Magazine, 14(3): 29-38 (2019) **JCR IF. 5.857/Q1]**[92]

8. Zafar Saeed, Rabeeh Ayaz Abbasi,Imran Razzak, Onaiza Maqbool, Abida Sadaf, Guandong Xu, "Enhanced Heartbeat Graph based Temporal Networks for Emerging Event Detection on Twitter" Vol, 136,pp 115,Äì132 , Expert System with Applications, **[JCR IF. 4.292/Q1]**[91]

9. Zafar Saeed, Rabeeh Abbasi, Imran Razzak, Guandong Xu, Text Stream to Temporal Network - A Dynamic Heartbeat Graph to Detect Emerging Events on Twitter, 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2018. **(Conference core rank: A)**[93]

10. Amina Naseer, Monahil Rani, Saeeda Naz, Muhammad Imran Razzak, Muhammad Imran, Guandong Xu, "Refining Parkinson's Neurological Disorder Identification Through Deep Transfer Learning" , Neural Computing and Applications", doi.org/10.1007/s00521-019-04069-0 **[JCR IF. 4.664/Q1]**[53]

11. Zafar Saeed, Rabeeh Ayaz Abbasi, Maqbool Onaiza, Sadaf Abida, Muhammad Imran Razzak, Daud Ali, Naif Radi Aljohani, Guandong Xu. "Whats Happening around the World? A Survey and Framework on Event Detection Techniques on Twitter", Journal of Grid Computing, Springer, doi.org/10.1007/s10723-019-09482-2 **[JCR IF. 3.288/Q1]** [90]

SUBMITTED/REVISION

12. Imran Razzak, Muhammad Imran, Michael Blumenstein, Guandong Xu, Robust Sparse Support Matrix Machines for Single Trial EEG Classification, Artificial Intelligence in Medicine *Minor Revision***[Core rank: A/JCR IF. 3.574/Q1]**[75]

13. Imran Razzak, Michael Blumenstein, Guandong Xu, Robust Support Matrix Machine for Classification of Corrupted Data, IEEE Transactions on Pattern Analysis and Machine Intelligence, *In Revision*, **[Core rank: A*/JCR IF. 17.730/Q1]** [78]

14. Imran Razzak, Muhammad Khurram, Muhammad Imran, Michael Blumenstein, and Guandong Xu, Randomized Nonlinear One-Class Support Vector Machines with Bounded Loss Function for Outliers Detection, Future Generation Computer Systems, *Minor Revision* **[Core rank: A/JCR IF. 5.768/Q1]** [33].

15. Imran Razzak, Raghib Abu Saris, Michael Blumenstein, Guandong Xu, Robust Two Dimensional Joint Sparse PCA with F-norm Minimization, IEEE Transactions on Image Processing **[Core rank: A*/JCR IF. 6.79/Q1]** [81]

16. Imran Razzak, Michael Blumenstein, Guandong Xu, Support Matrix Machine via Joint L2 and Nuclear Norm Minimization Under Matrix completion Framework for Classification of Corrupted Data, IEEE Transactions on Pattern Analysis and Machine Intelligence, **[Core rank: A*/JCR IF. 17.730/Q1]**[79]

17. Imran Razzak, Khurram Zafar, Michael Blumenstein, Guandong Xu, One-Class Support Tensor Machines with Bounded Hinge Loss Function for Classification of High-dimensional Data, Future Generation Computer Systems, *Minor Revision* **[Core rank: A/JCR IF. 5.768/Q1]** [77].

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Data acquisition has improved substantially over recent years, with devices acquiring data at faster rates and increased resolution. The interpretation process, however, has only recently begun to benefit from computer technology and still struggling especially for high dimensional and noisy data. We are still short of tools to convert all such data to useful information. Traditional support vector machines (SVMs) require data to reshape each matrix into a vectors, which ultimately results in losing the important structural information of the originally featured matrix. On the other-hand, the classification of high dimensional domains poses significant challenges. In contrast, modern classification approaches such as support matrix machine assume that all entities within each input matrix can serve as the explanatory features for its label. These methods are able to capture explanatory features by regularizing the regression matrix to be low-rank. However, in real-world, the data is noisy and most of the features may be redundant as well as may be useless, which in turn affect the classification performance. Thus it is important to perform robust feature selection under robust metric learning to filter out redundant features and ignore the noisy data points for more interpretable modelling. To overcome this challenge, in this work, we have adapted two different approaches. The first problem we address is the issue of dimensionality reduction. In our first approach, we introduce two-dimensional outliers-robust principal component analysis (ORPCA) by imposing the joint constraints on the objective function **(chapter 4)**. ORPCA relaxes the orthogonal constraints and penalizes the regression coefficient, thus, it selects most important features and in the meantime, it ignores the same features that have already been selected in other principal components. To overcome the data redundancy, we further extend ORPCA and introduced additional sparsity-inducing regularization that relaxes the orthogonal constraints resulting the joint features selection **(chapter 5)**. The introduced regularization terms penalizes all regression coefficients corresponding to single feature as a whole to features jointly. Hence, 2D-JSPCA approximates to high-dimensional data in flexible manner as it has more freedom to learn low-dimensional space efficiently.

Since the nuclear norm is the best convex approximation of the matrix rank over the unit ball of matrices, this makes it more tractable to solve the resulting optimization problem. Inspired by this, in our second approach, we propose a new model to address the classification problem of high dimensionality data by jointly optimizing the both regularizer terms ($||.||_{2,1}$ and $||.||_*$) and hinge loss. In our first approach **(chapter 6)**, we combine the hinge loss and regularization terms as spectral elastic net penalty. The regulariza-

tion term which promotes the structural sparsity and shares similar sparsity patterns across multiple predictors. It is a spectral extension of the conventional elastic net that combines the property of low-rank and joint sparsity together, to deal with complex high dimensional noisy data. Furthermore, it also leverages the structural information as well as the intrinsic structure of data and avoids the inevitable upper bound. The optimization problem for the RSMM is convex, non-smooth and non-differentiable, however, the combination of hinge loss, $\ell_{2,1}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To tackle this issue, we split the problem into sub-problems with the *Generalized Forward-Backward* (GFB) splitting approach to solve the optimization problem efficiently.

Support matrix machine is fragile to the presence of outliers: even few corrupted data points can arbitrarily alter the quality of the approximation, What if a fraction of columns are corrupted? Combining the recovery along with feature selection and classification could significantly improve the performance. We assume that the data consists of a low rank clean matrix plus a sparse noise matrix. We extended our work and present support matrix machine **(chapter 7)** based on matrix recovery framework under the incoherence and ambiguity conditions and able to recover intrinsic matrix of higher rank and recover data with much denser corruption. We perform matrix recovery, feature selection and classification through joint minimization of $\ell_{2,1}$ and nuclear norm. We assume that the data consists of a low rank clean matrix plus a sparse noise matrix i.e. the data matrix can be decomposed as $X = L + S$. S is the column-sparse matrix that corresponds to corrupted columns, thus at most αn columns are non zeros, L corresponds to non corrupted matrix, thus $rank(L) = r$ and $(1-\alpha)n$ columns of matrix L are non zeros, corresponding to the outliers. Since the objective function is convex, non-smooth and non-differentiable, however, the combination of hinge loss, $\ell_{2,1}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To decouple the hinge loss and nuclear norm with respect to W in SMMRe, we have introduced an *auxiliary variable, and applied Lagrange multiplier*.

Although, above both methods takes full advantage of low rank assumption to exploit the strong correlation between columns and rows of each matrix and able to extract useful features, however, are originally built for binary classification problems. To improve the robustness against data that is rich in outliers, we further extend this problem and present a novel multiclass support matrix machine **(chapter 8)** by utilizing the maximization of the inter-class margins (i.e. margins between pairs of classes). The proposed model is a combination of binary hinge loss for models fitting, and elastic net penalty as a regularization on regression matrix. The binary hinge loss uses C matrices to simulate one-vs-one classifier of all classes rather than $\frac{c(c-1)}{2}$ models. The optimization problem is convex but non-smooth and non-differentiable, thus, stochastic gradient descent and the Nesterov methods cannot be applied (i.e. in convex optimization setting, sub gradient of the nuclear norm function cannot be used in standard descent approaches and as a result solving it directly is difficult). Thus, an alternative approach is required to solve it, we devise an alternating direction method (*GFB splitting*) that can handle an arbitrary non-differentiable with a proximal operator.

Several non-convex and bounded loss function has been presented to substitute the

hinge loss function in order to suppress the affect of outliers and improve the robustness of support vector machines. However, there is no work done for the improvement of one-class tensor machines. Furthermore, computational complexity of traditional support tensor machines is high and increases with the increase of training samples. Thus, it limits the applicability of OCSTM for large dataset. We consider one class support tensor machines and introduce a scalable algorithm for large dataset by replacing the traditional hinge loss with bounded loss function resulting in reduction of classification error caused by outliers **(chapter 9)**. For larger dataset, we further used randomized features rather than finding the optimized support tensors which results in not only improving the robustness against outliers as well as significantly reduces the training time. To solve the corresponding optimization problem, we have presented *half quadratic optimization* to transform the objective function to same like traditional OCSTM, followed by solving it like a typical OCSTM optimization problem.

We demonstrate the significance and advantage of our methods on different available benchmark datasets such as person identification, face recognition and EEG classification. Results showed that our methods achieved significantly better performance both in terms of time and accuracy for solving the classification problem of highly correlated matrix data as compared to state-of-the-art methods.

CHAPTER 1

INTRODUCTION

It is a capital mistake to theorize before one has data.

S. Holmes

Classification is one of the major fields in machine learning and pattern recognition with the aim to identify which set of entities belong to which class based on the set of seen observations. The classification is the process of constructing the decision boundary between classes (also called targets/ labels or categories) based on a set of a-priori known examples that helps to predict the class (Y) of unknown input sample (X) [82]. Some of the examples of classification are diagnosis of a disease in a patient based on the patient data (EEG signal, vital signs, gender, presence of certain symptoms, etc.), assigning a given email to the either "spam" or "non-spam" class based on its features (count of each word in the email, sender country, sender IP etc.), identifying person based on gesture or handwriting. There are two types of classification binomial (binary) and multiclass. Email classification into spam or non-spam is a binary class problem since there are only two target classes (spam or not spam). In this case, aim of the classifier is to segregate the new emails into a spam or not-spam emails. The classifier uses some trained data that consist of both not-spam and spam examples to learn how the given set of input variables relates to a particular class. Once the classifier is trained on both examples, it is able to classify unknown emails into spam and not spam emails.

Data analyst uses different types of machine learning methods to discover the hidden patterns in the data that provide actionable insight into the data. These methods

are classified into two groups based on the way they learn to predict something, are supervised learning and unsupervised learning. Support vector machines are one of the simplest and commonly used supervised machine learning algorithms for both classification or regression. It constructs a hyperplane or set of hyperplanes by implicitly mapping the training data into high dimensional or infinite-dimensional space, that can be used for data classification. Intuitively, we can achieve good separation by the only hyperplane that has the largest distance to the nearest data point of any class, since in general, larger the margin results lower generalization error of the classifier. Support vector machines construct hyperplane by bisecting the two classes in a way that maximizes the margins of separation.

With the advent of recent data acquisition devices, generally, data is diverse, noisy, and high dimensional in real-world applications such as face recognition, hyperspectral image classification, action recognition, and object categorization, whereas the underlying structure in many cases is based on a small set of features, hence poses several challenges. This complex nature of data poses a serious challenge especially with data of limited size. The data has to be reshaped into vectors for classification [10, 87, 88] which could ultimately destroy the structural information embedded in e.g. spatial relationship of a neighboring pixel in an image, that is a very important factor for certain classification tasks. Representation of such data in the form of a matrix can preserve its structural information i.e. EEG signals which consist of voltage fluctuations at several electrodes during a time period, has a strong correlation with respect to certain frequency band and channels. Furthermore, reshaping of high dimensional data to vector results in an increase in dimensionality [141].

Vector-based methods have been successively applied for the classification and shown good results. State of the art vector-based methods are linear discriminant analysis (LDA) [67, 104, 130], support vector machines (SVM) [18, 31, 131], K nearest neighbor (KNN) [17, 41, 56, 57]. For these methods, the data has to be reshaped into vectors for classification purpose which could in-turn destroy the structural information embedded in. An alternative solution for this problem is to concatenate the matrix into a vectors for classification. However, it results in an increase in dimensionality that leads to model over-fitting. Recently, some efforts have been made to suppress the matrix into vectors using common spatial patterns [2, 35, 39, 45, 107, 128]. Most of these methods ignore the topological structure embedded in the matrix data, whereas considering structural information is of great interest and helps to improve the classification. Moreover, one of the major disadvantages of these methods is that each new feature in a low-dimensional

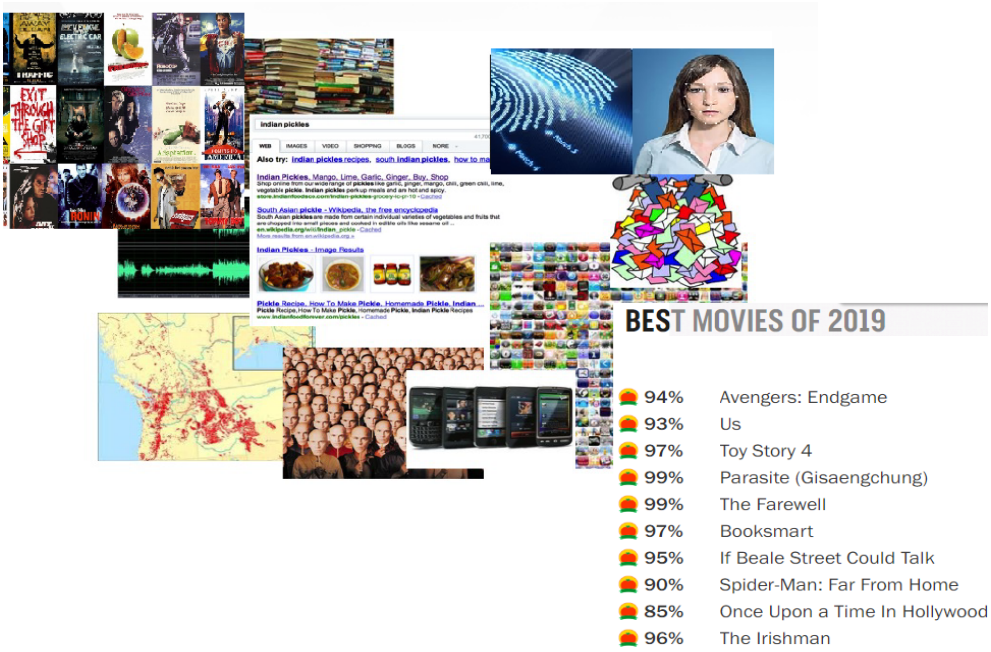

Figure 1.1: Real-world data in the form of matrix

subspace is the linear combination of all the original features in high-dimensional space. Thus, usually, it affects the classification performance due to the redundant features. Besides, it is often difficult to interpret new features.

The focus of this work is laid on solving the problem of high dimensional and noisy data classification. In this thesis, we address both of these challenges (dimensionality reduction and classification) for the analysis of high dimensional and corrupted data. In the following sections, we first present an introduction to the field of knowledge, basic problem with classification of high dimensional and noisy data followed by the motivation and aim behind this work. Following on from this, specific research questions aligned with the aims are presented to address the research and guide the investigation through certain objectives. Finally, given a separate list of key contributions, a high-level overview of each chapter is shown through a relational map as shown in figure 1.2, which illustrates how the research progress has been carried out through linking the contributions.

1.1 Background

Generally, the data is corrupt and high in dimension in the real-world. The complex nature of such data poses some serious challenges for its dimensionality reduction and

classification, especially with data of limited size. Such persistent or non-probabilistic data corruption may stem from failures of the sensor or malicious tampering. In addition to the corruption, some of the available data may not conform to the presumed low-dimensional model i.e. most of the columns are in low dimensional space, thus the corresponding matrix is low rank and a small number of columns are the outliers that correspond to column-sparse matrix [73]. Thus, usually, this type of data significantly affects the classification performance due to the redundant features and extensive noise.

Feature selection, the process of selecting the subset of discriminant patterns, is the key component for any machine learning problem, aiming to identify, to which set of categories, a new unseen observation belongs on the basis of a training set of data containing known observations. It plays an important role in many classification applications, as it does not only help to improve the classification performance but also speeds up the learning process, improves the generalization capability and alleviates the effect of the curse of dimensionality [95]. Conventional dimensional reduction methods such as PCA, LDA, etc. could be used for dimensionality reduction, however, they do not solve the problems as the features have natural meanings and cannot be projected.

Recently, several efforts have been made to classify the matrix directly without converting it into respective vectors, thus, exploiting the correlation between the columns or rows of matrix. Rank-k SVM models the regression matrix as a sum of k rank-one orthogonal matrix [119]. Pirsiavash et.al presented a bilinear classifier by applying the hinge loss for model fitting through factorization of the regression matrix into a low-rank matrix [65]. Zhang et. al. devised low-rank linearization to transform the non-linear SVM to corresponding linear SVM, through kernel map computed from the low-rank approximation of matrices [129]. One major disadvantage of these methods is that each new feature in a low-dimensional subspace is the linear combination of all the original features in high-dimensional space. Thus, usually, it affects the classification performance due to the redundant features. Besides, it is often difficult to interpret new features.

It is no surprise that most of the real-world data have such a high sparsity, i.e., only a small number of features are important for spam detection. An ad-hoc approach to deal with such problems is achieving the sparsity artificially by considering only those loadings that are greater than the threshold. however, in general, it is an inefficient approach. To tackle the challenge of robust feature selection, recently, the sparsity regularization in dimensionality reduction has been widely investigated for feature selection i.e. ℓ_1 [44, 102], ℓ_q [95], $\ell_{2,0}$ [60], $\ell_{2,1}$ [9]. Frobenius norm [37] has also been

applied to introduce the sparsity property in a regression matrix. These approaches work well and consider the correlation between columns and rows under the low-rank assumptions and provide satisfactory performance [142]. However, these approaches consider all the entities of the matrix as an explanatory factors, whereas in the real-world, features might be redundant or useless for certain classification tasks and only a small set of useful features could be used to classify the unseen data. For example, the low-rank nature of gene or human facial images, obtaining relevant features by removing irrelevant and redundant ones reduces the computational costs without significant loss of information or negative degradation of the learning performance.

It turns out that the nuclear norm can also be used as a convex relaxation of this optimization problem, which greatly simplifies the problem and allows further room for interesting applications such as accelerated algorithms for matrix completion (compressed sensing). Recently, classifier based on combination of hinge loss, nuclear norm and Frobenius norm [1, 46, 142], ℓ_1 [140, 141] has been presented. Although these methods showed excellent performance by taking advantage of the correlation between rows and columns of the regression matrix under the low-rank assumptions. But, they simply consider entities in the matrix as explanatory factors and do not consider the intrinsic group structure of data and are sensitive to outliers. Furthermore, they also tend to select the features without considering all classes.

1.2 Motivation

As discussed above, the data is noisy and high in dimension. Existing approaches could not deal well with nonlinear, high dimensional and noisy data efficiently. In result, it is quite difficult to explain the resulting features i.e. projection procedure involves all the original features and it may have redundant or irrelevant features. Furthermore, the outliers and non-standard noises make it a challenging task. For classification of high dimensional data, not only dimensional reduction but also important to find salient features that belong to specific part of image as projection procedure involves all the original features and it may have redundant or irrelevant features. To select such salient patterns, the projection matrix should consist of a sparse element with respect to such features. Thus, modeling sparsity into a support vector machine could help to encode semantic information, as well. An alternative way to model sparsity is an ad-hoc approach to deal achieve sparsity artificially by considering the loadings greater than threshold only but it was inefficient. In this work, our concern is the classification

problems on a set of the data matrix as structural information of the original features is very important for certain data analytic tasks. Input data is high in dimensions and noisy, hence, the complexity of the data motivated us to pay attention to regularizers that have the ability to promote sparsity and robustness against outliers, so that they can be used for selecting certain features. Moreover, our target is to endow the feature space that does not penalize the features individually.

1.3 Aims

The aims of this work are to:

- As the real-world data is high in dimensional that has to be reshaped into vectors for classification which could ultimately destroy the structural information embedded in, which is a very important factor for certain classification tasks. Representation of such data in the form of a matrix can preserve its structural information. The main objective is to overcome challenges of high dimensional sensitive data where structural information is important factor.

- Data in real-world is noisy. To overcome this limitation, we aim to develop an efficient approach that simultaneously deals with outliers and selects useful features across all data points resulting in improvement classification performance.

1.4 Objectives

To achieve the aim, the key objective of this work is to develop a robust support vector machine for high dimensional and noisy data, that leverages the structural information within matrices and able to select useful features by avoiding redundancy and ignoring the outliers. To achieve our objectives, we adopted two approaches: (1) dimensionality reduction followed by support matrix machine (briefly explained in part-I); (2) joint dimensionality reduction and support matrix machine in one objective function (briefly described in part-II and Part II).

1.5 Research Question

This research is structured to answer the following research questions:

Q1: Reshaping the high dimensionality data into vectors ultimately destroys the structural information embedded in it that results in affecting the classification accuracy. How can this structural information be utilized to promote the classification performance?

Q2: In the real-world, features might be redundant, noisy or useless for certain classification tasks. Sparsity regularization could be used in dimensionality reduction. How do sparsity regularization terms help to improve the classification for corrupted and high dimensional data?

Q3: The outliers and non-standard noises make the classification task challenging. What are the appropriate methods for addressing the robustness (both in terms of accuracy and time) against outliers for efficient classification of corrupted and high dimensional data?

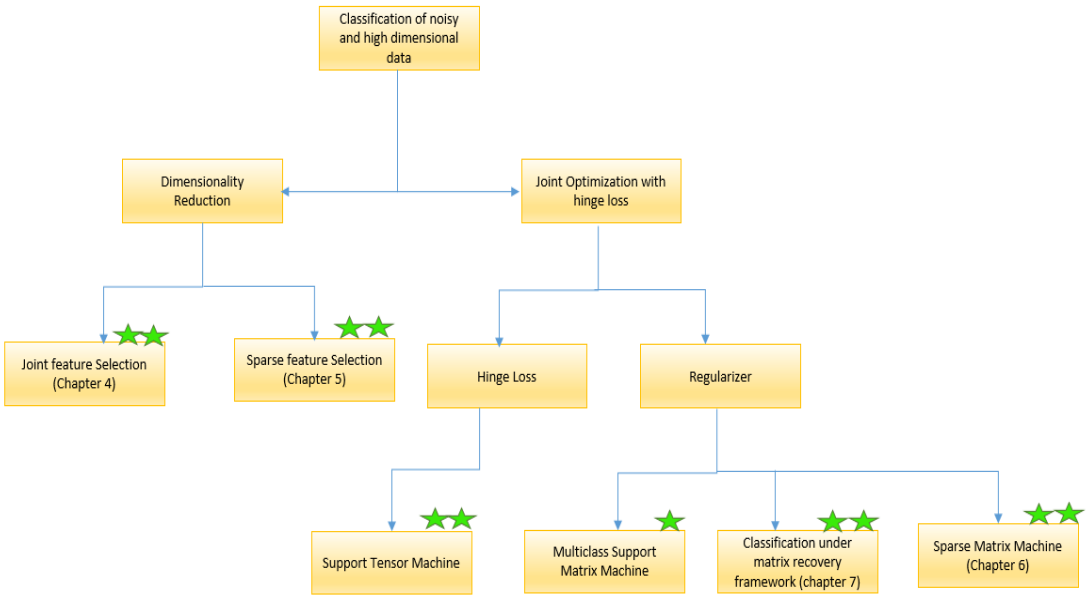

Figure 1.2: Organisation of thesis, key contributions (publications) are marked with ⋆

1.6 Organization and Contributions

Compared to the state-of-art dimensionality reduction and feature selection methods, we can describe the theoretical and empirical **key contributions** of this work as follows:

1. We proposed an effective feature extraction approach by effectively combining the robustness of 2DPCA and the sparsity-inducing lasso regularization that relaxes the orthogonal constraint that has more freedom to jointly select low-dimensional space features. Moreover, its joint sparse constraints to select features and learn the optimal transformation matrix simultaneously. This approach is described in chapter 4 and chapter 5.

2. We propose a novel classifier that works by effectively combining the hinge loss function for model fitting, and the elastic net penalty for regularization on the regression matrix. We achieve the goal stated above, by employing the regularizer term which promotes structural sparsity . The regularization term helps to avoid the inevitable upper bound for the number of selected features occurring in $\ell_{2,1}$-norm SVM. The linear combination of the nuclear norm, $\ell_{2,1}$ inherits the property of low-rank and sparsity together which not only helps to deal with outliers but also selects features across all data points with joint sparsity **(Q1 and Q2)**. Since the optimization is convex and one of the major challenges is, how to efficiently solve non-smooth optimization, we devised an efficient algorithm to solve the proposed objective function based on the Generalized Forward-Backward (GFB) splitting framework. The approach is described in chapter 6.

3. We propose a novel classifier effectively combining the hinge loss function for model fitting, low-rank matrix recovery and the elastic net penalty for regularization on the regression matrix. We performed a simultaneous matrix recovery and classification, which first performs matrix recovery followed by clean feature extraction and classification. SMMRe is able to classify data with denser corruptions ($L \leq \frac{C_r n}{log(n)}$ and $S \leq C_s n$, C_s and C_r are numerical constant) through exact recovery of intrinsic matrix of higher rank based on the incoherence conditions. Since the convex optimization cannot perform an exact recovery of the corrupted matrix, thus, we used an Oracle Problem for matrix recovery. As a result, convex optimization-based SSMRe performs correct matrix recovery as well as the identification of outliers, which improves the classification performance **(Q1 and Q2)**. We achieve the goal stated above, by employing the regularizer term (a combination of low rank and $\ell_{2,1)}$) which promotes structural sparsity and matrix recovery as well as selects features across all data points with joint sparsity. The low-rank matrix recovery helps to recover the unobserved entities as well as to avoid the inevitable upper bound for the number of selected features occurring in $\ell_{2,1}$-norm SVM. Since the

optimization is convex but non-smooth and one of the major challenges is, how to efficiently solve non-smooth optimization, we devised an efficient algorithm to solve the proposed objective function. The approach is described in chapter 7 in detail.

4. We present a novel classifier M-SMM which works by effectively combining the binary hinge loss function (to maximize the inter-class hyperplane margin for model fitting) and elastic net penalty (to promote low-rank plus sparsity), as a regularization on regression matrix. Unlike one vs one classification strategy, we have used C matrices to simulate the binary classification that not only helps to overcome the complexity issue but also maximizes the inter-class margin. Since the optimization is convex and one of the major challenges is how to efficiently solve non-smooth optimization problem?. Thus, in this chapter, we devised an efficient algorithm for solving the proposed objective functions **(Q2 and Q3)**.

5. We present novel support tensor machines with bounded hinge loss which is monotonic, bounded and nonconvex, thus robust to outliers by limiting the loss due to outliers. We use a randomized non-linear set of features rather than finding the support vectors, thus, eliminates the need to deal with large kernel matrices for large datasets resulting in a reduction in time and space complexity. To solve the non-convex objective function, we devised an iterative approach using the half quadratic optimization **(Q2 and Q3)**.

1.7 Thesis Organization

We have divided the thesis into four sections. In the earlier section, we present an introduction (chapter 1), background (chapter 2) and related work (chapter 3) of the study. In part I, we present the dimensionality reduction methods. It consists of chapter 4 and chapter 5. In the second part, we worked on the optimization of regularizer terms to improve the robustness of traditional support vector machines. We first present robust support matrix machines, support matrix machines based on matrix recovery framework followed by multiclass support matrix machines in chapter 6, chapter 7 and chapter 8 respectively. Finally in the third part, we focused on the optimization of hinge loss term for the classification of noisy tonsorial data. In chapter 9, we present support tensor machines with bounded hinge loss function. This thesis is organized as follows:

- *Chapter 2*: This chapter presents the notation and preliminaries used throughout this thesis. We further briefly describe the basic concept of principal component analysis and support vector machines.

- *Chapter 3*: This chapter presents a survey of recent efforts to reduce to effect of outliers and improves the effectiveness of dimensionality and support matrix machines. This chapter is divided into two subsections. In the first section, we highlighted the recent contribution and problem definition for dimensionality reduction based on the principal component analysis. in the second section, we describe the related work on optimization of support vector machine, support matrix machine and support tensor machine.

  Part-I:- *In this section, we mainly targeted dimensionality reduction methods (variants of PCA) analysis by relaxing the orthogonal constraints of the transformation matrix and imposing a penalty function on regularization term. Proposed method have the freedom to jointly select the important features and rejecting the redundant or irrelevant features, thus, only few features could represent the whole data efficiently, which in results will help to improve the robustness of PCA against outliers.*

- *Chapter 4*: Since the principal component analysis and its variants are sensitive to corrupted variables or observations that affect its performance and applicability in the real-world. This chapter presents a dimensionality reduction method for matrix data by introducing lasso regularization that relaxes the orthogonal constraint and has more freedom to jointly select low-dimensional space features. ORPCA relaxes the orthogonal constraints and penalizes the regression coefficient, thus, it selects important features and ignores the same features that exist in other principal components. Experimental results on four publicly available benchmark datasets show the effectiveness of joint feature selection and provide better performance as compared to state of the art dimensionality reduction methods ORPCA address the the research question 1 and 2.

- *Chapter 5*: Data redundancy makes it a good candidate for sparse representation. Most of the existing dimensionality reduction methods try to preserve a certain kind of linear representation after projection. However, these methods either fail to select useful features or are not that efficient in the presence of outliers. This chapter introduces a novel approach called two-dimensional joint sparse principal

component analysis by effectively combining the robustness of 2DPCA and sparsity-inducing regularization. The proposed approach relaxes the orthogonal constraints resulting in the joint features selection, besides avoiding the selection of the same features in different principal components. In addition to providing sparse solution, the regularization term in the proposed objective function improves the robustness against outliers. This chapter address the the research question 1 and 2.

Part-II:- *In this section, we focused on regularization terms to improve the robustness of support matrix machines against outliers by utilizing the low rank property of data as discriminant features exist in sparse structure and images are low rank. The objective functions are the spectral extension of the conventional elastic net that combines the property of matrix recovery along with low-rank and joint sparsity together, to deal with complex high dimensional noisy data.*

- *Chapter 6*: In many real-world classification problems of supervised tensor learning, high-dimensional data is represented as a matrix, also referred to as second-order tensors. Traditional support vector machines (SVMs) require data to reshape each matrix into vectors, thus, resulting in loss of structural information of the originally featured matrix. This chapter describes the proposed sparse support tensor machines by combining the elastic net and nuclear norm along with hinge loss function which helps to deal with outliers and selects useful features across all data points. The regularization term which promotes the structural sparsity and shares similar sparsity patterns across multiple predictors that is able to select useful features jointly, which is a combination of $\ell_{2,1}$ and nuclear norms. It is a spectral extension of the conventional elastic net that combines the property of low-rank and joint sparsity together, to deal with complex high dimensional noisy data. Furthermore, it also leverages the structural information as well as the intrinsic structure of data and avoids the inevitable upper bound. This chapter address the the research question 1 and 2.

- *Chapter 7*: In this chapter, we consider the problem of high dimensional data classification, when a number of the columns are arbitrarily corrupt. We proposed an efficient Support Matrix Machine by simultaneously performing matrix recovery, feature selection, and classification through joint minimization of $\ell_{2,1}$ and nuclear norm. We assume that the data consists of a low-rank clean matrix plus a sparse noise matrix. We provide convex optimization formulation of the proposed objective function and the sufficient conditions under which it classifies corrupted data

efficiently through low-rank feature recovery process. The proposed approach works under the incoherence and ambiguity conditions and able to recover the intrinsic matrix of higher rank and recover data with much denser corruption. This chapter address the the research question 1,2 and 3.

- *Chapter 8*: This chapter extends the proposed approach to a multiclass classification problem by using C simulated metrics to simulate the binary classification that not only helps to overcome the complexity issue but also maximizes the inter-class margin. In this chapter, we present multiclass support matrix machine from the perspective of maximizing the inter-class margins. The objective function is a combination of binary hinge loss that works on $C4$ matrices and spectral elastic net penalty as a regularization term. This regularization term is a combination of Frobenius and nuclear norm, which promotes structural sparsity and shares similar sparsity patterns across multiple predictors. It also maximizes the inter-class margins that help deal with complex high dimensional noisy data. This chapter address the the research question 2 and 3.

  Part-III:- *In this section, we consider the problem of one class classification and replaced the traditional hinge loss term with bounded hinge loss on tensor data.*

- *Chapter 9*: In this chapter, we consider of classification of tensor data and present a novel anomaly detection approach for large scale tensor data. We first present novel one-class support tensor machines with bounded loss function rather than finding optimized support vectors with an unbounded loss function. This results in improving the classification performance by limiting the loss caused by outliers. We further extend our approach by leveraging the randomness to design a scalable approach that can also be used for large scale anomaly detection. To solve the corresponding optimization problem, we have presented half quadratic optimization and transform the problem into typical OCSTM optimization problem. This chapter address the the research question 2 and 3.

- *Chapter 10:* In final chapter, we summarize and conclude the key contributions of this work.

CHAPTER 2

BACKGROUND KNOWLEDGE

If you torture the data long enough, it will confess.

R. Coase

In this chapter, we first start by establishing the notation and preliminaries used throughout this thesis followed by further discussion on basic concept of principal component analysis and support vector machines.

2.1 Notations

Table 2.1 list the basic symbols used through out this thesis.

Definition 1. Scalar: A scalar x , generally speaking, is a quantity that can be described by a real number, often accompanied by units of measurement. It is physical quantity having only magnitude, not direction. For example 6, -6, 0.381, etc.

Definition 2. Vector: A vector x of dimension n is an ordered collection of n elements, which are called components. Unlike scalar, the vector is a quantity consisting of both direction and magnitude. For example: $x = [4 \quad 6 \quad 9]$ is a vector of dimension 3.

Definition 3. Matrix: A matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$ is an array of numbers with one or more rows and one or more columns. For example $X = \begin{bmatrix} 4 & 8 \\ 2 & 6 \end{bmatrix}$ is matrix of dimension 2×2.

Definition 4. Tensor: Just as vectors (are n-dimensional represented by one-dimensional array), a tensors $\mathscr{X} \in \mathbb{R}^{I_1, \times \ldots, \times I_M}$ are a multidimensional array of real numbers that is

Table 2.1: Notations and their description

Symbols	Description
x	Lowercase letter represents a scalar
$\boldsymbol{x}$	Boldface lowercase letter represents a vector
$\boldsymbol{X}$	Boldface uppercase letter represents a matrix
$\mathscr{X}$	Calligraphic letter represents a tensor
I_p	$p \times p$ Identity matrix.
$\mathscr{R}$	Rank of tensor
y_i	$y_i \in 1, -1$ are the corresponding class labels
$[1:M]$	Set of integers in the range of 1 to M inclusively
$vec(\cdot)$	Denotes column stacking operation
$\langle \cdot, \cdot \rangle$	Denotes inner product
$\otimes$	Denotes tensor product
δ	Denotes delta function
$\mathscr{E}$	Erro rate
$C\sum\xi$	Hinge loss
$W \in \mathbb{R}^{pq}$	Vector of regression coefficients
$b \in \mathbb{R}^{pq}$	an offset term
$\mathscr{L}$	Lagrangian multiplier
$\mathscr{K}(\cdot, \cdot)$	Denotes kernel function
$\|\|.\|\|_{2,1}$	$\ell_{2,1} - norm$
$\|\|.\|\|_1$	$\ell_1 - norm$
$\|\|.\|\|_F$	Frobenius $norm$
$\|\|.\|\|_2$	$\ell_2 - norm$
$\|\|.\|\|_*$, $nuclear - norm$	
$prox\|\|X\|\|_*$	Proximal operator for nuclear nor
$prox\|\|X\|\|_{2,1}$	Proximal operator for $\ell_{2,1} - norm$

higher-order generalization of vectors (first-order tensors) and matrices (second-order tensors). Tensor is a geometric object that maps in a multilinear manner geometric vectors, scalars, and other tensors to a resulting tensor. Let $\mathscr{X} = [\boldsymbol{x_1, ..., x_N}]$ be the $\boldsymbol{M \times N}$ tensor containing $\boldsymbol{n}$ training samples such that $\mathscr{X} \in \mathbb{R}^{\boldsymbol{I_1 \times ... \times I_M}}$ (means $\mathscr{X}$ is real Mth order tensor and numbers $\boldsymbol{I_1, ..., I_M}$ are called the dimensions of the tensor). Their elements are denoted by indices ranging from 1 to capital letter $\boldsymbol{N}$ i.e. An element of tensor is denoted by $\boldsymbol{x_{i_{1,...,n}}}$ where $\boldsymbol{1 \leq n \leq N}$ and $\boldsymbol{1 \leq i_n \leq I_n}$.

Definition 5. Tensor Product: The product of tensors also know as outer product of two tensor $\mathscr{X} \in \mathbb{R}^{\boldsymbol{I_1 \times ... \times I_P}}$ and $\mathscr{Z} \in \mathbb{R}^{\boldsymbol{I'_1 \times ... \times I'_M}}$ can be represented as

$$(\mathscr{X} \otimes \mathscr{Z})_{\boldsymbol{i_1,...,i_P,i'_1,...,i'_M}} = \boldsymbol{x_{i_1,...,i_P} z_{i'_1,...,i'_M}} \tag{2.1}$$

for all values of the indices.

Definition 6: Inner Produce of Tensor The inner produce also know as scalar product of two tensors of same size ($\mathscr{X}, \mathscr{Z} \in \mathscr{R}^{I_1 \times \cdots \times I_M}$) is defined as the sum of products of their entries

$$\langle \mathscr{X}, \mathscr{Z} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_M=1}^{I_M} x_{i_{1,\ldots,M}} z_{i_{1,\ldots,M}} \tag{2.2}$$

Definition 7: Rank One Tensor A Mth order tensor $\mathscr{X}$ has rank one if it is the tensor product of N vectors $\mathbf{u}_i \in \mathbb{R}^{Ii}$, where $1 \leq i \leq M$

$$\mathscr{X} = \mathbf{u}^i \otimes \cdots \otimes \mathbf{u}^M = \prod_{n=1}^{N} \otimes \mathbf{u}^n \tag{2.3}$$

The rank R of Mth order tensor $\mathscr{X}$ is determined by the minimum number of rank one tensors that produces $\mathscr{X}$ in a linear combination. Storing the component vectors a $u^1, \ldots, u^M$ instead of the whole tensor U significantly reduces the required number of storage elements, however, rank-1 tensor is rare in real-world applications.

Definition 8: Tensor Factorization A tensor decomposition represent a d-way tensor X as a d third order tensor. It can be factorized if it can be decomposed as a rank-one tensor of length R.

$$\mathscr{X} = \sum_{r=1}^{R} x_r^1 \otimes \cdots \otimes x_r^M \tag{2.4}$$

Definition 9: Frobenius Norm/ ℓ_2-norm The Frobenius norm also called the Euclidean norm is a matrix norm of an $p \times q$ matrix X and can be defined as the square root of the sum of the absolute squares of its elements. For example for vector $x = [3 \quad 4]$, $||x||_2 = \sqrt{|3|^2 + |4|^2} = \sqrt{9+16} = \sqrt{25} = 5$

$$||X||_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{q} X_{i,j}{}^2} \tag{2.5}$$

It is also equal to the square root of the matrix trace of $\boldsymbol{XX^H}$, where $\boldsymbol{X^H}$ is the conjugate transpose.

$$||\boldsymbol{X}||_{\boldsymbol{F}} = \sqrt{\boldsymbol{Tr(XX^H)}} \tag{2.6}$$

The Frobenius norm of a tensor $\mathscr{X} \in \mathbb{R}^{\boldsymbol{n_1 \times \cdots \times n_d}}$ is defined as

$$||\mathscr{X}||_{\boldsymbol{F}} = \sqrt{\langle \mathscr{X}, \mathscr{X} \rangle} \tag{2.7}$$

Definition 10: ℓ_1 Norm It is also known as Taxicab norm or Manhattan Distance . $\boldsymbol{\ell_1}$-norm is the sum of the magnitudes of the vectors in a space. $\boldsymbol{\ell_1}$-norm is the most natural way to measure the distance between vectors and can be defined by the sum of absolute difference of the components of the vectors. For example: for vector $\boldsymbol{x} = \mathbf{[3 \quad 4]}$, $||\boldsymbol{x}||_{\mathbf{1}} = |\mathbf{3}| + |\mathbf{4}| = \mathbf{7}$

$$||\boldsymbol{X}||_{\mathbf{1}} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} |\boldsymbol{x^i}| \tag{2.8}$$

Definition 11: $\ell_{2,1}$ Norm For a matrix $\mathbf{X} \in \mathbb{R}^{\boldsymbol{p \times q}}$, $\boldsymbol{\ell_{2,1}}$, norm of matrix is denoted as

$$||\boldsymbol{X}||_{\mathbf{2,1}} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} ||\boldsymbol{x^i}||_{\mathbf{2}} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} \sqrt{\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}} ||\boldsymbol{x^2_{i,j}}||} \tag{2.9}$$

It is rotational invariant for rows for any rotational matrix R i.e. $||\boldsymbol{X_R}||_{\mathbf{2,1}} = ||\boldsymbol{X}||_{\mathbf{2,1}}$. It can be used for tensor factorization and multitask learning. The $\boldsymbol{\ell_{2,1}}$ norm can be generalized to $\boldsymbol{r,p}$-norm. Generally methods based on $\boldsymbol{\ell_{2,1}}$ are robust than that of based on $\boldsymbol{\ell_1}$ norm due to its special definition

Definition 12: Randomized Nonlinear Projection Suppose $\boldsymbol{\phi}$ is the feature map $\mathscr{X} \longrightarrow \mathscr{H}$ such that the dot product in $\mathscr{H}$ can be computed using some kernel function as $\mathscr{K}\boldsymbol{(x,x')} = < \boldsymbol{\phi}(\mathscr{X}), \boldsymbol{\phi}(\mathscr{X}') >$ i.e. $\mathscr{X}$ is mapped from input space $\mathbb{R}^{\boldsymbol{M}}$ to feature space $\mathbb{R}^{\boldsymbol{H}}$ via nonlinear function $\boldsymbol{\phi}(\mathscr{X}) = \mathbb{R}^{\boldsymbol{M}} \longrightarrow \mathbb{R}^{\mathscr{H}}$.

Definition 13: Sparse Matrix It is a matrix that consist of only few non-zero elements. In case of 2-dimensional array,as there are only few non zero entries, thus most of the space is a wasted i.e. consider matrix of size 50 x 50 with only 6 observed non-zero elements.

Definition 14: Hinge Loss. It is a loss function used for training classifiers. The hinge loss is used for "maximum-margin" classification, most notably for support vector machines.

2.2 Proximal Algorithm

A proximal algorithm is very useful in machine learning.It is an algorithm for solving a convex but non-smooth optimization problem that uses the proximal operators of the objective terms. Here, we first introduce the proximal operators, which serve as an important components in proximal algorithms.

Definition:Proximal Operator Consider

$$f: \mathbb{R}^{m\times n}[-\infty,+\infty]$$

be a lower semi continuous convex function. The proximal operator

$$prox \quad f: \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$$

of f at point Z is defined by

$$prox \quad f(Z) = arg\min_{W}(f(W) + \frac{1}{2}||W-Z||_F^2)$$

2.2.1 Proximal Operator Nuclear Norm

Specifically, if $f = \tau||W||_*$, the proximal operator for the trace norm can be derived as follows:

$$prox||X||_* = \mathbb{D}_\tau(Z)$$

$$prox \quad ||X||_* = US_\tau(Z)V^T \tag{2.10}$$

2.2.2 Proximal Operatior $\ell_{p,q}norm$

For a matrix $X \in \mathbb{R}^{m\times n}$, its $\ell_{p,q}$ norm is

$$||X||_{p,q} = \left(\sum_1 n\left(\sum_1^m |x_{i,j}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}$$

As our objective function is based on $\ell_{2,1}$, the in the following formalization, we consider $\ell_{2,1}$ if $f = \gamma||X||_{2,1}$, the proximal operator for the $\ell_{2,1}$ norm is given as

$$prox_{\gamma||.||_{2,1}} = arg\min_{X\in\mathbb{R}^{m\times n}}\left(||Y||_{2,1} + \frac{1}{2}||Y-X||_F^2\right)$$

$$= \boldsymbol{prox}_{\gamma||.||_2}(\boldsymbol{X_1}), \boldsymbol{prox}_{\gamma||||_2}(\boldsymbol{X_2}) ...., \boldsymbol{prox}_{\gamma||||_2}(\boldsymbol{X_n})$$

Here each $\boldsymbol{prox}_{\gamma||.||_2}$ for $\boldsymbol{i} = \mathbf{1,2....}$ is given as

$$\boldsymbol{prox}_{\gamma||.||_2} = \begin{cases} \frac{||\boldsymbol{x}||_2 - \gamma}{||\boldsymbol{X}||_2}\boldsymbol{x}, & \boldsymbol{if} ||\boldsymbol{x}||_2 > \boldsymbol{\gamma} \\ \mathbf{0} & \boldsymbol{otherwise}, \end{cases} \tag{2.11}$$

2.3 PCA

Principal Component Analysis is a commonly applied dimensionality reduction method that uses orthogonal transformation to reduce a large set of variables to a small set of variables without losing much information. In other words, it reduces data by geometrically projecting it to much lower-dimensional space with the aim to find the best representation of the original data point using a small number of principal components.

PCA converts the set of observations of possibly correlated variables into linearly uncorrelated variables called principal components. Assume that $\boldsymbol{X_1, X_2 ..., X_N}$ are a set of training matrix (mean centered) with size $\boldsymbol{m \times n}$, where $\mathbf{N}$ is the number of training matrix in the dataset. $\boldsymbol{V = [v_1, v_2, ...v_d]} \in \mathbb{R}^{\boldsymbol{n \times d}}$ is the projection matrix, where $\boldsymbol{v_1}$ is the first basis vector of 2DPCA that maximizes the $\boldsymbol{\ell_1}$-norm-based dispersion of projected samples.

$$\boldsymbol{V}^* = \text{argmin}_{\boldsymbol{V^T V = I_d}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} ||\boldsymbol{X_i} - \boldsymbol{X_i V V^T}||_{\boldsymbol{F}}^{\mathbf{2}}$$

Where $\|\cdot\|_{\boldsymbol{F}}$ denotes the Frobenius norm of matrix and is the sum of square of $\boldsymbol{\ell_2}$-norm of row/column vectors of matrix. Above objective function is equivalent to the following objective function based on the fact $\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} ||\boldsymbol{X_i} - \boldsymbol{X_i V V^T}||_{\boldsymbol{F}}^{\mathbf{2}} + \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} ||\boldsymbol{X_i V}||_{\boldsymbol{F}}^{\mathbf{2}} = \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} ||\boldsymbol{X_i}||_{\boldsymbol{F}}^{\mathbf{2}}$

$$\boldsymbol{V}^* = \text{argmax}_{\boldsymbol{V^T V = I_d}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} ||\boldsymbol{X_i V}||_{\boldsymbol{F}}^{\mathbf{2}}$$

Where $\boldsymbol{tr}(\cdot)$ is the trace function of matrix. As $\boldsymbol{V}^* = \text{argmax}_{\boldsymbol{V^T V = I_d}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} ||\boldsymbol{X_i V}||_{\boldsymbol{F}}^{\mathbf{2}} = \boldsymbol{tr}(\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{V^T A_i^T X_i V})$, we let $\boldsymbol{S_t} = \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{X_i^T X_i}$ denotes the co-variance matrix. By finding the orthogonal eigenvector of $\boldsymbol{S_t}$ corresponding to the first $\mathbf{d}$ largest eigenvalues. 2DPCA is sensitive to noise and outliers, thus optimal projection matrix of objective function mentioned above is not **roubut** in the sense that outlying measurement can skew the

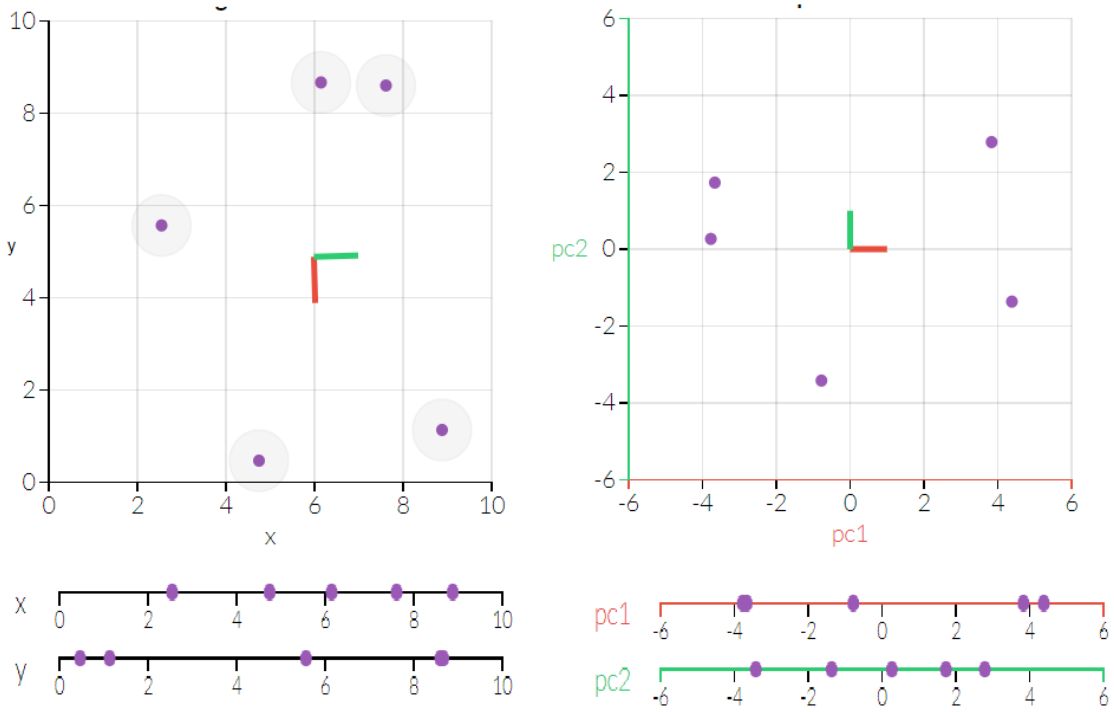

Figure 2.1: (a) Original dataset (b)PC1 vs PC2 (c) Original data along a pair of lines (d) PC2 plotting shows small loss since it it contributes the least to the variation in the data set. We can notice more variation in PC1 as compared to PC2

solution. To overcome this issue, 2DPCA-L1 was proposed which finds the basis vectors that maximizes the dispersion of the projected image in terms of ℓ_1 norm.

$$\boldsymbol{V}^* = \text{argmax}_{\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I_d}} \sum_{n=1}^{N} ||\boldsymbol{X_i}\boldsymbol{V}||_{\ell_1} \tag{2.12}$$

$$\text{subject to } \|\boldsymbol{V}\|_{\ell_2} = \mathbf{1}$$

where $\|\cdot\|_{\ell_1}$ denotes the ℓ_1 norm. Results showed that 2DPCA based on $\ell_1 - \boldsymbol{norm}$ is robust to outliers than 2DPCA.

Although PCA and its variants are able to minimize the effect of outliers to some extent, however one of the major disadvantage of these methods is the redundancy of features. Moreover, these methods are not able to extract useful features, however, the selection of unique and useful features is quite important especially in a case, when features have the physical meaning in many high dimensional data analysis

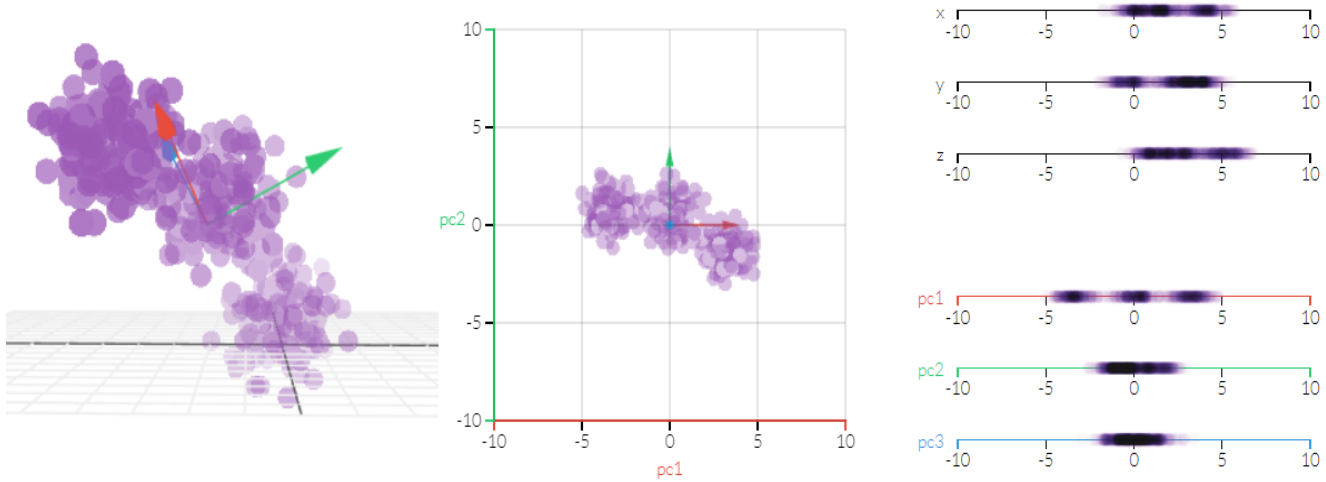

Figure 2.2: 3D Visualization of PCA. We can notice more variatin in PC1 as compared to PC2 and PC3

applications. Considering the only loadings that are greater than threshold could help to achieve sparsity somehow, however, it is an inefficient approach. It could be obtained by imposing the ℓ_0 coefficient on the regression coefficient which penalizes the number of non-zero coefficient whereas the loss term helps to minimize the reconstruction error simultaneously.

$$\underset{A,B}{arg\,min} = \underset{A,B}{arg\,min} \|X - A^T BX\|_F^2 + \lambda_1 \|\beta_j\|_0 \quad \text{subject to} \quad A^T A = I_k \tag{2.13}$$

The above objective function is able to determine informative features individually, however, it does not consider the structural relationship among mulitple features. SCoTLASS successfully derives sparse loadings using the lasso constraint in PCA, yet it is computationally inefficient, and lacks a good rule to pick tuning parameter [32].

Structured-sparsity regularization is popular for sparse learning because of its flexibility of encoding the feature structures. To derive principal components with sparse loadings, several methods have been proposed to achieve the sparseness goal. Sparse PCA produces modified principal components with sparse loadings that are obtained by imposing the lasso constraint on the regression coefficients [145].

$$\underset{A,B}{arg\,min} = \underset{A,B}{arg\,min} \sum_{i=1}^{n} \|x_i - AB^T x_i\|^2 + \lambda_1 \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{2,j} ||\beta_j||_1 \tag{2.14}$$

However, SPCA does not jointly select the useful features as ,Ñì1-norm is imposed on each transformation vector which is not able to select consistent features. In addition, another regularization ℓ_2-norm is imposed on loss term, which makes it sensitive to outliers. Yi et al. presented joint sparse principal component analysis (JSPCA) that select useful features jointly which helps to enhance the robustness of objective function against outliers [127] by imposing the joint sparse constraints ($\ell_{2,1}$-norm is imposed on both loss term and the regularization term) to improve the robustness of algorithms.

$$\underset{B,A}{arg\,min}\, J(B,A) = \underset{A,A}{arg\,min} \|X - AB^T X\|_{2,1} + \lambda \|B\|_{2,1} \tag{2.15}$$

Khan et al. presented joint group sparse PCA (JGSPCA) that ensure the group sparsity and forces the basic coefficient corresponding to a group of features to be jointly

sparse [36]. The group sparsity ensures that the structural integrity of the features. JGSPCA is able to select important features jointly and ensure the group sparsity, however, it is sensitive to outliers due to sensitivity of F-norm against outliers.

$$\underset{A,B}{arg\min} = \underset{A,B}{arg\min} \|X - \sum_{i=1}^{g} X_{\mathbb{G}} A^T B^{\mathbb{G}}\|_F^2 + \lambda \sum_{i=1}^{g} \eta_i \|B^{\mathbb{G}}\|_F \tag{2.16}$$

2.4 Support Vector Machines

Support Vector Machine is a discriminative classifier introduced in the 1990s. Since then it has been successfully applied for classification and regression to many engineering-related applications. It is formally defined by a separating hyperplane in boundless dimensional space by implicitly mapping the training data into high dimensional or infinite-dimensional space. There exist many possible candidate hyperplanes that could be chosen to successfully separate the data. Intuitively speaking, our target is to achieve good separation by the only hyperplane that has the largest distance to the nearest data point of any class, since in general, larger the margin results in lower generalization error of the classifier. A simple example is a yellow line that marks the center road for two-way traffic. Maximizing the hyperplane margin provides partial reinforcement that helps to classify the future with confidence.

Suppose, we have given a set of training samples $T = \{X, y_i\}_{i=1}^{n}$, where $X_i \in \mathbb{R}^{p \times q}$ is the the i^{th} input sample matrix and $y_i \in \{1, -1\}$ is its corresponding class label. Generally, the data needs to be transformed into corresponding vector. In order to fit a classifier, matrix X is needed to be stacked into vector.

Let $x_i = vec(X_i^T) = ([X_i]_{11}, [X_i]_{12}, ...[X_i]_{1q}, [X_i]_{2,1}, [i]_{22}, ...[X_i]_{pq})^T \in \mathbb{R}^{pq}$.

The classical soft margin SVM is defined as

$$arg\min \frac{1}{2} tr(w^T w) + C \sum 1 - y_i [tr(2^T x_i) + b]_+ \tag{2.17}$$

Where $1 - y_i[tr(W^T X_i) + b]_+$ is the hinge loss, $W \in \mathbb{R}^{pq}$ is the vector of regression coefficients, $b \in \mathbb{R}^{pq}$ is an offset term and C is a regularization parameter.

2.4.1 Support Matrix Machine

In equation 2.17, we need to reshape the matrix into vectors which result in losing the correlation among columns or rows in the matrix. By directly transforming the equation 2.17 for matrix, we get

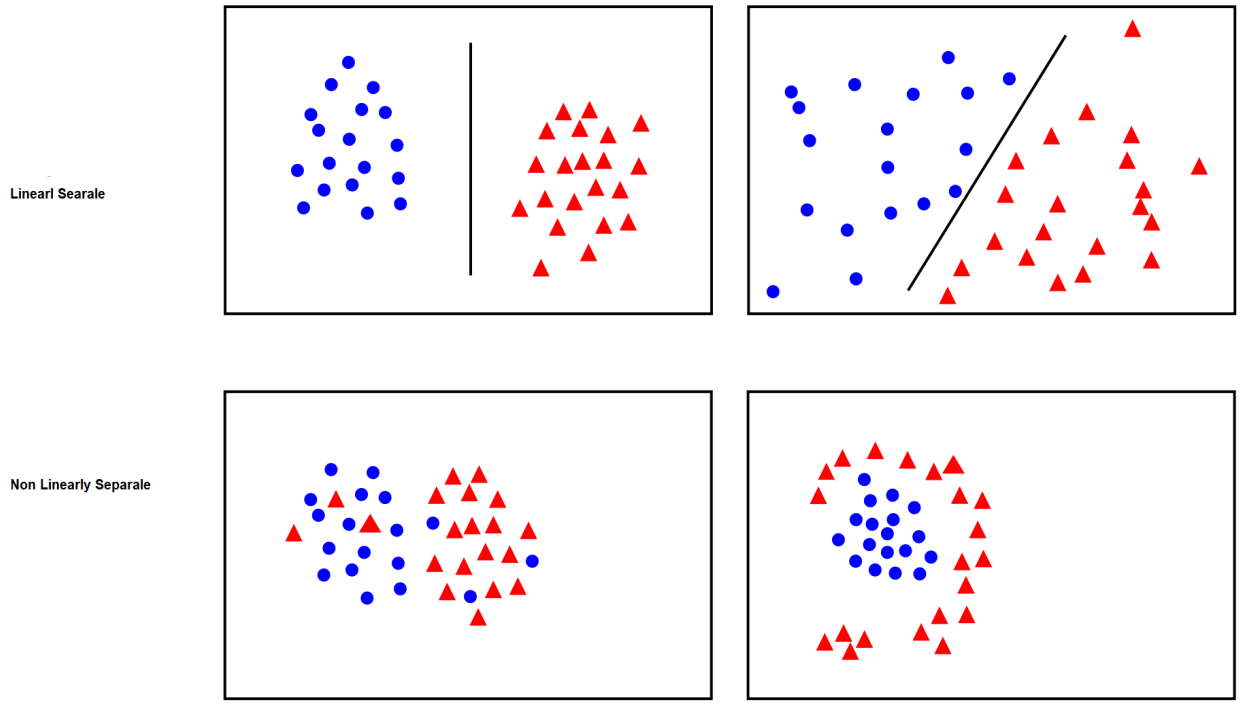

Figure 2.3: (top) Linearly separable data and (bottom) non linearly separable data

$$\boldsymbol{arg\min \frac{1}{2} tr(W^T W) + C \sum 1 - y_i [tr(W^T X_i) + b]_+} \tag{2.18}$$

It is known that $\boldsymbol{tr(WW^T) = vec(W)vec(W^T)}$ and $\boldsymbol{tr(W^T X_i) = vec(W)^T vec(X_i)}$, thus the above objective function cannot capture the intrinsic structure of each input matrix efficiently, due to the loss of structural information during the reshaping process. To take the advantage of intrinsic structural information within each matrix, one intuitive way is to capture the correlation within each matrix through low-rank constraints on the regression parameter.

2.4.2 One-Class Support Vector Machines

One class support vector machines aims to identify suitable region that includes most of the input samples from unknown probability distribution and correctly classify the samples that resembles with training data. OCSVM can be used to identify outliers by finding hypersphear with minimal radius. Consider $\boldsymbol{X = \{x_i : i = 1, ..., N\}}$ are the training

samples with y_i corresponding labels such that $y_i \in \{1,0\}$. Traditional one class support matrix can be formulated as the following quadratic optimization

$$\min_{w,\zeta} \frac{1}{2}||w||_2^2 - \frac{1}{vN}\sum_{i=1}^{N}\zeta_i - p \tag{2.19}$$

$$\text{subject to} \quad \langle w.\phi(x_i)\rangle \geq p - \zeta_i$$

$$\zeta_i \geq 0 \quad \forall i = 1,2,...,N$$

where ζ are the slack variables for penalizing the outliers by segmenting them (some of the data vectors that are outliers) to lie on the other side of the hyperplane. $v \in (0,1]$ is the regularization parameter which controls the fraction of anomalies and support vectors thus, enables the analyzing of noisy data points. Since non-zero slack variables are penalized in the objective function, the decision hyperplane that maximizes the distance data points from the hyperplane is given by the equation

$$\langle w, \phi(x_i) - p\rangle = 0 \tag{2.20}$$

Here, the weight vector w defines the hyperplane in the feature space separating the projections of data from the hyperplane to the origin. A positive definite kernel function k is defined as $k(x,x') \leq \langle \phi(x), \phi(x')\rangle$, that implicitly maps data x into a high dimensional feature space. By introducing the Lagrange multiplier and setting weight vector w, slack variable ζ and offset to zero, the quadratic program can be derived as the dual of the primal program in Eq.2.23

$$\min_{\alpha} \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j k(x_i,x_j) \tag{2.21}$$

$$s.t. \quad 0 \leq \alpha - i \leq \frac{1}{Mv}, \sum_i \alpha_i = 1$$

where α_i are the Lagrange multiplier. Finally, the decision function for input data space X can be defined as

$$f(x) = sgn(w.\phi(x) - p) \tag{2.22}$$

$$f(x) = sgn\left(\frac{1}{2}\sum_{i}^{M} \alpha_i k(x_i,x_j) - p\right) \tag{2.23}$$

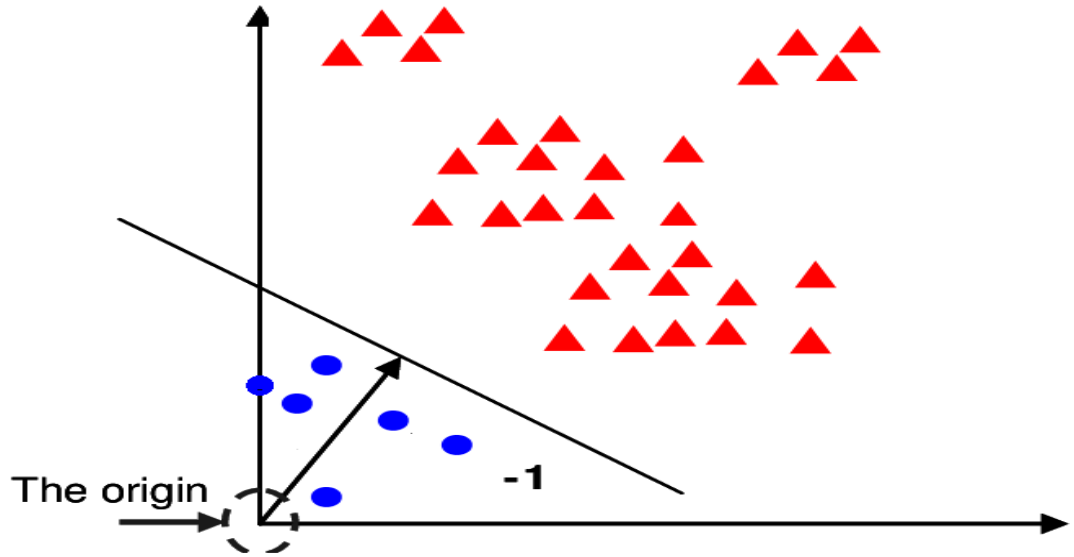

Figure 2.4: Simulation of one class support vector machine on linearly separable data point

2.4.3 Multiclass Support Vector Machine

Support vector machine is inherently two-class classifier. The traditional approach to deal with multiclass classification problem with support vector machines is to break the multiclass problem into series of binary class classification problem such as one-vs-rest (OvR) or one-vs-one (OvO) strategies (e.g. In OvsR, the mutli-class problem is solved by splitting it into $\boldsymbol{n}$ binary class classification problems, whereas OvsO approach splits the problem into $\frac{\boldsymbol{c(c-1)}}{\boldsymbol{2}}$ binary classification problems.) but are computationally expensive and may results in unbalanced distribution of input samples.

$$\boldsymbol{arg\min_{w_j,b_j} \frac{1}{2} tr(w_j^T w_j) + C \sum_{i=1}^{n} \xi_i^j} \tag{2.24}$$

such that

$$\boldsymbol{w_j^T x_i + b \geq 1 - \xi_i^j,\ if\ y_i = j}$$
$$\boldsymbol{w_j^T x_i + b \leq -1 + \xi_i^j,\ if\ y_i \neq j}$$
$$\boldsymbol{\xi_i^j \geq 0}$$

Where $\boldsymbol{\xi_i^j = 1 - y_i[tr(W^T X_i) + b]_+}$ is the hinge loss, $\boldsymbol{W \in \mathbb{R}^{pq}}$ is the vector of regression coefficients, $\boldsymbol{b \in \mathbb{R}^{pq}}$ is an offset term and $\boldsymbol{C}$ is a regularization parameter. This problem is considered unbalanced even though the number of training samples in class

are balanced due to one-vs-all strategy. This property affects the classification performance, and was resolved through one-vs-one classification strategy. To classify unseen data, voting strategy is used and the class with maximum votes is considered as output. In result, it is required to build $\frac{c(c-1)}{2}$ number of classification models in total and can be defined as follow

$$arg \min_{w_{jk},b_{jk}} \frac{1}{2} tr(w_{jk}^T w_{jk}) + C \sum_{i=1}^{n} \xi_i^{jk} \tag{2.25}$$

such that

$$w_{jk}^T x_i + b_{jk} \geq 1 - \xi_i^{jk}, \; if \; y_i = j \; w_{jk}^T x_i + bjk \leq -1 + \xi_i^{jk}, \; if \; y_i \neq k$$
$$\xi_i^{jk} \geq 0$$

Later on, Guermeur formulated a theoretical SVM framework for multiclass classification [25] which can be written as

$$arg \min_{w^{d \times c}, b^c} \frac{1}{2} \sum_{j=1}^{c-1} \sum_{k=j+1}^{c} ||w_j - w_k||_2^2 + \sum_{j=1}^{c} ||w||_2^2 + C \sum_{i=1}^{n} \sum_{j \neq y_i} \xi_i^{jk} \tag{2.26}$$

such that

$$w_{y_i}^T x_i + b_{y_i} \geq w_j^T x_i + bj + 1 - \xi_i j$$
$$\xi_i j \geq 0, \forall i \in 1, ...c_i$$

2.4.4 One-Class Support Tensor Machines

Consider input samples in the dataset $D = \{\mathcal{X}_i, y_i\}_{i=1}^N$ are the Mth-order tensors $\mathcal{X}_i \in \mathbb{R}^{I_1 \times \cdots \times I_M}$ with $y_i \in \{1, 0\}$ corresponding class labels for $i = 1, 2, ..., N$. One-class support tensor machines with traditional loss function can be formulated as the following quadratic optimization

$$\min_{\mathcal{W}, p, \zeta} \quad \frac{1}{2} ||\mathcal{W}||_F^2 + \frac{1}{Nv} \sum_{i=1}^{N} \zeta_i - p \tag{2.27}$$

$$s.t. \quad \left(\langle \mathcal{W}, \phi(\mathcal{X}_i \rangle + b \right) \geq p - \zeta_i,$$

$$\zeta_i \geq 0, \forall i = 1, ..., N$$

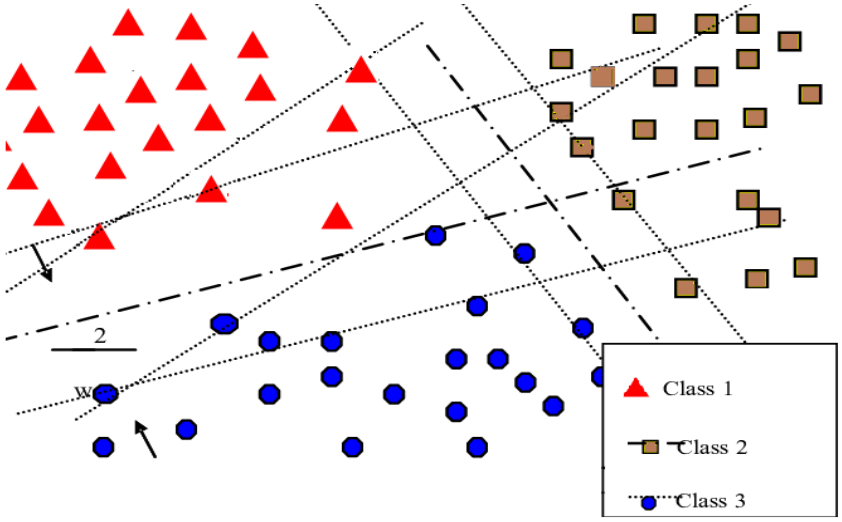

Figure 2.5: Simulation of multiclass support vector machine

where $\mathcal{W}$ tensor is a weight of the separating hyper-plane, $\boldsymbol{v} \in \mathbf{(0,1]}$ is the regularizer that controls the fraction of anomalies and fraction of support vectors. Let $\boldsymbol{\phi}$ is the mapping function that maps the dataset into Hilbert space H and can be formulated as $\boldsymbol{\phi}: \mathscr{X} \longrightarrow \boldsymbol{\phi}(\mathscr{X}) \in \mathbb{R}^{\boldsymbol{H_1} \times \boldsymbol{H_2} \times \ldots \times \boldsymbol{H_{M'}}}$. $\boldsymbol{\zeta_i}$ are the slack variables that allow some of the data point on the other side of hyperplane. By applying the Lagrange multiplier and solving Eq 2.27, we arrive at following quadratic problem

$$\min_{\alpha_1,\ldots,\alpha_N} \frac{\mathbf{1}}{\mathbf{2}} \sum_{i,j} \boldsymbol{\alpha_i \alpha_j} \mathcal{K}(\mathscr{X}_i, \mathscr{X}_j) \tag{2.28}$$

$$\boldsymbol{s.t.} \quad \mathbf{0} \leq \boldsymbol{\alpha - i} \leq \frac{\mathbf{1}}{\boldsymbol{Nv}}, \sum_i \boldsymbol{\alpha_i} = \mathbf{1}$$

The decision function for tensor can be written as

$$\boldsymbol{f}(\mathscr{X}) = \boldsymbol{sgn}\left(\frac{\mathbf{1}}{\mathbf{2}} \sum_{i}^{N} \boldsymbol{\alpha_i} \mathcal{K}(\mathscr{X}_i, \mathscr{X}_j) - \boldsymbol{p}\right) \tag{2.29}$$

The solution Eq.2.29 is characterized by the parameter $\boldsymbol{v}$ that sets lower bound on the number of training examples used as support vectors and upper bound on the fraction of anomalies. Using the Karush-Kuhn-Tucker optimality condition, the input tensor data can be classified based on the its projection below, above or on the hyper-plane boundary in the feature space based on the support tensors.

CHAPTER 3

RELATED WORK

The price of light is less than the cost of darkness.

A. C. Nielsen

In this chapter, we provide the brief description of the recent development for dimensionality reduction and classification problem of high dimensional data. We have divided the below discussion into two subsections. In first section, we describe the recent development for dimensionality reduction for matrix data followed by discussion on the problem of classification for high dimensional data in the second seciton.

3.1 Dimensionality Reduction

The big data era exacerbates the curse of dimensionality from the computational perspective. Even when data dimensions are moderately high, many existing algorithms find it difficult to handle such computational complexity. Furthermore, the underlying structure in many cases is based on a small set of features, hence poses several challenges. Thus, there is a need to develop fast and reliable algorithms without compromising the accuracy. In order to handle such real-world data adequately, dimensionality reduction plays important roles with the aim to transform the high-dimensional data into meaningful representation of the low-dimensional data by preserving the quality of the data, so that it could be classified efficiently. An intuitive example of dimensionality reduction can be discussed through a spam e-mail detection. Only few of features are involved in

classification (i.e. title, structure and contents of the email etc.) while most of the features are either irrelevant or overlapping. Hence, we can reduce the number of features in such problems [83].

Ideally, the low dimensional representation has exhibited a dimensionality that corresponds to intrinsic dimensionality of data, which is the minimum number of parameters required to account for the observed properties of data. The other challenges with modern data analytics are the presence of outliers and missing values. One might argue that this is not not major issues as modern challenge as it was the case before. However, high dimensionality, outliers and missing value problems, evolve with the modern data formats and pose serious challenges that affect the accuracy significantly.

In past ten years, dimensionality reduction has seen much activity contribution, primarily due the current high dimensional data. For example, DNA micro-array data, measures the expression level of thousands of genes in single experiment. An alternative solution is to use the deep learning methods, however, it is not necessary that we could get access to large amount of annotated data i.e gene data usually is high dimensional but consists of small number of samples [84]. Thus, dimensionality reduction tools are used to extract low dimensional manifold and are common solution to deal with such data. The low representation of the data describes the structure of original data.

In order to deal with the challenge of high dimensionality, several vector-based methods are in use especially during the last two decades, such as PCA [108], LDA[5], LPP [28], SPP [68] and NPE [28] etc. Among these, PCA is one of the most extensively used statistical modeling approach, associated with multivariate analysis since its introduction by Pearson [62] and Hotelling [30]. PCA projects the high-dimensional input data into linear orthogonal space. The main objective of which is to, sequentially extract uncorrelated orthogonal features, eventually maximizing data variability thus guaranteeing minimal information loss. However, one of the major drawbacks is that PCA is linear combination of all variables and loadings are typically non-zero. Each sensed feature may have additional cost of acquisition, processing, and storage [36]. Moreover, not all extracted features are important for potential application. This makes PCA data interpretation difficult, and it is still sensitive to outliers (as its co-variance matrix is derived from $\boldsymbol{\ell_2}$-norm that affects its performance). Furthermore, reshaping the data into vectors could ultimately destroy the structural information embedded in, which is a very important factor for certain classification tasks. For example, EEG signals which consists of voltage fluctuations at several electrodes during a time period, has a strong correlation with certain frequency band and channels [73]. Thus, it fails to deal

with outliers which often appears in real-world data. Moreover, before applying PCA and LDA, there is a need to convert the image into one-dimensional vector, consequently, it may not exploit the spatial structural information very well which is very important for image representation. To overcome these issues, several variants of PCA have been proposed to improve the effectiveness of dimensionality reduction and robustness against outliers [59].

Due to limitation of traditional PCA for high dimensional data, tensor based subspace learning methods have been widely applied for dimensionality reduction. Results showed that methods directly based on tensors are far more efficient than one-dimensional PCA, due to their direct formulation based on high-dimensional space rather than one dimensional vector. For example, two-dimensional subspace learning methods directly calculate the class scatter metrics from images, hence can reveal the spatial structural information of image, an important factor for the classification [85]. $\boldsymbol{\ell_1}$ norm based subspace learning methods have shown great performance against outliers for tensor data classification [112]. Ke and Kanade presented matrix factorization as a $\boldsymbol{\ell_1}$ norm minimization problem that is able to handle missing data straightforwardly. Wang et al. presented robust 2DPCA with non-greedy $\boldsymbol{\ell_1}$-norm maximization in which all projection directions are optimized simultaneously [115]. Luo et al. extended it by learning the optimization matrix by maximizing the sum of the projected difference between each pair of instances, rather than the difference between each instance and the mean of the data [47]. Although, $\boldsymbol{\ell_1}$ based methods provided great performance, however, these methods do not relate to co-variance matrix which characterizes the geometric structure of the data and works as a robust measure for sample dispersion, not the regularizing basis vectors. Several efforts have been made to utilize F-norm as subspace learning such as 2DPCA [124, 125], 2D-PCA [105], F-norm 2DPCA [42], NM-2DPCA [12], Angle 2DPCA [22], $\boldsymbol{R_1}$-2DPCA [23], Optimal 2DPCA [116],$\boldsymbol{\ell_1}$-2PCANet [43]. However, the limitation of 2DPCA is the dense basis which makes it difficult to explain the resulting features. As such, it is desirable to select the most relevant or salient elements from a large number of features.

Since outliers do not have a precise mathematical meaning, the problem of robust PCA is still not well- defined. Several classical heuristics have been proposed to improve the robustness against outliers. Compared to the traditional PCA, $\boldsymbol{\ell_1}$ and $\boldsymbol{\ell_{2,1}}$ based on matrix recovery based methods effectively improve the robustness of algorithms [109, 114, 118, 135]. Some work suggest that means, in the least squared sense, is not optimal of distance metrics such as $\boldsymbol{\ell_1}$, $\boldsymbol{\ell_{2,1}}$ and nuclear norm [28, 52, 111, 115, 133].

To improve their performance, simultaneously optimizing mean and projection matrix, the criterion function has been introduced [116]. Later, Song et al. presented robust PCA by simultaneously optimizing global mean and projection matrix [101]. Recently, a novel robust PCA (RPCA-AOM) is presented by by maximizing the sum of projected differences between each pair of data based on the ℓ_1-norm distance by avoiding the mean computation in solving the projection matrix [47]. However, RPCA-AOM does not well characterize the geometric structure of data and it is computationally expensive as well as difficult to solve the local optimal solution of RPCA-AOM.

Combination of nuclear norm with other ($\ell_1, \ell_{2,1}$) has shown great performance by providing sparse but also low-rank solution. Zhang combined nuclear norm and $\ell_{2,1}$-norm to extract neighborhood preserving features by minimizing reconstruction error due to Frobenius norm that is very sensitive to outliers [133, 134]. $\ell_{2,1}$ ensures the projection to be sparse in rows so that discriminative features are learned in the latent subspace whereas the nuclear-norm ensures the low-rank property by projecting data into their respective subspaces. The addition of nuclear norm with $\ell_{2,1}$-norm results not only sparse but also low-rank feature representation. Zhao et al. presented Local and global information (LLGDI) for effective semi-supervised dimensionality reduction [137]. LLGDI adopts a set of local classification functions in order to preserve local geometrical as well as discriminative information. Moreover, it also adopts global classification function that preserve the global discriminative information by solving the regression and dimensionality reduction simultaneously. 2DPCA and its variations cannot reveal the spatial structural information which is one of the core components in image representation [85, 116]. Moreover, features in low-dimensional subspace are linear combination of all features in high-dimensional space, thus, it usually consists of redundant features that affect the classification performance. However, it is quite difficult to interpret new feature set whereas it is quite important to extract new features especially when they have spatial meaning [73].

It is no surprise that most of the real world data have such a high sparsity i.e. only small number of features are important for spam detection. An adhoc approach to deal with such problems, is achieving the sparsity artificially by considering only those loadings that are greater than threshold. however, in general, it is an inefficient approach. Recently, addition of sparsity constraint into PCA is widely explored to overcome the dimensionality issues and to reduce the number of explicitly used variables. Sparse principal component (SPCA) analysis is presented to learn sparse projection matrix, yet it can not jointly select the features. To overcome aforementioned issue, recently, joint

sparse PCA is presented that relaxes the orthogonal constraints of transformation matrix to select features jointly and can effectively integrate selection of the features process into subspace to exclude redundant features. Though, it is able to select the features jointly with robustness against outliers, however, it is based on 1D and as a result can not utilize the structural information embedded in data. Whereas, direct computation of 2D-joint sparse PCA is not effective due to the sensitivity of F-norm against outliers, as the square of F-norm remarkably enlarges the distance in criterion function that affects its performance based on criterion function. Smallman et. al. developed a sparse method for data from an exponential-family distribution by adding ℓ_1 and SCAD penalties to introduce sparsity [100]. However, it suitable for only vectors. To address aforementioned challenges, we have presented a novel approach called ORPCA and 2D robust Joint Sparse PCA (2D-JSPCA) that combines the subspace learning and feature selection together in order to exclude the effect of redundant patterns besides avoiding the selection of same features in different principal components.

3.2 Support Matrix Machines

In this section, we provide a brief description of matrix classification problem. Practically, it has been noticed that the selection of features and model designs, is far more important than the choice of loss [51]. Hence, in this coherence, we focused the regularization term in dealing the feature selection approach embedded in.

The classical soft margin SVM is defined as

$$arg\min \frac{1}{2} tr(w^T w) + C \sum 1 - y_i [tr(2^T x_i) + b]_+ \tag{3.1}$$

Where $1 - y_i[tr(W^T X_i) + b]_+$ is the hinge loss, $W \in \mathbb{R}^{pq}$ is the vector of regression coefficients, $b \in \mathbb{R}^{pq}$ is an offset term and C is a regularization parameter.

In equation 3.1, we need to reshape the matrix into vector which results in losing the correlation among columns or rows in the matrix. By directly transforming the equation 3.1 for matrix, we get

$$arg\min \frac{1}{2} tr(W^T W) + C \sum 1 - y_i [tr(W^T X_i) + b]_+ \tag{3.2}$$

It is known that $\boldsymbol{tr(WW^T) = vec(W)vec(W^T)}$ and $\boldsymbol{tr(W^T X_i) = vec(W)^T vec(X_i)}$, thus the above objective function cannot capture the intrinsic structure of each input matrix efficiently, due to the loss of structural information during the reshaping process. To take the advantage of intrinsic structural information within each matrix, one intuitive way is to capture the correlation within each matrix through low-rank constraints on the regression parameter.

As the hinge loss enjoys the large margin principle, it also embodies sparseness and robustness, which are two desirable properties for a good classifier. Motivated by this, recently Luo et. al. presented sparse matrix machine shown in equation 3.3 [46]. The objective function in equation 3.3 consists of hinge loss plus nuclear norm and Frobenius norm as a reguarlizer.

$$\boldsymbol{arg\min \frac{1}{2} tr(W^T W) + \tau ||W||_* + C \sum 1 - y_i [tr(W^T X_i) + b]_+} \tag{3.3}$$

The spectral elastic net regularization $\boldsymbol{\frac{1}{2} tr(W^T W) + \tau ||W||_*}$ captures the correlation within each matrix. In addition, the nuclear norm in the reguarlizer is used to control the rank of $\boldsymbol{W}$ that is NP-hard problem. In this scenario, it provides the best approximation of rank of the matrix $\boldsymbol{W}$. The objective function shown in equation 3.3 is capable of capturing the latent structure within each matrix and further perform the classification based on all entities of each matrix which effect the classification performance, thus, making the model complicated . To overcome this challenge, Zheng et. al. presented sparse support matrix machine that consists of loss plus nuclear norm and $\boldsymbol{\ell_1}$ as reguarlizer term [141].

$$\boldsymbol{arg\min \gamma ||W||_1 + \tau ||W||_* + C \sum 1 - y_i [tr(W^T X_i) + b]_+} \tag{3.4}$$

The classification function in equation 3.4 incorporates the loss and constraints on the regression matrix which is a linear combination of $\boldsymbol{\ell_1}$ norm and nuclear norm. $\boldsymbol{\ell_1}$ norm encourages matrix $\boldsymbol{W}$ to be sparse by serving as a convex surrogate for non zeros entries. The regularizer term in equation 3.4 is combination of $\boldsymbol{\ell_1}$ norm and nuclear norm which provides structural sparsity. A common features of approach based on Frobenius norm [46] and $\boldsymbol{\ell_1}$ norm [141] is that they treat both indices (row and column) in the same way. However, they have different meanings i.e. $\boldsymbol{i}$ and $\boldsymbol{j}$ run through data points and spatial dimension respectively. This subtle distinction makes it easy to get loss for the matrix, whereas, $\boldsymbol{\ell_{2,1}}$ norm captures this subtle distinction and provides structural sparsity. Furthermore, studies have shown that $\boldsymbol{\ell_{2,1}}$ is sparser than $\boldsymbol{\ell_1}$-regularization as

it finds the joint solutions and encourages multiple predictors to share similar sparsity patterns.

To tackle the challenge of robust feature selection, recently, the sparsity regularization in dimensionality reduction has been widely investigated for feature selection i.e. ℓ_1 [44] , ℓ_q [95], $\ell_{2,0}$ [60], $\ell_{2,1}$ [9]. Forbenius norm [37] has also been applied to introduce the sparsity property in a regression matrix. These approaches work well and consider the correlation between columns and rows under the low-rank assumptions and provide satisfactory performance [142]. However, these approaches consider all the entities of the matrix as explanatory factor, whereas in real world, features might be redundant or useless for certain classification tasks. Furthermore, it turns out that the nuclear norm can also be used as a convex relaxation of this optimization problem, which greatly simplifies the problem and allows further room for interesting applications such as accelerated algorithms for matrix completion (compressed sensing). Recently, classifier based on combination of hinge loss, nuclear norm and Forbenius norm [46, 142], ℓ_1 [140, 141] has been presented. Although these methods showed excellent performance by taking advantage of correlation between rows and columns of the regression matrix under the low-rank assumptions. But, they simply consider entities in matrix as explanatory factors and do not consider the intrinsic group structure of data and are sensitive to outliers. Furthermore, they tend to select features without considering all classes.

3.2.1 Support Tensor Machines

Generally, outliers detection algorithms can be classified into three categories that are supervised, unsupervised, and semi-supervised. Unsupervised anomaly detection methods detect the anomalies in an unlabeled data under the assumption that the majority of the instances are normal and small fraction of data show anomalous behaviour. To improve the robustness of anomaly detection, recently, Xiao et al. utilized non-convex ramp loss function into OCSVM optimization to reduce the affect of outliers [120]. Similar to [120], Yingjie et.al. [106] presented robust and sparse anomaly detection approach by replacing the hing loss with non-convex ramp loss function to make robust and sparse semi-supervised algorithm and used concave-convex procedure to solve the model that is a non-differentiable non-convex optimization problem. Recently, to improve the robustness of traditional OCSVM against outliers, Xing et.al. presented replace the hinge loss with rescaled hinge loss function [121]. Experimental results showed that these methods can effectively reduce the influence of outliers to some extent, however, are computationally complex.

To improve the robustness of OCSTM against outliers and computational efficiency, recently, many researchers have focused on the improvement of the loss function and kernel methods respectively. He et al. presented a structure-preserving kernel for non-linear tensor learning by deriving the kernel based on structure-preserving feature mapping [27]. Erfani et al presented a randomized kernel support tensor machine based on nonlinear randomized projection, however, it is sensitive to outliers [20]. Anaissi et al. presented sparse and smooth representations by replacing with ℓ_1 regularized tensor decomposition to overcome the sensitivity of OCSTM against outliers [3]. Yanyan et al. developed Linear Support Tensor Domain Description (LSTDD) based on a linear tensor-based algorithm to find a closed hypersphere with the minimal volume in the tensor space [15]. Traditional support tensor machine is not robust to outliers as unboundedness of the loss function results in larger loss due to outliers and the decision boundary may deviate from the optimal hyperplane [72]. Several non-convex substitution of hinge loss function has been presented in order to suppress the effect of outliers and improve the robustness for support vector machines. It is well known that methods based on tensors are better in term of both computational complexity as well as accuracy [1, 14, 20]. However, according to our knowledge, no work has been done so for on the improvement of one-class tensor machines[1]. Extensive experimental analysis shows that proposed bounded one-class support tensor machines considerably improves the robustness against outliers and significantly reduces the computational complexity as compared to state of the art anomaly detection methods.

3.3 Summary

This chapter presents the related work followed by the problem. We have divide the discussion into two sections. In first section, we describes the recent development for dimensionality reduction to improve the performance of traditional PCA. In the second section, we focus on support matrix machine and briefly describe the recent approaches for tensor data classification. Practically, it has been noticed that the selection of features and model designs, is far more important than the choice of loss [51]. Hence, in this coherence, we focused the regularization term in promoting the structural sparsity and leveraging the intrinsic structure of data. Moreover, features in low-dimensional subspace are linear combination all features in high-dimensional space, thus, it usually consists of redundant features that effect the classification performance. Thus, modeling sparsity into feature extraction or classification function could help to encode semantic

information, as well. Thus, it is required to have classification function that promotes sparseness as well as preserve the structural information.

Part I

Dimensionality Reduction and Feature Selection

CHAPTER 4

JOINT FEATURE SELECTION

If we have data, let's look at data. If all we have are opinions, let's go with mine.
J. Barksdale

The aim of dimensionality reduction is to transform the high-dimensional data into low-dimensional representation by preserving the quality of the data so that it could be classified efficiently. To deal with curse of dimensionality, recently several vector-based methods are in use during the last two decades such as Principal Component Analysis (PCA) [108], Linear Discriminant Analysis (LDA) [5, 83, 126], LPP [28], SPP [68], SPPE [136], Isomap[132] and NPE [28]. Principal Component Analysis is one of the extensively used unsupervised dimensionality reduction method that projects high-dimensional representation into linear orthogonal space. However, one of the major drawbacks is that PCA is linear combination and loading are non-zero. This makes PCA data interpretation difficult, and it is still sensitive to outliers (as its covariance matrix is derived from ℓ_2-norm that affects its performance. Thus, it fails to deal with outliers that often appears in real-world data. Moreover, before applying PCA and LDA, there is need to convert the image into one-dimensional vector, thus it may not exploit image's spatial structural information very well [21, 28, 55, 108, 109, 122, 127, 145] which is very important for image representation. To overcome these issues, several variants of PCA have been proposed to improve the effectiveness of dimensionality reduction and robustness against outliers. Since the principal component analysis and its variants are sensitive to outliers that affect their performance and applicability in real world. To overcome the issue

of sensitivity of PCA against outliers, in this chapter, we introduce two dimensional outliers robust principal component analysis (ORPCA) by imposing the joint constraints on the objective function. ORPCA relaxes the orthogonal constraints and penalizes the regression coefficient, thus, it selects important features and ignores the same features that have already been selected (exist) in other principal components. It is commonly known that square Frobenius norm is sensitive to outliers, in order to deal with the sensitivity challenge for Frobenius norm issue, we have devised an alternative way to derive objective function. Experimental results show the effectiveness of joint feature selection and provides better performance as compared to state of the art dimensionality reduction methods.

4.1 Motivation

As the aforementioned analysis in chapter 2 and chapter 3, for the classification of high dimensional noisy data, it is always important to find salient features that belong to specific part of image. To select such salient patterns, projection matrix should consist of important features that could contribute in the classification and reconstruction. Most of the PCA based methods are sensitive to outliers and are unable to select optimal set of features. Which feature is important or ignoreable? selecting or rejecting it, could helps to improve the performance. Moreover, integrating feature selection into subspace learning could help to encode semantic information that helps to approximates high-dimensional data in a flexible way. Based on these above hypothesis, we have imposed the joint constraint on the objective and added a penalty term which helps to avoid redundant feature selection by avoiding selection of same features in different principal components. Furthermore, sensitivity of F-norm is another challenge. To overcome this issue, we have devised an alternative approach to derive objective function. Compared with traditional PCA based on Frobenius norm, ORPCA not only select feature jointly, but also weaken the effect of large distance and has rotational invariance properties.

4.2 Outliers Robust 2DPCA

In this section, we present outliers robust dimensionality reduction approach (ORPCA) in detail. As described in earlier sections, the projection procedure consist of all the original features, thus, it may also have irrelevant and redundant features which could influence the performance of dimensionality reduction, in result affecting the classification perfor-

mance. Furthermore, outliers strongly affect the feature selection which depresses the classification performance. In this work, we present a novel formulation for PCA that combines the subspace learning and feature selection together in order to exclude the effect of redundant patterns and joint feature selection. We employed Frobenius norm as distance metric learning and seeks the projection matrix by joint minimization of regularizer and penalty terms. We relaxes the orthogonal constraints of transformation matrix and introduces another transformation that helps to jointly select important features and discard the features that already selected in other principal components. To overcome the sensitivity issue due to squared Frobenius norm, we devised an efficient way to compute F-Norm, as a result, ORPCA has more freedom to select robust features jointly for low dimensional representation that helps to minimizes the affect of outliers as well as redundancy. However, it does not guarantee fully sparse solution. We present the spare solution by adding additioanl regualizer term in chapter 5.

4.2.1 Objective Function

Considering the appearance of outliers in the input data, we propose the following objective function

$$\min_{P,Q} J(P,Q) = \min_{P,Q} \sum_{j=1}^{N} \left\| A_j - A_j Q P^T \right\|_F^2 + \lambda \|Q\|_F^2 \tag{4.1}$$

where $P, Q \in \mathbb{R}^{n \times d}$. Matrix Q is used to transforms each sub-image into low-dimensional subspace and matrix P is used to recovers the matrix A such that $A = [A_1, ..., A_N]$, where $A_j \in \mathbb{R}^{m \times n}$. Furthermore, while we require the matix P to be orthogonal ($P^T P = I_d$), we do not require the orthogonality of the matrix Q,thus ORPCA has more freedom to learn low dimensional space. In addition, the regularization parameter $\|Q\|_F^2$ reduces the constraints and enables the ORPCA to select important features jointly select. The penalty term penalizes the regression coefficient to makes PCA possible to select features jointly and discard those features that have already been selected in other principal components. Moreover, regularization term $\|Q\|_F^2$ is convex that can be easily optimized. The parameter $\lambda \geq 0$ balances the loss and regularization terms. In short, we relaxed the orthogonal constraint of transformation matrix Q, introduce another transformation matrix P and added an additional regularization parameter $\|Q\|_F^2$ to make the objective function robust and able to select features jointly.

4.2.2 Optimization

Squared F-norm is not robust in the sense that outlying measurements can arbitrarily skew the solution from the desired. We devised an efficient way to compute F-Norm to overcome its sensitivity challenge. Although the objective function is shown in Eq 4.1 is based on square F-norm, however, computation of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are not squared. Compared with squared F-norm, the proposed derivation can weaken the effect of large distance but also has rotational invariance. ORPCA sees the projection matrix that makes the value of objective function small. The objective function has two main unknown terms $\boldsymbol{P}$ and $\boldsymbol{Q}$. The following two theorems play a key role in determining the minimizers of the optimization problem 4.1.

Theorem 4.1. *The minimizers of the objective function given in the Equation 4.1 satisfy the following equation*

$$\boldsymbol{Q} = \left[\sum_{j=1}^{N}\left(\lambda \boldsymbol{I_n} + \boldsymbol{A_j^T A_j}\right)\right]^{-1}\left[\sum_{j=1}^{N}\boldsymbol{A_j^T A_j}\right]\boldsymbol{P} \tag{4.2}$$

Proof. According to the definition of Frobenius norm, the linearity and cyclic properties of trace function, and orthogonality of matrix $\boldsymbol{P}$, the above objective function can be written in a more computationally traceable way as

$$\boldsymbol{J(P,Q)} = \sum_{j=1}^{N}\left\|\boldsymbol{A_j - A_j Q P^T}\right\|_{\boldsymbol{F}}^{\boldsymbol{2}} + \lambda\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\boldsymbol{2}} \tag{4.3}$$

$$= \sum_{j=1}^{N}\mathrm{tr}\left[\left(\boldsymbol{A_j^T - P Q^T A_j^T}\right)\left(\boldsymbol{A_j - A_j Q P^T}\right)\right] + \lambda\mathrm{tr}\left(\boldsymbol{Q^T Q}\right) \tag{4.4}$$

$$= \sum_{j=1}^{N}\mathrm{tr}\left(\boldsymbol{A_j^T A_j - A_j^T A_j Q P^T - P Q^T A_j^T A_j + P Q^T A_j^T A_j Q P^T}\right) + \lambda\mathrm{tr}\left(\boldsymbol{Q^T Q}\right) \tag{4.5}$$

$$= \sum_{j=1}^{N}\mathrm{tr}\left(\boldsymbol{A_j^T A_j - 2A_j^T A_j Q P^T + A_j^T A_j Q Q^T}\right) + \lambda\mathrm{tr}\left(\boldsymbol{Q^T Q}\right) \tag{4.6}$$

Now, differentiation Eq (5.4),

$$\frac{\boldsymbol{\partial J}}{\boldsymbol{\partial Q}} = \sum_{j=1}^{N}\left(\boldsymbol{-2A_j^T A_j P + 2A_j^T A_j Q}\right) + \boldsymbol{2\lambda Q}. \tag{4.7}$$

Therefore,

$$\frac{\partial J}{\partial Q} = 0 \Rightarrow \sum_{j=1}^{N} \left(-2A_j^T A_j P + 2A_j^T A_j Q \right) + 2\lambda Q = 0 \tag{4.8}$$

Simplifying the above equation, we get

$$\sum_{j=1}^{N} (A_j^T A_j P) = \sum_{j=1}^{N} (A_j^T A_j Q) + \lambda Q \tag{4.9}$$

The above equation can be rewritten as

$$\left(\sum_{j=1}^{N} A_j^T A_j \right) P = \left(\sum_{j=1}^{N} (\lambda I_n + A_j^T A_j) \right) Q \tag{4.10}$$

Hence, we can write

$$P = \left(\sum_{j=1}^{N} (\lambda I_n + A_j^T A_j) \right) Q \left(\sum_{j=1}^{N} A_j^T A_j \right)^{-1} \tag{4.11}$$

■

Once, matrix Q is known, we can optimize matrix P with respect to matirx Q.

Theorem 4.2. *If UDV^T is the singular value decomposition (SVD) of $\sum_{j=1}^{N} A_j^T A_j Q$, then*

$$P = U I_{n \times d} V^T \tag{4.12}$$

is orthogonal and minimizes the Eq. (5.4) for a given matrix Q.

Proof. As we know that the matrices V and U are orthogonal matrices of sizes $d \times d$ and $n \times n$, respectively. As such,

$$P^T P = V I_{n \times d}^T U^T U I_{n \times d} V^T = I_d$$

■

The orthogonal constraints on matrix P reduces the feature redundancy and forces the objective function to be small. Below in table 4.1, we describe an iterative algorithm of ORPCA for training samples $A_1, ..., A_n$ of size $m \times n$, and regularization parameter λ.

Table 4.1: Algorithmic procedure of ORPCA

Input: $A_j \in \mathbb{R}^{m \times n}$ for $j = 1,...,N$ where A is centralized, and parameter λ. **Output:** Matrix P and Matrix Q

Step-I: Randomly initialize the matrix P
While do not converge do
Step-II: Minimize the objective function with respect to matrix Q by finding the matrix Q using Eq.(4.2)
Step-III: Compute the Singular Value Decomposition of $\sum_{j=1}^{N} A_j^T A_j Q$
Step-IV: Update the matrix P using Eq.(5.10) to minimize the objective function with respect to matrix P
end while

4.2.3 Numerical Algorithm

Table 4.1 describe an iterative algorithm of ORPCA for training samples $A_1,...,A_n$ of size $m \times n$, and regularization parameter λ.

4.2.4 Convergence Analysis

First, we provide to the following lemma

Lemma 4.1. *For any nonzero matrix $P,Q \in R^{n \times d}$, the following results hold:*

$$\|P\|_F - \frac{\|P\|_F^2}{2\|Q\|_F} \leq \|Q\|_F - \frac{\|Q\|_F^2}{2\|Q\|_F} \tag{4.13}$$

Proof. We start with an obvious inequality $(\sqrt{S} - \sqrt{S_t})^2 \geq 0$, we have

$$(\sqrt{S} - \sqrt{S_t})^2 \geq 0$$

$$\Rightarrow S - 2\sqrt{SS_t} + S_t \geq 0$$

$$\Rightarrow \sqrt{S} - \frac{S}{2\sqrt{S_t}} \leq \frac{1}{2} S_t$$

$$\Rightarrow \sqrt{S} - \frac{S}{2\sqrt{S_t}} \leq \sqrt{S_t} - \frac{S_t}{2\sqrt{S_t}}$$

Now substituting S and S_t by $||P||_F$ and $||Q||_F$ respectively, we arrive at Eq. 4.13. ■

Based on the above lemma 4.1, we provide the following convergence theorem.

Theorem 4.3. *Given all the variables in objective function equation 4.1, the iterative scheme of proposed ORPCA described in table 4.1 shows that objective function value is monotonically decreasing thus converges to local optima.*

Proof. For given initial value of matrix $\boldsymbol{P}$, say $\boldsymbol{P_0}$, we can compute the matrix $\boldsymbol{Q_0}$ by minimizing the objective function $\boldsymbol{J(P_0,Q)}$. Consequently,

$$\boldsymbol{J(P_0,Q_0) \leq J(P_0,Q)}$$

We can calculate matrix $\boldsymbol{P_1}$ by minimizing the objective function $\boldsymbol{J(P,Q_0)}$. Hence,

$$\boldsymbol{J(P_1,Q_0) \leq J(P_0,Q_0)}$$

Since the matrix $\boldsymbol{Q_1}$ minimizes the objective function $\boldsymbol{J(P_1,Q)}$, we have

$$\boldsymbol{J(P_1,Q_1) \leq J(P_1,Q_0) \leq J(P_0,Q_0).}$$

That is

$$\begin{aligned} & \sum_{j=1}^{N} \left\| \boldsymbol{A_j - A_j Q_1 P_1^T} \right\|_F^2 + \boldsymbol{\lambda} \|\boldsymbol{Q_1}\|_F^2 \\ \leq & \sum_{j=1}^{N} \left\| \boldsymbol{A_j - A_j Q_0 P_1^T} \right\|_F^2 + \boldsymbol{\lambda} \|\boldsymbol{Q_0}\|_F^2 \\ \leq & \sum_{j=1}^{N} \left\| \boldsymbol{A_j - A_j Q_0 P_0^T} \right\|_F^2 + \boldsymbol{\lambda} \|\boldsymbol{Q_0}\|_F^2 \end{aligned}$$

Iteratively, we obtain

$$\boldsymbol{J(P_{t+1},Q_{t+1}) \leq J(P_t,Q_t) \quad for\ t = 0,1,2,.....}$$

Since the singular value decomposition (SVD) provides optimal $\boldsymbol{P_t}$ which decreases the value of objective function further. In other-words, the algorithms attains the optimal solution of the objective function in each iteration. Once, we compute the optimal value of matrix $\boldsymbol{Q}$ and $\boldsymbol{P}$, in the following iteration, the matrix $\boldsymbol{P_t}$ converges to local optima. Moreover, the objective function is convex. The sequence $\boldsymbol{J(P_t,Q_t)}$ is monotonically decreasing in each iteration. Thus, by the Monotonic Convergence Theorem, the objective function $\boldsymbol{J(P_t,Q_t)}$ converges to a local optimal value.

$$\sum_{j=1}^{N} \mathrm{tr}\left[\left(\boldsymbol{A_j^T - P_\infty Q_\infty^T A_j^T}\right)\left(\boldsymbol{A_j - A_j Q_\infty P_\infty^T}\right)\right] + \boldsymbol{\lambda}\,\mathrm{tr}\left(\boldsymbol{Q_\infty^T Q_\infty}\right)$$

■

4.2.5 Connections to Other PCA algorithm

In this section, we analyze the relations between our model and other PCA based method ($\ell_{2,1}$ norm). Below theorem validates our claim that proposed objective function provide robust and stable solution as compared to $\ell_{2,1}$ norm.

Theorem 4.4. *If $\boldsymbol{A}$ is an $\boldsymbol{m} \times \boldsymbol{n}$ matrix, then $\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_{2,1}$.*

Proof. Recall that

$$\begin{aligned} \|\boldsymbol{A}\|_F &= \sqrt{\mathrm{tr}\left(\boldsymbol{A}^T\boldsymbol{A}\right)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \\ \|\boldsymbol{A}\|_{2,1} &= \sum_{j=1}^{n} \|\mathbf{a}_j\|_2 = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} a_{i,j}^2} \end{aligned}$$

where $\mathbf{a}_j$ is the jth column of $\boldsymbol{A}$. With that in mind,

$$\begin{aligned} \|\boldsymbol{A}\|_{2,1}^2 &= \sum_{j=1}^{n} \|\mathbf{a}_j\|_2^2 + \underbrace{2\sum_{r=1}^{n}\sum_{s=1,s\neq r}^{n} \|\mathbf{a}_r\|_2 \, \|\mathbf{a}_s\|_2}_{\text{nonegative term}} \\ &\geq \sum_{j=1}^{n}\sum_{i=1}^{m} a_{i,j}^2 = \|\boldsymbol{A}\|_F^2 \end{aligned}$$

■

From Theorem 4.4, we can deduce that

$$\boldsymbol{arg\min_{Q,P}} \|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{Q}^T\boldsymbol{X}\|_{2,1} + \boldsymbol{\lambda}\|\boldsymbol{Q}\|_{2,1} \geq \boldsymbol{arg\min_{Q,P}} \|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{Q}^T\boldsymbol{X}\|_F + \boldsymbol{\lambda}\|\boldsymbol{Q}\|_F \tag{4.14}$$

The above Eq. 4.14 shows that the objective function is robust and provide stable solution as compare to $\ell_{2,1}$. In other words, ℓ_1 and $\ell_{2,1}$ penalizes the coefficients more than ℓ_F, however, robust solution can be obtained by selecting joint features using ℓ_F norm.

Theorem 4.5. *Notice that, if regression coefficient $\boldsymbol{\lambda} = \mathbf{0}$, then $\boldsymbol{Q} = \boldsymbol{P}$.*

$$\left(\sum_{j=1}^{N} \boldsymbol{A}_j^T\boldsymbol{A}_j\right)\boldsymbol{P} = \left(\sum_{j=1}^{N} \boldsymbol{A}_j^T\boldsymbol{A}_j\right)\boldsymbol{Q}$$

$$\boldsymbol{Q} = \left[\sum_{j=1}^{N} \boldsymbol{A}_j^T \boldsymbol{A}_j\right]^{-1} \left[\sum_{j=1}^{N} \boldsymbol{A}_j^T \boldsymbol{A}_j\right] \boldsymbol{P} = \boldsymbol{P}.$$

Moreover, the equation 4.11 simplifies to

$$\boldsymbol{J}(\boldsymbol{Q},\boldsymbol{P}) = \sum_{j=1}^{N} \left\| \boldsymbol{A}_j - \boldsymbol{A}_j \boldsymbol{P} \boldsymbol{P}^T \right\|_F^2$$

Hence, we can say that the proposed objection function degenerates to traditional 2DPCA. As such, the proposed objective function generalizes the 2DPCA. In this case, the optimal solution in Eq. 4.1 aims to find robust feature matrix.

4.3 Experimental Results

In order to evaluate the performance of proposed ORPCA, in this section, we have discussed and compared the performance of proposed ORPCA on four commonly used image dataset including AR [50], Yale B [98] , ORL and CMU PIE [24]. We have used k-nearest neighbour (where $\boldsymbol{k} = \mathbf{1}$) for classification. The main contribution of this work is introducing joint feature selection in order to select useful features by effectively combining the robustness of traditional two dimensional principal component analysis and the lasso regularization. Furthermore, we have introduced penalty term introduced in the objective function to exclude redundant features and provide robustness against outliers. Thus, to validate the our claims against outliers, we have corrupted the datasets with outliers to visualize the robustness of proposed approach in the presence of outliers. In addition, since 2D-RPCA is unsupervised method, we only compare its performance with unsupervised methods including PCA, 2DPCA, PCA$\boldsymbol{\ell_1}$, 2DPCA-$\boldsymbol{\ell_1}$ OMF-2DPCA and F-2DPCA on contaminated and non-contaminated benchmark datasets.

To validate the performance of of dimensionality reduction both persuasively and objectively, we have conducted several experiments on both original (non-contaminated) dataset and contaminated datasets. We have performed several of ORPCA at different $\boldsymbol{\lambda}$ value ($\mathbf{0} < \boldsymbol{\lambda} < \mathbf{1}$ to find optimal $\boldsymbol{\lambda}$. Once we have optimal value of $\boldsymbol{\lambda}$, we have performed 10-fold validation.

4.3.1 Datasets

AR face dataset consist of 120 individual, 26 images per individual taken in two session, with total images 3120 [49]. The dataset was captured in two different session at different

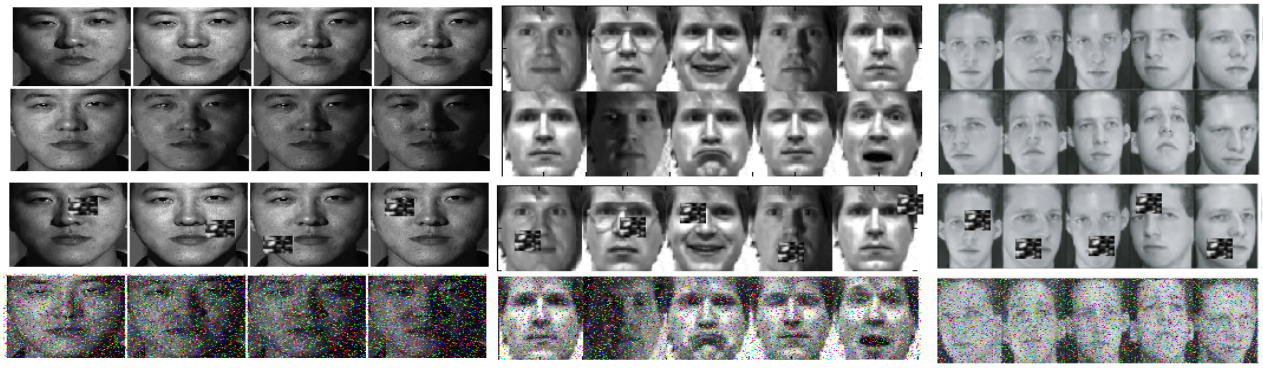

Figure 4.1: Sample images of CMU PIE, ORL, Yale and AR First two rows real dataset, Row 3 contaminated with block and Row 4 is contaminated with salt and chapter noise 15%

lightning condition and variable expressions. Face portion is cropped from their main images and then normalized to 32x32. Moreover, AR dataset consists of few images that are occluded with sunglasses, scarf or towels as shown in figure 4.1. In this experiment, we have considered face images with occlusion considered as noise images. Yale dataset consists of 64 images(except few 11-17,59-63), per subject with in total 2414 images under different lightning conditions from 38 individuals whereas half the dataset is corrupted by reflection or shadow. Figure 4.1 shows some reference of of Yale B dataset [139]. The database contains 5850 single light source images of 10 subjects (9 poses x 64 illumination conditions). For every subject in a particular pose, an image with ambient (background) illumination was also captured. ORL is face datset of 40 individuals with 10 images of each individual [94]. It consists of frontal views of faces with different expression and lightning conditions. CMU PIE dataset consists of 2856 frontal face images of 68 individual, 42 image per individual (with variation in lighting condition. We have selected 26 images randomly for training that consist of 7 noisy images [99].

We have resized the images in each dataset to 32 x 32 pixel. For training and evaluation purpose on non-contaminated datasets, we have divided 70%/30% and 80%/20% into training/testing. In order to validate the robustness of proposed method against outliers, **20**% images have been selected randomly and various types of noise (i.e. block occlusions, salt and peeper etc). We have added random noise (salt and peeper) with intensity of 10%, 15% on randomly selected images in each dataset as shown in figure 4.1. Similarly, we have added block occlusion of variable sizes at random locations with variable size (5x5, 10x10, 10x15) as shown in figure 4.1. In order to evaluate performance of proposed ORPCA on corrupted datasets, we have randomly selected 60% and 70% and

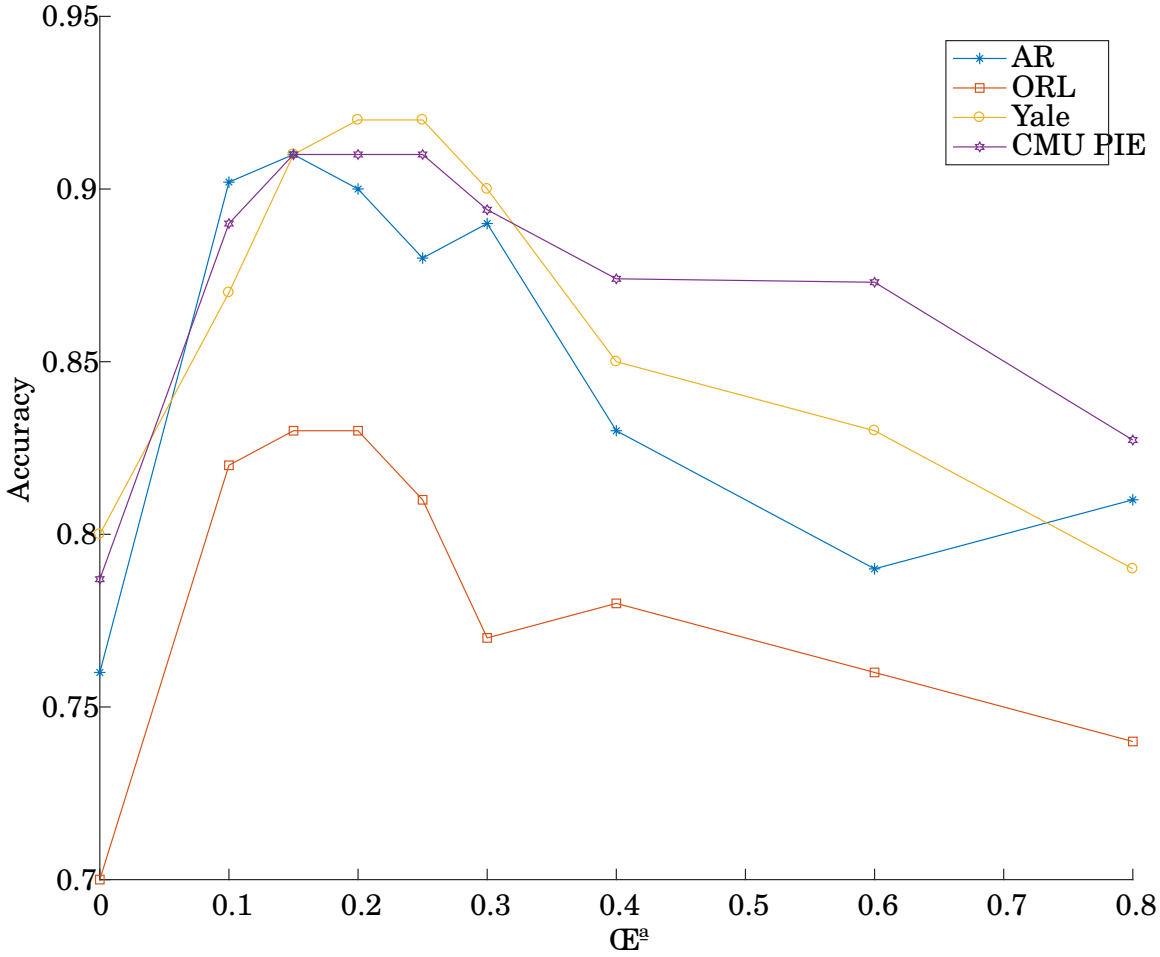

Figure 4.2: Classification performance at different value of λ for real (left) and contaminated (right) datasets

80% samples for each subject form each dataset as training set.

4.3.2 Parameter Selection

The objective function in equation 4.1 has only one parameter λ required to be be optimal. λ controls the regression coefficient. The greater value of λ could results in heavy penalty on regression coefficient that could affect the structural information, similarly smaller value of λ leads to 2DPCA. In order to find optimal λ, we have performed several experiments with different λ value with $\mathbf{0} \leq \lambda \leq \mathbf{4}$ and narrow down its range after few experiments based on its convergence and better accuracy. Firstly, we evaluated

Table 4.2: Average classification accuracy (accuracy ±corresponding standard deviation) on real dataset at optimal result of ORPCA

Dataset	PCA	RPCA	2DPCA	PCA2Dℓ_1	OMF-2DPCA	F-2DPCA	ORPCA
AR	0.6832 ± 0.005	0.6459 ± 0.008	0.7589 ± 0.0071	0.8477 ± 0.0023	0.8577 ± 0.0011	0.8782 ± 0.021	0.8932 ± 0.003
ORL	0.7891 ± 0.0028	0.8009 ± 0.0091	0.8843 ± 0.0411	0.8637 ± 0.0071	0.8623± 0.019	0.8754 ± 0.023	0.9254 ± 0.0091
Yale B	0.6886 ± 0.0031	0.5976 ± 0.0061	0.7911 ± 0.0091	0.7305 ± 0.0071	0.6743 ± 0.021	0.6643 ± 0.019	0.6934 ± 0.0131
CMU PIE	0.7445 ± 0.0091	0.7666 ± 0.0027	0.8987 ± 0.0026	0.8607 ± 0.0015	0.8608 ± 0.018	0.8522 ± 0.025	0.8947 ± 0.0041

Table 4.3: Comparative evaluation based on average classification accuracy ((accuracy ±corresponding standard deviation)) on contaminated datasets at optimal result of ORPCA

Dataset	PCA	RPCA	2DPCA	PCA2Dℓ_1	OMF-2DPCA	F-2DPCA	ORPCA
AR	0.5741 ± 0.0023	0.5387 ± 0.0022	0.6576 ± 0.0049	0.6277 ± 0.0053	0.781 ± 0.019	0.773 ± 0.021	0.8121 ± 0.014
ORL	0.6385 ± 0.0012	0.7411 ± 0.00321	0.8161 ± 0.0094	0.838 ± 0.0021	0.832 ± 0.016	0.856 ± 0.019	0.8892 ± 0.013
Yale B	0.5153 ± 0.0034	0.4865 ± 0.0083	0.5983 ± 0.0043	0.621 ± 0.0091	0.8109± 0.0031	0.80 ± 0.0017	0.82892 ± 0.0071
CMU PIE	0.577 ± 0.0032	0.5981 ± 0.0007	0.7181 ± 0.0091	0.6886 ± 0.0083	0.836± 0.021	0.8221± 0.012	0.8513 ± 0.008

on difference of **0.5** to find optimal interval where it provided better result followed by several experiments in selected interval. We have noticed that $\boldsymbol{\lambda}$ provided good accuracy between 0.15 to 0.25 for original datasets whereas it provided good accuracy between 0.1 to 0.3 for corrupted datasets. ORPCA achieved better performance over reasonable range of $\boldsymbol{\lambda}$. The value of $\boldsymbol{\lambda}$ marginally varies for different datasets, however, it provided best accuracy on interval [0.1,0.3], ideally when $\boldsymbol{\lambda}$ is close to 0.2. We have also noticed that accuracy was reduced when $\boldsymbol{\lambda}$=0 or $\boldsymbol{\lambda} \to \mathbf{0}$. Furthermore, as claimed in earlier section, ORPCA is a special case of 2DPCA, accuracy of ORPCA is same as 2DPCA when $\boldsymbol{\lambda} = \mathbf{0}$ which validates the claim "ORPCA is a special case of 2DPCA, it degenerates to 2DPCA when $\boldsymbol{\lambda} = \mathbf{0}$". Moreover, it indicates that $\boldsymbol{\lambda}$ is very important to achieve better robustness. Table 4.2 and Table 4.3 show that ORPCA achieved better accuracy over reasonable range of $\boldsymbol{\lambda}$ and robust to different setting of $\boldsymbol{\lambda}$ as long as it is in the range mentioned above. After selection of range of optimal $\boldsymbol{\lambda}$ generically, we performed experiment for

each dataset to find optimal $\boldsymbol{\lambda}$ explicitly for that datasets.

4.3.3 Evaluation on Original Dataset

In order to compare the performance of proposed objective function both persuasively and objectively, the classification is performed based on nearest neighbour. We have performed 10 fold validation on each dataset. We performed several experiments with variable sample size per individual i.e 60% and 70% and 80% samples for each individual subject and rest of samples are used for validation. The classification performance with different subspace dimensionality at optimal value of $\boldsymbol{\lambda} = \mathbf{0.18}$ is shown in table 4.2. Notice that, due to the dataset complexity (variations, pose, illumination and occlusions), getting high accuracy is quite challenging. Table 4.2 shows that proposed ORPCA achieved better classification in comparison to state of the art methods as shown in table 4.2, 4.3. Furthermore, we have notice that ORPCA selected important features that plays important role in classification.

4.3.4 Evaluation on Corrupted Dataset

In order to validate the robustness of proposed ORPCA against outliers and joint selection of features, we corrupted the dataset with outliers. In this experiment, we have randomly selected **70%** of images for corrupted datasets as a training set and consider rest of the images as a validation datasets. We have performed several experiments with different subspace dimensionality. Experimental results showed that the proposed ORPCA achieved much better performance as compared to state of the art methods in the presence of outliers that validate the robustness of proposed approach against outliers. Notice that ORPCA performed well for corrupted data however, it partially suffer from random corruption due to its joint feature selection ability.

4.3.5 Computational Complexity

Computation complexity of ORPCA has 3 steps in each iteration, First step is to compute Q using equation $\boldsymbol{Q} = \left[\sum_{j=1}^{N}\left(\boldsymbol{\lambda I_n} + \boldsymbol{A_j^T A_j}\right)\right]^{-1}\left[\sum_{j=1}^{N}\boldsymbol{A_j^T A_j}\right]\boldsymbol{P}$. Computational complexity of Q is $\boldsymbol{O(n^3)}$ as $\boldsymbol{A_j^T A_j}$ is the core step in computation of Q. The second step is to compute the SVD of $\sum_{j=1}^{N}\boldsymbol{A_j^T A_j Q}$, whose computational complexity is also $\boldsymbol{O(n^3)}$. Third step is to computation $\boldsymbol{P} = \boldsymbol{U I_{n\times d} V^T}$. Computation complexity of $\boldsymbol{P}$ is also $\boldsymbol{O(n^3)}$. Thus, computational complexity of one iteration is $\boldsymbol{O(n^3)}$. If the algorithm need $\boldsymbol{t}$ iteration to converge, it computation complexity will be $\boldsymbol{O(tn^3)}$.

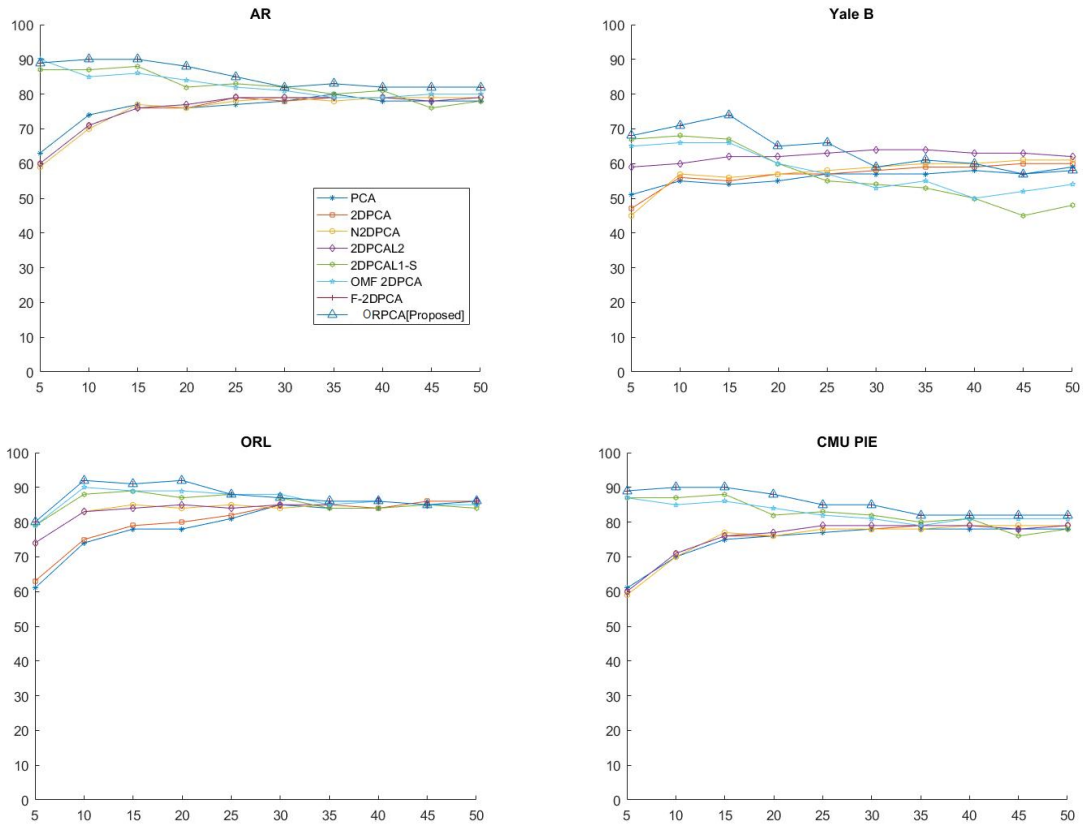

Figure 4.3: Comparative evaluation on real datasete (AR, Yale, ORL, and CMUIPIE)

4.3.6 Convergence Verification

To verify the convergence of algorithm 4.1, we tested different variations of parameters on all four datasets. The convergence of proposed ORPCA is shown in figure 4.5. It shows the convergence of objective function 4.1 along with each iteration. It can be found that objective function is non decreasing functions of iterations. As theorem 4.3 proves that ORPCA converges to local optima so does the case in figure 4.5 that shows that algorithm converges to local optima.

4.4 Discussion

We notice that methods based on matrix perform better as compared to vector based methods. Results shows that proposed ORPCA finds the representative features from high-dimensional space that are used for classification. Unlike 2DPCA based on ℓ_1-norm, ORPCA has rotational invariance property and has the freedom to jointly select the

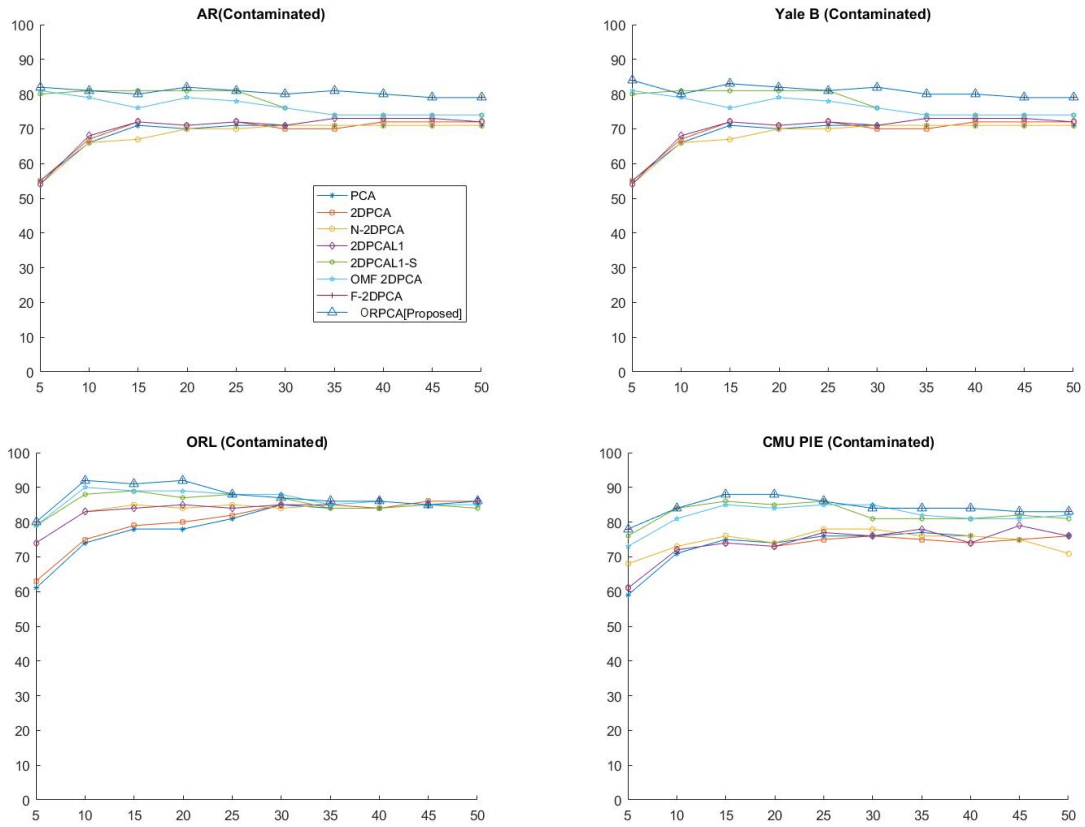

Figure 4.4: Comparative evaluation on corrupted datasete (AR, Yale, ORL, and CMUIPIE

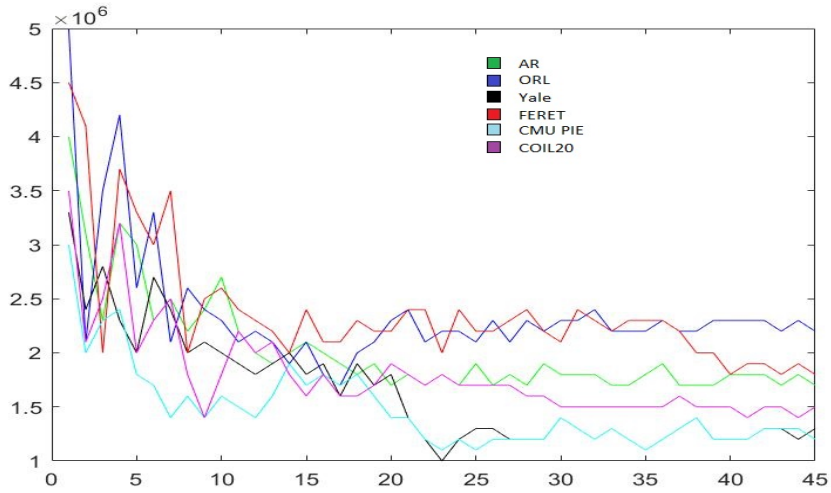

Figure 4.5: Convergence curve of ORPCA on four datasets

important and contributive features such as nose, eyes, lips in case of face image, while contours of different objects in non-facial datasets. Traditional methods are not able to interpret new features whereas it is quite important to interpret new features especially when they have spatial meaning. Results showed that ORPCA outperforms other PCA-based methods especially in the presence of outliers. This shows that proposed approach suppress the role of outliers. The proposed approach reveals the geometric structure due to the fact that it select the features by maintaining the spatial structural information of the image. It is due to the fact, that the solution of ORPCA relates to the weighted image covariance matrix which characterizes the spatial structure. We notice that the performance drop significantly with the increase in projection vectors.

4.4.1 Reconstruction Error

PCA can also be used for minimizing the reconstruction error with a few principal components. Reconstruction error is used as a parameter to measure the expressive capacity of the principal component. Let $\boldsymbol{X_i^*}...\boldsymbol{X_k^*}$ are the k contaminated images (1/4th of dataset) and $\bar{\boldsymbol{A}}$ is the mean of all images. On noisy dataset, reconstruction error is calculated for 2D-JSPCA as mentioned in equation 4.15 and for 2DPCA, 2DPCAL1, 2DPCAL1-S etc as mentioned in equation 4.16. Table 4.4 shows the comparative analysis of reconstruction error on four detest. Notice that, ORPCA has marginally poor reconstruction error as compared to others. This is due to the joint feature selection and ignoring the features that existing in other principal components.

$$\boldsymbol{E} = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}} ||\boldsymbol{X_i^*} - ((\boldsymbol{X_i^*} - \bar{\boldsymbol{X}})\boldsymbol{P}\boldsymbol{Q}^{\boldsymbol{T}} + \bar{\boldsymbol{X}})||_{\boldsymbol{F}}^{\mathbf{2}} \tag{4.15}$$

$$\boldsymbol{E} = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}} ||\boldsymbol{X_i^*} - ((\boldsymbol{X_i^*} - \bar{\boldsymbol{X}})\boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} + \bar{\boldsymbol{X}})||_{\boldsymbol{F}}^{\mathbf{2}} \tag{4.16}$$

4.4.2 Observations

Comparing with aforementioned experimental evaluation, we have the following interesting observations.

(I) The Objective function of the ORPCA degenerates into 2DPCA in case of $\boldsymbol{P}$ is equal to $\boldsymbol{Q}$ and $\boldsymbol{\lambda} = \mathbf{0}$. Thus, optimal $\boldsymbol{Q}$ in this case is the transformation matrix to accommodate the robustness against outliers in 2DPCA.

(II) Penalty term introduced in the objective function excludes redundant features and provides robustness against outliers, i.e., the regularization parameter $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\mathbf{2}}$ reduces the constraints and enables our method to jointly select features. In other-words, penalty term penalizes all regression coefficients corresponding to single feature as a whole to make PCA possible to select discriminant features jointly.

(III) Theoretical analysis shown in theorem 4.13 indicates that ORPCA is convergent to local optima as shown in figure 4.5.

(IV) We have noticed that discrimianant features selected by ORPCA are those important and contributive features such as nose, eyes, lips in case of face image, while contours of different objects in non-facial datasets.

Table 4.4: Average Reconstruction Error ($\times\mathbf{10^{-3}}$) and corresponding standard deviation of each approach on the Extended Yale B,AR, and CMU PIE databases

Methods/ Dataset	2DPCA	N-2DPCA	2DPCA-L1	2DPCAL1-S	OMF-2DPCA	F-Norm 2DPCA	ORPCA
AR	118.03 ± 2.98	119.21 ± 2.43	118.37 ± 3.47	118.02 ± 2.99	116.88 ± 2.16	117.97 ± 3.0	118.15 ± 2.11
Yale B	177.53 ± 1.6	176.59 ± 2.31	178.03 ± 2.5	177.11 ± 1.7	177.25 ± 1.78	176.93 ± 1.89	177.43 ± 1.66
CMU PIE	107.41 ± 2.1	106.41 ± 1.09	107.19 ± 1.46	107.37 ± 1.91	107.21 ± 1.33	106.87 ± 2.21	108.32 ± 1.45
CMU PIE	74.12 ± 0.80	80.47 ± 0.52	74.12 ± 0.80	80.15 ± 0.59	73.78 ± 1.21	73.80 ± 0.85	75.41 ± 1.43

4.5 Summary

In this chapter, we presented a robust dimensionality reduction method that by relaxing the orthogonal constraints of the transformation matrix and imposing a penalty function on regularization term. In contrast to previous work on robustness in PCA, we jointly select the important features. Introduction of penalty function results in the robustness against outliers by reducing their impact in projection matrix. Compared with state of the art methods, our evaluation results show the

improvement in effectiveness of ORPCA for image reconstruction and classification. In conclusion, the numerical results suggest that our method is superior to previous approaches. However, we observe that the objective function does not guarantee sparse solution.

CHAPTER 5

JOINT DIMENSIONALITY REDUCTION AND SPARSE FEATURE SELECTION

Data! Data! Data! I can't make bricks without clay!

A. C. Doyle

Data redundancy makes it a good candidate for sparse representation. Most of the existing dimensionality reduction methods try to preserve a certain kind of linear representation after projection. However, these methods either fail to select useful features or are not that efficient in the presence of outliers. In chapter 4, we present joint feature selection approach, whose factors themselves are not necessarily guaranteed to be sparse. To overcame the aforementioned issues of data redundancy in chapter 4, in this chapter, we introduce a novel approach called two dimensional joint sparse principal component analysis by effectively combining the robustness of 2DPCA and the sparsity-inducing regularization. The proposed approach relaxes the orthogonal constraints resulting in the joint features selection, besides avoiding the selection of same features in different principal components. In addition to provide sparse solution, the regularization term in the proposed objective function, improves the robustness against outliers. We demonstrate the significance and advantage of our methods on six publicly available benchmark data sets. Results showed that 2D-JSPCA provided better performance as compared to non-sparse methods (2DPCA and 2DPCA-L1) and

sparse methods (SPCA, JSPCA).

5.0.1 Motivation

As discussed in earlier sections, 2DPCA attempts to retain the spatial structural information. However, it could not deal with nonlinear data efficiently, and still dense in feature representation. As a result, it makes difficult to explain the resulting features i.e. projection procedure involves all the original features and it may have redundant or irrelevant features that considerably affect the performance of classification algorithm. For image classification, not only dimensional reduction, it is also important to find salient features that belong to specific part of image as projection procedure involves all the original features and it may have redundant or irrelevant features. To select such salient patterns, projection matrix should consist of sparse element with respect to such features. Thus, modeling sparsity into 2DPCA could help to encode semantic information, as well. Based on the above hypothesis, we have modeled sparsity into 2DPCA by imposing joint sparse constraint on the objective and added a penalty term. The objective function is shown in equation 5.1. The elastic net penalty is a convex combination of the ridge and lasso penalties as a result 2D-JSPCA not only jointly select useful features efficiently but also learn the transformation with sparsity.

The equation 5.1 integrates the ability of feature selection into subspace learning and provides sparse basis which accounts for joint feature selection as well as subspace learning. Imposing the additional penalty terms results in reduced sparsity and robustness against outliers. To provide some immediate motivation and application of matrix norms, we begin with an example that clearly brings out the issue sparse loadings. We have generated 350 data points including 20% outliers around straight line. We applied PCA, SPCA, JSPCA and 2D robust JSPCA and figures 5.1 shows the importance of proposed objective function against outliers which validate our claim.

5.1 2D Robust Joint Sparse PCA

Redundancy in the high dimensional data makes it a good candidate sparse representation. Most of the dimensionality reduction methods try to preserve a certain

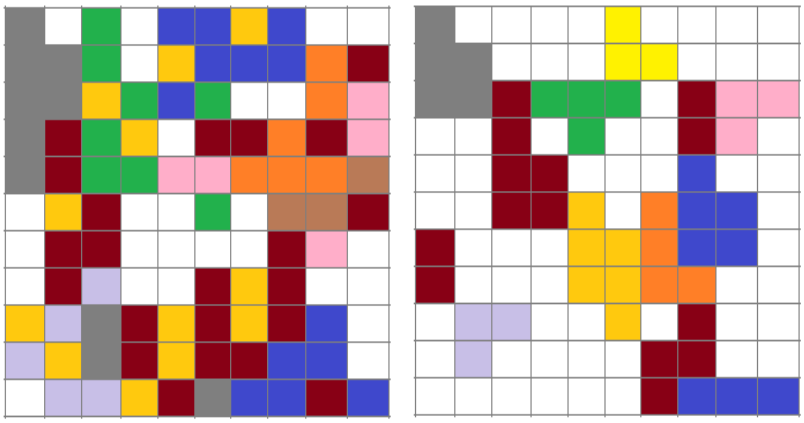

Figure 5.1: Illustration of SPCA (left) and 2D-JSPCA (right) using 10 x 11 matrix: white block represents zero loadings and color block represents different features

kind of linear representation after projection however, either these methods fail to select useful features or inefficient in the presence of outliers. Without sparsity, most of the loadings in high dimensional data are typically non-zero and redundant, however, we can reduce the number of non-zero components as well as data redundancy to manageable numbers without compromising the data reconstruction. In this section, we propose the two-dimensional robust joint sparse principal component analysis in detail. First, we present the motivation behind the proposed objective function, followed by the objective function and its derivation. Finally, we present an iterative optimal solution to solve the proposed objective function.

5.1.1 Objective Function

As described earlier, projection procedure involves all the original features and it may have redundant or irrelevant features. Considering the outliers appearance and consistent selection of optimal features, in this section, we propose two-dimensional joint sparse principal component analysis (2D-JSPCA) for reconstructing the data matrix that has more freedom to jointly select the useful features from low-dimensional representation. The objective of the proposed function is to effectively combine the robustness of 2DPCA and the sparsity-inducing regularization by imposing jointly sparse constraints on its objective function as well as introducing the additional penalty term. The addition of penalty term makes the

objective function robust against outliers as it penalizes all regression coefficient correspond to single feature as a whole.

Considering the data matrix $\boldsymbol{A}$ with outliers, where

$$\boldsymbol{A} = [\boldsymbol{X_1}, ..., \boldsymbol{X_N}]$$

where $\boldsymbol{X_j} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{n}}$.

We propose the following objective function

$$\min_{\boldsymbol{Q},\boldsymbol{P}} \boldsymbol{J}(\boldsymbol{Q},\boldsymbol{P}) = \min_{\boldsymbol{Q},\boldsymbol{P}} \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N}} \left\| \boldsymbol{X_j} - \boldsymbol{X_j} \boldsymbol{Q} \boldsymbol{P}^{\boldsymbol{T}} \right\|_{\boldsymbol{F}}^{\boldsymbol{2}} + \boldsymbol{\lambda_a} \|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\boldsymbol{2}} + \boldsymbol{\lambda_b} \|\boldsymbol{Q}\|_{\boldsymbol{2,1}} \tag{5.1}$$

where the matrix $\boldsymbol{Q} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{d}}$ transforms each sub-image into lower-dimensional subspace and the matrix $\boldsymbol{P} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{d}}$ recovers the data matrix. Furthermore, while we require $\boldsymbol{P}$ to be orthogonal ($\boldsymbol{P^T P} = \boldsymbol{I_d}$), we do not require the orthogonality of the matrix $\boldsymbol{Q}$. This enables the 2D-JSPCA has more freedom to learn low dimensional space that approximate to high dimensional data in flexible manner. In addition, the regularization parameter $\|\boldsymbol{Q}\|_{\boldsymbol{2,1}}$ reduces the constraints and enables our method to jointly select features whereas the regularization parameter $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\boldsymbol{2}}$ penalizes all regression coefficient correspond to single feature as a whole to makes PCA possible to select features jointly. Moreover, both the regularization terms $\|\boldsymbol{Q}\|_{\boldsymbol{2,1}}$ and $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\boldsymbol{2}}$ are convex and can easily be optimized iteratively. The parameter $\{\boldsymbol{\lambda_a}, \boldsymbol{\lambda_b}\} \geq \boldsymbol{0}$ balances the loss and regularization terms.

There are two core aims of this chapter: sparsity and outliers. Note that feature loadings across all the subspace dimensionality can not be ignored. The regularization term $\|\boldsymbol{Q}\|_{\boldsymbol{2,1}}$ learn the transformation with sparsity and extract only the important features. The other core issue is outliers. It is commonly known that squared F-norm is not robust in the sense that outlying measurements can arbitrarily skew the solution from the desired solution. We have proposed an efficient way to compute F-Norm to overcome its sensitivity challenge. Although the objective function is shown in equation 5.1 is based on square F-norm however, computation of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are not squared. Compared with squared F-norm, the proposed derivation can weaken the effect of large distance but also has rotational invariance. Moreover, the penalty terms $\|\boldsymbol{Q}\|_{\boldsymbol{2,1}}$ and $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\boldsymbol{2}}$ further enhances the robustness of

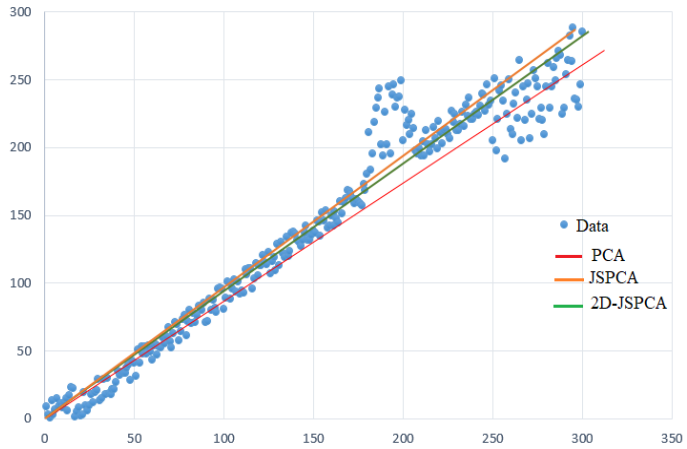

Figure 5.2: Illustration of 2D-JSPCA, JSPCA, PCA on 350 data points including 70 outliers: Results shows robustness of 2D-JSPCA against outliers

objective function against outliers through joint feature selection and discarding the features that already exist in other PCs to avoid redundancy. Although, the above objective function ensure sparseness and joint feature selection, simultaneous minimization of both $\boldsymbol{P}$ and $\boldsymbol{Q}$ make the problem non-convex and non-smooth, even though, all the sub terms in the objective function in Eq. 5.1 are convex. As a result, it cannot be approximated directly, however, the objective function is convex over the other term, if one of both terms is known. Iterative minimization of $\boldsymbol{P}$ and $\boldsymbol{Q}$ could result in locally convex solution. In order to optimize the objective function 5.1, we have divided the problem into two sub-problems and solved it iteratively. Firstly, we optimize the solution for $\boldsymbol{Q}$ with respect to $\boldsymbol{P}$ (at first, $\boldsymbol{P}$ is randomly initialize). Once, we the solution of $\boldsymbol{Q}$ obtained, the next step is to minimize the objective function with respect to $\boldsymbol{Q}$. The following derivation play a key role in determining the minimizers of the optimization problem 5.1. First, utilizing definition of the Frobenius norm, $\ell_{2,1}$-norm, the cyclic and linearity properties of the trace function, and the orthogonality of $\boldsymbol{P}$, we rewrite the objective function $\boldsymbol{J}$ in a more computationally tractable way.

$$J(P,Q) = \sum_{j=1}^{N} \left\| X_j - X_j Q P^T \right\|_F^2 + \lambda_a \|Q\|_F^2 + \lambda_b \|Q\|_{2,1} \tag{5.2}$$

$$= \sum_{j=1}^{N} \mathrm{tr}\left[\left(X_j^T - P Q^T X_j^T \right) \left(X_j - X_j Q P^T \right) \right] + \lambda_a \mathrm{tr}\left(Q^T Q \right) + 2\lambda_b \mathrm{tr}\left(Q^T D Q \right)$$

$$= \sum_{j=1}^{N} \mathrm{tr}\left(X_j^T X_j - X_j^T X_j Q P^T - P Q^T X_j^T X_j + P Q^T X_j^T X_j Q P^T \right) + \lambda \mathrm{tr}\left(Q^T Q \right) + 2\lambda_b \mathrm{tr}\left(Q^T D Q \right) \tag{5.3}$$

$$= \sum_{j=1}^{N} \mathrm{tr}\left(X_j^T X_j - 2 X_j^T X_j Q P^T + X_j^T X_j Q Q^T \right) + \lambda \mathrm{tr}\left(Q^T Q \right) + 2\lambda_b \mathrm{tr}\left(Q^T D Q \right) \tag{5.4}$$

Where D is $m \times m$ is diagonal matrix and can be computed as

$$D = \begin{bmatrix} \frac{1}{2\|Q\|_{2,1}} & & \\ & \ddots & \\ & & \frac{1}{2\|Q\|_{2,1}} \end{bmatrix} \tag{5.5}$$

Larger the value of D tends to force the objective function to small value. After several iteration, the objective function may tends to 0, which results in sparseness.

Now, differentiation Equation (5.4),

$$\frac{\partial J}{\partial Q} = \sum_{j=1}^{N} \left(-2 X_j^T X_j P + 2 X_j^T X_j Q \right) + 2\lambda_a Q + 2\lambda_b D Q.$$

$$\frac{\partial J}{\partial Q} = 2 \sum_{j=1}^{N} \left(-X_j^T X_j P + X_j^T X_j Q + \lambda_a Q + \lambda_b D Q \right)$$

Therefore,

$$\frac{\partial J}{\partial Q} = 0 \Rightarrow 2 \sum_{j=1}^{N} \left(-X_j^T X_j P + X_j^T X_j Q + \lambda_a Q + \lambda_b D Q \right)$$

$$\sum_{j=1}^{N} \left(X_j^T X_j P \right) = \sum_{j=1}^{N} \left(X_j^T X_j Q \right) + \lambda_a Q + 2\lambda_b D Q \tag{5.6}$$

$$\left(\sum_{j=1}^{N} \boldsymbol{X}_j^T \boldsymbol{X}_j\right) \boldsymbol{P} = \left(\sum_{j=1}^{N} (\lambda_a \boldsymbol{I}_n + \boldsymbol{X}_j^T \boldsymbol{X}_j) + \lambda_b \boldsymbol{D}\right) \boldsymbol{Q} \tag{5.7}$$

The above equation 5.7 is written as

$$\boldsymbol{Q} = \left(\sum_{j=1}^{N} \boldsymbol{X}_j^T \boldsymbol{X}_j\right) \boldsymbol{P} \left(\sum_{j=1}^{N} (\lambda_a \boldsymbol{I}_n + \boldsymbol{X}_j^T \boldsymbol{X}_j) + \lambda_b \boldsymbol{D}\right)^{-1} \tag{5.8}$$

$$\left(\sum_{j=1}^{N} \boldsymbol{X}_j^T \boldsymbol{X}_j\right) \boldsymbol{P} = \left(\sum_{j=1}^{N} (\lambda_a \boldsymbol{I}_n + \boldsymbol{X}_j^T \boldsymbol{X}_j) + \lambda_b \boldsymbol{D}\right) \boldsymbol{Q} \tag{5.9}$$

Hence, the result follows.

Once, the solution of $\boldsymbol{Q}$ is known, the next step is to optimize the objective function with respect to $\boldsymbol{P}$. For known $\boldsymbol{Q}$, the regularization penalty become irrelevant for the optimization with respect to $\boldsymbol{P}$. Theorem 5.1 describe the minimization of $\boldsymbol{P}$.

Theorem 5.1. *If $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ is the SVD (singular value decomposition) of $\sum_{j=1}^{N} \boldsymbol{X}_j^T \boldsymbol{X}_j \boldsymbol{Q}$, then*

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{I}_{n \times d}\boldsymbol{V}^T \tag{5.10}$$

is orthogonal and minimizes Equation (5.4) for a given $\boldsymbol{Q}$.

Proof. Recall that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of sizes $n \times n$ and $d \times d$, respectively. As such,

$$\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{V}\boldsymbol{I}_{n \times d}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{I}_{n \times d}\boldsymbol{V}^T = \boldsymbol{I}_d$$

■

Note that, orthogonal constraints $\boldsymbol{I}_n$ and $\lambda_b \boldsymbol{D}$ imposed the sparseness and deals with feature redundancy whereas Q projects the weighted data matrix and P is used to recover it.

Theorem 5.2. *The objective function consist of three terms* $||.||_F$, $\lambda_a||.||_F$ *and* $\lambda_b||.||_{2,1}$. *Observe that, if* $\lambda_a = 0$ *and* $\lambda_b = 0$, *then* $P = Q$, *as a result, the objective function degenerates to 2DPCA. In such case* Q *is not sparse.*

$$\left(\sum_{j=1}^{N} X_j^T X_j\right) P = \left(\sum_{j=1}^{N} X_j^T X_j\right) Q$$

$$Q = \left[\sum_{j=1}^{N} X_j^T X_j\right]^{-1} \left[\sum_{j=1}^{N} X_j^T X_j\right] P = P.$$

Furthermore, the objective function simplifies to

$$J(Q,P) = \sum_{j=1}^{N} \left\| X_j - X_j P P^T \right\|_F^2$$

Hence, the proposed 2D-JSPCA degenerates to 2DPCA. As such, in some sense 2D-JSPCA generalizes the 2DPCA. In this case, optimal solution in equation 5.1 aim to find optimal non-sparse matrix.

Theorem 5.3. *Considering* $\lambda_b = 0$, *we are left with two terms* $||.||_F$ *and* $\lambda_a||.||_F$. *In such case* $P \approx Q$, *The objective function in equation 5.1 degenerate robust 2DPCA.*

$$\left(\sum_{j=1}^{N} X_j^T X_j\right) P = \left(\sum_{j=1}^{N} (\lambda_a I_n + X_j^T X_j)\right) Q$$

$$Q = \left[\sum_{j=1}^{N} (\lambda_a I_n + X_j^T X_j)\right]^{-1} \left[\sum_{j=1}^{N} X_j^T X_j\right] P \approx P.$$

Furthermore, the objective function simplifies to

$$J(Q,P) = \sum_{j=1}^{N} \left\| X_j - X_j P P^T \right\|_F^2 + \lambda_a \|Q\|_F^2$$

In this case, optimal solution in equation 5.1 aim to find optimal non-sparse matrix.

Theorem 5.4. *Observe that, if* $\lambda_b > 0$ *and* $\lambda_a = 0$, *then* Q *is sparse. The amount of sparseness is controllable by the coefficient of the* Q *, given by the parameter* λ_b.

$$\min_{Q,P} \sum_{j=1}^{N} \left\| X_j - X_j Q P^T \right\|_F^2 + \lambda \|Q\|_{2,1} = \sum_{j=1}^{N} (-2X_j^T X_j P + 2X_j^T X_j Q) + 2\lambda_b D) \tag{5.11}$$

The objective function in equation 5.1 is to find sparse matrix Q *and orthogonal matrix* P

$$\left(\sum_{j=1}^{N} X_j^T X_j \right) P = \sum_{j=1}^{N} \left(\lambda_b D + X_j^T X_j \right) Q$$

Here, λ_b *is sparsity controlling parameter. Higher the value of* λ_b *leads to sparser components and vice versa.*

Hence, the proposed 2D-JSPCA is sparse when $\lambda > 0$.

There are two regularization term in the objective function. $||Q||_F$ and $||Q||_{2,1}$ impose robustness and sparseness respectively. We further prove that the proposed objective function is sparse, robust and provide stable solution. Theorem 5.2, Theorem 5.4 validates our claim that both P and Q in objective function shown in equation 5.1 imposes sparseness and robustness. We further elaborate with lemma 5.1 that proves that sparsity and robustness of objective function by relating it with $\ell_{2,1}$ norm.

Lemma 5.1. *If* A *is an* $m \times n$ *matrix, then* $\|A\|_F \leq \|A\|_{2,1}$.

Proof. Recall that

$$\|X\|_F = \sqrt{\mathrm{tr}\left(X^T X\right)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2}$$

$$\|X\|_{2,1} = \sum_{j=1}^{n} \|\mathbf{x}_j\|_2 = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} x_{i,j}^2}$$

where $\mathbf{a}_j$ is the j^{th} column of $\boldsymbol{A}$. With that in mind,

$$\begin{aligned}\|\boldsymbol{X}\|_{2,1}^2 &= \sum_{j=1}^{n} \|\mathbf{x}_j\|_2^2 + \underbrace{2\sum_{r=1}^{n}\sum_{s=1,s\neq r}^{n} \|\mathbf{x}_r\|_2 \, \|\mathbf{x}_s\|_2}_{\text{no-negative term}} \\ &\geq \sum_{j=1}^{n}\sum_{i=1}^{m} x_{i,j}^2 = \|\boldsymbol{X}\|_F^2\end{aligned}$$

■

5.1.2 Convergence Analysis

Before showing the convergence of 2D-JSPCA, we need to give the following lemma (Lemma 5.2) that plays important role in determining the proof of convergence of proposed objective function. The convergence of proposed objective function is explained in theorem 5.5.

Lemma 5.2. *For any nonzero vector $\boldsymbol{P},\boldsymbol{Q} \in \boldsymbol{R}^c$, the following results hold:*

$$\|\boldsymbol{P}\|_F - \frac{\|\boldsymbol{P}\|_F^2}{2\|\boldsymbol{Q}\|_F} \leq \|\boldsymbol{Q}\|_F - \frac{\|\boldsymbol{Q}\|_F^2}{2\|\boldsymbol{Q}\|_F}$$

Theorem 5.5. *Given all the variables in objective function equation 5.1, iterative scheme of 2D-JSPCA described in table 1 shows that objective function value is monotonically decreasing thus converge to local optima.*

Proof. Given the initial value of $\boldsymbol{P}$, say $\boldsymbol{P_0}$, we calculate $\boldsymbol{Q_0}$ by minimizing $\boldsymbol{J(P_0,Q)}$.
Hence,

$$\boldsymbol{J(P_0,Q_0) \leq J(P_0,Q)}$$

We calculate $\boldsymbol{P_1}$ by minimizing $\boldsymbol{J(P,Q_0)}$. Therefore,

$$\boldsymbol{J(P_1,Q_0) \leq J(P_0,Q_0)}$$

Since $\boldsymbol{Q_1}$ minimizes $\boldsymbol{J(P_1,Q)}$, we have

$$\boldsymbol{J(P_1,Q_1) \leq J(P_1,Q_0) \leq J(P_0,Q_0).}$$

That is

$$\begin{aligned} & \sum_{j=1}^{N} \left\| X_j - X_j Q_1 P_1^T \right\|_F^2 + \lambda_a \|Q_1\|_F^2 + \lambda_b \|Q_1\|_{2,1} \\ \leq & \sum_{j=1}^{N} \left\| X_j - X_j Q_0 P_1^T \right\|_F^2 + \lambda_a \|Q_0\|_F^2 + \lambda_b \|Q_0\|_{2,1} \\ \leq & \sum_{j=1}^{N} \left\| X_j - X_j Q_0 P_0^T \right\|_F^2 + \lambda_a \|Q_0\|_F^2 + \lambda_b \|Q_0\|_{2,1} \end{aligned}$$

Iteratively, we obtain

$$J(P_{t+1}, Q_{t+1}) \leq J(P_t, Q_t) \quad for\ t = 0, 1, 2, ...$$

Since the SVD as shown in step-III of table 5.1 gives the optimal P^t that further reduces the objective value. Once we obtained optimal value of P and Q, the next iteration further converge P to local optima. Hence, the sequence $J(P_t, Q_t)$ is monotonically decreasing. Thus, by the Monotonic Convergence Theorem, $J(P_t, Q_t)$ converges to a local optimal value.

$$\sum_{j=1}^{N} \mathrm{tr}\left[\left(X_j^T - P_\infty Q_\infty^T X_j^T \right) \left(X_j - X_j Q_\infty P_\infty^T \right) \right] + \lambda \mathrm{tr}\left(Q_\infty^T Q_\infty \right)$$

■

5.1.3 Numerical Algorithm

To optimize the objective function, we first minimize the objective function with respect to Q by computing the matrix Q using Equation (5.8), followed by SVD of $\sum_{j=1}^{N} X_j^T X_j Q$ and finally, minimizing the objective function with respect to P by updating the matrix. We have three terms in the objective function are $\left\| X_j - X_j Q P^T \right\|_F^2$, $\lambda_a \|Q\|_F^2$ and $\lambda_b \|Q\|_{2,1}$. Note that the regularization parameters λ_a an λ_b are used to balance the regularization terms and control the sparseness. The derivations in Theorem 5.2 shows the objective function is a generalization of 2DPCA. Below in table 5.1 we describe an iterative algorithm of 2D-JSPCA for training samples $X_1, ..., X_n$ of size $m \times n$, and regularization parameter λ. The minimization process alternating between P and Q until convergence.

Table 5.1: Algorithmic procedure of 2D Joint PCA

Input: $\boldsymbol{X_j} \in \mathbb{R}^{\boldsymbol{m \times n}}$ for $\boldsymbol{j = 1,...,N}$ where $\boldsymbol{A}$ is centralized, and parameter $\boldsymbol{\lambda}$. **Output:** Matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$

Step-I: Initialize the matrix $\boldsymbol{P}$
While not converge do
Step-II: Minimize the objective function with respect to $\boldsymbol{Q}$ by computing the matrix
$\boldsymbol{Q}$ using Equation $\boldsymbol{Q} = \left(\sum_{j=1}^{N} \boldsymbol{X_j^T X_j}\right) \boldsymbol{P} \left[\sum_{j=1}^{N} \left(\boldsymbol{\lambda_b D} + \boldsymbol{X_j^T X_j}\right)\right]^{-1}$
Step-III: Compute the SVD, $\boldsymbol{[U,S,V]} = \boldsymbol{SVD}(\sum_{j=1}^{N} \boldsymbol{X_j^T X_j Q})$
Step-IV: Minimize the objective function with respect to $\boldsymbol{P}$ by updating the matrix
$\boldsymbol{P}$ using Equation $\boldsymbol{P} = \boldsymbol{U I_{n \times d} V^T}$
end while

5.2 Results and Analysis

To evaluate the performance of proposed joint sparse 2DPCA, in this section, we have discussed and compared the performance of proposed 2D-JSPCA on six popular image dataset including AR, ORL FERET, Yale, COIL20 and CMU PIE. The main contribution is introducing two-dimensional joint sparse PCA by effectively combining the robustness of 2DPCA and the sparsity-inducing regularization. Perhaps most significantly, robustness is a strong property that can itself be used as an avenue to investigate different properties of the solution. We show that robustness of the solution can explain why the solution is sparse. Furthermore, we have introduced penalty term introduced in the objective function to exclude redundant features and provide robustness against outliers. We have performed evaluation of 2D-JSPCA at different $\boldsymbol{\lambda_a}$ value ($\boldsymbol{0 < \lambda < 1}$) and $\boldsymbol{\lambda_b}$ value (0.2, 0.4, 0.65) to find their optimal. In addition, since 2D-JSPCA is unsupervised method, we only compare its performance with unsupervised methods including PCA, PCAL1, JSPCA, 2DPCA, R2DPCA, PCA2DL1 and PCA2DL1-S on corrupted and non-corrupted benchmark datasets.

In order to compare the performance of dimensionality reduction both objectively

and persuasively, we evaluated using nearest neighbour classifier. Experiment is divided into two groups: Experiment-I is on non-contaminated (original) dataset and experiment-II is on contaminated datasets in order to validate the robustness against outliers. We have corrupted small portion of all dataset by adding random noise (salt and pepper or block of different sizes) as shown in figure 5.3. It shows some original images in first two rows and third row is corrupted by random blocks whereas forth row is contaminated by 10% and 15% salt and pepper noise.

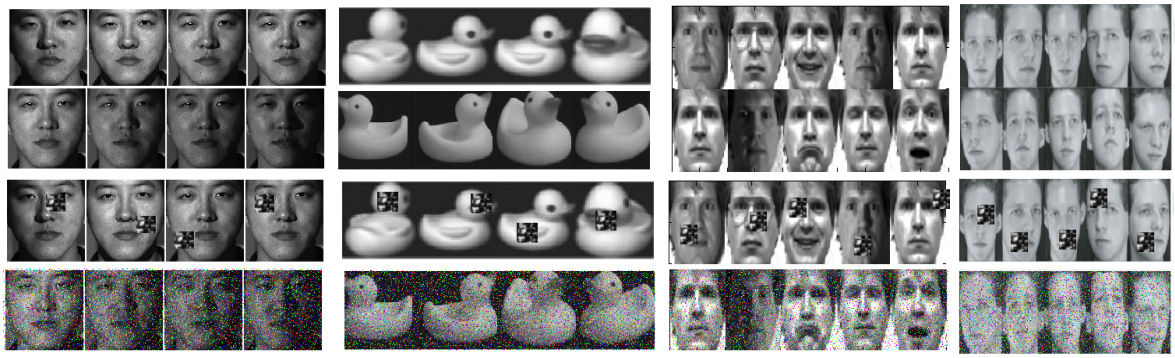

Figure 5.3: Sample images of CMU PIE, COIL20, Yale [98] and AR [50] First two rows real dataset, row 3 contaminated with block and Row 4 is contaminated with salt and pepper noise 15%

5.2.1 Datasets

FERET dataset consists of 7 images each of 200 individuals with total images 1400 [64]. Image size is 80 by 80 with variable expression as shown in figure 5.3. AR face dataset consist of 120 individual, 26 images per individual taken in two session, with total images 3120 [49]. The dataset was captured in two different session at different lightning condition and variable expressions. Face portion is cropped from their main images and then normalized to 32x32. Moreover, AR dataset consists of few images that are occluded with sunglasses, scarf or towels as shown in figure 5.3. In this experiment, we have considered face images with occlusion considered as noise images. Yale dataset consists of 64 images(except few 11-17,59-63), per subject with in total 2414 images under different lightning conditions from 38 individuals whereas half the dataset is corrupted by reflection or shadow. Figure 5.3 shows some reference of Yale B dataset [139]. The database contains 5850 single light source images of 10 subjects (9 poses x 64 illumination conditions). For every

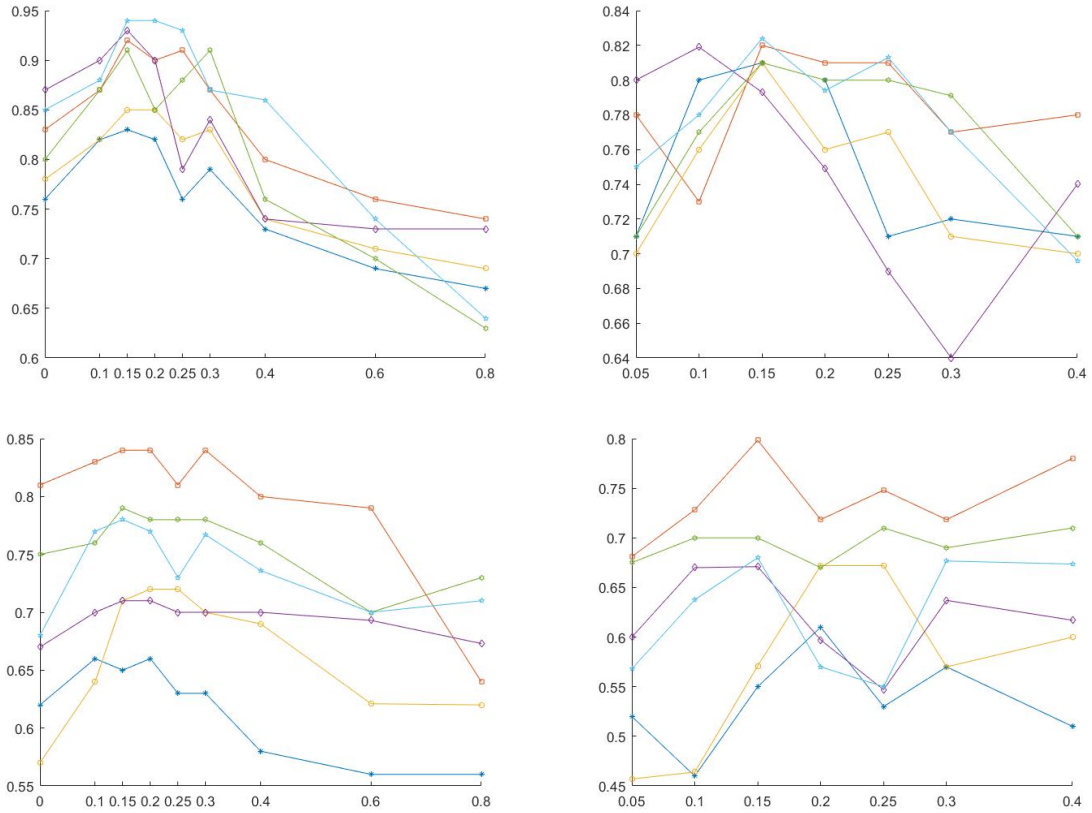

Figure 5.4: Comparative evaluation at different value of $\boldsymbol{\lambda_a}$ (left column) and $\boldsymbol{\lambda_b}$ (right column) for real (top row) and contaminated (bottom row) datasete

subject in a particular pose, an image with ambient (background) illumination was also captured. ORL is face datset of 40 individuals with 10 images of each individual [94]. It consists of frontal views of faces with different expression and lightning conditions. CMU PIE dataset consists of 2856 frontal face images of 68 individual, 42 image per individual (with variation in lighting condition. We have selected 26 images randomly for training that consist of 7 noisy images [99]. In this chapter, we have also considered non-facial dataset COIL100 and converted it into grey scale [54]. It consists of 7200 images of 100 individual captured at pose interval of $\mathbf{5}^{\circ}$.

All six datasets images are re-sized to 32 x 32 pixel. For training and evaluation purpose on non-contaminated datasets, we have divided 70%/30% and 80%/20% into training/testing. In order to validate the robustness of proposed method against outliers, we have randomly selected **20%** images to add noise in the datasets. We

corrupted the dataset, i.e., random noise well as block occlusions. Random noise is salt and pepper noise spread randomly at 10%, 15% on random selection of images from dataset as shown in figure 5.3. Similarly, block occlusion is added block of different sizes at random locations with variable size 5x5, 10x10, 10x15 as shown in figure 5.3. For evaluation on contaminated datasets, we have selected 60% and 70% and 80% samples per individual for each dataset as training dataset (3,5 and 7 ; 8, 13 and 18 ; 22, 27 and 32 ; for ORL, AR and Yale datasets respectively). We have conducted various number of experiments on each dataset and average classification accuracy is computed as shown in figure 5.5 and 5.6 , table 5.2 and 5.3.

Table 5.2: Comparative evaluation based on average classification accuracy on real dataset at optimal result of 2DJSPCA

Dataset	PCA	RPCA	SPCA	JSPCA	2DPCA	PCA2DL1	2DJSPCA
AR	0.6832 ± 0.005	0.6459 ± 0.008	0.7345 ± 0.0221	0.7896 ± 0.006	0.7589 ± 0.0071	0.8477 ± 0.0023	0.8541 ± 0.003
ORL	0.7891 ± 0.0028	0.8009 ± 0.0091	0.8322 ± 0.0011	0.8981 ± 0.0032	0.8843 ± 0.0411	0.8637 ± 0.0071	0.9254 ± 0.0091
Yale	0.6886 ± 0.0031	0.5976 ± 0.0061	0.5723 ± 0.0009	0.7563 ± 0.0021	0.7911 ± 0.0091	0.7305 ± 0.0071	0.8634 ± 0.0531
FERET	0.8400 ± 0.0039	0.8322 ± 0.0039	0.8409 ± 0.0014	0.9222 ± 0.0022	0.9112 ± 0.0042	0.8900 ± 0.0022	0.9461 ± 0.0009
CMU PIE	0.7445 ± 0.0091	0.7666 ± 0.0027	0.8334 ± 0.0091	0.9011 ±**1** 0.0011	0.8987 ± 0.0026	0.8607 ± 0.0015	0.9347 ± 0.0041
COIL20	0.7923 ± 0.0023	0.7523 ± 0.0044	0.7744 ± 0.0012	0.8587 ± 0.0042	0.8639 ± 0.0036	0.8688 ± 0.0032	0.9245 ± 0.0007

5.2.2 Parameter Selection

The objective function in equation 5.1 has only two parameter $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$ required to be be optimal to make the solution robust and sparse. In order to find optimal $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$, we have performed several experiments with different $\boldsymbol{\lambda_a}$ value with $\mathbf{0} \leq \boldsymbol{\lambda_a} \leq \mathbf{4}$ and narrow down its range after few experiments based on its convergence and better accuracy. Similarly, we performed several experiment to find optimal $\boldsymbol{\lambda_b}$ for each value of $\boldsymbol{\lambda_a}$ to find best values of both parameters. Firstly, we evaluated on difference of $\mathbf{0.5}$ in $\boldsymbol{\lambda_a}$ value to find optimal interval where it provided better result followed by several experiments in selected interval. For each value of $\boldsymbol{\lambda_a}$, we performed six different experiment on $\boldsymbol{\lambda_b} \in \{\mathbf{0.05, 0.1, 0.15, 0.25, 0.3 and 0.4}\}$

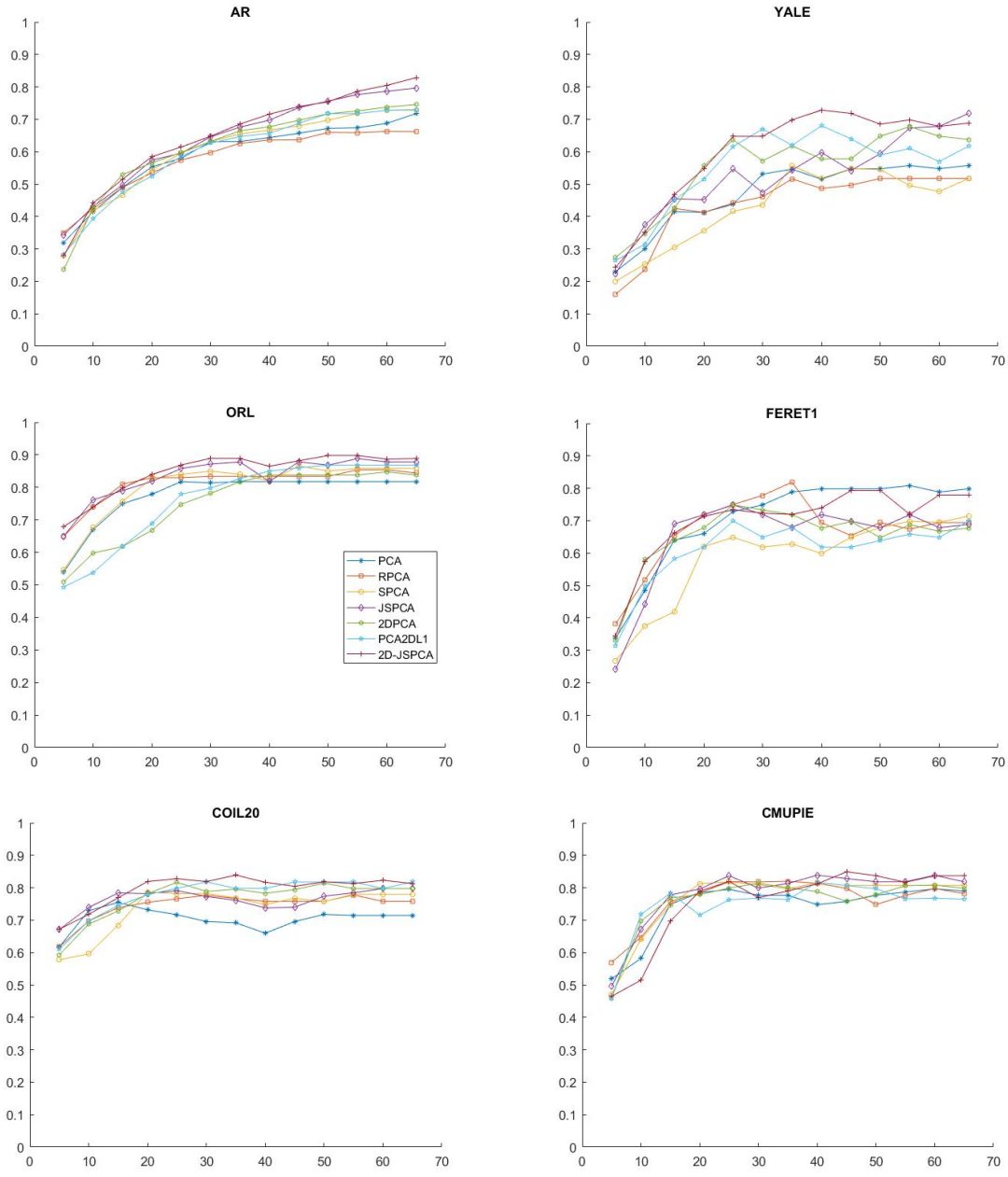

Figure 5.5: Comparative evaluation on real datasete (AR, Yale, ORL, FERET, COIL20 and CMUIPIE

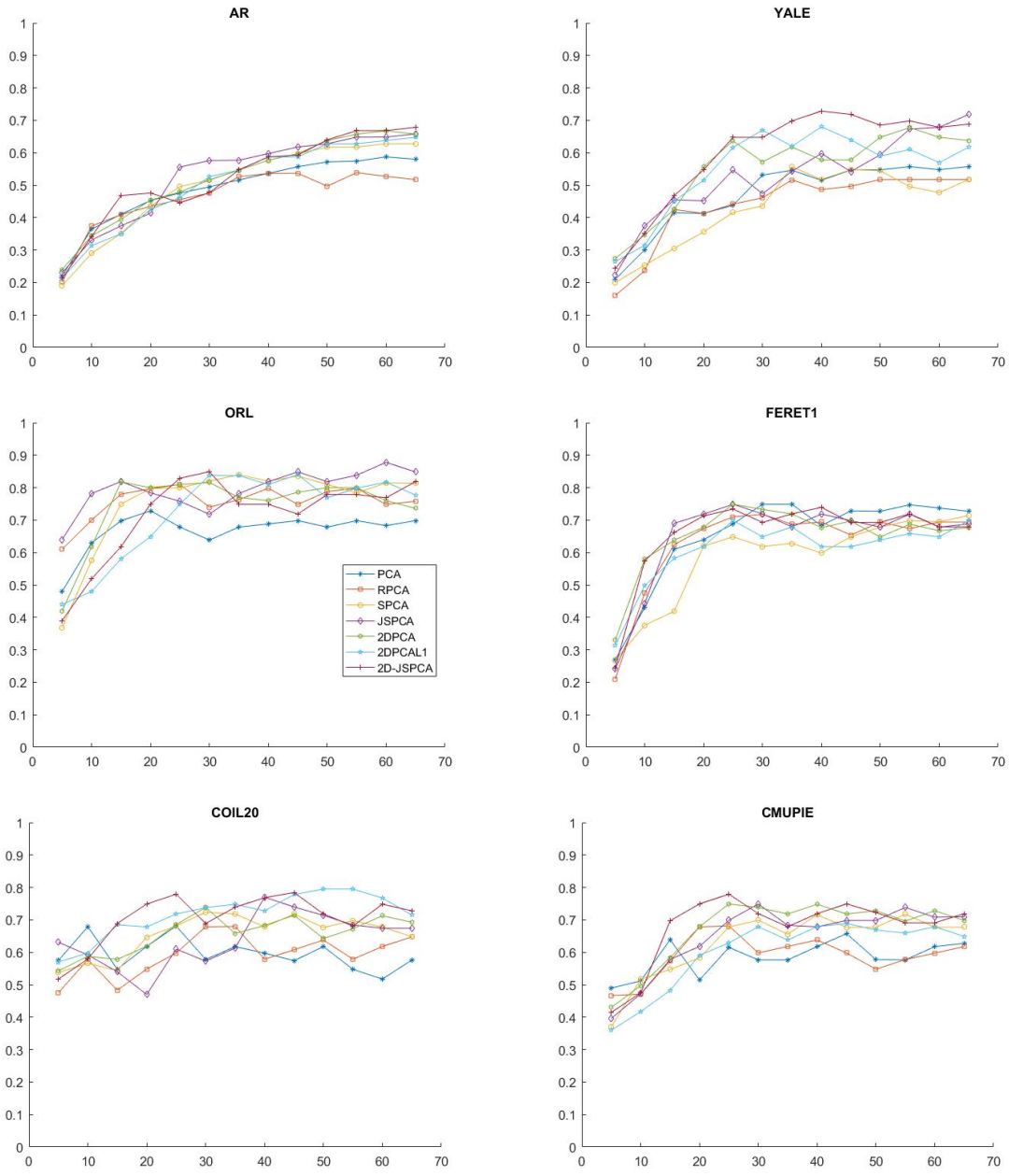

Figure 5.6: Comparative evaluation on corrupted datasete (AR, Yale, ORL, FERET, COIL20 and CMUIPIE

Table 5.3: Comparative evaluation based on average classification accuracy on contaminated dataset at optimal result of 2DJSPCA

Dataset	PCA	RPCA	SPCA	JSPCA	2DPCA	PCA2DL1	2DJSPCA
AR	0.5741 ± 0.0023	0.5387 ± 0.0022	0.6178 ± 0.0091	0.6481 ± 0.0091	0.6576 ± 0.0049	0.6277 ± 0.0053	0.6621 ± 0.0203
ORL	0.6385 ± 0.0012	0.7411 ± 0.00321	0.8201 ± 0.0081	0.7181 ± 0.0087	0.8161 ± 0.0094	0.838 ± 0.0021	0.8492 ± 0.00221
Yale	0.5153 ± 0.0034	0.4865 ± 0.0083	0.5177 ± 0.0073	0.5978 ± 0.0065	0.5983 ± 0.0043	0.621 ± 0.0091	0.72892 ± 0.0091
FERET	0.6831 ± 0.0029	0.6948 ± 0.0042	0.59851 ± 0.0065	0.7186 ± 0.0043	0.6771 ± 0.0054	0.6184 ± 0.0087	0.7391 ± 0.0065
CMU PIE	0.577 ± 0.0032	0.5981 ± 0.0007	0.6771 ± 0.0054	0.6987 ± 0.0054	0.7181 ± 0.0091	0.6886 ± 0.0083	0.7513 ± 0.0088
COIL20	0.5743 ± 0.0024	0.6081 ± 0.0032	0.7179 ± 0.0049	0.7474 ± 0.0076	0.7144 ± 0.0077	0.7786 ± 0.0053	0.7844 ± 0.0141

to find approximate value of $\boldsymbol{\lambda_a}$. Once, we have approximated $\boldsymbol{\lambda_b}$, we performed different experiment around that value to find its optimal value. Figure 5.4 shows the influence of different $\boldsymbol{\lambda_a}$ values. In general, we have noticed that $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$ provided good accuracy between [0.15-0.25] and [0.045-0.15] for original datasets whereas it provided good accuracy between [0.1-0.3] and [0.05-0.1] for corrupted datasets respectively. As shown in figure 5.4, 2D-JSPCA achieved better performance over reasonable range of $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$. The value of $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$ marginally varies for different datasets, however, it provided best accuracy on interval [0.1,0.3] and [0.04, 0.08] for $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$ respectively. Ideally, it provided the best results when $\boldsymbol{\lambda_a}$ is close to $\mathbf{0.2 \pm 0.1}$ and $\boldsymbol{\lambda_b}$ is close to $\mathbf{0.05 \pm 0.03}$ as shown in figure 5.4.a and figure 5.4.b.

We have also noticed that accuracy was reduced when $\boldsymbol{\lambda_b}$=0 as shown in figure 5.4 or $\boldsymbol{\lambda_a} = \mathbf{0}$ or $\{\boldsymbol{\lambda_a}, \boldsymbol{\lambda_b}\} = \mathbf{0}$. Furthermore, as claimed in earlier section, 2D-JSPCA is a special case of 2DPCA, Figure 5.4 shows that accuracy of 2D-JSPCA is same as 2DPCA when $\boldsymbol{\lambda} = \mathbf{0}$ which validates the claim "2D-JSPCA is a special case of 2DPCA, it degenerates to 2DPCA when $\boldsymbol{\lambda_b} = \mathbf{0}$". Moreover, it indicates that $\boldsymbol{\lambda_b}$ is very important to achieve better sparseness. Figure 5.5 and 5.6 show that 2D-JSPCA achieved better accuracy over reasonable range of $\boldsymbol{\lambda_b}$ and $\boldsymbol{\lambda_b}$. Similarly, it is robust against outliers at different setting of $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$ as long as it is in the range mentioned above. After selection of range of optimal $\boldsymbol{\lambda_a}$ generically, we performed

experiment for each dataset to find optimal λ_b explicitly for that datasets.

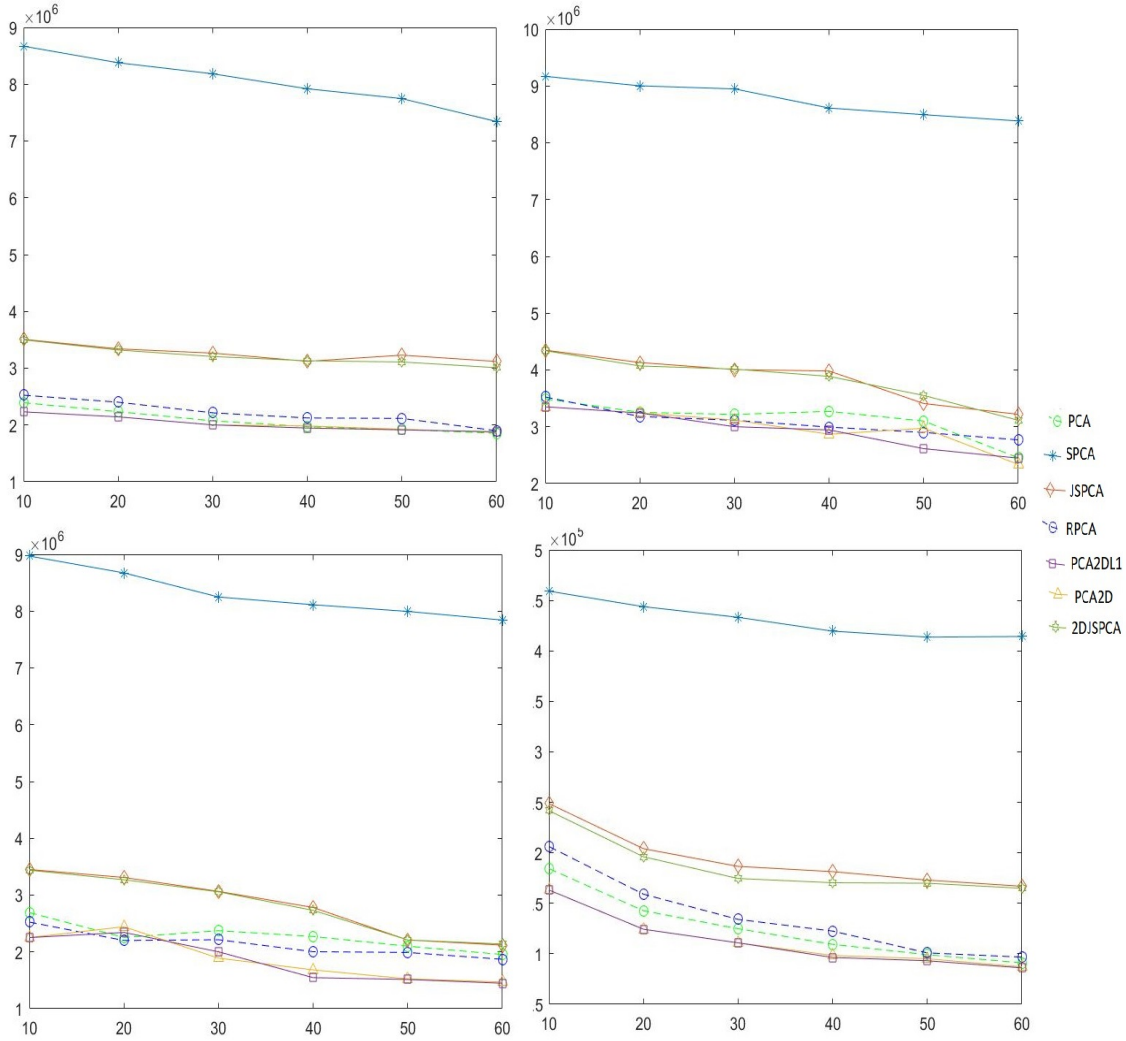

Figure 5.7: Comparison:Reconstruction Error versus features numbers (a) AR (b) ORL (c) Yale (d) COIL20

5.2.3 Evaluation on Original Datasets

In order to compare the performance of proposed objective function both objectively and persuasively, we have used nearest neighbor to obtain classification accuracy. We have repeated experiment on each dataset ten times and average evaluation results show that as classifier, 2D-JSPCA achieved better accuracy as compared to

JSPCA [127], SPCA [16], R2DPCA [115] and 2DPCAL1-S [112] as shown in table 5.2, 5.3 and figure 5.5 and 5.6.

In first experiment, we have selected datasets with original but re-sized to 32x32. Table 2 shows the variation of classification accuracy with different subspace dimensionality at optimal $\boldsymbol{\lambda_a} = \mathbf{0.18}$ and $\boldsymbol{\lambda_b} = \mathbf{0.05}$. For evaluation purpose, we have selected 60% and 70% and 80% samples per individual for each dataset as training dataset and rest of the datasets for validation.

Due to the complexity of datasets such as illumination, variations and occlusions etc, it is quite challenging to obtain high classification accuracy, however, the experimental results show that 2D-JSPCA obtained better classification accuracy as compare to the PCA, 2DPCA, SPCA, and JSPCA. It is due to sparsity in two dimension, selection of robust features as well as discarding the redundant patterns. Furthermore, it enables the 2D-JSPCA has more freedom to learn low dimensional space that approximate to high dimensional data in a flexible manner. In addition, regularization term $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\mathbf{2}}$ is convex and can easily be optimized as it can gradually trending to smaller value iteratively. Moreover, it reduces the constraints and enables our method to jointly select features.

5.2.4 Robustness against Outliers

To investigate the performance of 2D joint sparse PCA, 25% of the dataset are contaminated with random noise and block occlusions. Rectangular noise located at the different position of different size (10x10 and 20 x 20) is added as shown in figure 5.3 whereas random noise is salt and pepper noise spread randomly at 10%, 15% on random selection of images from datasets. After dataset corruption, we have selected **70%** of corrupted images for training and rest of the images are part of validation datasets. Results show that 2D-JSPCA performed well for corrupted data as compared to other PCA-based methods as shown in figure 5.5 and table 5.3, however, it suffers from random corruption due to its feature selection ability. In conclusion, we can say that 2D-JSPCA is robust to slight variations rather than random variations in the datasets.

5.2.5 Reconstruction Error

Form the figure 5.7, we can notice that 2D-JSPCA provided poor in term of reconstruction as compared to non-sparse methods due to the loss of extensive

information, however, in comparison to sparse methods, 2D-JSPCA reconstruction is better and able to select those features for that are effective for reconstruction as shown in figure 5.7.

In conclusion, we can say that 2D-JSPCA finds the representative features from high-dimensional space that are good for classification. Results showed that 2DJSPCA outperforms other PCA-based methods especially SPCA and Joint SPCA in term of classification as it selected features jointly by maintaining the images spatial structural information.

5.2.6 Computational Complexity

Computation complexity of 2D-JSPCA has 3 steps in each iteration, First step is to compute $\boldsymbol{Q}$, $\boldsymbol{Q} = \boldsymbol{Q} = \left[\sum_{j=1}^{N}(\lambda_a \boldsymbol{I_n} + \boldsymbol{X_j^T X_j})\right]^{-1}\left[\sum_{j=1}^{N}\boldsymbol{X_j^T X_j}\right]\boldsymbol{P} \approx \boldsymbol{P}$. Computational complexity of $\boldsymbol{Q}$ is $\boldsymbol{O(n^3)}$ as $\boldsymbol{X_j^T x_j}$ is the core step in computation of $\boldsymbol{Q}$. The second step is to compute the SVD of $\sum_{j=1}^{N}\boldsymbol{X_j^T X_j Q}$, whose computational complexity is also $\boldsymbol{O(n^3)}$. Third step is to computation $\boldsymbol{P} = \boldsymbol{U I_{n\times d} V^T}$. Computation complexity of $\boldsymbol{P}$ is also $\boldsymbol{O(n^3)}$.

Thus, computational complexity of one iteration is $\boldsymbol{O(n^3)}$. If the algorithm need $\boldsymbol{t}$ iteration to converge, it computation complexity will be $\boldsymbol{O(tn^3)}$.

5.2.7 Observations

In this chapter, we introduced penalty terms to accommodate sparseness and robustness in the 2D principal component analysis. By mean of case study on benchmark dataset and simulating the outliers, the experiment showed excellent performance against outliers, with better construction as compared to state of the art sparse methods. Comparing with aforementioned experimental evaluation, we have the following interesting observations.

(A) According to the Theorem 5.4, the Objective function of the 2D-JSPCA degenerates into 2DPCA in case of $\boldsymbol{P}$ is equal to $\boldsymbol{Q}$ and $\{\boldsymbol{\lambda_a, \lambda_b}\} = \boldsymbol{0}$. Thus, optimal $\boldsymbol{Q}$ in this case is the transformation matrix to accommodate the sparsity in 2DPCA.

(B) Penalty terms introduced in the objective function excludes redundant features and provides robustness against outliers, i.e., the regularization parameter

$\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\mathbf{2}}$ and $\|\boldsymbol{Q}\|_{\mathbf{2,1}}$ reduces the constraints and enables our method to jointly select features. In other-words, penalty terms penalizes all regression coefficients corresponding to single feature as a whole to make PCA possible to select discriminant features jointly.

(C) Eventually, 2D-JSPCA has poor reconstruction error because it suffers from loss of information. However, it provides better reconstruction error with respect to SPCA and JSPCA. It might be due to the selection of important features that helps to reproduce the image.

(D) Theoretical analysis shown in theorem 5.5 indicates that 2D-JSPCA is convergent to local optima. Furthermore, we noted that higher sparsity leads to slower convergence.

(D) We have noticed that discrimianant features selected by 2D-JSPCA are those important and contributive features such as nose, eyes, lips in case of face image, while contours of different objects in non-facial datasets.

There is significant scope of the extension of this work. First, one can look on to multiple value of $\boldsymbol{P}$ and $\boldsymbol{Q}$ i.e. having more than one $\boldsymbol{P}$ and one $\boldsymbol{Q}$, offers more flexibility in accommodating the discriminant features locally and could achieve better sparseness.

5.3 Summary

In this chapter, we present a new subspace learning method, robust joint sparse solution for two dimensional principal component analysis by relaxing the orthogonal constraints of the transformation matrix and imposing a penalty function on regularization term. Results validate the claims that proposed approach is robust against outliers and able to select important features. 2D-robust JSPCA has the freedom to jointly select the important features, thus, only few features could represent the whole data efficiently. This property makes it suitable for compressed sensing. We have noticed that discriminant features selected by 2D-JSPCA are those important and contributive features such as nose, eyes, lips in case of face image, while contours of different objects in non-facial datasets. Evaluation results on benchmark datasets contaminated with outliers show the improvement in effectiveness of 2D-robust JSPCA for image reconstruction and classification.

We expect that proposed method can be used in various applications especially in field of compressed sensing.

Part II

Regualizer Optimization

CHAPTER 6

SUPPORT MATRIX MACHINE

Classification of mathematical problems as linear and non-linear is like classification of the universe as bananas and non-bananas

In many real-world classification problems of supervised tensor learning, high-dimensional data is represented as a matrix, also referred to as second order tensors. Traditional support vector machines (SVMs) require data to reshape each matrix into vectors, thus, resulting in loss of structural information of the originally featured matrix. In this chapter, we propose Robust Sparse Support Matrix Machine (RSSM) which is defined as hinge loss and regularization term as spectral elastic net penalty. The regularization term which promotes the structural sparsity and shares similar sparsity patterns across multiple predictors that is able to select useful features jointly, is a combination of $\boldsymbol{\ell_{2,1}}$ and nuclear norm. It is a spectral extension of the conventional elastic net that combines the property of low-rank and joint sparsity together, to deal with complex high dimensional noisy data. Furthermore, it also leverages the structural information as well as the intrinsic structure of data and avoids the inevitable upper bound. A comprehensive experimental study on the publicly available data set is carried out to validate the proposed approach. The experiment results, supported by the theoretical analysis and statistical test, show the effectiveness of the RSSM for solving classification problems while keeping a reasonable number of support vectors.

6.1 Motivation

In this chapter, our concern is the classification problems on a set of data matrix as structural information of the original features is very important for certain data analytic tasks. Input data is high in dimensions and noisy, hence, we focus our attention on regularizers that have the ability to promotes sparsity and robustness against outliers, so that they can be used for selecting certain features. Moreover, our target is to endow the feature space that does not penalize the features individually as in the case of the ℓ_1 norm. To leverage the structural information as well as dimensionality challenge, we employed the regularizers term into SVM which promote structural sparsity and model the intrinsic structure. As a result, the regularization term $\ell_{2,1}$ norm along with the nuclear norm and loss not only helps to avoid the inevitable upper bound for the number of selected features but also combines the property of low-rank and sparsity together. Furthermore, the loss function based on $\ell_{2,1}$ and the nuclear norm could help to overcome the outliers as methods based on ℓ_2 [46] and ℓ_1 [141] are sensitive to outliers. From figure 6.1, we can notice that sparse and the low-rank can leverage the topological structural information of a matrix, similarly, ℓ_1 norm does not consider the intrinsic group structure whereas Figure 1(d) shows that RSSM helps to jointly select useful features for low-dimensional representation.

6.2 The proposed RSSM

In this section, we introduce the proposed RSSM, which, as a matter of fact, is a novel classifier, that not only removes the redundant information but also selects the discriminant patterns as well as considers the strong correlation of rows and columns in the matrix. Although the objective function in equation 3.3 is the combination of sparse and low-rank properties but the ℓ_1 norm regularizer term provides structural sparsity and ignores the intrinsic structure as it tends to select the features without considering all the classes, which can be obtained using the $\ell_{2,1}$ norm.

6.2.1 Objective Function

It is well known that hinge loss enjoys a large margin as it provides a tight and convex upper bound on the indicator function which penalizes misclassifications. It

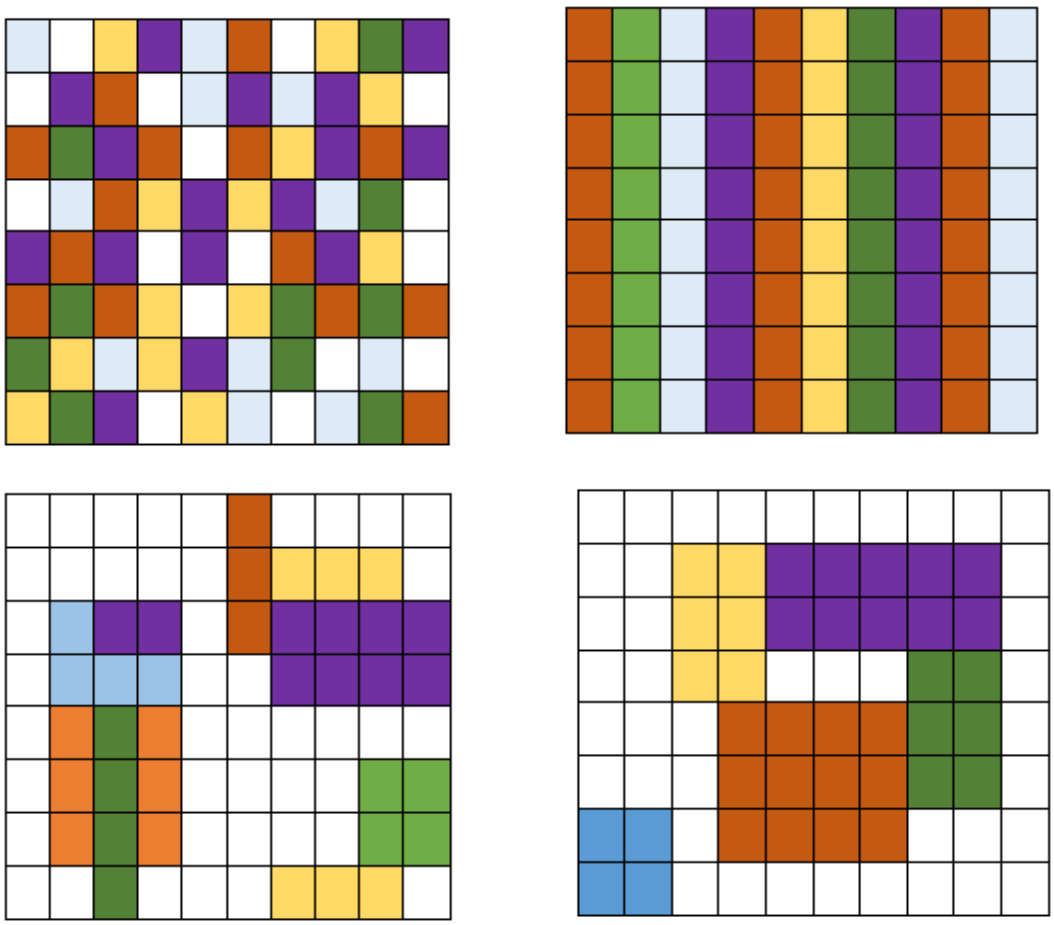

Figure 6.1: Four matrices with special structures: (a) sparse; (b) low-rank; (c) sparse and low-rank using ℓ_1. (d) sparse and low-rank using $\ell_{2,1}$ (proposed). Various colors denote different numerical values and white color represents zero.

embodies sparseness and robustness as it acts like a regularizer which induces joint sparsity (in term of support vectors, SVM is sparse as compared to least-squares SVM). In this regard, we adopt the loss function and proposed a robust approach that efficiently impose sparseness as well as preserves the structural information. The proposed objective function is the combination of hinge loss for model fitting plus the elastic net penalty as the regularization on the regression matrix that is a linear combination of the $\ell_{2,1}$ norm and nuclear norm. The spectral elastic net follows the property of group effect to select robust structural features i.e. strong correlation of rows and columns.

To this end, we have the objective function

$$arg\min \gamma ||W||_{2,1} + \tau ||W||_* + C\sum \xi \tag{6.1}$$

$$w_j^T x_i + b \geq 1 - \xi_i^j, \; if \; y_i = j \; w_j^T x_i + b \leq -1 + \xi_i^j, \; if \; y_i \neq j$$
$$\xi_i^j \geq 0$$

where $\xi_i^j = 1 - y_i[tr(W^T X_i) + b]_+$ is the hinge loss, $W \in \mathbb{R}^{pq}$ is the vector of regression coefficients, $b \in \mathbb{R}^{pq}$ is an offset term and C is a regularization parameter.

The above Eq. 6.1 is a combination of the hinge loss function, $\ell_{2,1}$ and nuclear norm thus it inherits the low-rank and sparsity together into the objective function which helps to deal with outliers as well. The regularizers term in Eq. 6.1 is able to encode the prior knowledge and guides the selection of features by modeling the structure of the feature space.

Rewriting the above problem as

$$arg\min \gamma ||W||_{2,1} + \tau ||W||_* + C\sum 1 - y_i[tr(W^T X_i) + b]_+ \tag{6.2}$$

The objective function is shown in Eq. 6.2 is convex but not smooth. The loss function and the regularization terms are convex and admit a globally optimal solution. In this chapter, we presented an efficient efficient iterative algorithm that quickly converge and obtain global optimum.

The ℓ_1 norm regularizer term provides structural sparsity however, we are also looking to model the group intrinsic structure since the ℓ_1-norm does not consider it and tends to pick features without considering all the classes which can be obtained using the $\ell_{2,1}$ norm. It includes group features detection, joint sparsity, hierarchical group features, etc. Common features of an approach based on Frobenius norm [46] and ℓ_1 norm [141] that they treat both indices (row and column) in the same way, however, they have different meaning i.e. i and j runs through data points and spatial dimension respectively. This subtle distinction is easy to get the loss for matrix, whereas the $\ell_{2,1}$ norm captures this subtle distinction. In conclusion, $\ell_{2,1}$-norm regularization is performed to select robust features across all data points with joint sparsity, i.e. each feature (gene expression or EEG signal from

different channels) either have small scores for all data points or has large scores over all data points.

We will show in the next section that the problem can be solved using a simple yet efficient algorithm.

Table 6.1: Algorithmic procedure of sparse support matrix machine

Input: : Labeled Training dataset: $[\boldsymbol{X_i}, \boldsymbol{y_i}]$ where $\boldsymbol{X_j} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}}$ for $\boldsymbol{j} = \boldsymbol{1,...,N}$, low-rank co-efficient $\boldsymbol{\tau}$, sparsity coefficient $\boldsymbol{\gamma}$, smoothing parameter $\boldsymbol{\alpha}$, weights $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$
Output: Matrices $\boldsymbol{W}$ and bias $\boldsymbol{b}$

Step-I: Initialize the matrix $\boldsymbol{W, Z_{2,1}, Z_*} = \boldsymbol{0}$

While not converge do
Step-II: Compute $\nabla_{\boldsymbol{W}} \boldsymbol{h_\alpha}$ using Eq 6.11.

$$\partial_{\boldsymbol{W}} \boldsymbol{h_\alpha} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \begin{cases} -\boldsymbol{y_i X_i}, & \text{if } \boldsymbol{z_i \leq 0}, \\ \boldsymbol{y_i X_i (z_i^\alpha - 1)}, & \text{if } \boldsymbol{0 < z_i < 0}, \\ \boldsymbol{0} & \text{if } \boldsymbol{z_i \geq 1}, \end{cases}$$

Step-III: Evaluate the resolvent $\boldsymbol{Z_{2,1}}$ for regularizer $\boldsymbol{f_{\ell_{2,1}}}$ using Eq. 6.14

$$\boldsymbol{Z_{2,1} = Z_{2,1} + \lambda_t (prox_{\frac{\theta}{w_1}\gamma||\cdot||_{2,1}} (2W - Z_{2,1} - \theta \nabla_W h_\alpha) - W)}$$

Step-IV: Evaluate the $\boldsymbol{Z_*}$ for regularizer $\boldsymbol{f_{\ell_*}}$ using Eq. 6.15

$$\boldsymbol{Z_* = Z_* + \lambda_t (prox_{\frac{\theta}{w_1}\gamma||\cdot||_*} (2W - Z_* - \theta \nabla_W h_\alpha) - W)}$$

Step-V: Update $\boldsymbol{W}$ using Eq. 6.16

$$\boldsymbol{W = \omega_{2,1} Z_{2,1} + \omega_* Z_*}$$

Step-VI: Update $\boldsymbol{b}$ using Eq. 6.17

$$\boldsymbol{b_{t+1} = b_t - \theta \nabla_b h_\alpha}$$

end while
Step-VII: Return $\boldsymbol{W}$ and $\boldsymbol{b}$

6.2.2 Theoretical Justification

In this section, we theoretically analyze and show how RSSM possesses some elegant features as compared to conventional SVM, conventional elastic net SMM [46] and SSMM [141]. Although, several other regularizations could be possible, why have we selected $\ell_{2,1}$? The following brief discussion provides a comprehensive discussion on the reason for the selection of the $\ell_{2,1}$-norm along with nuclear norm and hinge loss.

With the development of sparsity regularization, dimensionality reduction has been extensively explored and even has been applied for the selection of discriminant patterns i.e. ℓ_1 is used for feature selection [141] for support matrix machines. The ℓ_1 norm regularizer term has some limitations due to the fact that the selected features are upper bounded by the data sample size. Hence, it provides structural sparsity and does not discover the intrinsic group structure, resulting in the selection of features without considering all the classes. However, we are looking for structural sparsity as well as we modeling the group intrinsic structure that can be obtained using the $\ell_{2,1}$ norm. The regularization term helps to select the features across all data points with joint sparsity i.e. each feature either has small scores or large scores over all data points. The objective function includes group features detection, jointly vector sparsity, hierarchical group features, etc. In results, the proposed objective function selects the intrinsic structural patterns for all the classes. A common features of an approach based on the Frobenius norm [46] and the ℓ_1 norm [141] that both treat both indices (row and column) in the same way, however, they have different meaning i.e. i and j run through the data points and spatial dimension respectively. This subtle distinction is easy to get the loss for the matrix, whereas the $\ell_{2,1}$ norm captures this subtle distinction. Furthermore, the objective function based on $\ell_{2,1}$ and nuclear achieve better classification performance especially in the presence of outliers and it also helps to avoid the inevitable upper bound for the selected features by the data sample size, hence there is more flexibility for feature selection.

6.2.3 Empirical Risk Minimization

We further analyze the theoretical excess risk bounds of RSSM. In the learning framework of RSSM, the input lies in a separable Hilbert space H and each entity obeys the standard Gaussian distribution [66]. A classical and intuitive learning

strategy is empirical risk minimization. The objective function in equation 6.1 can be rewritten as

$$\arg\min_{W,b} \sum_{i=1}^{n} h(W,b,X_i,y_i) \quad s.t. \quad ||W||_{2,1} \leq c_0 \text{ and } ||W||_* \leq c_1 \tag{6.3}$$

where c_0 and c_1 are the constants.
The loss function can be written as

$$h(W,X_i',y_i) = \left\{1 - y_i\left[tr(W^T X_i') + \frac{1}{n}\sum_{j=1}^{n} y_j\right]\right\}_+ \tag{6.4}$$

The loss function in equation is L-lipschitz continuous with $X_i' = X_i - \frac{1}{n}\sum_{j=1}^{n} X_j$.

For observed output y_i, the incurred loss is

$$\ell(\langle W,X_i'\rangle, y_i)$$

where ℓ is the loss function on objective function in equation 6.3 and assumed to have value [0, 1].

Our target is to choose W so as to minimize the total average risk $R(W)$. The expected loss of of weight vector W without bias term for loss function can be written as

$$R(W) = \mathbb{E}_{(X,y)\ \mu} h'(W,X_i',y_i),$$

where μ is the probability distribution that data points are sampled. For optimal solution W, the expected risk is minimized as

$$W^0 = \arg_w \min R(W) \quad s.t. ||W||_{2,1} \leq c_0 \text{ and } ||W||_* \leq c_1 \tag{6.5}$$

The average empirical risk can be minimized as

(6.6) $$W^{'} = \mathbf{arg}_{w}\,\mathbf{min} R^{'}(W)$$

$$s.t. ||W||_{2,1} \leq c_0 \text{ and } ||W||_{*} \leq c_1$$

6.2.4 Numerical Algorithm

The optimization problem for the RSSM is convex, non-smooth and non-differentiable, however, the combination of hinge loss, $\ell_{2,1}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To tackle this issue, we split the problem into sub-problems with the Generalized Forward-Backward (GFB) splitting approach [69], and develop a novel and effective algorithm to solve the optimization problem efficiently.

The objective function in Eq. 6.1 consists of three terms, all of which are convex. The $\ell_{2,1}$-norm and Nuclear norm are convex as both satisfy the triangle and homogeneity properties whereas the other term is linear functions thus it is also convex. Although the objective function in Eq. 6.1 is convex but non-differentiable and non-smooth due to the $\ell_{2,1}$-norm and the nuclear norm, thus, stochastic gradient descent and the Nesterov methods cannot be applied (i.e. In convex optimization setting, subgradient of the nuclear norm function cannot be used in standard descent approaches and as a result solving it directly is difficult). Thus, an alternative approach is required to update W.

(6.7) $$arg\min \sum_{k=1}^{2} f_k(W) + C\sum \xi$$

$$w_j^T x_i + b \geq 1 - \xi_i^j,\ if\ y_i = j\ w_j^T x_i + b \leq -1 + \xi_i^j,\ if\ y_i \neq j$$
$$\xi_i^j \geq 0$$

where $f_1 = \tau||W||_*$ and $f_2 = ||W||_{2,1}$, $C\sum\xi$ is the hinge loss, $W \in \mathbb{R}^{pq}$ is the vector of regression coefficients, $b \in \mathbb{R}^{pq}$ is an offset term and C is a regularization parameter.

We rewrite the above problem in Eq. 6.7 as

(6.8) $$arg\min_{W,b} F = arg\min_{W,b} h(W,b) + \sum_{k=1}^{2} f_k(W)$$

where $\boldsymbol{h} = \sum_{i=1}^{n}\{\mathbf{1} - \boldsymbol{y_i}[\boldsymbol{tr}(\boldsymbol{W^T X_i}) + \boldsymbol{b}]\}_+$,

The Eq. 6.8 is the sum of three lower semi-continuous, convex function and is non-smooth with respect to matrix $\boldsymbol{W}$ and $\boldsymbol{b}$ in real Hilbert space. Thus, we can say that there exist an optimal solution of $\boldsymbol{F}$ and its minimizer verifies

$$\mathbf{0} \in \partial \boldsymbol{h} + \sum_{k=1}^{2} \partial \boldsymbol{f_k}$$

Where ∂ is the sub-differential operator and $\partial \boldsymbol{f_k}$ determine the regression matrix $\boldsymbol{W}$ through the linear combination of both $\boldsymbol{\ell_{2,1}}$ and the nuclear norm.

As we know, the objective function in Eq. 6.8 is non-differential, as a result, it cannot be approximated directly. For example, ADMM can be used for at most two non-differential terms optimization, whereas a direct extension of ADMM does not necessarily converge for such problems. The Nesterov methods can be applied for one non-negative term whereas, GFB splitting can handle an arbitrary non-differentiable with a proximal operator. However, it requires the gradient of the loss function to be Lipschitz-continious, thus, we smooth the loss function $\boldsymbol{h}$ by approximating it with generalized smooth hinge loss $\boldsymbol{h_\alpha}$ with

$$\sum_{i=1}^{n} \boldsymbol{h}_\alpha(\boldsymbol{z_i}) = \begin{cases} \frac{\alpha}{\alpha+1} - \boldsymbol{z_i}, & \text{if } \boldsymbol{z_i} \leq \mathbf{0}, \\ \frac{1}{\alpha+1}(\boldsymbol{z_i})^{\alpha+1} + \frac{\alpha}{\alpha+1} - \boldsymbol{z_i}, & \text{if } \mathbf{0} < \boldsymbol{z_i} < \mathbf{0}, \\ \mathbf{0} & \text{if } \boldsymbol{z_i} \geq \mathbf{1}, \end{cases} \tag{6.9}$$

where $\boldsymbol{z_i} = \boldsymbol{y_i}[\boldsymbol{tr}(\boldsymbol{W^T X_i}) + \boldsymbol{b}]$.
To compute the gradient of $\boldsymbol{h_\alpha}$, we applied the chain rule with respect to $\boldsymbol{W}$

$$\partial_{\boldsymbol{W}} \boldsymbol{h}_\alpha = \frac{\partial \boldsymbol{h}_\alpha}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} \tag{6.10}$$

$$\partial_{\boldsymbol{W}} \boldsymbol{h}_\alpha = \sum_{i=1}^{n} \begin{cases} -\boldsymbol{y_i X_i}, & \text{if } \boldsymbol{z_i} \leq \mathbf{0}, \\ \boldsymbol{y_i X_i}(\boldsymbol{z_i^\alpha} - \mathbf{1}), & \text{if } \mathbf{0} < \boldsymbol{z_i} < \mathbf{0}, \\ \mathbf{0} & \text{if } \boldsymbol{z_i} \geq \mathbf{1}, \end{cases} \tag{6.11}$$

Now, GFB learning can be applied iteratively. ∂f_k at various points of Hilbert space ćould be evaluated individually in each iteration. We introduce two auxiliary variables $Z_{2,1}$ and Z_* for $\ell_{2,1}$ and nuclear norm respectively.

$Z_{2,1}$ is updated as,

$$Z_{2,1}^{t+1} = Z_{2,1}^{t} + \lambda_t(j_{\frac{\theta}{\omega_{2,1}}\partial f_{2,1}}(2W_t - Z_{21,t} - \theta\nabla_W h_\alpha) - W_t) \tag{6.12}$$

where $\theta > 0$ is the step size, λ is the relaxation parameter, $t \in N$, $\omega_k \in [0,1]$ is the weight of $Z_{2,1}$ and $j_{\frac{\theta}{\omega_k}\partial f_{2,1}}$ is the resolvent of $\frac{\theta}{\omega_k}\partial f_{2,1}$.
Similarly, Z_* is updated as

$$Z_*^{t+1} = Z_*^{t} + \lambda_t(j_{\frac{\theta}{\omega_*}\partial f_*}(2W_t - Z_{*,t} - \theta\nabla_W h_\alpha) - W_t) \tag{6.13}$$

The resolvent maximal monotone operators $\partial f_{2,1}$ and ∂f_* of $\ell_{2,1}$ and nuclear norm are equal to proximal operator [7] $prox_{f_{2,1}}$ and $prox_{f_*}$ respectively. Proximal algorithms are are computationally efficient for non-smooth and convex optimization problems (detail of Proximal algorithm implementation of $||X||_*$ and $||X||_{2,1}$ is explained in Chapter 2).

$Z_{2,1}$ and Z_* is can be computed using proximal algorithms as

$$Z_{2,1}^{t+1} = Z_{2,1}^{t} + \lambda_t(prox_{\frac{\theta}{w_1}\gamma||.||_{2,1}}(2W_t - Z_{21,t} - \theta\nabla_W h_\alpha) - W_t) \tag{6.14}$$

and

$$Z_*^{t+1} = Z_*^{t} + \lambda_t(prox_{\frac{\theta}{w_1}\gamma||.||_*} - (2W_t - Z_{*,t} - \theta\nabla_W h_\alpha) - W_t) \tag{6.15}$$

Finally, regression matrix W can be computed by linear combination of $Z_{2,1}$ and Z_*

$$W_{t+1} = \omega_{2,1}Z_{21,t+1} + \omega_* Z_{*,t+1} \tag{6.16}$$

Where $Z_k \in [Z_{2,1}$ and $Z_*]$.

Similarly, we can compute the bias b using the gradient descent algorithm as

$$b_{t+1} = b_t - \theta\nabla_b h_\alpha \tag{6.17}$$

In Table 6.1, we describe an iterative algorithm of RSSM for training samples $X_1,...,X_n$ of size $m \times n$, where y_i is the label of X_i, w_1 and w_2 are weights, τ, λ and α low-rank co-efficient, regularization parameter, and smoothing parameter respectively.

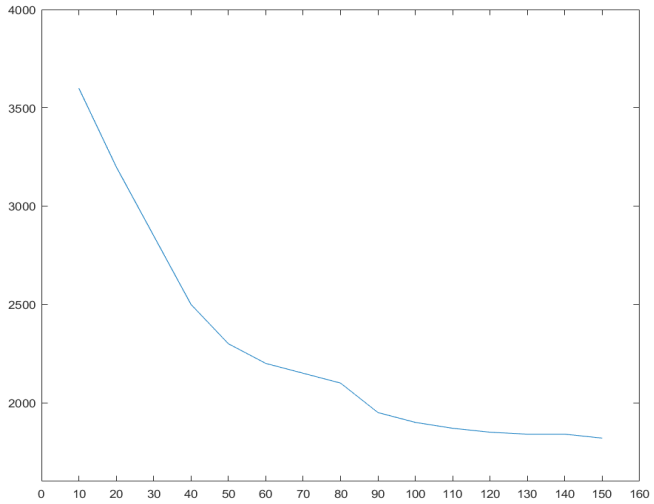

Figure 6.2: Convergence curve of RSSM

6.2.5 Convergence Analysis

The convergence of proposed objective function is explained in theorem 6.1. It shows that the algorithm converges to local optima and enjoy a faster convergence rate. Further detail of GBF splitting convergence proof can be found in [69] section 4.

Theorem 6.1. *Suppose* $\mathscr{E}_{1,t}$ *and* $\mathscr{E}_{2,t}$ *are the error rates for* $\nabla_{w}\boldsymbol{h}_{\alpha}$ *and* $\boldsymbol{j}_{\frac{\theta}{\omega_k}}\partial f_{2,1}$ *at* $\boldsymbol{i^{th}}$ *iteration. if following condition are satisfied then* $\boldsymbol{W_{t+1}} = \boldsymbol{\omega_{2,1}Z_{21,t+1}} + \boldsymbol{\omega_{*}Z_{*,t+1}}$ *converges weekly towards the solution of objective function as well as its robustness to errors.*

- $\nabla \boldsymbol{h}_{\alpha} + \sum_{k=1}^{2} \partial \boldsymbol{f}_{k}$
- $\{\sum_{t=0}^{\infty} ||\mathscr{E}_{1,t}||, \sum_{t=0}^{\infty} ||\mathscr{E}_{2,t}||\} \langle \infty$
- $\boldsymbol{\lambda_t} \in [0,2]$ *and* $\sum_t \boldsymbol{\lambda_t}(2-\boldsymbol{\lambda_t}) = \infty$ *where* $\boldsymbol{t} \in \boldsymbol{N}$

Figure 6.2 shows the convergence curve of proposed objective function.

6.2.6 Computational Complexity

In this section, we provided the asymptotic computational complexity analysis of algorithm 6.1. In step II, the gradient of the hinge loss with respect to $\boldsymbol{W}$ is

computed. For a given n number of samples with dimension $\boldsymbol{x} \times \boldsymbol{y}$, its complexity is $\boldsymbol{O(nxy)}$ as it compute n dot product. Step III is the resolvent $\boldsymbol{Z_{2,1}}$ for regularizer $\boldsymbol{f_{\ell_{2,1}}}$. Step IV evaluates the $\boldsymbol{Z_*}$ for regularizer $\boldsymbol{f_{\ell *}}$. Its complexity is $\boldsymbol{O(xy)}$. Step V is the computation of matrix $\boldsymbol{W}$ and it has complexity $\boldsymbol{O(xy)}$. Final step is the computation of b, that is the computation of hinge loss with respect to $\boldsymbol{b}$ and its complexity is $\boldsymbol{O(nxy)}$.

Thus,the time required for single iteration is $\boldsymbol{O(nxy)}$. For, N number of iteration, the computational complexity of algorithm 6.1 is $\boldsymbol{N} \times \boldsymbol{O(nxy)}$. The computational cost for SVM is $\boldsymbol{O(n^3)}$ and computational cost of SMM is $\boldsymbol{O(n^2xy)}$, that are computationally expensive than RSSM. Our method has same computational complexity as of SSMM.

6.3 Experimental Evaluation

In this section, we describe the experimental setup and evaluate the proposed approach on two important empirical applications such as image classification and EEG classification. As our objective is matrix data classification, thus, for evaluation purposes, we have used datasets where the data is naturally in the form of a matrix and structural information is very important such i.e. voltage fluctuations of EEG signal have a very strong correlation with respect to certain frequency band and channels. We used two different types of publicly available benchmark real-world datasets for image classficiation namely Caltech face dataset and INRIA person dataset. For EEG classification, we have used two-three EEG classification datasets BCI-III IVa, BCI-VI 2a and BCI-VI 2b. The summary of datasets is described as table 6.3. Notice that, the dimension of data is much higher than the number of images within the training set for vector classification due to reshaping the matrix data into vectors. This makes the data classification task not only complex but also affect the classification accuracy.

To validate the effectiveness of the proposed classifier, we extensively evaluate the proposed approach and compare it with both vector based classifiers (i.e. SVM [11, 29], Sparse SVM (SSVM) [144] , LSVM [48], BSVM [34]) as well as with state of the art matrix based classifiers (i.e. SSMM [141],Robust SMM [140], SMM [46]) and regularized matrix regression (RGLM) [143].

Table 6.2: Summary of dataset.

Dataset	subject	Dimension	Train	Test
Caltch Face	435	320×280	218	217
BCI-III IVa	5	120×300	140	140
BCI-VI 2a	54	240×150	72	72
BCI-VI 2b	9	150×24	200	160

6.3.1 Image Classification

We evaluated proposed approach on one of the most fundamental application of image classification. We have applied RSSM on two important datasets (Caltech face dataset and INRIA dataset). The detail of datasets is shown in table 6.3.

6.3.1.1 Caltech Face Dataset

It is a gender recognition dataset of 435 individuals that consist of various facial expressions of size 592×896 captured under different illumination conditions and backgrounds. We have divided the dataset into a training dataset (147 male and 71 female) and test dataset (131 male and 86 female). Images are converted to greyscale, cropped the face using Viola-Jones face detector. We have re-sized the face to size 320×280 and used the pixel values as an input matrix without any advanced feature extraction techniques. Figure 6.3 shows sample images of Caltech face dataset. Notice that, the images share similar features in terms of face outlines and structure, however, gender can be differentiated from small detail such as persons' eyes and hair, etc. The dataset is challenging due to huge variations in facial expression, face appearance, lighting conditions, and backgrounds as shown in figure 6.3.

From the result, as shown in figure 6.4, we observed that classifiers based on the matrix data provided better results as compared to those methods based on data as a vector, which shows that vector-based methods ignore the structural information thus, they are very sensitive to the curse of dimensionality. However, matrix-based approaches leverage the structural information of the data which is greatly beneficial to the improvement of the classification performance. The other main reason is low rank property as discriminant features exist in sparse structure and images are low rank. In comparison to matrix-based methods, RSSM outperforms both sparse (i.e. SSVM) and low rank methods (i.e. BSVM, SMM and SSMM) which validate the claim that RSSM promotes the structural sparsity and

Figure 6.3: Sample images from Caltech Face dataset. Face images shows that the dataset is challenging due to different face appearance, expressions and lighting conditions etc

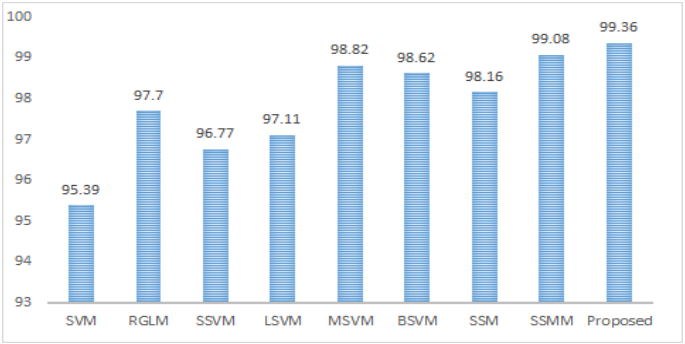

Figure 6.4: Comparative evaluation (accuracy) based on average classification accuracy on Caltech Face dataset

shares similar sparsity patterns across multiple predictors. To further validate the robustness against outliers, we have contaminated the dataset set with random noise i.e. we have randomly selected **20%** images to add noise in the datasets. We corrupted the dataset, i.e., random noise well as block occlusions. Random noise is salt and pepper noise spread randomly at 10%, 15% on random selection of images from dataset. Similarly, block occlusion is added block of different sizes at random locations with variable size 5x5, 10x10, 10x15. For evaluation on contaminated datasets, we have selected 60% and 70% and 80% samples per individual for each dataset as the training dataset. Results showed in figure 6.4 that RSSM is showed better performance against outliers or challenging conditions.

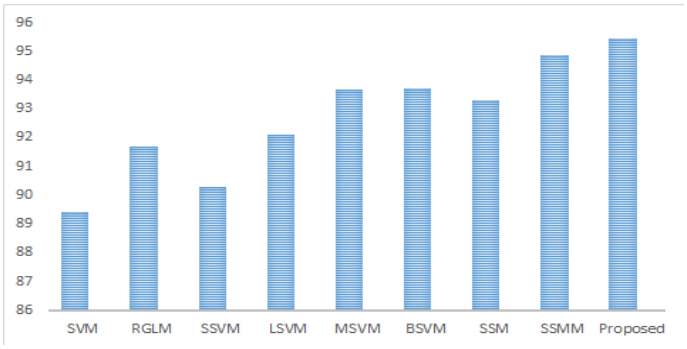

Figure 6.5: Comparative evaluation (accuracy) based on average classification accuracy on contaminated Caltech Face dataset

6.3.1.2 INRIA person dataset

It is collected to detect the existence of person in an image or video. INRIA person dataset is divided in two formats original images with corresponding annotation files and positive images in normalized $\mathbf{64 \times 128}$ pixel format. It consist of **2416** images with people and **1218** people-free ones for training, and **1126** images with people and **453** people-free samples for testing. Person detection is challenging task due to similar background and arbitrary appearance of human in the image. Figure 6.6 shows sample image of dataset. In this experiment, we have converted each image into gray-scale with dimensions ($\mathbf{160 \times 96}$). For person detection, we have used gray-scale image as it is without feature extraction to show the structural correlation of pixels, thus, we have converted the input image into gray level of size $\mathbf{160 \times 96}$.

Figure 6.6: Sample images from INRIA person dataset. The human detection is challenging due to similar appearance of persons and human statues

Figure 6.6 describes the evaluation results on INRIA person dataset. Results

showed that matrix based methods (SMM, SSMM and RSMM etc) outperform the vector based methods that shows that leveraging the correlation of matrix data is meaningful empirically that can not be achieved through vectors due to loss of information. Among matrix based methods, RSSM and achieve competitive results due to its feature selection performance and maintaining the structural information. Results showed that combining low rank property with feature selection.

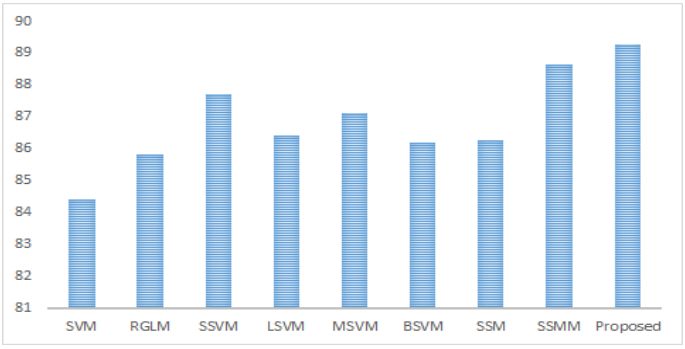

Figure 6.7: Comparative evaluation (accuracy) based on average classification accuracy on INRIA person dataset

To evaluate the robustness of RSSM against outliers, we further contaminated the datasets with random noise by randomly selected **20%** images to add random noise (**10%**, **15%** salt and pepper noise on random selection of images) well as block occlusions. Similarly, block occlusion is added block of different sizes at random locations with variable size $\mathbf{5 \times 5}$, $\mathbf{10 \times 10}$, $\mathbf{10 \times 15}$. Figure 6.5 and figure 6.8 shows that RSSM is showed better performance against outliers or challenging conditions.

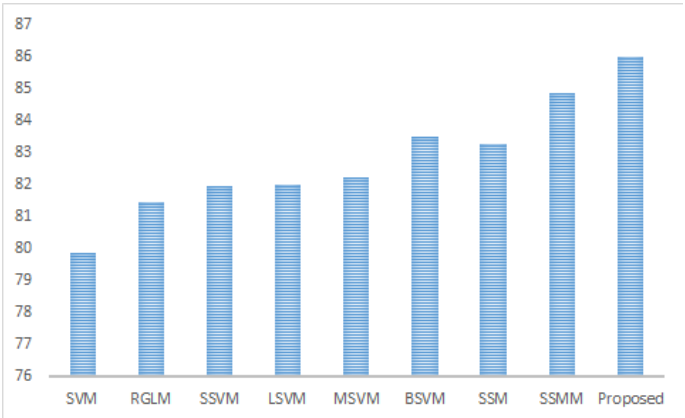

Figure 6.8: Comparative evaluation (accuracy) based on average classification accuracy on contaminated INRIA person dataset

6.3.2 EEG Classification

We further evaluate the proposed approach, we applied RSSM on to the application of electroencephalogram (EEG) data classification. An EEG tracks and records brain wave patterns. EEG brain computer interface is a modern way to communicate with machine as a potential communication system without any requirement of peripheral muscular activity. EEG test can be used to find problems related to electrical activity of the brain. EEG signals data consist of two-dimensional matrices that have high correlation among the rows and columns within each sample, which could be effectively captured by the matrix classification methods [141]. In this experiment, three EEG data of BCI competition-IV namely BCI III-IVa[1], BCI IV-IIa[2] and BCI IV-IIb[3] are used to evaluate the performance of proposed approach. Table 6.3 describes the detail of datasets. Both datasets consist of small number of samples with redundant data, that makes EEG classification challenging. To evaluate the RSSM performance, we have compared the state of the art matrix classification methods such as SSM, SSMM and RSSM etc.

6.3.2.1 BCI-IV EEG dataset

We first evaluted the performance on BCI-IV-2a EEG dataset of BCI competition-IV. BCI-IV 2a dataset consists of EEG data from 9 healthy subjects recorded in two different sessions performing four classes of motor imagery (left-hand, right-hand, foot and tongue labeled as class 1, 2, 3 and 4 respectively). There are 72 trials per motor imagery task and 288 trials in total per session for each individuals.

Motor imagery-based BCI, which translates the mental imagination of movement to commands, is the huge inter-subject variability with respect to the characteristics of the brain signals [4]. Furthermore, poor characteristics of EEG data such as measurement artifacts, outliers and non-standard noises make it challenging task. In order to reduce the variations, spatial filtering has prevent itself as an effective method for extraction of features has been used as a preprocessing technique to explore the discriminative spatial patterns and eliminate uncorrelated informations. In this chapter, we have used Filter Bank Common Spatial Pattern (FBCSP) algorithm [4] to filter out the artifacts and unrelated sensorimotor rhythms by performing autonomous selection of discriminative subject-specific frequency range

[1]http://www.bbci.de/competition/III/download
[2]http://www.bbci.de/competition/iv/dataset2a
[3]http://www.bbci.de/competition/iv/dataset2b

for band-pass filtering of the EEG measurements. To select dominant channels for each motor imagery task, we have applied CSP [39] followed by Time domain parameters for feature selection [63] due to its robust performance [110, 141, 142]. As RSSM is a binary class classifier, thus , to evaluate, we transformed the multi-class classification problem into binary class problem and generated $\boldsymbol{C_4^2 = 6}$ binary subjects namely, L-vs-R, L-vs-F, L-vsT, R-vs-F, R-vs-T and F-vs-T. We have fed the time domain parameters to RSSM for classifications and averaged the classification accuracy of nine subjects for each subset. Table 6.4 shows the comparative evaluation on BCI EEG dataset. Results showed its strong efficiency in the task of EEG signal classification by outperforming state of the art matrix based classification methods. This is due to the fact that EEG signals are strong correlated and sparse. RSSM leverages the structural information as well as dimensionality challenge and promote structural sparsity and model the intrinsic structure. As a result, the regularization term $\boldsymbol{\ell_{2,1}}$ norm along with the nuclear norm and loss not only helps to avoid the inevitable upper bound for the number of selected features but also combines the property of low-rank and sparsity together. Furthermore, the loss function based on $\boldsymbol{\ell_{2,1}}$ and the nuclear norm could help to overcome the outliers as methods based on $\boldsymbol{\ell_2}$ [46] and $\boldsymbol{\ell_1}$ [141] are sensitive to outliers.

We further evaluated the performance of RSSM on BCI-IV 2b EEG dataset used for the detection of motor imagery with left and right hand from nine healthy subjects. For each subject, five sessions are recorded, first two sessions (feedback are not considered) are used for training and last three session (recorded with feedback) are used for classification. The evaluation results of all algorithms on the testing set are reported in table 6.3 and table 6.4. Results showed that RSSM provided better classification accuracy as compared to state of the art matrix classification methods that shows that RSSM is powerful in selection of robust features.

6.3.2.2 BCI III

We further evaluated the performance of RSSM on BCI III-IVa dataset. The BCI III-IVa dataset consist of 118-channel EEG signals recorded from five subjects (aa, al, av, aw and ay) sampled at 100Hz. The signals are sampled with 250 Hz and band-pass filtered between 0.5 Hz and 100 Hz. For preoprocessing and feature extraction, we performed same techniques that are applied on BCI-IV 2a EEG dataset in section 6.3.2.1. The evaluation results of all algorithms on the testing set are reported in figure 6.9 and table 6.5. Results showed that matrix based classifier

such as SSM, SSMM consistently outperform vector based classifiers on all subjects. In comparison to matrix based methods, RSSM achieves best performance.

The high scores for all these measures indicate that proposed approach has high prediction quality. Therefore, it again validates the benefits of leveraging the structural information and multiclass hinge loss for the EEG classification problem.

6.3.3 Parameter Selection

There are several parameter $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$, $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ required to set to compute the objective function. $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ control the trade-off between hinge loss and regularization terms i.e. $\boldsymbol{\gamma}$ captures the feature selection behaviour whereas $\boldsymbol{\tau}$ captures the correlation of data matrix. We observe that the the RSSM degenerates to $\boldsymbol{\ell_{2,1}}$-SVM [9], when $\boldsymbol{\tau} = \mathbf{0}$. Similarly, fixing $\boldsymbol{\gamma} = \mathbf{0}$, degenerate the model to BSVM [34]. Thus, we conclude that the proposed model is a generalization of SVM and possess sparse and low-rank properties. As a result, it consider correlation among matrices as well as perform feature selection simultaneously. $\boldsymbol{\alpha}$ manages the trade-off between computational complexity and smoothness. The objective function requires these parameter to be be optimal to make the solution robust and sparse. In order to find optimal $\boldsymbol{\lambda_a}$ and $\boldsymbol{\lambda_b}$, we have performed several experiments with different value of $\boldsymbol{\gamma}$ and fixed $\boldsymbol{\tau}$. Afterwards, we fix the $\boldsymbol{\gamma}$ and find optimal value of $\boldsymbol{\tau}$. In this experiment, we have set $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$, $\boldsymbol{\alpha} = \mathbf{4}$, $\boldsymbol{\lambda} = \mathbf{1}$, $\boldsymbol{w_1} = \mathbf{0.5}$ and $\boldsymbol{w_2} = \mathbf{0.5}$.

6.4 Discussion

In this chapter, we have compared the performance of proposed RSSM with both vector based classifiers (SVM, SSVM, RGLM,LSVM, BSVM) and matrix based classifiers (SSM, RSSM, SSMM, RSMM). We have performed evaluation on two different matrix classification task image classification and EEG classification. For image classification, we have used Caltech Face dataset and INRIA person identification task. For EEG classification, we have evaluation performance on BCI competition III (IVa) and BCI competition IV (2a and 2b). The EEG signals are two-dimensional matrices, with high correlation among the rows and columns within each sample, which could be effectively captured by the matrix classification methods.

Table 6.3: Classification performance (accuracy) of different algorithms on dataset BCI 2b.

Subject	BCI IV Winner	SVM	SSVM	RGLM	LSVM	BSVM	SSM	RSMM	SSMM	RSSM
S1	0.60	0.68	0.73	0.69	0.69	0.68	0.68	0.73	0.74	0.78
S2	0.40	0.50	0.53	0.51	0.51	0.51	0.52	0.56	0.55	0.61
S3	0.21	0.52	0.54	0.53	0.51	0.53	0.53	0.56	0.56	0.59
S4	0.95	0.91	0.91	0.92	0.87	0.93	0.93	0.97	0.94	0.97
S5	0.86	0.8	0.83	0.82	0.80	0.84	0.83	0.88	0.87	0.89
S6	0.61	0.73	0.82	0.76	0.79	0.74	0.75	0.79	0.82	0.85
S7	0.56	0.69	0.76	0.75	0.72	0.71	0.72	0.78	0.77	0.81
S8	0.85	0.82	0.91	0.87	0.85	0.86	0.83	0.92	0.92	0.94
S9	0.74	0.74	0.84	0.77	0.78	0.76	0.76	0.83	0.86	0.87
Avg.	0.67	0.71	0.76	0.74	0.72	0.73	0.73	0.78	0.78	0.82

Table 6.4: Comparative evaluation based on average classification accuracy on BCI 2a

Motor Imagery	SVM	SSVM	RGLM	LSVM	BSVM	SSM	RSMM	SSMM	RSSM
LvsR	0.80	0.82	0.82	0.81	0.81	0.81	0.83	0.83	0.85
LvsF	0.87	0.89	.89	0.88	0.89	0.88	0.90	0.90	0.91
LvsT	0.86	0.88	.88	0.88	0.88	0.88	0.91	0.90	0.92
RvsF	0.88	0.87	.87	0.88	0.89	0.89	0.89	0.90	0.90
RvsT	0.87	0.87	0.86	0.89	0.89	0.88	0.90	0.90	0.91
FvsT	0.80	0.80	0.81	0.80	0.81	0.81	0.84	0.84	0.86

Table 6.5: Comparative evaluation based on average classification accuracy on BCI III-IVa

Subject	SVM	SSVM	RGLM	LSVM	BSVM	SSM	RSMM	SSMM	RSSM
aa	0.73	0.74	0.71	0.72	0.75	0.74	0.76	0.77	0.79
a1	0.98	1	0.98	0.99	1	1	1	1	1
av	0.68	0.67	0.66	0.68	0.68	0.67	0.7	0.7	0.71
aw	0.7	0.75	0.71	0.71	0.72	0.74	0.83	0.81	0.83
ay	0.7	0.71	0.7	0.69	0.7	0.69	0.76	0.76	0.78
avg	0.76	0.77	0.75	0.76	0.77	0.76	0.81	0.81	0.82

Results on two different datasets (Cletch face dataset and INRIA person dataset) for image classification task showed that RSSM outperform state of the art matrix based methods. Figure 6.4, 6.5, figure 6.7 and figure 6.8 show the comparative analysis with state of the art methods. Results showed that RSSM provide better performance compared to both vector and matrix based methods. It is due to the fact that illumination and background have low rank property and discriminant features exist sparse structures, thus, the intrinsic structural horizontal pattern of $\boldsymbol{\ell_{2,1}}$-norm along with low rank property of nuclear norm is able to capture the joint sparse structure to select features across all the classes.

We have further evaluated on contaminated data by synthetically adding the noise. Results showed that proposed approach is not able to select useful features but also robust against outliers. Similarly, results of all three EEG datasets are shown in figure 6.9. It shows that RSSM outperform state of the art matrix based methods. This is due to the fact that EEG signals are usually highly correlated, and the useful features are rather sparse. RSSM consider the both low-rank and sparse properties and can extract robust features. Taking the advantage of robust feature

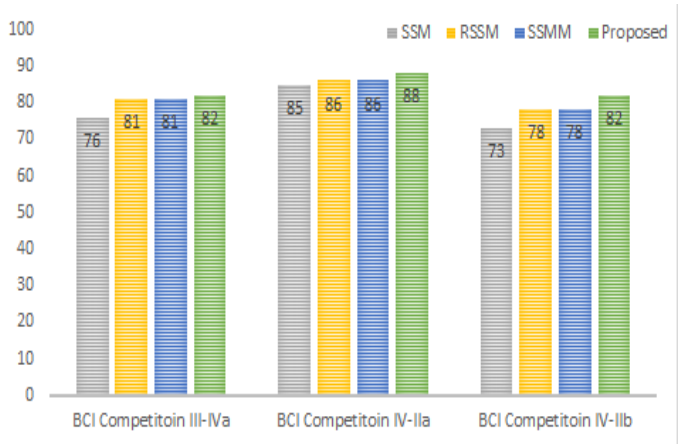

Figure 6.9: Comparative evaluation (accuracy) based on average classification accuracy on BCI dataset

selection, RSSM is a powerful approach even in the presence of outliers.

Classification of such task with small number of data samples, requires low rank plus sparse as well as selection of robust features that can capture the intrinsic and structural properties. Thus, in this case, sparse models (SSVM), low rank (BSVM, SMM) or low rank plus sparse (SSMM) are not sufficient to capture the underlying structural and intrinsic property of the data entirely. ℓ_1 regularizer term has some limitations due to the fact that the selected features are upper bounded by the data sample size. Hence, it provides structural sparsity and does not discover the intrinsic group structure, resulting in selection of features without considering all the classes. Furthermore, there could be outliers in the data that could affect the classification performance. RSSM modeled the group intrinsic structure. The regularization term helps to select the features across all data points with joint sparsity i.e. each feature either has small scores or large scores over all data points. The results on contaminated data shows that RSSM provided better results as compared to state of the art methods, which validate our claim that RSSM is robust against outliers and able to model the intrinsic property of the data entirely.

6.5 Summary

In this chapter, we presented a novel classifier RSSM which is a combination of hinge loss and regularization term as the spectral elastic net penalty. The regularization term promotes structural sparsity and share similar sparsity pattern

across multiple predictor, is a combination of $\ell_{2,1}$ and nuclear norm is a spectral extension of conventional elastic net that combines the property of low-rank and joint sparsity to select features across all the classes. Furthermore, it also leverages the structural information and avoid the inevitable upper bound that simultaneously promotes a good fit to the data as well as combines the property of low-rank and sparsity together. A comprehensive experimental study on publicly available datasets is carried out to validate the proposed approach. The experiment results supported by theoretical analysis and statistical test, show the effectiveness of RSSM approach for solving classification problems while keeping reasonable number of support vectors. In conclusion, the numerical results suggest that our method is superior to previous approaches and demonstrates the promise of RSSM for real-world applications. However, this calls for further analysis and variations in parameters values to control the low-ranks and sparseness properties.

CHAPTER 7

SUPPORT MATRIX MACHINE WITH MATRIX RECOVERY FRAMEWORK

Traditional support vector machines are extremely fragile to the presence of outliers: even a single corrupted data point can arbitrarily alter the quality of the approximation if a fraction of columns are corrupted then the quality may be poor. This chapter considers the problem of high dimensional data classification when a number of the columns are arbitrarily corrupt. We proposed an efficient **S**upport **M**atrix **M**achine by simultaneously performing matrix **Re**covery (SSMRe), feature selection and classification through joint minimization of $\ell_{2,1}$ and nuclear norm. We assume that the data consists of a low-rank clean matrix plus a sparse noise matrix. SSMRe works under the incoherence and ambiguity conditions and able to recover the intrinsic matrix of higher rank and recover data with much denser corruption. The objective function is a spectral extension of the conventional elastic net that combines the property of matrix recovery along with low-rank and joint sparsity together, to deal with complex high dimensional noisy data. Furthermore, it also leverages the structural information as well as the intrinsic structure of data and avoids the inevitable upper bound. The experiment results, supported by the theoretical analysis and statistical test, show the effectiveness of our approach for solving classification problems especially in the presence of outliers while keeping a reasonable number of support vectors.

7.1 Motivation

In this chapter, our concern is the classification problem on a set of the corrupted data matrix. Input data is high in dimension and noisy, hence, we focus our attention on regularizers that have the ability to recover the corrupted data and promote structural sparsity to find robust solutions against outliers. Moreover, our target is to endow the feature space that does not penalize the features individually as in the case of the ℓ_1 norm. Recently, low-rank matrix recovery has shown tremendous performance for the recovery of unobserved noisy data. Inspired by this performance, we intend to combine the matrix recovery into support matrix machines through simultaneous optimization. As a result, SMMRe is not only able to recover the unobserved entities, but also combines the property of low-rank and sparsity together.

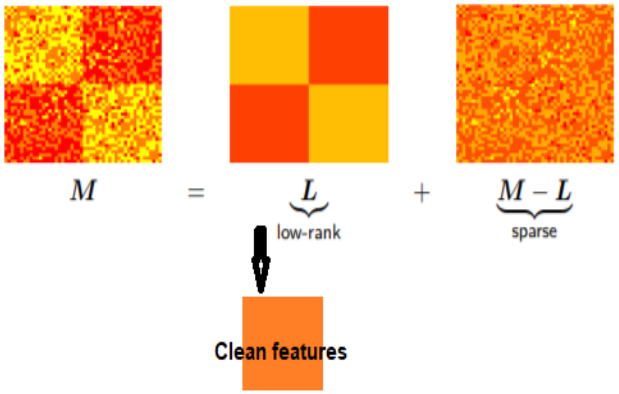

Figure 7.1: Motivation for joint low rank plus matrix recovery based classification for missing plus corrupted data

7.2 Problem Formulation

Suppose, we are given data $\boldsymbol{X}$ with dimension $\boldsymbol{p} \times \boldsymbol{q}$ to classify. A fraction of these columns spans r-dimensional subspace while the rest of the columns are arbitrarily corrupted. We are given only a partial set of observations and our goal is to classify such type of data based on the partial set of observations. The data matrix can be decomposed as $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$. $\boldsymbol{S}$ is the column-sparse matrix that corresponds to corrupted columns, thus at most $\boldsymbol{\alpha n}$ columns are nonzeros. $\boldsymbol{L}$ corresponds to none corrupted matrix, thus $\boldsymbol{rank(L) = r}$ and $\boldsymbol{(1-\alpha)n}$ columns of matrix$\boldsymbol{L}$are nonzeros,

corresponding to the outliers. Better performance can not be guaranteed in all cases because there could be completely unobserved rows or columns resulting in no hope of selecting features belonging to the missing data, in such case missing value can not be recovered. Notice that we have a fraction of observed. Suppose $\boldsymbol{\Omega \subset [p] \times [q]}$ are observed entities, and $\boldsymbol{P(\Omega)}$ is the orthogonal projection onto the linear subspace of matrices supported on $\boldsymbol{\Omega}$ i.e. $\boldsymbol{P_{\Omega}(M) = M_{i,j}}$ if $\boldsymbol{i,j \in \Omega}$ and $\boldsymbol{P_{\Omega}(M) = 0}$ if $\boldsymbol{i,j \notin \Omega}$. We intend to classify the corrupted data efficiently through a matrix recovery framework. Thus, we propose to optimize matrix recovery and classification with an additional objective of low-rank feature representation. We assume that the matrix $\boldsymbol{L}$ satisfies the incoherence conditions ($\boldsymbol{max|Ue_i|^2 \leq \mu \frac{r}{p}}$ and $\boldsymbol{max|Ve_j|^2 \leq \mu \frac{r}{(1-\alpha)n}}$), where $\boldsymbol{e}$ is the unit matrix.

7.3 SMM with Matrix Recovery Framework

In this section, we introduce the proposed SMMRe, which, as a matter of fact, is a novel classifier. SMMRe simultaneously recovers the corrupted matrix whilst removing the redundant information. The classifier also selects the discriminant patterns and considers the strong correlation of rows and columns in the matrix. It is well known that hinge loss enjoys a large margin as it provides a tight and convex upper bound on the indicator function which penalizes misclassifications. It embodies sparseness and robustness as it acts like a regularizer which induces joint sparsity (in term of support vectors, SVM is sparse as compared to least-squares SVM). In this regard, we adopt the loss function and propose a robust approach that efficiently performs matrix recovery, clean feature extraction from the recovered matrix, imposes sparseness as well as preserves the structural information. The proposed objective function is joint optimization of low-rank matrix recovery, hinge loss for model fitting plus the regularization on the regression matrix. To this we end, we have the objective function

$$\begin{aligned} \min_{\boldsymbol{W,b,\{L_i,S_i\}_{i=1}^n}} & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{\alpha_1 ||L_i||_* + \alpha_2 ||S_i||_{2,1}}) + \boldsymbol{\tau ||W||_*} \\ & + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{1 - y_i [tr(W^T L_i) + b]_+} \\ & \text{such that } \forall \boldsymbol{i, X_i = L_i + S_i} \end{aligned} \tag{7.1}$$

where $\boldsymbol{L_i} \in \mathbb{R}^{\boldsymbol{p}\times\boldsymbol{q}}$, $\boldsymbol{S_i} \in \mathbb{R}^{\boldsymbol{p}\times\boldsymbol{q}}$ and $\boldsymbol{W} \in \mathbb{R}^{\times\boldsymbol{q}}$ are the low rank matrix corresponding to non-corrupted columns, sparse matrix corresponding to corrupted columns and regression matrix respectively. $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$ and $\boldsymbol{\tau}$ are positive scalars that penalize the sparse matrix, the nuclear norm of low rank and nuclear norm of regression matrix respectively.

The above Eq. 7.1 is a combination of four terms, hinge loss function, matrix recovery ($\boldsymbol{\ell_{2,1}}$, nuclear norm of $\boldsymbol{L}$) and nuclear norm of $\boldsymbol{W}$. In results, the objective function not only inherits the properties of matrix recovery and identifies the corrupted column with high probability but also holds the properties of low-rank and sparsity together which helps to deal with outliers and corrupted data. Moreover, the regularizer terms in Eq. 7.1 are able to encode the prior knowledge and guide the selection of features by modeling the structure of the feature space.

The objective function in Eq. 7.2 consists of four terms, all of which are convex. The $\boldsymbol{\ell_{2,1}}$-norm and Nuclear norm are convex as both satisfy the triangle and homogeneity properties whereas the other term is a linear function thus it is also convex. The optimization problem for the SMMRe is convex, non-smooth and non-differentiable, however, the combination of hinge loss, $\boldsymbol{\ell_{2,1}}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To decouple the hinge loss and nuclear norm with respect to $\boldsymbol{W}$ in SMMRe, we have introduced an auxiliary variable and applied Lagrange multiplier. The above equation can be written as

$$\begin{aligned}(7.2)\quad \min_{\boldsymbol{W},\boldsymbol{b},\{\boldsymbol{L_i},\boldsymbol{S_i}\}_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}}} & \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} (\boldsymbol{\alpha_1}||\boldsymbol{L_i}||_* + \boldsymbol{\alpha_2}||\boldsymbol{S_i}||_{\boldsymbol{2,1}}) + \boldsymbol{\tau}||\boldsymbol{W}||_* \\ & + \boldsymbol{C}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{h}(\boldsymbol{W},\boldsymbol{b},\boldsymbol{L_1}) \\ \text{s.t. } \forall \boldsymbol{i}, & \boldsymbol{X_i} = \boldsymbol{L_i} + \boldsymbol{S_i} \text{ and } \boldsymbol{W} = \boldsymbol{Z} \text{ where } \boldsymbol{Z} \text{ is auxiliary variable}\end{aligned}$$

Now the constrained problem in Eq. 7.2 can be efficiently solved using Augmented Lagrangian Multiplier algorithm (ALM). The key of ALM method is to search for a saddle point of the augmented Lagrangian function instead of solving the original constrained optimization problem. The augmented Lagrangian function of Eq. 7.3 is given as

$$\mathscr{L}(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{b},\boldsymbol{L_i},\boldsymbol{S_i},\boldsymbol{V},\boldsymbol{M}) = \sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{W},\boldsymbol{b},\boldsymbol{L_i}) + \boldsymbol{\tau_1}||\boldsymbol{Z}||_* + \boldsymbol{tr}[\boldsymbol{V}^T(\boldsymbol{Z}-\boldsymbol{W})] + \frac{\boldsymbol{\mu_1}}{\boldsymbol{2}}||\boldsymbol{Z}-\boldsymbol{W}||_F^2 + \sum_{i=1}^{n}\{\boldsymbol{\alpha_1}||\boldsymbol{L_i}||_* + \boldsymbol{\alpha_2}||\boldsymbol{S_i}||_{2,1} + \boldsymbol{tr}[\boldsymbol{M}_i^T(\boldsymbol{X_i}-\boldsymbol{L_i}-\boldsymbol{S_i}) + \frac{\boldsymbol{\mu_2}}{\boldsymbol{2}}||\boldsymbol{X_i}-\boldsymbol{L_i}-\boldsymbol{S_i}||_F^2\} \tag{7.3}$$

where $\boldsymbol{h}(\boldsymbol{W},\boldsymbol{b},\boldsymbol{L_i}) = \boldsymbol{1} - \boldsymbol{y_i}[\boldsymbol{tr}(\boldsymbol{W}^T\boldsymbol{L_i}) + \boldsymbol{b}]_+$, $\boldsymbol{M},\boldsymbol{V} \in \mathbb{R}^{\boldsymbol{pq}}$ are the Lagrange multiplier. $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the positive penalty parameters. $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$ and $\boldsymbol{\tau}$ control the trade-off between hinge loss and regularization terms i.e. $\boldsymbol{\alpha_1},\boldsymbol{\alpha_2}$ controls the recovery process and clean feature selection whereas $\boldsymbol{\tau}$ captures the correlation of data matrix. Updating Lagrange multipliers as

$$(\boldsymbol{W}^{\boldsymbol{k}},\boldsymbol{Z}^{\boldsymbol{k}},\boldsymbol{b}^{\boldsymbol{k}}) = \min_{\boldsymbol{W},\boldsymbol{Z},\boldsymbol{b}} \mathscr{L}(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{b},\boldsymbol{L}_i^{\boldsymbol{k-1}},\boldsymbol{V}^{\boldsymbol{k-1}}) \tag{7.4}$$

$$(\boldsymbol{L}^{\boldsymbol{k}},\boldsymbol{S}^{\boldsymbol{k}}) = \min_{\boldsymbol{L_i},\boldsymbol{S_i}} \mathscr{L}(\boldsymbol{W}^{\boldsymbol{k}},\boldsymbol{b}^{\boldsymbol{k}},\boldsymbol{L_i},\boldsymbol{S_i},\boldsymbol{M}_i^{\boldsymbol{k-1}}) \tag{7.5}$$

$$\boldsymbol{V}^{\boldsymbol{k}} = \boldsymbol{V}^{\boldsymbol{k-1}} + \boldsymbol{\mu_1}(\boldsymbol{Z}^{\boldsymbol{k}} - \boldsymbol{W}^{\boldsymbol{k}}) \tag{7.6}$$

$$\boldsymbol{M}_i^{\boldsymbol{k}} = \boldsymbol{M}_i^{\boldsymbol{k-1}} + \boldsymbol{\mu_2}(\boldsymbol{X_i} - \boldsymbol{L}_i^{\boldsymbol{k}} - \boldsymbol{S}_i^{\boldsymbol{k}}) \tag{7.7}$$

Notice that, the Eq. 7.4 estimates the model parameter for matrix classification, Eq. 7.5 performs the matrix recovery and clean feature selection simultaneously. Thus, it validates our core objective of clean feature extraction through matrix recovery. As Eq. 7.4 is difficult to solve directly, thus, we solved (described in Theorem 7.1) by minimizing $\mathscr{L}$ against $\boldsymbol{W}$, $\boldsymbol{Z}$ and $\boldsymbol{b}$.

To compute Z, minimizing Eq. 7.3 ($\mathscr{L}(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{b},\boldsymbol{L_i},\boldsymbol{S_i},\boldsymbol{V},\boldsymbol{M})$) with respect to $\boldsymbol{Z}$, we get

$$\boldsymbol{f}(\boldsymbol{Z}) = \boldsymbol{\tau_1}||\boldsymbol{Z}||_* + \boldsymbol{tr}(\boldsymbol{V}^T\boldsymbol{Z}) + \frac{\boldsymbol{\mu}}{\boldsymbol{2}}||\boldsymbol{Z}-\boldsymbol{F}||_F^2 \tag{7.8}$$

$\boldsymbol{Z}$ can be updated base on the following theorem.

Theorem 7.1. *For any positive scalars* α *and* μ_1*, consider* $f(Z)$ *denotes* $\tau_1||Z||_* + tr(V^T Z) + \frac{\mu}{2}||Z-F||_F^2$
We have $\partial f(Z) = 0$
Minimizing $f(Z)$ *with respect to Z, we reach the following optimal solution.*

$$Z = \frac{1}{\mu_1}\mathbb{D}_\xi(\mu_1 W - V) \tag{7.9}$$

$\mathbb{D}_\xi$ *can be computed as*

$$\mathbb{D}_\xi = U\mathcal{S}_\tau(\Sigma)V^T$$

Where $\mathcal{S}_\tau$ *is the entry-wise soft thresholding operator.*

Proof. The equation 7.8 consist of quadratic terms, thus $f(Z)$ is convex. There exist an optimal minimizer Z' such that $Z = \frac{1}{\mu_1}\mathbb{D}_\xi(\mu_1 W - V)$. Z' minimizes $f(Z)$ only if subgradient of $f(Z')$ is 0. We can write

$$0 \in \partial||Z'||_* + V + \mu_1(Z' - W) \tag{7.10}$$

Where $\partial||Z||_*$ is the set of subgradients of nuclear norm.

Consider Z is an arbitrary matrix, we can write

$$\partial||Z||_* = UV^T + M \text{ s.t. } M \in \mathbb{R}^{p\times q}, U^T M = 0, MV = 0, ||M||_F \leq 1 \tag{7.11}$$

To prove $Z = \frac{1}{\mu_1}\mathbb{D}_\xi(\mu_1 W - V)$ satisfies equation 7.10, we decompose $\mu_1 W - V$ into following components

$$\mu_1 W - V = U_0\Sigma_0 V_0^T + U_1\Sigma_1 V_1^T$$

From the above equation, we can write

$$\begin{aligned}\mu_1(W - Z') - V &= \mu_1 W - V - \mu_1 Z' \\ &= \tau(U_0 V_0^T + \frac{1}{\tau}U_1\Sigma_1 V_1^T\end{aligned} \tag{7.12}$$

Comparing Eq. 7.11, we can define $M = \frac{1}{\tau}U_1\Sigma_1 V_1^T$. Thus, it can be verified that $U_0 M_0 = 0$ and $MV_0 = 0$ and $||M||_F \leq 0$. Thus, we have $\mu_1(W - Z') - V \in \tau\partial||Z'||_*$, Hence proved.

■

Similarly to compute W and b, we can rewrite the Eq 7.3 as

$$\min_{W,b} \sum_{i=1}^{n} h(W,b,L_i) \quad - \quad tr(V^T W) \quad + \quad \frac{\mu_2}{2}||Z \quad - \quad W||_F^2 \tag{7.13}$$

W is computed as

$$W = \frac{1}{\mu}(\mu Z + V + \sum_{i=1}^{n} a_i y_i L_i)$$

$$a = \max_{\alpha} -\frac{1}{2}\alpha^Y K\alpha + q^T\alpha$$

$$K = \frac{1}{\alpha_1} y_i y_j tr(L_i^T, L_j)$$

and

$$q = 1 - \frac{1}{\alpha_1} y_i tr(\alpha_1 Z + V)^T L_i)$$

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - tr(W^T L_i))$$

Finally, to compute Lagrange multipliers, differentiating Eq 7.3, we get

$$\min_{L_i} h(W,b,L_i) + \alpha_1||L_i||_* - tr(M_i^T L_i) + ||X_i - L_i - S_i||_F^2 \tag{7.14}$$

$$L_i = \mathbb{D}_\xi(y_i W + \alpha_1(X_i - S_i) + M_i) \tag{7.15}$$

$$\min S = \alpha_2||S_i||_{2,1} - tr(M^T S_i) + \frac{\alpha_1}{2}||X_i - S_i) + M_i)$$

The above equation can be computed using column wise soft thresolding

Now updating the Lagrange multipliers and coefficient

$$M^K = M^{k-1} + \alpha_2(X_i - L_i - S_i) \tag{7.16}$$

$$V^K = V^{k-1} + \alpha_1(Z - W) \tag{7.17}$$

$$\alpha_1 = p\alpha_1$$

$$\alpha_2 = p\alpha_2$$

Table 7.1: Algorithmic procedure of proposed sparse support matrix machine under matrix recovery framework (**SMMRe**)

Input: : Labeled Training dataset: $[X_i, y_i]$ where $X_j \in \mathbb{R}^{m \times n}$ for $j = 1,...,N$, low-rank co-efficient τ, sparsity coefficient α_1 α_2 α_3,; smoothing parameter α, weights w_1 and w_2

Output: Matrices W, L, S and bias b

Step-I: Initialize the matrix $W, = 0$, $L_i = X_i$, $S_i = X_i - L_i$, $M = 0$, $V = 0$

While not converge do

Step-II: Compute $Z = \frac{1}{\mu_1}\mathbb{D}_\xi(\mu_1 W - V)$

Step-III: Compute $W = \frac{1}{\mu}(\mu Z + V + \sum_{i=1}^{n} \alpha_i y_i L_i)$

Step-IV: Compute $\mathbf{min}_S = \alpha_3||S_i||_{2,1} - tr(M^T S_i) + \frac{\alpha_2}{2}||X_i - S_i) + M_i)$

Step-V: Update S=$\alpha_3||S_i||_{2,1} - tr(M^T S_i) + \frac{\alpha_2}{2}||X_i - S_i) + M_i)$

Step-VI: Update $b = \frac{1}{n}\sum_{i=1}^{n}(y_i - tr(W^T L_i))$

Step-VII: Update $M = M^{k-1} + \alpha_2(X_i - L_i - S_i)$

Step-VIII: Update V=$V^{k-1} + \alpha_1(Z - W)$

end while

Step-VII: Return W, L, S and b

The above convex optimization cannot recover the matrix correctly. To overcome this challenge, We use an oracle problem that is defined by the structure we are interested in recovering. Thus, Oracle-based convex optimization-based SSMRe algorithm is able to recover the corrupted columns correctly as well as can identify the outliers.

For the algorithm to succeed it is sufficient for the recovered pair (L', S') to have the right column space and correct column of non-corrupted matrix L, Similarly, it

requires right column support for sparse matrix S. To identify such a solution, we consider the Oracle Problem. $\boldsymbol{\alpha}$ denotes the space of matrices supported on the set of all entries in the non-corrupted columns plus the observed entries in the corrupted columns. We are required to minimize $\mathbf{min}||\boldsymbol{L}||_* + ||\boldsymbol{S}||_{\mathbf{2,1}}$ subject to $\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{L}+\boldsymbol{C}) = \mathbb{P}_{\boldsymbol{\alpha}}\boldsymbol{X}$, $\mathbb{P}(\boldsymbol{L}) = \boldsymbol{L}$ and $\mathbb{P}_{\boldsymbol{I}}(\boldsymbol{S}) = \boldsymbol{S}$. Consider $(\boldsymbol{L},\boldsymbol{S})$ is the solution for the Oracle Problem as we know it is feasible due to the feasibility of true pair $(\boldsymbol{L}',\boldsymbol{S}')$. Now, we must satisfy the conditions such as $(\boldsymbol{L}',\boldsymbol{S}')$ is an optimal solution to Algorithm 7.1 and it must have correct column space and column support. $\boldsymbol{Q}$ is a dual certificate as long as it satisfies the following conditions (I) $\boldsymbol{Q}' \in \boldsymbol{\Omega}$; (II) $\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{Q}') - \boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}} = \mathbf{0}$; (III) $\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{Q}') < \mathbf{1}$; (IV) $|\boldsymbol{P}_{\boldsymbol{I}}(\boldsymbol{Q}')|_{\infty,\mathbf{2}}$ and (V) $\boldsymbol{P}_{\boldsymbol{I}}(\boldsymbol{Q}') \in \boldsymbol{\lambda}\boldsymbol{H}$ *s.t* $\boldsymbol{H} \in \mathbb{R}^{\boldsymbol{pq}} | \boldsymbol{P}_{\boldsymbol{I}}(\boldsymbol{H}) = \mathbf{0}$). The next step is to consider any feasible perturbation, $(\boldsymbol{L}' + \boldsymbol{\Delta}_{\boldsymbol{L}}, \boldsymbol{S}' + \boldsymbol{\Delta}_{\boldsymbol{S}})$. For a given $\boldsymbol{Q}'$, if it satisfies the above conditions shows that $(\boldsymbol{L}' + \boldsymbol{\Delta}_{\boldsymbol{L}}, \boldsymbol{S}' + \boldsymbol{\Delta}_{\boldsymbol{S}})$ is sub-optimal solution.

$$\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{\xi} + \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\boldsymbol{\alpha_1}||\boldsymbol{L_i}||_* + \boldsymbol{\alpha_2}||\boldsymbol{S_i}||_{\mathbf{2,1}}) + \boldsymbol{\tau}||\boldsymbol{W}||_* \leq \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{\xi} + \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\boldsymbol{\alpha_1}||\boldsymbol{L_i} + \boldsymbol{\Delta_L}||_* + \boldsymbol{\alpha_2}||\boldsymbol{S_i} + \boldsymbol{\Delta_S}||_{\mathbf{2,1}}) + \boldsymbol{\tau}||\boldsymbol{W}||_* \tag{7.18}$$

The next step is construction of dual certificate that satisfy the following conditions (I) $\boldsymbol{Q}' \in \boldsymbol{\Omega}$; (II) $\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{Q}') - \boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}} = \mathbb{P}_{\boldsymbol{\alpha}}\mathbb{R}^{-\mathbf{1}}(\mathbb{B})$ s.t. $\mathbb{B} = \sqrt{\frac{\boldsymbol{m}}{\mathbf{2}\boldsymbol{pn}}\boldsymbol{\lambda}}$; (III) $\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{Q}') \leq \mathbf{0.5}$; (IV) $|\boldsymbol{P}_{\boldsymbol{I}}(\boldsymbol{Q}')|_{\infty,\mathbf{2}}$ and (V) $\boldsymbol{P}_{\boldsymbol{I}}(\boldsymbol{Q}') \in \frac{\boldsymbol{\lambda}}{\boldsymbol{H}}$ *s.t* $\boldsymbol{H} \in \mathbb{R}^{\boldsymbol{pq}} | \boldsymbol{P}_{\boldsymbol{I}}(\boldsymbol{H}) = \mathbf{0}$). Ignoring the requirement of $\boldsymbol{Q}' \in \boldsymbol{\Omega}$, is a more manageable problem that allows to consider the fully observed problem of separating the low-rank matrix from a column-sparse matrix. The final step is the sampling i.e. compute $\boldsymbol{Q}$ from $\boldsymbol{Q}'$ that is performed by modified batched sampling-with replacement scheme [86].

7.4 Dataset

We evaluated the proposed approach on the most fundamental applications of classification. We have applied SMMRe on important datasets (Caltech face dataset and INRIA dataset) and BCI competition (III-IVa and BCI IV-IIa).

7.4.1 Caltech Face Dataset

It is a gender recognition dataset of 435 individuals that consists of images containing various facial expressions of size 592×896 captured under different illumination conditions and backgrounds shown in figure 6.6. We have divided the dataset into

Table 7.2: Summary of dataset.

Dataset	subject	Dimension	Train	Test
Caltch Face	435	320×280	218	217
BCI-III IVa	5	120×300	140	140
BCI-VI 2a	54	240×150	72	72
BCI-VI 2b	9	150×24	200	160

a training dataset (147 male and 71 female) and test dataset (131 male and 86 female). Images are converted to greyscale and the face in the image has been cropped using Viola-Jones face detector. We have re-sized the face to 320×280 and used the pixel values as an input matrix without any advanced feature extraction techniques. Figure 6.6 shows sample images of Caltech face dataset. Notice that, the images share similar features in terms of face outlines and structure, however gender can be differentiated from small detail such as persons eyes and hair etc.

7.4.2 INRIA person dataset

It is collected to detect the existence of a person in an image or video. INRIA person dataset is divided into two formats, original images with corresponding annotation files and positive images in normalized 64x128 pixel format. It consists of 2416 images with people and 1218 people-free images for training, and 1126 images with people and 453 people-free samples for testing. Person detection is a challenging task due to the similar background and arbitrary appearance of humans in the image. Figure 6.6 shows sample image of dataset. In this experiment, we have converted each image into a gray-scale with dimensions ($\mathbf{160 \times 96}$). For person detection, we have used the gray-scale image as it is without feature extraction to show the structural correlation of pixels, thus, we have converted the input image into the gray level of size 160×96.

7.4.3 BCI Competition

We further evaluate the SMMRe on the application of electroencephalogram (EEG) data classification. EEG signals consist of two-dimensional matrices that have a high correlation among the rows and columns within each sample, which could be effectively captured by matrix classification methods [141]. In this experiment,

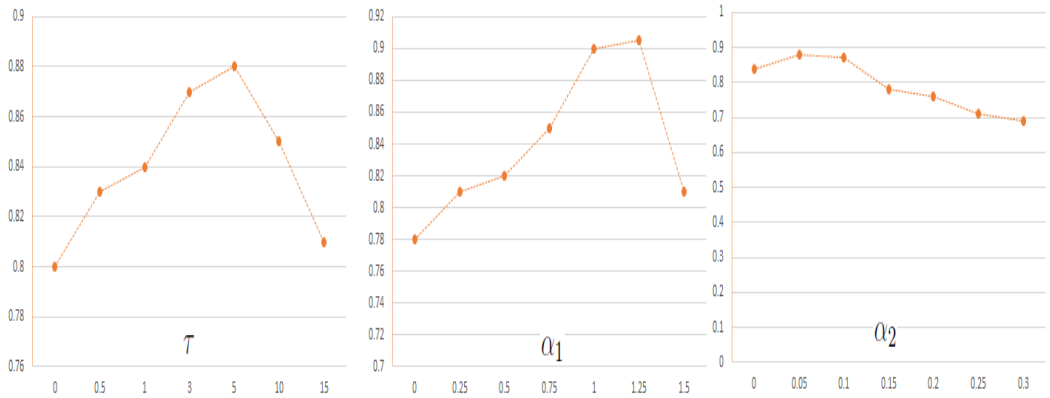

Figure 7.2: Effect of different parameters (τ, α_1 and α_2) values

Table 7.3: Classification performance (accuracy) of different algorithms on dataset BCI 2b.

Sub.	BCI-Win	SVM	SSVM	RGLM	LSVM	BSVM	SSM	RSMM	SSMM	SMMRe
S1	0.60	0.68	0.73	0.69	0.69	0.68	0.68	0.73	0.74	0.797
S2	0.40	0.50	0.53	0.51	0.51	0.51	0.52	0.56	0.55	0.64
S3	0.21	0.52	0.54	0.53	0.51	0.53	0.53	0.56	0.56	0.622
S4	0.95	0.91	0.91	0.92	0.87	0.93	0.93	0.97	0.94	0.975
S5	0.86	0.8	0.83	0.82	0.80	0.84	0.83	0.88	0.87	0.906
S6	0.61	0.73	0.82	0.76	0.79	0.74	0.75	0.79	0.82	0.871
S7	0.56	0.69	0.76	0.75	0.72	0.71	0.72	0.78	0.77	0.828
S8	0.85	0.82	0.91	0.87	0.85	0.86	0.83	0.92	0.92	0.952
S9	0.74	0.74	0.84	0.77	0.78	0.76	0.76	0.83	0.86	0.886
Avg.	0.67	0.71	0.76	0.74	0.72	0.73	0.73	0.78	0.78	0.878

Table 7.4: Comparative evaluation based on average classification accuracy on BCI 2a

Motor Imagery	SVM	SSVM	RGLM	LSVM	BSVM	SSM	RSSM	SSMM	SMMRe
LvsR	0.80	0.82	0.82	0.81	0.81	0.81	0.83	0.83	0.858
LvsF	0.87	0.89	.89	0.88	0.89	0.88	0.90	0.90	0.924
LvsT	0.86	0.88	.88	0.88	0.88	0.88	0.91	0.90	0.933
RvsF	0.88	0.87	.87	0.88	0.89	0.89	0.89	0.90	0.915
RvsT	0.87	0.87	0.86	0.89	0.89	0.88	0.90	0.90	0.922
FvsT	0.80	0.80	0.81	0.80	0.81	0.81	0.84	0.84	0.882

three EEG data observations from BCI competition-IV, namely BCI III-IVa[1], BCI

[1]http://www.bbci.de/competition/III/download

IV-IIa[2] and BCI IV-IIb[3], are used to evaluate the performance of the proposed approach. Table 7.2 describes the detail of the datasets. Both datasets contain a small number of samples with redundant data a property that makes EEG classification challenging.

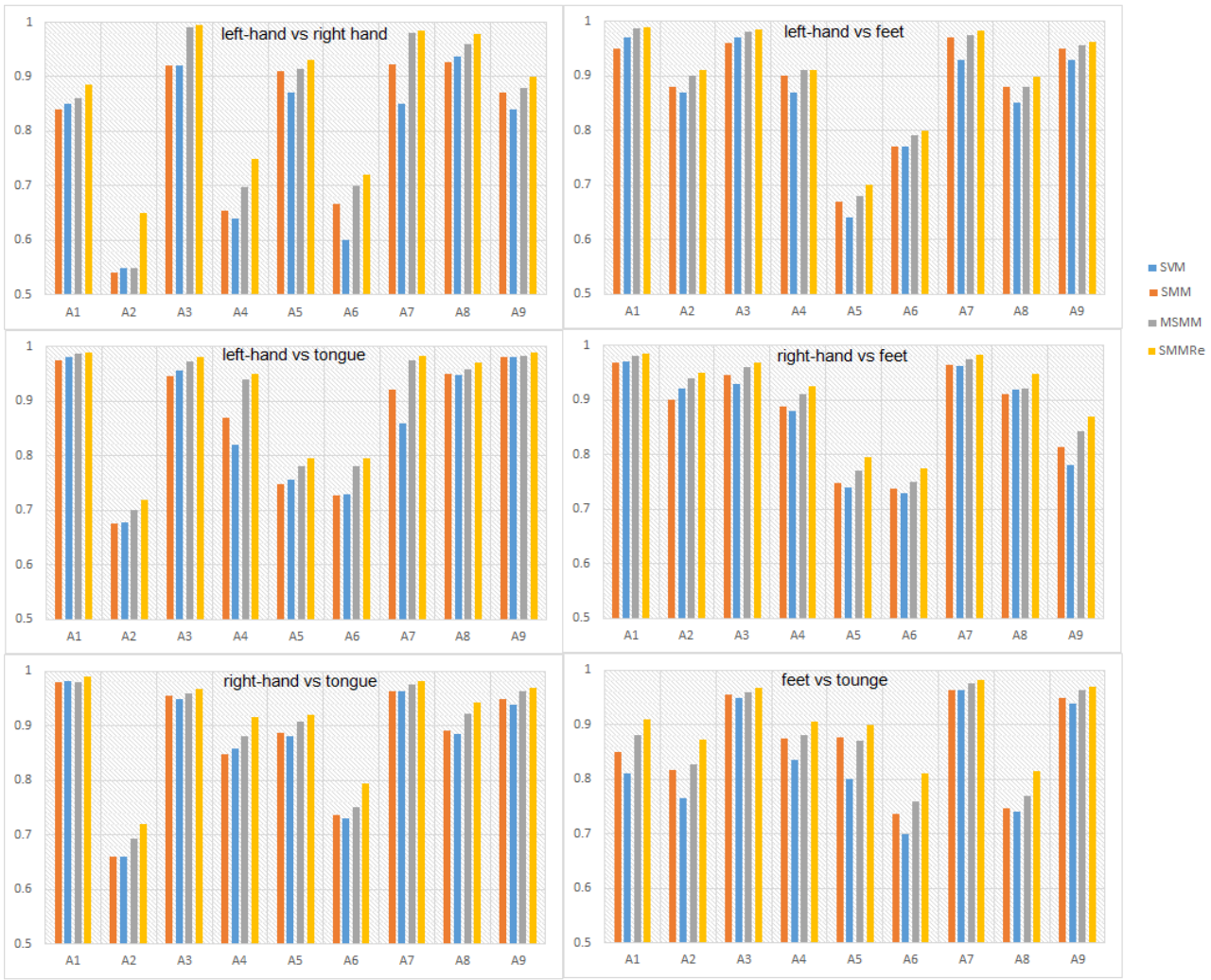

Figure 7.3: Comparative evaluation of SVM, SMM, MSMM and SMMRe on IVa:top left to bottom right (left-hand vs right hand, left-hand vs feet, left-hand vs tongue, right-hand vs feet, right-hand vs tongue, feet vs tongue)

[2]http://www.bbci.de/competition/iv/dataset2a
[3]http://www.bbci.de/competition/iv/dataset2b

7.5 Result and Discussion

To evaluate the SMMRe performance, we have compared the state of the art vector based methods methods such as SVM [11], Sparse SVM (SSVM) [144] , LSVM [48], BSVM [34]) as well as with state of the art matrix based classifiers (i.e. SSMM [141],RSMM [140], SMM [46]) and regularized matrix regression (RGLM) [143] on benchmark face recognition, person identification and EEG datasets.

Surprisingly, we can simultaneously perform matrix recovery, low-rank feature extraction, identification of non-corrupted columns and their position and classification based on a set of a fraction of observed entries. Figure 7.4.2 shows the effect of different parameter values on the classification. Table 7.3, table 7.4 and figure 7.5 show the classification results on EEG datasets (BCI 2a, BCI 2b and IVa). From figure 7.5, we can notice that SMMRe considerably performed better against challenging conditions (A2 and A5 in left-hand vs right-hand, A2, A5 and A6 in left-hand vs tongue, A5, A6 in left-hand vs feet and right-hand vs feet) in comparison to others.

Results showed that support matrix machines based on matrix recovery outperform state of the art methods. Similar results can be noticed in figure 7.5 and figure 7.5 for face recognition and person identification on Caltech Face and INRIA datasets respectively. Furthermore, we can observe that classifiers based on the matrix data provided better results as compared to those methods based on data as a vector, which shows that vector-based methods ignore the structural information thus, they are very sensitive to the curse of dimensionality. However, matrix-based approaches leverage the structural information of the data which is greatly beneficial to the improvement of the classification performance. The other main reason is low-rank property as discriminant features exist in sparse structure and images are low rank.

In comparison to matrix based methods, SMMRe outperforms both sparse (i.e. SSVM) and low rank methods (i.e. BSVM, SMM and SSMM) which validate the claim that SMMRe promotes the structural sparsity and shares similar sparsity patterns across multiple predictors. To further validate the robustness against outliers, we have contaminated both Caltech face dataset and INRIA dataset with random noise, specifically we have randomly selected **20%** images to add noise in each dataset. We corrupted both datasets via the addition of random noise well as block occlusions. Random noise is salt and pepper noise spread randomly at 30%,

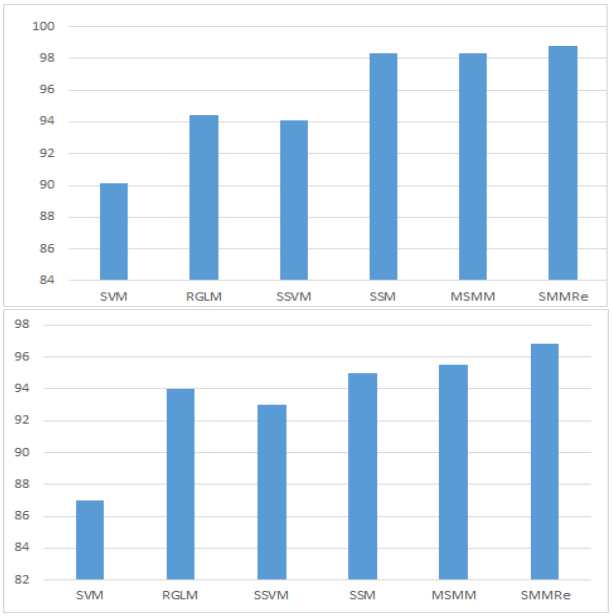

Figure 7.4: Comparative evaluation (accuracy) based on average classification accuracy on real (top) contaminated (bottom) Caltech Face dataset

50% on random selection of images from dataset. Similarly, block occlusion is added by placing blocks of different sizes at random locations with variable size 5x5, 10x10, 10x15. For evaluation on contaminated datasets, we have selected 60% and 70% and 80% samples per individual for each dataset as training dataset and add blocks of variable sizes. Figure 7.5 (b) and figure 7.5 (b) shows the comparative evaluation on caltech face dataset and INRIA dataset. Notice that SMMRe considerably performed better against outliers or challenging conditions in comparison to others. This is due to the simultaneously matrix recovery through identification of non-corrupted columns, low rank robust feature extraction, and classification. It shows that SMMRe is robust even from partially observed matrix which validate our claim that SMMRe is able to classify data with denser corruptions through exact recovery of intrinsic matrix of higher rank based on the incoherence conditions.

We also consider the influence of parameters ($\boldsymbol{\tau}, \boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$) on the performance of SMMRe. $\boldsymbol{\tau}$ is the penalty on nuclear norm of regression matrix that controls the sparseness. $\boldsymbol{\alpha_1}$ is the penalty term on nuclear norm that controls the recovery process. $\boldsymbol{\alpha_2}$ is the penalty on $\boldsymbol{\ell_{2,1}}$ norm to overcome the affect of outliers in feature matrix, as a result, it helps to extract robust features from cleaned matrix. The objective function degenerates to traditional support matrix machine for $\boldsymbol{\tau, \alpha_1, \alpha_2} = \mathbf{0}$, that shows that SSMRe is the special case of support matrix machines. Similarly,

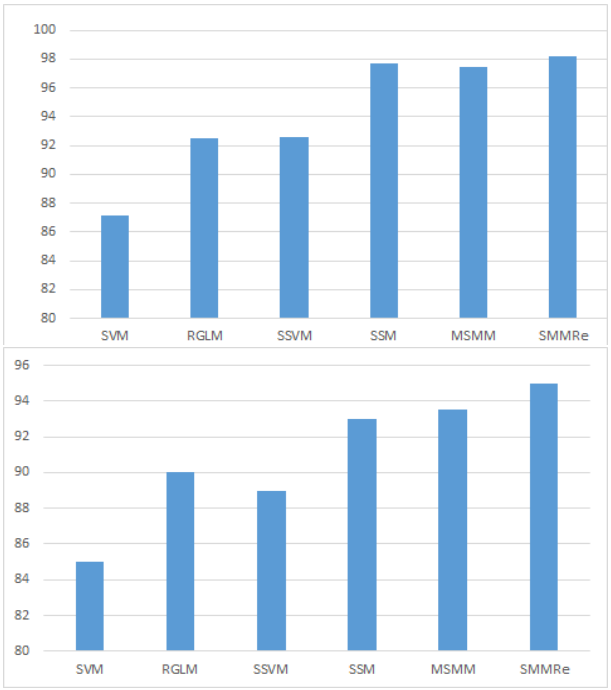

Figure 7.5: Comparative evaluation (accuracy) based on average classification accuracy on real (top) Corrupted (bottom) INRIA person dataset

fixing $\boldsymbol{\alpha_1} = \mathbf{0}$, degenerate the model to SMM. To study the influence of parameter, we fix $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ and find the best optimum value of $\boldsymbol{\tau}$ to control the sparseness. Once we have sparseness control, we repeated the process for other two terms. Figure 7.4.2 shows the effect of different parameter setting of $\boldsymbol{\tau}$, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$.

7.6 Conclusion

In this chapter, we have integrated the matix recovery and support matrix machines for the classification of dense corrupted data. SMMRe is simultaneously able to performs matrix recovery, low rank feature representation and classification, thus able to classify data with denser corruptions through exact recovery of intrinsic matrix of higher rank based on the incoherence conditions. The regularization term promotes the low rank matrix recovery and structural sparsity as well as shares similar sparsity pattern across multiple predictor. Furthermore, it also leverages the structural information and avoids the inevitable upper bound that simultaneously promotes a good fit to the data. A comprehensive experimental study on four publicly available datasets of image classification and EEG classifi-

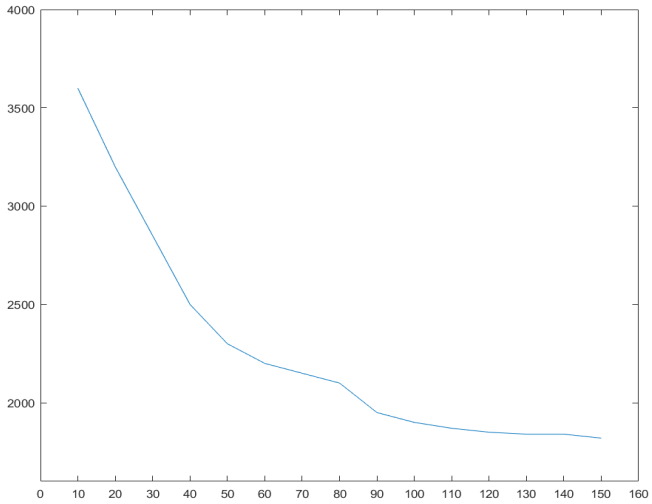

Figure 7.6: Convergence curve of SSMRe (objective function value (y-axis) vs iteration (x-axis)

cation was carried out to validate the proposed approach. The experiment results showed the effectiveness of SMMRe approach for solving classification problems even fraction of columns are corrupted while keeping reasonable number of support vectors.

CHAPTER 8

MULTICLASS SUPPORT MATRIX MACHINES

The alchemists in their search for gold discovered many other things of greater value.

A. Schopenhauer

One of the most important challenges, associated with the classification of EEG signals is how to design an efficient classifier consisting of strong generalization capability. Aiming to improve the classification performance, in this chapter, we propose a novel multiclass Support Matrix Machine (M-SMM) from the perspective of maximizing the inter-class margins. The objective function is a combination of binary hinge loss that works on C matrices and spectral elastic net penalty as a regularization term. This regularization term is a combination of Frobenius and nuclear norm, which promotes structural sparsity and shares similar sparsity patterns across multiple predictors. It also maximizes the inter-class margins that help deal with complex high dimensional noisy data. The extensive experiment results supported by theoretical analysis and statistical tests show the effectiveness of the M-SMM for solving the problem of classifying EEG signals associated with motor imagery in Brain-computer Interface (BCI) applications.

8.1 Motivation

Initially, support vector machines (SVM) were designed for binary classification. How to effectively use it for the multiclass problem, it is still ongoing research. In short, several extensions of SVM has been proposed to deal with multiclass classification problem and can be classified into two types splitting the multiclass classification into many binary classification problems or solving the multiclass problem in a single optimization. Existing SVM classifiers are either single optimization model (i.e. regression-like formulation [58]) or multiple optimization (i.e. OvsR or OvsO). In OvsR, the multiclass classification problem is split into $\boldsymbol{n}$ binary class classification problem whereas OvsO split the problem into $\frac{\boldsymbol{c(c-1)}}{\boldsymbol{2}}$ binary classification problems.

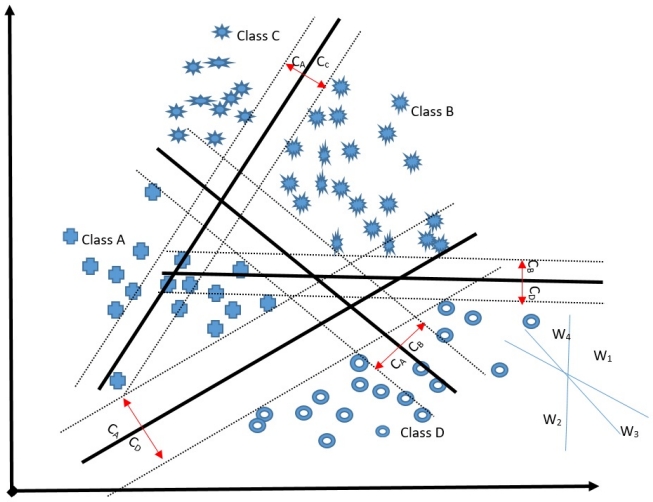

Figure 8.1: Illustration of multiclass support matrix machine: For four classes, we need three parameters $\boldsymbol{W_1}$,$\boldsymbol{W_2}$, $\boldsymbol{W_3}$, and $\boldsymbol{W_4}$ to maximize the inter-class margins

8.2 Maximizing Inter-Class Margins for SMM

In this section, we introduce the proposed approach for maximizing the inter-class margin for the support matrix machine. Figure 8.2 illustrates the proposed multiclass support matrix machines for the classification of EEG signals. It is in principal novel classifier being able to, maximize the inter-class margins, select the discriminant patterns by removing the redundant information, and to consider the strong correlation of rows and columns in the matrix. Figure 8.1 shows the

motivation of M-SMM. The objective function in Eq. 8.1 is a combination of sparse and low-rank properties aiming at efficient capture of the correlations with each input matrix and further maximization of the inter-class hyperplane margin for better multiclass classification.

8.2.1 Objective Function

Given a c-class ($c \geq 2$) matrix form training data $\{X_i, y_i\}_{i=1}^{n} \in \{X, Y\}$, where $X_i \in \mathbb{R}^{pq}$ is the i^{th} feature matrix and $y_i \in \{1, 2, 3....c\}$ is the corresponding class label. The support matrix classifier ($arg\min \frac{1}{2}||W||_F^2 + C\sum_{i=1}^{n} \xi$) focuses on binary classification and hence incapable of dealing with multiclass problems. We devised a novel objective function that maximizes the margin between inter-class.

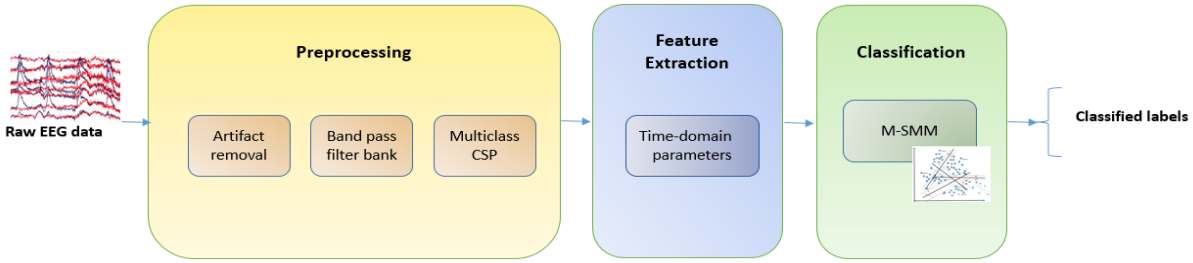

Figure 8.2: Illustration of proposed framework equipped with M-SSM for EEG signal classification

To maximize the inter class margin,

$$arg \min_{w^{d\times c}, b^c} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||W_j - W_k||_F^2 + \tau \sum_{j=1}^{c} ||W_j - W_k||_* \quad (8.1)$$
$$C\sum_{j=1}^{c}\sum_{k=j+1}^{c}\sum_{y_i \in j,k} \xi_i^{jk}$$

such that

$$y_i^{jk} f_{jk}(X_i) \geq 1 - \xi_i^{jk}, \forall y_i \in j,k$$
$$\xi_i^{jk} \geq 0$$

Where $W \in R^{pq}$ denotes the regression parameter in the form of tensor and $||X||_F$ is the Frobenius norm of W. The objective function in Eq. 8.1 resulted in multiple

optimal solutions. In order to reach the objective function which provides single global optima, we further added constraints in the objective function as follow

$$\begin{aligned}(8.2)\quad \boldsymbol{arg}\min_{\boldsymbol{W}^{d\times c},\boldsymbol{b}^{c}} \frac{\mathbf{1}}{\mathbf{2}}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}||\boldsymbol{W_j}-\boldsymbol{W_k}||_F^2+\boldsymbol{\tau}\sum_{j=1}^{c}||\boldsymbol{W_j}-\boldsymbol{W_k}||_* \\ +\frac{\mathbf{1}}{\mathbf{2}}\sum_{j=1}^{c}\boldsymbol{b}_j^2+\boldsymbol{C}\sum_{j=1}^{c}\sum_{k=j+1}^{c}\sum_{y_i\in j,k}\boldsymbol{\xi}_i^{jk}\end{aligned}$$

such that

$$\begin{gathered}\boldsymbol{y}_i^{jk}\boldsymbol{f}_{jk}(\boldsymbol{x}_i)\geq \mathbf{1}-\boldsymbol{\xi}_i^{jk},\ \forall \boldsymbol{y}_i\in \boldsymbol{j},\boldsymbol{k} \\ \boldsymbol{\xi}_i^{jk}\geq \mathbf{0}\end{gathered}$$

Whereas as $\boldsymbol{f}_{jk}\boldsymbol{x}_i=(\boldsymbol{W}_j-\boldsymbol{W}_k)^T\boldsymbol{x}_i+\boldsymbol{b}_j-\boldsymbol{b}_k$ and $\boldsymbol{y}_i^{jk}=\{\mathbf{1},-\mathbf{1}\}$.

For classification of unseen data object, we follow the same voting strategy as in one-vs-one multiclass classification and simulated using C matrices. Thus, M-SMM does not require to compute $\frac{\boldsymbol{c(c-1)}}{\mathbf{2}}$ decision function, it only needs to compute the decision function c times and decided based on the largest value. Surprisingly, the problem could be solved using a simple yet efficient algorithm as shown in table 8.2.2.

8.2.2 Learning Algorithm

The objective function in Eq.8.2 consists of four terms and all of them are convex i.e. the Nuclear and Frobenius norm, which satisfies the triangle and homogeneity properties. The other two terms are linear functions, hence, they are also convex. In conclusion, the objective function in Eq.8.2 is convex but non-differentiable and non-smooth. In a convex optimization setting, the sub-gradient of the nuclear norm function cannot be used in standard descent approaches, as a result solving it directly is difficult. Thus the alternative approach is required to update $\boldsymbol{W}$. As we know, the dependency of matrix $\boldsymbol{W}$ can be revealed by its $\boldsymbol{rank(W)}$, so we can impose rank on $\boldsymbol{W}$. To conclude, rank matrix minimization is non-convex and NP-hard and can be solved as

$$(8.3)\quad \boldsymbol{arg}\min_{\boldsymbol{W}^{d\times c},\boldsymbol{b}^{c}} \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c-1}}\sum_{\boldsymbol{k=j+1}}^{\boldsymbol{c}}||\boldsymbol{W_j}-\boldsymbol{W_k}||_{\boldsymbol{F}}^{\boldsymbol{2}}+\boldsymbol{\tau}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c}}||\boldsymbol{W_j}-\boldsymbol{W_k}||_{*} + \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c}}\boldsymbol{b_j^2}+\boldsymbol{C}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c}}\sum_{\boldsymbol{k=j+1}}^{\boldsymbol{c}}\sum_{\boldsymbol{y_i\in j,k}}[\boldsymbol{1}-\bar{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{jk}}\boldsymbol{f_{jk}(x_i)}]_{+}$$

whereas as $\boldsymbol{W}\in\mathbb{R}^{\boldsymbol{d\times c}}$, $\boldsymbol{b}\in\mathbb{R}^{\boldsymbol{c}}$ and decision function $\boldsymbol{f_jk(x_i)}=(\boldsymbol{W_j}-\boldsymbol{W_k})^{\boldsymbol{T}}\boldsymbol{x_i}+(\boldsymbol{b_j}-\boldsymbol{b_k})\boldsymbol{.y_n^{jk}}$ is the resultant class that is classified for unlabeled data.

We select the training sample randomly in each iteration. The objective function in Eq. 8.3 can be rewritten as

$$(8.4)\quad \boldsymbol{arg}\min_{\boldsymbol{W}^{d\times c},\boldsymbol{b}^{c}} \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c-1}}\sum_{\boldsymbol{k=j+1}}^{\boldsymbol{c}}||\boldsymbol{W_j}-\boldsymbol{W_k}||_{\boldsymbol{F}}^{\boldsymbol{2}}+\boldsymbol{\tau}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c}}||\boldsymbol{W_j}-\boldsymbol{W_k}||_{*} + \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c}}\boldsymbol{b_j^2}+\boldsymbol{C}\Bigg(\sum_{\boldsymbol{j=\bar{y}_i+1}}^{\boldsymbol{c}}[\boldsymbol{1}-\bar{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{jk}}\boldsymbol{f_{\bar{y}_i j}(x_i)}]_{+}+ \sum_{\boldsymbol{j=1}}^{\boldsymbol{\bar{y}_i-1}}[\boldsymbol{1}+\bar{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{jk}}\boldsymbol{f_{j\bar{y}_i}(x_i)}]_{+}\Bigg)$$

As all the terms in objective function in Eq. 8.4 are non-smooth and non-differential, thus, stochastic gradient descent and Nesterov methods can not be applied. Since the objective function is convex in all four terms, we have employed a widely used framework ADMM for the convex optimization problem, by breaking the objective function into sub-problems that are easier to optimize.

The problem in Eq.8.4 can be equivalently written as,

$$\boldsymbol{arg}\min_{\boldsymbol{W,b}}\boldsymbol{P(W)}+\boldsymbol{Q(S)}$$
$$\boldsymbol{s.t}\quad \boldsymbol{S}-\boldsymbol{W}=\boldsymbol{0}$$

Where $\boldsymbol{S}\in\mathbb{R}^{\boldsymbol{P\times Q\times k}}$ is an additional decision variable to split the primal problem into two sub problems.

$$(8.5)\quad \boldsymbol{P(W,b)}=\frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c-1}}\sum_{\boldsymbol{k=j+1}}^{\boldsymbol{c}}||\boldsymbol{W_j}-\boldsymbol{W_k}||_{\boldsymbol{F}}^{\boldsymbol{2}}+\frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{c}}\boldsymbol{b_j^2} + \boldsymbol{C}\Bigg(\sum_{\boldsymbol{j=\bar{y}_i+1}}^{\boldsymbol{c}}[\boldsymbol{1}-\bar{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{jk}}\boldsymbol{f_{\bar{y}_i j}(x_i)}]_{+}+\sum_{\boldsymbol{j=1}}^{\boldsymbol{\bar{y}_i-1}}[\boldsymbol{1}+\bar{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{jk}}\boldsymbol{f_{j\bar{y}_i}(x_i)}]_{+}\Bigg)$$

and

$$\boldsymbol{Q(S)} = ||\boldsymbol{W_j} - \boldsymbol{W_k}||_* \tag{8.6}$$

where $\boldsymbol{P(W)}$ is the hinge loss function obtained from negative likelihood, $\boldsymbol{Q(S)}$ is an additional penalty function defined on singular value of matrix. For simplicity, we used term $\boldsymbol{W}$ instead of $\boldsymbol{W_j} - \boldsymbol{W_k}$ and $\boldsymbol{W_i}$. To solve the Eq. 8.4, we applied augmented Lagrangian method and obtained

$$\boldsymbol{L(W,b,S,)} = \boldsymbol{P(W)} + \boldsymbol{G(S)} + \frac{\boldsymbol{p}}{\boldsymbol{2}}||\boldsymbol{S} - \boldsymbol{W}||_{\boldsymbol{F}}^{\boldsymbol{2}} + \langle \mathscr{L}, (\boldsymbol{S} - \boldsymbol{W}) \rangle \tag{8.7}$$

where $\boldsymbol{p > 0}$ is the hyperparameter and $\mathscr{L}$ is the Lagrange multiplier.

We have divided the optimization problem in Eq. 8.4 into two sub-problems $\boldsymbol{W}$ and $\boldsymbol{S}$. Solving it iteratively, we first needed to minimize $\boldsymbol{S}$ and $\boldsymbol{W}$ followed by updating the Lagrangian multiplier accordingly as,

$$\boldsymbol{S^{t+1}} = \boldsymbol{arg}\min_{\boldsymbol{S}} \boldsymbol{L(S, W^t, \mathscr{L}^t)} \tag{8.8}$$

$$\boldsymbol{W^{t+1}} = \boldsymbol{arg}\min_{\boldsymbol{S}} \boldsymbol{L(S^{t+1}, W, \mathscr{L}^t)} \tag{8.9}$$

$$\mathscr{L} = \mathscr{L}^{\boldsymbol{t}} + \boldsymbol{p(S^{t+1} - W^{t+1})} \tag{8.10}$$

Where t and t+1 are the $\boldsymbol{t^{th}}$ and $\boldsymbol{(t+1)^{th}}$ iterations respectively.

Minimizing the objective function in Eq. 8.8 with respect to $\boldsymbol{S}$ by fixing $\boldsymbol{W}$, is to minimize the sum of all terms $\boldsymbol{S}$ term. Assuming $\boldsymbol{W}$ is fixed, we get

$$\min_{\boldsymbol{S}} \boldsymbol{L_S} = \boldsymbol{G(S)} + \langle \mathscr{L}, \boldsymbol{S} \rangle + \frac{\boldsymbol{p}}{\boldsymbol{2}}||\boldsymbol{W} - \boldsymbol{S}||_{\boldsymbol{F}}^{\boldsymbol{2}} \tag{8.11}$$

To update S, Eq. 8.11 can be solved by minimizing $\boldsymbol{L_s}$. As $\boldsymbol{L_s}$ is non-differential but convex, the sub-gradient of $\boldsymbol{L_s}$ is computed as (see the proof in theorem 1)

$$\boldsymbol{S^{t+1}} \frac{\boldsymbol{1}}{\boldsymbol{p}} \mathbb{D}_{\boldsymbol{\tau}}(\boldsymbol{pW} - \mathscr{L}) = \frac{\boldsymbol{1}}{\boldsymbol{p}} \boldsymbol{U_0}(\boldsymbol{\Sigma_0} - \boldsymbol{\tau I}) \boldsymbol{V_0^T} \tag{8.12}$$

Where $\mathbb{D}$ is the singular value threshold operator.

Theorem 1. For $\tau \geq 0$, one optimal solution for the following problem

$$\min_{S} L_S = G(S) + \langle \mathscr{L}, S \rangle + \frac{p}{2} ||W - S||_F^2$$

is

$$S_c^{t+1} = \frac{i}{1+p} \mathbb{D}_\tau (pW_c - V_c)$$

Where $\mathbb{D}_\tau$ is the singular value thresholding operator (defined in chapter 2).

Similarly, fixing S and minimizing the objective function with respect to W

$$\min_{S} L_W = H(W)_+ + \langle -\mathscr{L}, W \rangle + \frac{p}{2} ||S - W||_F^2 \tag{8.13}$$

The Eq. 8.13 is the non-negative convex sum of term $H(W)$ (combination of hinge loss, Frobenius norm and penalty term) and linear and square functions.

Here, we have two different cases i.e. $j = y_i$ and $j \neq y_i$. Considering $j = y_i$ first, we have

$$W^{t+1} = \frac{1}{p+1} \Big(\mathscr{L} + pS + \begin{cases} -Cx_i & if \quad 1 - f_{jk}(x_i) > 0 \\ 0 & if \quad 1 - f_{jk}(x_i) \leq 0 \end{cases} \tag{8.14}$$

Similarly, when $j = y_i$, W is updated as

$$W^{t+1} = \frac{1}{p+1} \Big(\mathscr{L} + pS + \begin{cases} Cx_i & if \quad 1 - f_{jk}(x_i) > 0 \\ 0 & if \quad 1 - f_{jk}(x_i) \leq 0 \end{cases} \tag{8.15}$$

Finally the Lagrangian multiplier can be updated as

$$\mathscr{L} = \mathscr{L}^t + p(S^{t+1} - W^{t+1}) \tag{8.16}$$

$$p^{t+1} = \beta p^t \tag{8.17}$$

Theorem 2. Optimal solution of S^{t+1} such that

$$\min_{S} L_S = G(S) + \langle \mathscr{L}, S \rangle + \frac{p}{2}||W - S||_F^2$$

satisfies $0 = \partial L_s(S_c^{t+1})$, now we are required to find one S_c subject to

$$0 \in S_c + \tau \partial ||S_c||_* + \mathscr{L} + p(S_c - W_c)$$

Let $U_c \Sigma_c V_C^T$ denotes the singular decomposition of an arbitrary matrix S_c.

Sub gradient of nuclear norm (defined in chapter 2) $\partial ||S_c||_*$ is

$$\partial ||S_C||_* = U_C V_c^T + Z : Z \in \mathbb{R}^{l_1 \times l_2}, U_c^T$$

The above equation can be rewritten as

$$\partial ||S_C||_* = 0, V_c$$

Which can be simplified as

$$\partial ||S_C||_* = 0, ||Z||_F < 1$$

Let Y denotes $PW_c - \mathscr{L}_c$ and decompose it as $Y = U_1 \Sigma_1 V_1^T + U_2 \Sigma_2 V_2^T$ where U and V are the singular vectors associated with singular values greater than τ (smaller than or equal).

If $S_c = \frac{U_1(\Sigma_1 - \tau I)V_1^T}{1+p}$; according to $0 \in S_c + \tau \partial ||S_c||_* + \mathscr{L} + p(S_c - W_c)$, we have the following relation

$$\partial ||S_c||_* = \frac{1}{\tau}[Y - (1+p)S_c]$$

The above Eq. can be simplified as

$$\partial ||S_c||_* = U_1 V_1^T + \frac{1}{\tau} U_2 \Sigma_2 V_2$$

Consider $Z = U_2 \Sigma_2 V_2$, $U_c = U_1$ and $V_c = V_1$, we have $0 \in \partial L_s$ when $S_c^* = S_c$

8.2.3 Theoretical Justification

In this section, we theoretically analyze and illustrate how M-SMM possesses some elegant features as compared to conventional SVM, conventional elastic net SMM [46] and MSMM [142]. As discussed earlier, data is unbiased in real-world, thus one versus the rest class problems will not work and will affect the performance. Similarly, one versus one strategy has high space and time complexity, especially in the case of matrix data, since it requires training of $\frac{c(c-1)}{2}$ SVM classifiers. M-SMM works the same as one vs one fashion and does not use voting strategy and compute decision function for each class. We build a classifier for every two classes, however, different from one versus one strategy, it uses C matrices to simulate all these binary classifiers, thus it does not need to use vote strategy $\frac{c(c-1)}{2}$ times. It just needs to compute decision function c times. This results in reduction of space complexity to same level as one-vs-rest strategy and finds the largest value.

We now show that S-SMM is the generalization of multiclass SMM. Considering the hinge loss of proposed objective function as shown in Eq.8.2.

$$C \sum_{j=1}^{c-1} \sum_{k=j+1}^{c} \sum_{y_i \in j,k} \xi_i^{jk}$$

The above equation can be written as

$$\begin{aligned}
& C \sum_{j=1}^{c-1} \sum_{k=j+1}^{c} \sum_{y_i \in j,k} [1 - \bar{y}_i^{jk} f_{jk}(X_i)]_+ \\
&= C \sum_{j=1}^{c-1} \sum_{k=j+1}^{c} \left(\sum_{y_i \in j} [1 - f_{jk}(x_i)]_+ + \sum_{y_i \in k} [1 - f_{kj}(X_i)]_+] \right) \\
&= C \sum_{j=1}^{c-1} \sum_{y_i \in j} \sum_{k=j+1}^{c} [1 - ((W_j - W_k)^T X_i + (b_j - b_k)))]_+ \\
&= C \sum_{i=1}^{n} \sum_{k \neq y_i} [1 - ((W_{y_i} - W_k)^T X_i + (b_{y_i} - b_k)))]_+
\end{aligned}$$

The objective function in Eq. 8.4 can be written as below, which is MSMM [142].

$$\begin{aligned}
(8.18) \quad arg \min_{W^{d \times c}, b^c} & \frac{1}{2} \sum_{j=1}^{c-1} \sum_{k=j+1}^{c} ||W_j - W_k||_F^2 + \sum_{j=1}^{c} ||W_j - W_k||_* \\
& + \frac{1}{2} \sum_{j=1}^{c} b_j^2 + C \sum_{j=1}^{c} \sum_{i=1}^{n} [1 - \bar{y}_i^{jk} f_{jk}(X_i)]_+
\end{aligned}$$

The proposed objective function degenerate to SMM.

8.3 Experimental Evaluation

In this section, we described the experimental setup and evaluation of the proposed approach. To validate the effectiveness of proposed classifier, we extensively evaluated the proposed M-SMM and compared it with MSMM [142], SMM [46], BSMM [37], MSVM [34], KNN [6] and SCSSP[2] as well as winners of BCI competitions on benchmark EEG datasets (IIIa and IIa) using four different evaluation metrics (recall, prevision, F-measure and kappa coefficient).

8.3.1 Dataset

In this experiment, we have used two publicly available benchmark data-sets namely IIIa (BCI competition III)[1] and IIa[2] (BCI competition IV). IIIa consisted of 60 channel single-trial EEG signals obtained from three subjects(k3b, k6b, and l1b) while performing four classes of motor imagery (left-hand, right-hand, foot and tongue labeled as class 1, 2, 3 and 4 respectively). IIIa consisted of 45, 30, 30 trials per class for subject k3b, k6b and l1b respectively. Similarly, IIa data-set collected in two sessions from nine subjects performing four classes of motor imagery (left-hand, right-hand, foot and tongue). IIIa consisted of 288 in total (72 trails per motor imagery). It consisted of 22 EEG channels and 3 monopolar EOG channels. IIIa and IIa are sampled with 250 Hz and band-pass filtered between 0.5 Hz and 100 Hz. In this experiment, we have considered two subjects (k6b and l1b) for IIIa data-set and EEG channel for IIa data-set.

We have conducted k-fold ($\boldsymbol{k} = \mathbf{5}$) cross-validation to analyze the generalization of the results to an independent dataset. The reason behind k-fold cross-validation is that it guarantees that each sample eventually becomes part of training as well as testing sets. For this purpose, we have divided the trials of each subject into 5 sets. We have repeated the experiments five times and each time a different test set is selected while the other four sets are considered as the training dataset.

[1] http://www.bbci.de/competition/iii/download
[2] http://www.bbci.de/competition/iv/dataset2a

8.3.2 Evaluation Metrics

In order to evaluate the performance of the proposed classifier, we employed different evaluation metrics such as kappa coefficient, precision, recall, and F-measure. Furthermore, we have also compared the training time with state of the art approaches. Kappa measure provides evaluation comparison as it considers the accuracy occurring by chance better. Higher the value of k means the gain in classification performance and $\boldsymbol{k > 0}$ shows the gain is better than a random guess. It is defined as $\boldsymbol{k = \frac{accuracy - p_o}{1 - p_o}}$. Here, $\boldsymbol{p_o}$ is the random guess i.e. for a k-class dataset with balanced sample sizes among different classes, we have $\boldsymbol{p_o = \frac{1}{k}}$. The other evaluation measures we have used are precision, recall and F measure. Precision also referred to as positive predictive value (PPV) is the true positive relevant measure and is calculated as $\boldsymbol{P = \frac{tp}{tp + fp}}$. Recall is referred to as the true positive rate or sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual class. Recall is calculated as $\boldsymbol{R = \frac{t_p}{t_p + f_n}}$. F1 score takes both false positives and false negatives into account is the weighted average of precision and recall. It is needed when we are seeking a balance between precision and recall. It is calculated as $\boldsymbol{F_1 = 2\frac{R \times P}{R + P}}$.

8.3.3 EEG Preprocessing and Feature Extraction

Motor imagery-based BCI, which translates the mental imagination of movement to commands, is the huge inter-subject variability with respect to the characteristics of the brain signals [4]. Furthermore, poor characteristics of EEG data such as measurement artifacts, outliers, and non-standard noises make it a challenging tasks. In order to reduce the variations, spatial filtering has presented itself as an effective method for the extraction of features has been used as a preprocessing technique to explore the discriminative spatial patterns and eliminate uncorrelated information. In this chapter, we have used Filter Bank Common Spatial Pattern (FBCSP) algorithm [4] to filter out the artifacts and unrelated sensorimotor rhythms by performing autonomous selection of discriminative subject-specific frequency range for band-pass filtering of the EEG measurements. To select dominant channels for each motor imagery task, we have applied CSP [39] followed by Time domain parameters for feature selection [63] due to its robust performance [110, 141, 142]. We have fed the time domain parameters to multiclass support matrix machines for classifications.

8.3.4 Results

The main goal of this work is to elucidate the best comparable performance as compared to state of the art approaches followed by computational complexity. In this experiment, we have used four evaluation measures to compare the performance of the proposed approach with seven states of the art approaches on two publicly EEG data-sets. As our contribution is on the classification of matrix data, thus, to compare the proposed classifier for a fair comparison, we employed the same preprocessing and feature extraction approach for other approaches. The evaluation results on data-set IIIa shown in table 8.2 and 8.4 obtained the highest score in the validation procedure. We have transformed the matrix into vectors followed by PCA for dimensionality reduction for vector-based methods such as BCI competition winner, MSVM, KNN, and SCSSP. To compare the performance on the multiclass problem, we have extended the approaches using OvR strategy except for MSMM.

We have also computed the error rate in Kappa measure for better comparison. The evaluation results on data-set IIa are shown in table 8.3, and 8.5. From the results of both data-sets, we observed that classifiers based on maximizing the inter-class hyperplane margin for matrix data provided better results as compared to those methods based on vectors. It further validated that leveraging the structural information of data is greatly beneficial to the improvement of the classification performance.

Notice that, the objective function consist of $\boldsymbol{\tau} \sum_{j=1}^{c} ||\boldsymbol{W_j} - \boldsymbol{W_k}||_*$. Here $\boldsymbol{\tau}$ manages the penalty by controlling the number of low rank of the regression parameter. It determines the structural information. Large value of $\boldsymbol{\tau}$ imposes a heavy penalty that sets most of the singular values in the regression parameter to zero which results in losing most structural information embedded in data. Figure 8.3 shows the convergence process of M-SMM on subjects k3b and l1b of IIIa dataset. We have used ADMM for the convex optimization problem, by breaking the objective function into sub-problems that are easier to optimize. Notice that M-SMM converges to the global optimum in only a few iterations. Similar trends also occur IIa dataset.

8.3.5 Parameter Setting

The objective function is a combination of Frobenius norm, nuclear norm, and hinge loss function, thus there are several parameters $\boldsymbol{\tau}$, $\boldsymbol{p}$, learning rate $\boldsymbol{\eta}$, $\boldsymbol{t}$ and $\boldsymbol{C}$, are

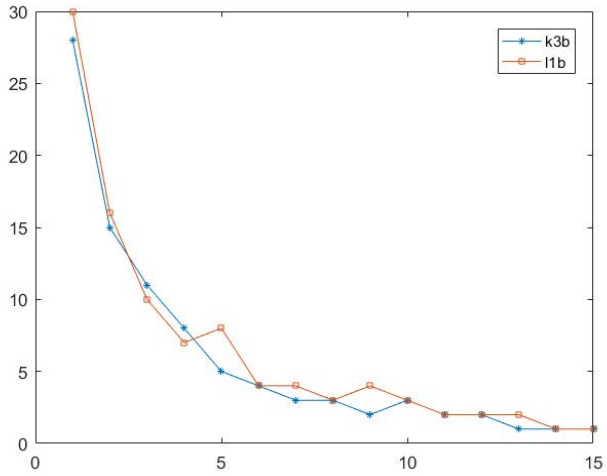

Figure 8.3: Convergence process of M-SMM on subject k3b and l1b of IIIa dataset

required to be adjusted, in order to compute the objective function. $\boldsymbol{\tau}$ is a penalty added on the nuclear norm that captures the correlation of the data matrix. Thus, it determines how much structural information is involved in the classification. We notice that the magnitude of $\boldsymbol{\tau}$ manages the penalty on nuclear norm by controlling the number of singular values (rank) of the regression parameter. The large value of $\boldsymbol{\tau}$ results powerful penalty on the structure information as a result most of the singular values in the regression parameter are set to zero which results in losing most structural information embedded in data. We observe that the proposed model degenerates to the problem [123] for vector data when $\boldsymbol{\tau} = \mathbf{0}$. Figure 8.4 validate the aforementioned claim. Notice that results are the same as of MSMM when $\boldsymbol{\tau} = \mathbf{0}$, similarly, the results start to degrade when $\boldsymbol{\tau}$ is larger. Thus, we concluded that the proposed model is a generalization of SVM and possess sparse and low-rank properties. As a result, it considers correlation among matrices and performs feature selection simultaneously. In this experiment, we have set the learning rate $\boldsymbol{\eta} = \mathbf{0.21}$, $\boldsymbol{\tau} = \mathbf{2.6}$ and $\boldsymbol{p} = \mathbf{3}$ for IIIa dataset and learning rate $\boldsymbol{\eta} = \mathbf{0.23}$, $\boldsymbol{\tau} = \mathbf{3}$ and $\boldsymbol{p} = \mathbf{3}$ for IIa dataset.

8.3.6 Computational Complexity

One of the major objectives of the proposed approach was computational efficiency. As discussed in earlier sections, the existing methods required $\frac{\boldsymbol{c(c-1)}}{\mathbf{2}}$ support matrix machines, that is computational complex In this work, we have used the same

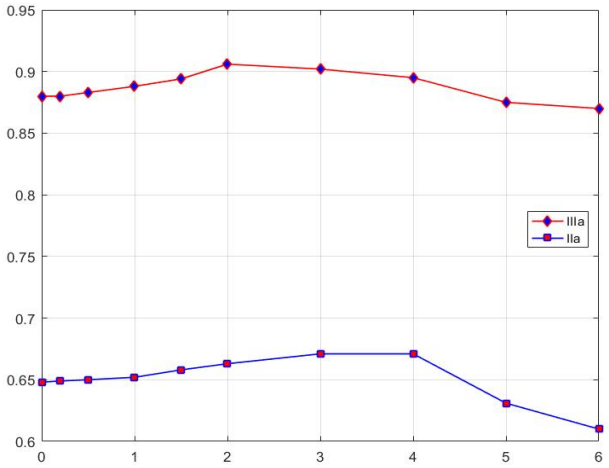

Figure 8.4: Behaviour of $\boldsymbol{\tau}$ on on the classification performance for IIa and IIIa datasets

strategy as of OvsO, however rather than computation of $\frac{\boldsymbol{c(c-1)}}{\boldsymbol{2}}$ support matrix machine, we simulated the OvsO strategy using $\boldsymbol{c}$ support vectors. To investigate the computational efficiency studies of the classification model. We have compared the run time of the algorithms based on matrix data. The experiments were conducted on Intel Xeon E5-1620,3.7GHz, 16GB RAM, Window 7. We compared the average training and testing time on both data-sets between different methods. We have only selected classifiers (i.e. SMM, BSMM, and MSMM) based on matrix data. The average training and testing time on both data-sets are shown in table 8.6. It can be depicted that M-SMM training time is comparable with other approaches however, in comparison to the testing time, it is much faster. The reason behind more training time and better testing time is that we have more number of parameters in training whereas we require C vector for testing respectively.

8.3.7 Discussion

In this section, we provide a comprehensive analysis of the proposed approach. Notice that, the M-SMM achieved better performance as compared to the state of the art methods. Results show that the proposed approach is able to find the representative features from high-dimensional space that are used for classification. The nuclear norm promotes structural sparsity and shares similar sparsity patterns across multiple predictors. $\boldsymbol{\tau}$ determines the level of structural information involved in the classification by controlling the number of singular value (rank) of the

regression parameter. This means greater the value of $\boldsymbol{\tau}$ could account more structural information encoded in the matrix results in improving the classification accuracy. M-SMM reveals the geometric structure embedded in the data due to the fact that it select the features by maintaining the spatial structural information of the matrix.

Comparing with aforementioned experimental evaluation, we have the following interesting observations

(I) M-SMM degenerates to the problem [123] for vector data, when $\boldsymbol{\tau} = \mathbf{0}$. Thus, it is a generalization of SVM and possess sparse and low-rank properties.

(II) Larger value of $\boldsymbol{\tau}$ results powerful penalty on the structure information. However, too large value of $\boldsymbol{\tau}$ results in decreasing the performance due to the fact that high value of $\boldsymbol{\tau}$ results in setting the singular values in the regression parameter to zero which discard the structural information embedded in matrix.

In this work, we have presented a multiclass support matrix machines with the perspective of maximizing the intra-class margins. As a case study, we solved one of the important problem of EEG classification to show the performance of the proposed approach. Results showing considerable improvement in accuracy as well as the computational complexity is also attractive. Although in this experiment, we have applied the EEG dataset for validation, however, the proposed approach is a general machine learning classifier and could be applied to any high dimensional data involving multiclass problem.

8.4 Summary

In this work, we presented a novel classifier name Multiclass Support Matrix Machine (M-SMM) from the perspective of maximizing the intra-class margins (maximizing the distance between training point and hyper-plane) for multiclass classification of high dimensional data such as EEG classification. We combined the hinge loss, nuclear and Frobenius norm and followed the idea of maximizing the margin between two-class problems and use $\boldsymbol{c}$ support matrices to simulate all binary classifiers rather than computing support vector between every two classes. The objective function not only maximized the inter-class margins but was

a spectral extension of the conventional elastic net that combines the property of low rank and joint sparsity together to deal with complex high dimensional noisy data. Hence resulted in an improved classification performance supported by the experimental evaluation. The M-SMM has achieved 0.916 $\boldsymbol{k}$ value for IIIa in comparison to 0.88 and 0.782 for MSMM and SSM respectively. Similarly, 0.671 $\boldsymbol{k}$ value for IIa in comparison to 0.648 and 0.519 for MSMM and SSM respectively. In conclusion, the numerical results suggest that our method is superior to previous approaches and demonstrates the promise of M-SMM for real-world applications.

Table 8.1: Algorithmic procedure of sparse support matrix machine

Input: : Labeled Training dataset: $[X_i, y_i]$ where $X_j \in \mathbb{R}^{m \times n}$ for $j = 1,...,N$, Lagrangian multiplier $\mathscr{L}$, learning rate $\eta \in \{0,1\}$, $p > 0, t = 1$, C, τ
Output: Matrix W

Step-I: Initialize the $W, S, \mathscr{L} = 0$

While not converge **do**
Step-II Minimize S with respect to W

$$\min_S L_S = G(S) + \langle \mathscr{L}, S \rangle + \frac{p}{2} ||W - S||_F^2$$

Step-III Minimize W with respect to S

$$\min_S L_W = H(W)_+ + \langle -\mathscr{L}, W \rangle + \frac{p}{2} ||S - W||_F^2$$

for $i = 1$ to c do
Step-IV: **if** $1 - f_{jk}(x_i) \leq 0$

$$\nabla W = \frac{1}{p+1}(\Lambda + pS - Cx_i)$$

Step-V: **if** $1 - f_{jk}(x_i) \leq 0$

$$\nabla W = \frac{1}{p+1}(\Lambda + pS + Cx_i)$$

end for
for $i = 1$ to n **do**
Pick $i_t \in 1,2,...n$ randomly and update parameter
for $i = 1$ to c **do**

$$W = W - \eta \nabla W$$

$$S = \frac{1}{p} \mathbb{D}_\tau (pW - \Lambda)$$

$$\mathscr{L} = \mathscr{L}^t + p(S^{t+1} - W^{t+1})$$

$$p^{t+1} = \beta p^t$$

end while

Table 8.2: kappa/error rate %: classification performance of different algorithms on data-set IIIa

Subject	BCI Com.	KNN	MSVM	SCSSP	SMM	BSMM	MSMM	M-SMM
k3b	0.83/18.6	0.81/14	0.89/8.3	0.71/22.3	0.852/11.1	0.94/4.4	0.948/3.9	0.961/3.6
l1b	0.74/22.1	0.49/38	0.68/24.2	0.69/36.2	0.71/21.7	0.8/15	0.811/14.2	0.85/13.2
Avg	**0.78/19.8**	**0.65/26**	**0.78/16.3**	**0.64/23.6**	**0.78/16.4**	**0.87/9.7**	**0.88/9.0**	**0.89/9.2**

Table 8.3: kappa/error rate%: classification performance of different algorithms on dataset IIa

Sub	BCI Comp.	KNN	MSVM	SCSSP	BSMM	SMM	MSMM	M-SMM
S1	0.68/24	0.71/22	0.72/21	0.62/26	0.73/21	0.69/0.23	0.73/20	0.76/18
S2	0.42/44	0.4/45	0.37/47	0.28/54	0.4/45	0.23/0.58	0.43/43	0.44/39
S3	0.75/19	0.77/17	0.76/17	0.6/26	0.75/19	0.69/0.24	0.84/11	0 0.84/8.4
S4	0.48/39	0.45/41	0.36/48	0.33/51	0.51/37	0.54/0.35	0.59/31	0.64/28
S5	0.4/45	0.38/47	0.42/43	0.15/64	0.39/46	0.32/0.51	0.5/38	0.55/41
S6	0.27/55	0.24/57	0.19/61	0.25/56	0.32/51	0.15/0.63	0.41/44	0.45/39
S7	0.77/17	0.69/23	0.66/25	0.41/44	0.81/14	0.72/0.21	0.85/12	0.88/11
S8	0.76/18	0.62/29	0.45/41	0.6/31	0.71/22	0.71/0.22	0.77/17	0.81/13.7
S9	0.61//26	0.48//39	0.56/33	0.66/25	0.62/29	0.63/0.27	0.72/21	0.77/14
avg	**0.57/32**	**0.53/36**	**0.5/37**	**0.44/42**	**0.58/31**	**0.52/0.36**	**0.65/26**	**0.74/16**

Table 8.4: Comparative evaluation of classification performance of different algorithms on IIIa data-set

Method	Kappa	Precision	Recall	$\boldsymbol{F_1}$ Score
KNN	0.732	0.768	0.799	0.804
MSVM	0.784	0.85	0.838	0.844
BSMM	0.871	0.91	0.903	0.906
SMM	0.782	0.847	0.836	0.841
MSMM	0.880	0.916	0.91	0.913
M-SMM	**0.916**	**0.927**	**0.918**	**0.922**

Table 8.5: Comparative evaluation of classification performance of different algorithms on IIa data-set

Method	Kappa	Precision	Recall	F 1 Score
KNN	0.527	0.684	0.645	0.663
MSVM	0.499	0.689	0.624	0.653
BSMM	0.581	0.715	0.686	0.7
SMM	0.519	0.674	0.64	0.656
MSMM	0.648	0.751	0.736	0.744
M-SMM	**0.671**	**0.793**	**0.766**	**0.761**

Table 8.6: Comparison of average training and testing time (in seconds) on IIIa and IIa data-sets

Classifier	IIIa		IIa	
	Training	Testing	Training	Testing
SMM	18.995	0.0594	47.198	0.243
BSMM	20.381	0.0636	47.198	0.243
MSMM	22.257	0.0541	65.528	0.230
M-SMM	**24.366**	**0.0414**	**67.261**	**0.161**

Part III

Hinge Loss Optimization

CHAPTER 9

ONE CLASS SUPPORT TENSOR MACHINES

Errors using inadequate data are much less than those using no data at all.

C. Babbage

While a one-class support tensor machine has been proven an effective approach for anomaly detection, their ability to model large corrupted datasets is limited. In this section, we present a novel anomaly detection approach by using the randomized nonlinear features and replacing the hinge loss with a bounded loss function. This results in improving the performance against outliers and also reduces the training time significantly. As traditional loss function is unbounded which results in larger loss caused by outliers. Furthermore, finding support vectors is computationally expensive and does not work well for large datasets. Thus, instead of utilizing traditional hinge loss function and performing a search in high dimensional space, we first present a novel anomaly detection approach for large scale tensor data. We first present novel one-class support tensor machines with bounded loss function rather than finding optimized support vectors with an unbounded loss function. We further extend it by leveraging the randomness to design a scalable approach that can also be used for large scale anomaly detection. To solve the corresponding optimization problem, we have presented half quadratic optimization followed by solving it like a typical OCSTM optimization problem at each iteration. We demonstrate our algorithms through experiments on fourteen real-world benchmark datasets on which we compare against the state-of-the-

art. Experimental results show the robustness of the proposed approach against outliers while computational complexity remains very attractive for large datasets.

$$\begin{aligned} \min_{\mathcal{W},p,\zeta} \quad & \frac{1}{2}||\mathcal{W}||_F^2 + \frac{1}{Nv}\sum_{i=1}^{N}\zeta_i - p \\ s.t. \quad & \left(\langle \mathcal{W}, \phi(\mathcal{X}_i\rangle + b\right) \geq p - \zeta_i, \\ & \zeta_i \geq 0, \forall i = 1,...,N \end{aligned} \tag{9.1}$$

where $\mathcal{W}$ tensor is a weight of the separating hyper-plane, $v \in (0,1]$ is the regularizer that controls the fraction of anomalies and fraction of support vectors. Let ϕ is the mapping function that maps the dataset into Hilbert space H and can be formulated as $\phi : \mathcal{X} \longrightarrow \phi(\mathcal{X}) \in \mathbb{R}^{H_1 \times H_2 \times .... \times H_{M'}}$. ζ_i are the slack variables that allow some of the data points on the other side of the hyperplane.

9.1 Motivation

Traditional loss function in Eq.9.1 is unbounded which results in larger loss caused by outliers, thus is not able to efficiently identify anomalies. Furthermore, methods based on it work well for small datasets however, they are not scalable and computationally complex for larger datasets. Thus, it limits the applicability of one class support tensor machines for anomaly detection for large datasets especially when the datasets are heavily corrupted. Bounding the hinge function could in turn help to reduce the loss caused by outliers. Similarly, the computational complexity can be avoided by exploiting nonlinear random feature as random projection avoids the computational complexity of optimization methods required for nonlinear kernels. Thus, the aim of this work is to design a robust one-class support tensor machine for anomaly detection for large scale datasets.

9.2 Randomized Kernel Bounded One-Class STM

While a one-class support tensor machine has been proven an effective approach for anomaly detection, their ability to model large corrupted datasets is limited. In this section, we present a novel anomaly detection approach by using the randomized nonlinear features and replacing the hinge loss with a bounded loss function. This

results in improving the performance against outliers and also reduces the training time significantly. As traditional loss function is unbounded which results in larger loss caused by outliers. Furthermore, finding support vectors is computationally expensive and does not work well for large datasets. Thus, instead of utilizing traditional hinge loss function and performing a search in high dimensional space, we propose bounded loss function and randomized set of features which results in an improvement in anomaly detection and training time respectively. In the following discussion, we first presented support tensor machines with bounded loss function followed by the generation of nonlinear random features and their application for anomaly detection for tensor data.

9.2.1 Bounding Loss Function

One-class support tensor machines for anomaly detection tries to find an optimal hyperplane in high dimensional data that best separates the data from anomalies with maximum margin. However, the hinge loss of traditional one-class support vector machines is unbounded, which results in larger loss caused by outliers effecting its performance for anomaly detection. Bounding hinge loss function results in reducing the influence of outliers.

To this end, we can rewrite the optimization problem of OCSTM (given in equation 9.1) as

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \frac{1}{2}||\mathscr{W}||_F^2 - \frac{1}{vN}\sum_{i=1}^{N}\aleph_i - p \tag{9.2}$$

$$\text{subject to} \quad \langle \mathscr{W}.\text{œ}\ddot{\text{U}}(\mathscr{X}_i)\rangle \geq p - \aleph_i$$

$$\aleph_i \geq 0 \quad \forall i = 1,...,N$$

where $\aleph_i = max\{0, p - \mathscr{Z}_i\}$ is the hinge loss function with $\mathscr{Z}_i = \mathscr{W}\phi(\mathscr{X}_i)$.

Notice that the hinge loss in Eq. 9.2 is unbounded which results in larger loss occurred due to the outliers which in turn effect the performance of anomaly detection. To overcome the aforementioned challenge, we present the following objective function (Eq.9.3) with bounded loss function (Eq.9.4).

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \frac{1}{2}||\mathscr{W}||_F^2 - p + \frac{1}{vN}\sum_{i=1}^{N}\wp_i \tag{9.3}$$

$$\text{subject to} \quad \langle \mathscr{W}.\phi(\mathscr{X}_i)\rangle \geq p - \aleph_i$$

$$\aleph_i \geq 0 \quad \forall i = 1,...,N$$

$$\wp_i = \beta[1 - e^{-\eta\aleph_i}] \tag{9.4}$$

where $\beta = \frac{1}{1-e^{-\eta}}$ is the normalization constant and $\eta \geq 0$ is the scale constant. The normalization constant β ensures that $\wp_i = 1$. Here, the scale constant η controls the upper bound. For $\eta = 0$ the bounded loss function ($\wp$) degenerates to traditional hinge loss ($\aleph$), thus the traditional hinge loss function (Eq.9.2 is a special case of bounded loss function 9.3).

Eq. 9.2 shows that similar to the traditional one-class support tensor machines, the bounded loss function is also monotonic, bounded however non-convex. By simplifying the Eq. 9.2 and Eq. 9.3, We can rewrite the objective function as

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \frac{\beta}{vN}\sum_{i=1}^{N} e^{-\eta\zeta_i} + p - \frac{1}{2}||\mathscr{W}||_2^2 \tag{9.5}$$

9.2.2 Optimization

As discussed earlier, the objection function in Eq. 9.5 is non-convex due to non-convexity of hinge loss function, thus traditional optimization can not be applied directly. We can solve the above equation through half quadratic optimization by defining a convex function as

$$\mathscr{R}(u) = -ulog(-u) + u, u < 0 \tag{9.6}$$

By applying the conjugate function theory, we get

$$e^{-\eta\aleph} = \sup_{u<0} \eta\aleph u - g(u) \tag{9.7}$$

We can obtain the supermum of $e^{-\eta\aleph}$ at $u = -e^{-\eta\aleph} < 0$.

Now, we can rewrite the Eq. 9.5 as

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \frac{\beta}{vN}\sum_{i=1}^{N}\sup_{u_i<0}\{\eta\aleph_i u_i - g(u_i)\} + p - \frac{1}{2}||\mathscr{W}||_F^2 \tag{9.8}$$

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \frac{\beta}{vN}\sup_{u<0}\left\{\sum_{i=1}^{N}\eta\aleph_i u_i - g(u_i)\right\} + p - \frac{1}{2}||W||_F^2 \tag{9.9}$$

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \sup_{u<0}\left\{\frac{\beta}{vN}\sum_{i=1}^{N}\eta\aleph_i u_i - g(u_i) + p - \frac{1}{2}||W||_F^2\right\} \tag{9.10}$$

We can simplify the Eq. 9.9 as

$$\max_{\mathscr{W},u,p} J(\mathscr{W},u,p) = \frac{\beta}{vN}\sum_{i=1}^{N}\eta\aleph_i u_i - g(u_i)\Big\} + p - \frac{1}{2}||\mathscr{W}||_F^2 \tag{9.11}$$

Iteratively solving the above equation (9.10) using alternating methods to compute $\mathscr{W}$, u and p.

Finally, we can write the Eq. 9.9 as

$$\max_{\mathscr{W},p} J(\mathscr{W},p) = \frac{\beta}{vN}\sum_{i=1}^{N}\eta\aleph_i u_i + p - \frac{1}{2}||\mathscr{W}||_F^2 \tag{9.12}$$

We can rewrite the above Eq. 9.12 as

$$\min_{\mathscr{W},p} J_o(\mathscr{W},p) = \frac{1}{2}||\mathscr{W}||_F^2 + \frac{\beta}{vN}\sum_{i=1}^{N}\eta\aleph_i u_i - p \tag{9.13}$$

The above problem in Eq. 9.13 can be solved by applying Lagrange multiplier. By applying Lagrange multiplier on th above optimization problem, we get

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\mathscr{K}(\mathscr{X}_i,\mathscr{X}_j) \tag{9.14}$$

$s.t. \quad \sum_{i=1}^{N} \alpha_i = 1$ and $0 \geq \alpha_i \leq \frac{1}{vN} s_i$ for $i = 1,...,N$

where $\alpha = [\alpha_1,....,\alpha_N]^T$ is the vector of Lagrange multipliers, k is the kernel matrix.

After solving the dual optimization problem 9.14, the weight tensor $\mathscr{W}$ can be calculated as

$$\mathscr{W} = \sum_{i=1}^{N} \alpha_i \phi(\mathscr{X}_i) \tag{9.15}$$

Finally, the decision function is defined as

$$f(x) = sgn\Big(w\phi(x) - p\Big) \tag{9.16}$$

$$f(x) = sgn\Big(\sum_{i=1}^{N} \alpha_i (x_i, x) - p\Big) \tag{9.17}$$

The solution to the above quadratic problem in Eq. 9.17 is characterized by parameter v that sets an upper and lower bound on the fraction of anomalies and the number of training samples used as support vectors respectively, thus limiting the loss due to outliers.

To apply the kernel methods for tensor data, it has been converted into vectors or matrices [96, 97, 138] which in results in high dimensionality and destroy the structural information embedded in the tensor data. Thus, kernel learning is an important aspect for tensor data to keep the structural information embedded in the tensor data by sets of key structural features and design kernel on such sets. CANDECOMP/PARAFAC (CP) factorization has been employed to tensor to foster the use of kernel methods by extracting a structure-preserving kernel in tensor product feature space [27]. It provides a good approximation to the original tensor data. More specifically, in this way, each tensor can be represented as a sum of rank-one tensors in its original space following by mapping them to tensor product features space for kernel learning.

Let $\mathscr{X} = \sum_{r=1}^{R} \prod_{n=1}^{M} \otimes \boldsymbol{X}_r^m$ be the CP factorization of tensor $\mathscr{X}$ such that $\mathscr{X} \in \mathbb{R}^{I_1 \times \cdots \times I_M}$. The kernel of two same size tensor can be written as $\mathscr{K}(\mathscr{X}, \mathscr{Y}) =$

$\prod_{m=1}^{M} \mathcal{K}(x^m, y^m)$. Tensor data can be factorized in the feature space, similar to the original space. Feature space mapping on rank $R = 1$ feature mapping of a tensor can be defined as

$$\phi : \mathcal{X}^m \longrightarrow \phi(\mathcal{X}^m) \in \mathbb{R}^{H_1 \times \ldots \times H_M} \tag{9.18}$$

$$\phi : \prod_{m=1}^{M} \otimes x^{(n)} \longrightarrow \prod_{m=1}^{M} \otimes \phi(x^{(m)}) \tag{9.19}$$

The CP factorization of tensor in the feature space similar to the original sapce. The CP factorization of tensor $\mathcal{X}$ and $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ is given as

$$\mathcal{X} = \prod_{m=1}^{N} \otimes x^{(m)} \text{ and } \mathcal{Y} = \prod_{n=1}^{M} \otimes y^{(m)} \tag{9.20}$$

The kernel function of two same size tensors $\mathcal{X}$ and $\mathcal{Y}$ can be written as

$$\mathcal{K}(\mathcal{X}, \mathcal{Y}) = \prod_{m=1}^{M} \mathcal{K}(x^{(m)}, y^{(m)}) \tag{9.21}$$

The feature mapping of tensor $\mathcal{X}$ and $\mathcal{Y}$ can be derived as

$$\phi \quad : \quad \sum_{r=1}^{R} \prod_{m=1}^{M} \otimes x^{(m)} \quad \longrightarrow \quad \sum_{r=1}^{R} \prod_{m=1}^{M} \otimes \phi\left(x^{(m)}\right) \tag{9.22}$$

This transformation correspond to mapping the tensor data to high dimensional tonsorial feature space and performing the factorization in the high dimensional space. Then the kernel in the high dimensional space is the standard inner product of the tensor data in that feature space [27]. We can directly drive the naive tensor products kernels as

$$\mathcal{K}\left(\sum_{r=1}^{R} \prod_{m=1}^{M} \otimes \phi(x^m), \sum_{r=1}^{R} \prod_{m=1}^{M} \otimes \phi(y^m)\right) \tag{9.23}$$

$$\sum_{i=1}^{R} \sum_{j=1}^{R} \prod_{m=1}^{M} \mathcal{K}\left(x_i^m, y_j^m\right)$$

Although the objective function limits the affect of outliers, however, it is computationally complexity increases quadratically with the increase in a number of training samples. This issue can be solved using the linear kernel, however, it introduces biasness to the origin. Another alternative is RBF kernel however it results in high computational complexity for high dimensional kernels that makes it inefficient for the larger dataset. The use of randomization such as linear random projection showed itself a substitute to overcome the computational burden of kernel matrix construction [26]. Thus, to deal with the aforementioned challenge of computational and space complexity, we propose to use randomized nonlinear projections that serve as a good approximation of nonlinear kernel and eliminates the need to deal with large kernel matrices for larger datasets, consequently reduction in time complexity. Section 9.2.3 introduce the use of randomized non-linear projections into support tensor machines.

Table 9.1: Algorithmic procedure of OCSTM-BH

Input: : Training dataset: $\mathscr{X}_i{}_{i=1}^{N}$ where $\boldsymbol{X_j} \in \mathbb{R}^{\boldsymbol{m\times n}}$ for $\boldsymbol{j = 1,...,N}$, kernel function $\mathscr{K}(\mathscr{X}_i,\mathscr{X}_j$ trade-off parameter $\boldsymbol{\tau}$, scale constant $\boldsymbol{\eta}$, $\boldsymbol{T_{max}}$
Output: Lagrange multiplier $\boldsymbol{\alpha}$ and margin parameter p,

Step-I: Parameter Initialization: Auxiliary variable $\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{M}}$ such that $\boldsymbol{u_i < 0}$, Number of iteration T=0,
While T $\leq \boldsymbol{T_{max}}$ do
Step-II: Compute $\boldsymbol{\alpha^{T+1}}$ and margin parameter p by solving Eq. 9.14,
Step-III: Compute $\boldsymbol{u^{T+1} = -e^{-\eta\aleph}}$.
Step-IV: Increment T by 1 and repeat the step II-III until converges.
end while
Step-VI: Return $\boldsymbol{\alpha}$ and p

9.2.3 Randomized Feature Embedding

As discussed in earlier sections 9.1 and 9.2.2, the complexity of one-class support tensor machine (objective function in Eq.9.1) grows quadratically with the

increase of training samples, thus, is not efficient for larger datasets. To solve the aforementioned issues, in this section, we described the embedding of non-linear randomized features into robust one-class support tensor machines described in section 9.2.1. Random projections are extremely popular techniques in order to deal with the curse-of-dimensionality. We can randomly sample the parameters from a data-independent distribution and construct a d-dimensional randomized feature map. Thus, we applied one-class support vector machines with bounded loss function on the randomized nonlinear projection which results in reducing the computational complexity by eliminating the need of large kernel matrices for larger datasets consequently reducing the space and computational complexity considerably while outperforming anomaly detection performance in comparison to conventional nonlinear machines.

Our target is to find the optimal fitting function $\boldsymbol{f(x)}$ in order to minimize the empirical risk.

$$\boldsymbol{f(\mathcal{X}) = \min \frac{1}{N} \sum_{i=1}^{N} c(f(\mathcal{X}_i), y_i)} \text{ such that } \boldsymbol{y_i = 1} \tag{9.24}$$

where $\boldsymbol{c(f(\mathcal{X}_i), y_i)}$ is the bounded loss function that penalizes the deviation between label $\boldsymbol{y_i}$ and prediction.

Thus, the fitting function $\boldsymbol{f(\mathcal{X})}$ can be estimated by minimizing the regularized risk as

$$\boldsymbol{R_{Reg}[f(\mathcal{X})] = R_{Emp}[f(\mathcal{X})] + \frac{1}{2}||f(\mathcal{X})||_F^2} \tag{9.25}$$

where $\boldsymbol{R_{Reg}[f(x)]}$, $\boldsymbol{R_{Emp}[f(x)]}$ and $\frac{1}{2}||\boldsymbol{f(x)}||_2^2$ is the regularizer risk (average loss), empirical risk and regularizer respectively. The empirical risk can be calculated as

$$\boldsymbol{R_{Emp}[f(\mathcal{X})] = \frac{1}{N} \sum_{i-1}^{N} L_B(f(\mathcal{X}_i), y_i)} \tag{9.26}$$

where $\boldsymbol{L_B(f(\mathcal{X}_i), y_i)}$ is the bounded loss function (described in section 9.2.2) that penalizes the deviation between labels and predicted values.

The fitting function can be solved using random sampling $\boldsymbol{s_i} \in \mathbb{R}^{\boldsymbol{d}}$ from independent distribution of data and generating $\boldsymbol{d}$ dimensional features.

$$\boldsymbol{Z(\mathcal{X}) = [(\mathcal{Z}_1, \mathcal{Z}_2, \ldots \mathcal{Z}_d)n]}$$

where $\mathcal{Z}_i = [cos(s_i^T, x_1 + b_i), ..., cos(s_i^T, x_N + b_i)]$ and $e_j = [cos(s_j^T, y_1 + b_j), ..., cos(s_j^T, y_N + b_j)$ are Fourier based random features.

Now, replacing the nonlinear kernels with randomized features kernel by unitizing the randomize rank one tensor and CP factorization. We can rewrite the kernel in Eq. 9.23 as

$$\mathcal{K}\Big(\sum_{r=1}^{R}\prod_{m=1}^{M}\otimes\phi(x^m), \sum_{r=1}^{R}\prod_{m=1}^{M}\otimes\phi(y^m)\Big) = \sum_{i=1}^{R}\sum_{j=1}^{R}\prod_{m=1}^{M}\left(\mathcal{Z}_i^{(m)}\right)^2 e_j^{(m)} \tag{9.27}$$

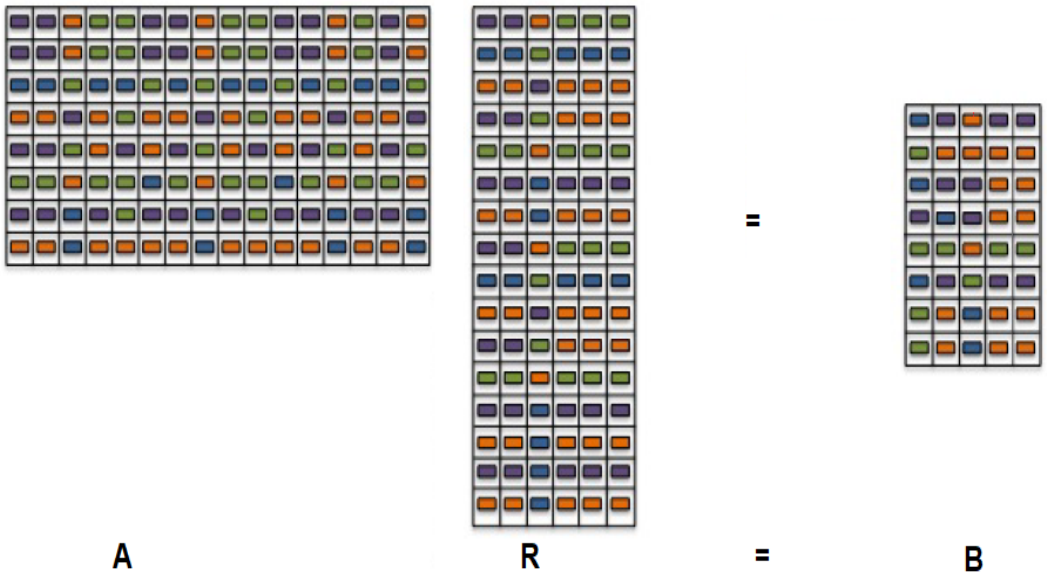

Figure 9.1: Randomized projection of matrix data

The above Eq. 9.27 randomized kernels,

$$\min_{\alpha\in\mathbb{R}^d} \frac{1}{N}\sum_i (\alpha^T z_i, y_i) \quad \text{s.t.} \quad ||\alpha||_\infty \leq \beta \tag{9.28}$$

where β is a regularization constant.

Thus, utilizing nonlinear randomized features, the above formalization remarkably simplifies the computation. Theorem 9.1 justifies this claim.

Theorem 9.1. *Let* $\boldsymbol{D}$ *is the distribution on* $\boldsymbol{\Omega}$ *and* $\boldsymbol{\phi(x;s) \leq 1}$. *Let* $\mathcal{F} = \{\boldsymbol{f(x)} = \int_{\delta} \boldsymbol{\alpha(s)\phi(x;s)ds : \alpha(s) \leq \beta D(s)}\}$. *Let* $\boldsymbol{l}$ *be the L-Lipschitz loss function and* $\boldsymbol{\lambda > 0}$. *Draw* $\boldsymbol{s_1,..s_i}$ *iid from distribution* $\boldsymbol{D}$. *We can write* $\{\boldsymbol{f^*(x) = \sum_{j=1}^{i} \alpha_j \phi(x;s_j)}$ *minimizes the empirical risk*

$$\boldsymbol{E_D[l(f^*(x),y)] - \min_{f \in F} E_D[l(f(x),y)]} \leq \boldsymbol{O\left(\left(\frac{LB}{\sqrt{N}} + \frac{LB}{\sqrt{d}}\right)\sqrt{log\frac{1}{\delta}}\right)} \tag{9.29}$$

with a probability of at least $\boldsymbol{1 - 2\delta}$.

9.2.4 Convergence

The convergence rate of proposed anomaly detection can be related to original one class support vector machines with its original kernel and can be expressed by the following theorem.

Theorem 9.2. *For the given data* $\mathcal{X} \in \mathbb{R}^{N \times M}$, *kernel matrix* $\mathcal{K}_{i,j} = \mathcal{K}(x_i, x_j)$ *and its approximation* $\hat{\mathcal{K}}$ *using* $\boldsymbol{d}$ *random features, the following condition holds*

$$\boldsymbol{E}||\hat{\mathcal{K}} - \mathcal{K}|| \leq \sqrt{\frac{3N^2 logN}{d}} + \frac{2N logN}{d} \tag{9.30}$$

Proof: $\hat{\mathcal{K}} : \frac{1}{d}\sum_{i=1}^{d} \mathcal{Z}_i \mathcal{Z}_i^T = \frac{1}{d}\sum_{i=1}^{d} \mathcal{K}^i$ is a $N \times N$ kernel matrix such that $\mathbb{E}[\hat{\mathcal{K}}] = \mathcal{K}$ where $\hat{\mathcal{K}} = \frac{1}{d}\sum_{i=1}^{d} \mathcal{K}^i$. Since the matrix X is defined to be constant and random features ($\boldsymbol{d}$) are sampled based on independent and identical distribution. We can considering the individual error matrices as

$$\boldsymbol{E} = \hat{\mathcal{K}} - \mathcal{K}$$

$$\boldsymbol{E} = \sum_{i=1}^{d} \boldsymbol{E_i} \qquad \boldsymbol{s.t.} \qquad \mathbb{E}[\boldsymbol{E_i}] = \boldsymbol{0} \quad \forall \boldsymbol{E_i; i = 1,...d} \tag{9.31}$$

where as

$$\boldsymbol{E_i} = \frac{\hat{\mathcal{K}}^{(i)} - \mathcal{K}}{\boldsymbol{d}}$$

Since, in our case, we have bounded features, thus it follows that there exist a constant $\boldsymbol{B}$ such that $||\boldsymbol{Z}||^2 \leq \boldsymbol{B}$. Thus, we can write

$$||\boldsymbol{E_i}|| = \frac{\mathcal{Z}_i\mathcal{Z}_i^T - \mathbb{E}[\mathcal{Z}\mathcal{Z}]}{\boldsymbol{d}} \leq \frac{||\mathcal{Z}_i||^2 - \mathbb{E}[||\mathcal{Z}||^2}{\boldsymbol{d}} \leq \frac{\boldsymbol{2B}}{\boldsymbol{d}}$$

because of the triangle inequality on the norm and Jensen's inequality on the expected value. In order to bound the variance of $\boldsymbol{E}$, we first bound the variance of each of its sumands $\boldsymbol{E_i}$

$$\mathbb{E}[\boldsymbol{E_i^2}] = \frac{\mathbb{E}[(\mathcal{Z}_i\mathcal{Z}_i^T - \mathcal{K})^2]}{\boldsymbol{d^2}}$$

where as $\mathcal{K} = \mathbb{E}[\mathcal{Z}_i\mathcal{Z}_i^T]$

$$\begin{aligned}\mathbb{E}[\boldsymbol{E_i^2}] &= \frac{\mathbb{E}[||\mathcal{Z}_i^2||\mathcal{Z}_i\mathcal{Z}_i^T - \mathcal{Z}_i\mathcal{Z}_i^T\mathcal{K} - \mathcal{K}\mathcal{Z}_i\mathcal{Z}_i^T + \mathcal{K}]^2}{\boldsymbol{d^2}} \\ &\geq \frac{\boldsymbol{1}}{\boldsymbol{d^2}}[\boldsymbol{B}\mathcal{K} - \boldsymbol{2}\mathcal{K}^{\boldsymbol{2}} + \mathcal{K}^{\boldsymbol{2}}] \\ &\geq \frac{\boldsymbol{B}\mathcal{K}}{\boldsymbol{d^2}}\end{aligned}$$

Now, taking all summands together, we get

$$||\mathbb{E}[\boldsymbol{E^2}]|| \leq ||\sum_{\boldsymbol{i=1}}^{\boldsymbol{d}} \mathbb{E}\boldsymbol{E_i^2}|| \leq \frac{\boldsymbol{B}||\mathcal{K}||}{\boldsymbol{d}} \tag{9.32}$$

Thus, we can conclude,

$$\boldsymbol{E}||\hat{\mathcal{K}} - \mathcal{K}|| \leq \sqrt{\frac{\boldsymbol{3B^{'}|K||logN}}{\boldsymbol{d}}} + \frac{\boldsymbol{2BlogN}}{\boldsymbol{d}} \tag{9.33}$$

Notice that both random feature (as $||\mathcal{Z}||^2 \leq \boldsymbol{B}$, where B is bounded) and kernel evaluation are upper bounded by 1, thus, we can conclude that both $\boldsymbol{B}$ and $||\mathcal{K}||$ are bounded by N, resulting Eq. 9.30.

9.3 Experiments

In this section, we evaluate and compare the performance of one class support tensor machines and the effect of randomized feature selection for the task of anomaly detection. To validate the gain in performance, we have performed k-fold (k=10) validation on both vector and tensor datasets downloaded from the UCI machine learning repository and compared the performance with state of the art

vector and tensor-based methods. As our core objective is the classification of the anomalies in large scale data, thus, in order to validate the effectiveness of R1STM-BH against the large datasets, in this section, we performed several experiments with various numbers of dimensions and records. To further validate the effect of bounding the hinge loss function against outliers, we have corrupted the datasets with anomalies.

9.3.1 Dataset

In order to validate the performance of the proposed approach against outliers and gain in computational complexity, we have conducted several experiments on both vectored and tensored dataset. In our first experiment, we have used vector data and transformed it into tensor form. For this purpose, we have download publicly available twelve datasets mostly from UCI machine learning repository that are Breast Cancer [103], Iris, Import, Ionospher, Lung, Sona, Delftpump AR, USPS, Daily and Sport Activity (DSA), Gas Sensor Array (GSA) and PAMAP2 Physical activity monitoring dataset (PAMAP). Most of these datasets are originally vector-based thus, we have transformed these datasets to tensorial representation. For the first eight dataset, we have generated tensor data by transforming the vector data [8] and select the tensor size based on [14]. The datasets (viii-xii) are time series, thus, we transform these datasets into 3rd order tensor as features x samples x times. In our second experiment, we have used tensor data. For this purpose, we have considered four CASIA gait recognition dataset (A dataset [113]). The size of Dataset A is about 2.2GB and the database includes 19139 images. Furthermore, we have also used the face recognition dataset (The ORL Database of Faces) and handwritten digits database (MNIST [38]).

We have normalized all the records in each dataset between [0,1]. For training and validation purposes, the datasets are divided into training and testing set by randomly selecting 80% and 20% records respectively. To validate the robustness of proposed approach, we have corrupted the datasets by 5% anomalies drawn from $\mathbb{U}(\mathbf{0,1})$. As, our approach is unsupervised anomaly detection, thus, during the training phase, we omitted the class labels, however, we have used label during testing.

Table 9.2: Average accuracy (%) and ACU (%) on Breast Cancer dataset with differnt training samples

Sample Size	Target Class	Accuracy	AUC
2	Class 1	82.86 ± 14.85	99.80 ± 0.15
	Class 2	84.06 ± 6.77	90.11 ± 16.78
4	Class 1	84.87 ± 14.34	99.82 ± 14.27
	Class 2	86.43 ± 10.11	97.86 ± 10.45
6	Class 1	89.91 ± 7.22	97.86 ± 6.75
	Class 2	91.65 ± 9.81	99.42 ± 4.65
8	Class 1	91.05 ± 11.24	94.45 ± 7.54
	Class 2	92.11 ± 4.95	98.76 ± 3.45
10	Class 1	92.11 ± 8.87	99.89 ± 2.65
	Class 2	94.67 ± 4.75	98.95 ± 4.50

Table 9.3: Average accuracy (%) and ACU (%) on corrupted Breast Cancer dataset with different training samples

Sample Size	Target Class	Accuracy	AUC
2	Class 1	80.13 ± 10.21	99.10 ± 0.17
	Class 2	80.46 ± 7.79	88.76 ± 10.95
4	Class 1	79.23 ± 9.56	98.51 ± 12.65
	Class 2	82.96 ± 10.56	96.45 ± 9.86
6	Class 1	86.46 ± 8.54	97.86 ± 6.75
	Class 2	88.43 ± 10.23	99.12 ± 2.54
8	Class 1	89.65 ± 8.56	93.22 ± 7.54
	Class 2	90.54 ± 7.55	98.12 ± 4.45
10	Class 1	91.51 ± 10.19	98.65 ± 4.62
	Class 2	90.24 ± 8.88	98.34 ± 6.56

9.3.2 Results and Discussion

The main goals of this work are to improve the robustness of anomaly detection and overcome the complexity issue of support tensor machines for the larger dataset. We have conducted several experiments on both vectored and tensored dataset and performed k-fold cross-validation. As described in section 9.3.1, we transformed the vector data into tensoral representation. Initially, we performed k-fold cross-validation for both vector (Breast cancer) and tensor (MNIST) dataset to find the optimal range of parameters followed by experiments on the rest of the datasets within that optimal range. To validate the robustness against outliers, we have contaminated the datasets with anomalies.

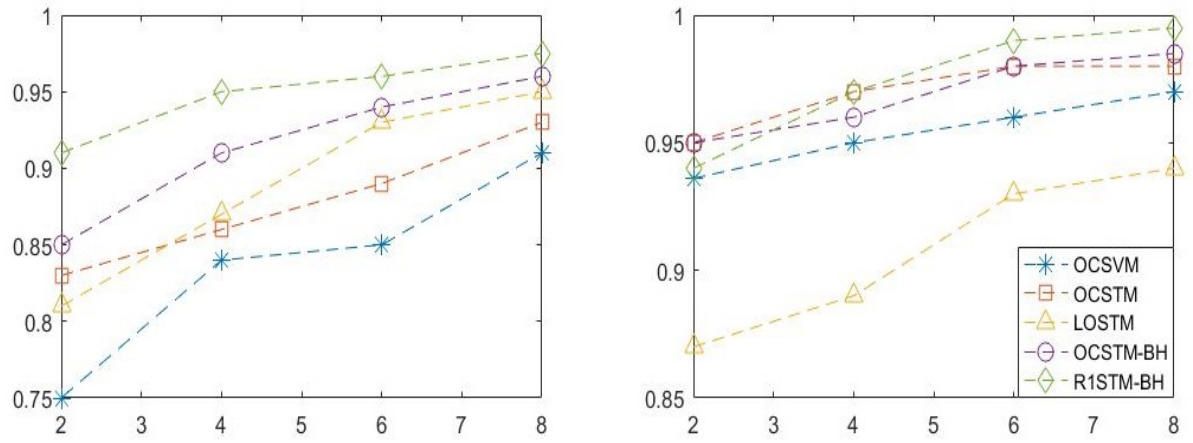

Figure 9.2: Performance comparison of proposed R1STM-BH with state of the art methods on Iris dataset

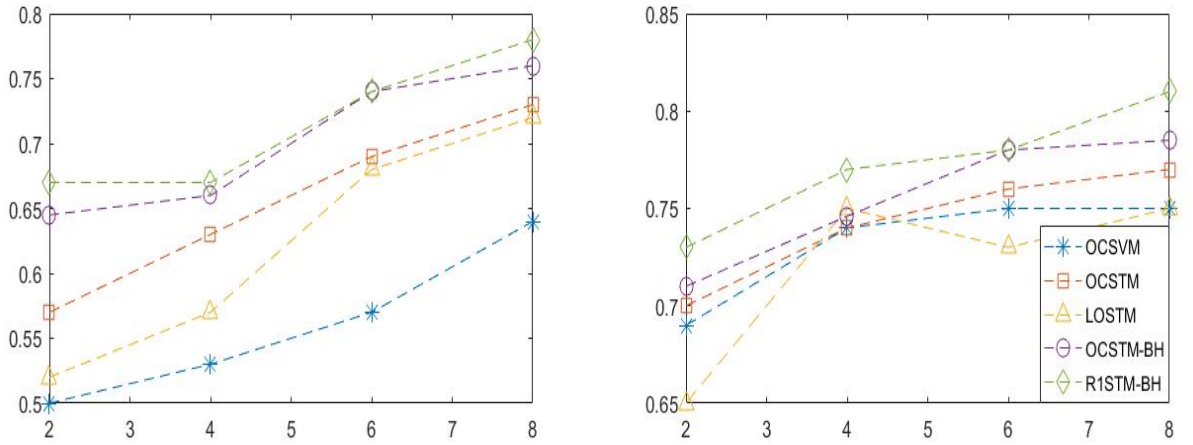

Figure 9.3: Performance comparison of proposed R1STM-BH with state of the art methods on Lungs dataset

In order to estimate the effects of random feature projection and bounded loss function in the construction of the projection matrices, we repeated our cross-validation experiments ten times for all datasets. We have randomly selected 30% of training data to form a validation dataset which we have used to tune parameters. The size of the training dataset is very important for efficient anomaly detection. Some of the methods in the literature showed good performance for large dataset however, are poor for small datasets. Similarly, some of the methods worked better performance for a small dataset, however, showed poor performance for the larger

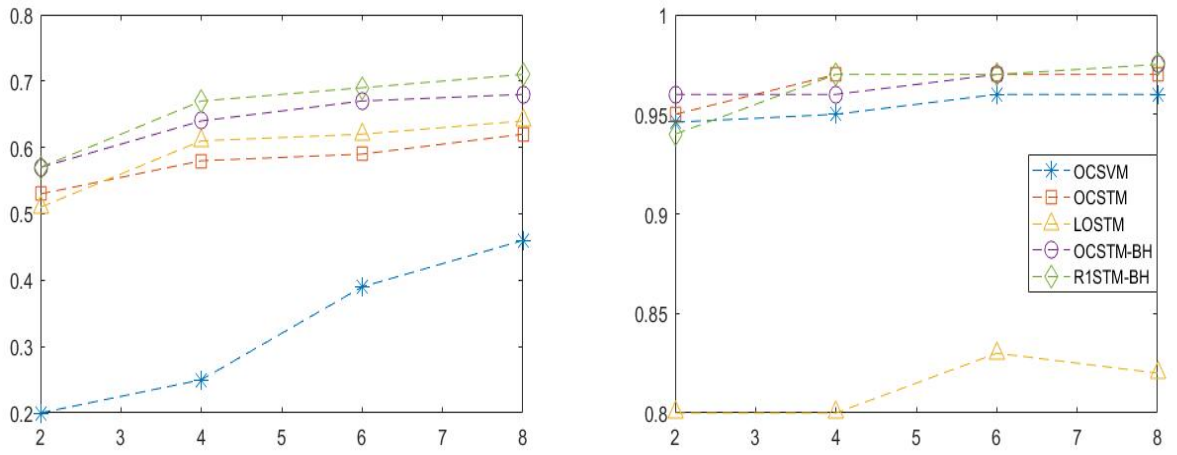

Figure 9.4: Performance comparison of proposed R1STM-BH with state of the art methods on the task of face recognition (ORL dataset)

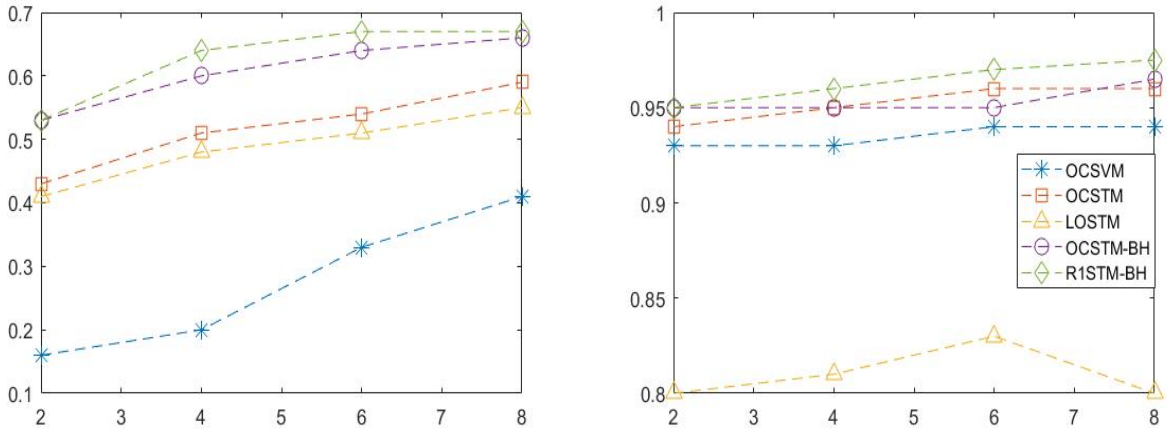

Figure 9.5: Performance comparison of proposed R1STM-BH with state of the art methods on contaminated ORL dataset)

dataset. The other major challenge is computational complexity which grows with the size of data i.e. complexity of kernel based methods can group quadartically. Although our major concern is anomaly detection for larger datasets, however, for generalization purposes and to observe its performance for small dataset, we have conducted several experiments on different sizes of datasets with a varying numbers of dimensions and records to validate the effectiveness of the proposed methodology. In the following the discussion, we provide the experimental results on both vectored (syntactically transformed to 2nd and 3rd order tensor) and

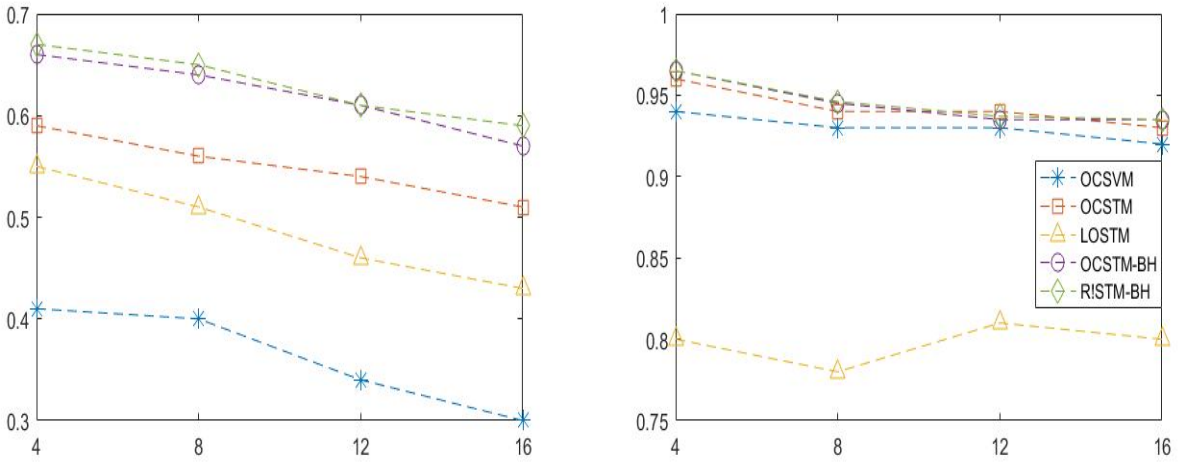

Figure 9.6: Performance comparison of proposed R1STM-BH with state of the art methods with different level of corruption on ORL dataset

tensored datasets. Table 9.2 and table 9.3 shows the results on real and corrupted breast cancer datasets with different subjects.

To elucidate the best comparable anomaly detection performance as compared to state of the art approaches such as OCSTM [14], R1STM [20], LOCSTM [14]), vector methods (OCSVM [11], LOCSVM [13], R1SVM [19]) and deep one-class classification methods (One-Class Deep SVDD [89], Soft Bounded Deep SVDD (SB Deep SVDD) [89]) on **fourteen** publicly available benchmark datasets. 9.7 shows the comparison of results on different numbers of training samples. We can notice that R1STM-BH not only showed better performance for a small number of training samples (2) but also better performance for a larger number of samples (8) per individual whereas the computational complexity remains very attractive. Table 9.4 shows the comparison of results on all datasets. We can notice that R1STM-BH showed significantly better performance in comparison to OCSVM, LOCSVM, LOCSTM, R1SVM, OCSTM, and LOCSTM however results are comparable to deep SVDD and SB deep SVDD. Figure 9.2, figure 9.3 and figure 9.4 show the results on Iris, Lungs and ORL datasets respectively. We can clearly notice that R1STM significantly outperforms all for small sample size (2), however, results are comparable to deep SVD and SB deep SVDD for large sample size. This shows that R1STM-BH is scale-able and works for both small samples and larger sample datasets. Notice that accuracy of anomaly detection is almost comparable when data is free from outliers, however, computationally, the proposed approach showed

better performance.

To validate the robustness against outliers, we have performed evaluation on corrupted data. Results on corrupted dataset are shown in table 9.3 and figure 9.5 on breast cancer and ORL face dataset respectively. We can notice that R1STM-BH showed significantly better performance as compared to the state of the art methods for corrupted data. This shows that bounding the hinge loss overcome the larger loss occurred due to the outliers which in turn effect the performance of anomaly detection. Notice that with the increase in a number of outliers, the proposed anomaly detection showed superiority over the state of the art methods.

Table 9.7 compare the training time, test time and the number of iterations to converge. In this comparison, we only compared the performance with methods based on support vector machines such as OCSVM and OCSTM. Results show that R1STM-BH is much faster both in terms of training and testing as compared to other methods. Furthermore, R1STM-BH converges with a low number of iterations as compared to OCSVM and OCSTM.

We have the following **key observations**

- For $\boldsymbol{\eta} = \mathbf{0}$, the proposed objective function degenerates into to traditional hinge loss ($\aleph$), thus the traditional hinge loss function (Eq.9.1) is a special case of bounded loss function (9.3).
- The objective function is monotonic, bounded and non-convex. Thus, we have used half quadratic optimization to transform the problem into a traditional support tensor machine.
- We observe that randomized projection eliminates the need to deal with large kernel matrices for large datasets result in a not only reduction in time and space complexity but also improving the anomaly detection performance.
- We noticed that a large dimension of random features results in high computational complexity, thus we suggest the smaller size of randomize features.

9.3.3 Parameter Setting

We performed several experiments with different values of parameters to find an optimal range on breast cancer and MNIST datasets. Once we have an optimal

Table 9.4: Average %age of test AUC on different datasets with sample size 2

	AUC								
Dataset	OCSVM	LOCSVM	SB-Deep SVDD	OC Deep SVDD	OCSTM	LOCSTM	R1STM	1STM-BH	R1STM-BH
Breast Cancer	90.17	87.65	95.32	94.22	94.29	88.74	96.02	96.35	96.55
SONAR	58.43	66.21	72.13	72.23	61.88	67.87	69.43	72.11	74.43
Lung	56.88	61.49	78.68	82.56	67.43	66.70	73.45	78.76	81.80
Iris	92.66	94.65	98.74	98.42	94.43	95.11	96.65	98.16	98.47
Delftpump AR	76.77	79.43	94.76	96.67	85.66	87.22	90.43	92.68	96.60
IONOSPHERE	70.45	73.45	86.70	88.22	75.43	77.43	81.20	84.76	88.44
Import	59.32	64.54	87.19	88.43	67.65	71.43	78.91	86.44	88.32
USPS	99.43	99.61	99.91	99.85	99.75	97.81	99.87	99.91	99.95
UHAD	83.42	89.41	98.67	99.13	95.12	97.11	98.47	99.06	99.23
ORL	96.12	73.87	97.21	97.58	96.43	69.43	96.89	97.58	98.01
DSA	79.43	83.47	98.57	99.12	98.24	98.12	99.17	99.2	99.30
PAMAP2	89.43	91.23	98.47	98.21	94.45	95.11	97.45	98.77	98.85
CASIAA	80.41	83.75	98.45	98.61	96.10	96.54	97.77	98.21	98.88
MNIST	74.67	81.43	94.76	95.43	90.32	90.21	93.47	94.54	95.16

Table 9.5: Performance comparision of proposed R1STM-BH with state of the art methods on the task of handwritten digit recognition (MNIST dataset)

Class	OCSVM	SB Deep SVDD	OC Deep SVDD	OCSTM	LOSTM	OCSTM-BH	R1STM-BH
0	66.6 ± 4.5	97.8 ± 0.7	98.0 ± 0.7	83.6 ± 1.9	80.3 ± 2.1	93.5 ± 1.7	98.2 ± 0.7
1	67.5 ± 3.4	99.6 ± 0.1	99.7 ± 0.1	87.4 ± 2.2	86.5 ± 2.6	92.4 ± 0.9	98.5 ± 1.3
2	59.5 ± 5.4	89.5 ± 1.2	91.7 ± 0.8	81.3 ± 5.7	82.2 ± 1.9	89.4 ± 0.7	92.4 ± 1.2
3	61.4 ± 3.8	90.3 ± 2.1	91.9 ± 1.5	87.4 ± 4.3	86.5 ± 3.1	90.7 ± 1.2	92.2 ± 0.7
4	64.6 ± 4.7	93.9 ± 1.5	94.9 ± 0.8	84.3 ± 2.4	85.4 ± 2.4	92.5 ± 0.9	93.4 ± 0.6
5	58.6 ± 6.3	85.8 ± 2.2	88.5 ± 0.9	82.7 ± 5.2	84.4 ± 3.5	87.6 ± 0.7	91.4 ± 0.9
6	67.4 ± 4.6	98.1 ± 0.5	98.3 ± 0.5	88.4 ± 3.2	88.4 ± 3.5	94.5 ± 1.7	97.6± 1.7
7	63.5 ± 5.2	92.8 ± 1.4	94.6 ± 0.9	87.6 ± 2.7	86.8± 3.4	93.5 ± 1.6	92.8 ± 0.9
8	65.4 ± 5.2	92.9 ± 1.4	93.9 ± 1.6	87.5 ± 2.2	87.4 ± 2.2	90.5 ± 0.9	94.2 ± 0.8
9	64.7 ± 3.8	95.1± 0.7	96.5 ± 0.3	88.2 ± 1.8	84.8 ± 2.6	92.4 ± 1.8	95.6 ± 1.1

Table 9.6: Computational and Space complexity analysis of proposed approach with state of the art methods

Approach	Computational Complexity	Space Complexity
OCSVM[40]	$\boldsymbol{O(dN^3)}$	$\boldsymbol{O(d+N^2}$
SVDD [20]	$\boldsymbol{O(dN^2)}$	$\boldsymbol{O(dN^2)}$
Auto-Encoder [20]	$\boldsymbol{O(dmN)}$	O(dq)
ROCSVM [120]	$\boldsymbol{O(dN^3)}$	$\boldsymbol{O(d+N^2}$
R1SVM [19, 117]	$\boldsymbol{O(kn)}$	$\boldsymbol{O(kn)}$
RSVM-RHHQ [106]	$\boldsymbol{O(IN^3)}$	$\boldsymbol{O(IN^3)}$
OCSTM-BH (RBF)	O($\boldsymbol{Bkn^2}$)	O($\boldsymbol{Bkn^2}$)
R1STM-BH	O($\boldsymbol{Bkn}$)	O($\boldsymbol{Bkn}$)

range of these parameters, There are four parameters (scale constant $\boldsymbol{\eta}$, width parameter $\boldsymbol{\sigma}$, trade-off parameter $\boldsymbol{v}$ and the dimension of random features k that are required to be optimal. Inappropriate selection of these parameters may result in poor anomaly detection, thus the value of these parameters should be selected carefully.

We performed k-fold validation (k=10) to find an optimal range of parameters on breast cancer and MNIST datasets. There are four parameters (scale constant $\boldsymbol{\eta}$, width parameter $\boldsymbol{\sigma}$, trade-off parameter $\boldsymbol{v}$ and the dimension of random features k that are required to be optimal. Inappropriate selection of these parameters may result in poor anomaly detection, thus the value of these parameters should be selected carefully. Once we found the optimal range of parameter values, we performed different experiments on that optimal range to find the optimal parameter for a specific dataset. We have observed that larger value of $\boldsymbol{k}$ results in high computational complexity, thus we suggest the smaller size of randomize features.

Similarly, the best performance of proposed anomaly detection we have achieved at
$\boldsymbol{\sigma}$ = {$\mathbf{10,14,14,27,20,15,9,21,24,28,43,31,36}$},
$\boldsymbol{\eta}$ = {$\mathbf{0.3,0.4,0.2,.25,0.25,0.25,0.2,0.5,0.3,0.4,1.45,1.65,1.25}$}
and
$\boldsymbol{v}$ = {$\mathbf{0.2,0.25,0.2,0.3,0.2,0.25,0.3,0.3,0.25,0.2,0.3,0.25,0.35}$} for Breast Cancer, Iris, Import, Ionospher, Lung, Sona, Delftpump AR, USPS, Daily and Sport Activity (DSA), Gas Sensor Array (GSA) and PAMAP2 Physical activity monitoring dataset (PAMAP), CASIA, ORL and MNIST dataset respectively.

9.3.4 Computational Complexity

In this section, we discuss the computational complexity of R1STM-BH. Considering $\boldsymbol{N}$ is the number of training samples and d is the dimension of features. The computational complexity of solving the dual optimization problem imposed by one-class SVM is $\boldsymbol{O(dN^3)}$ and the computational complexity of OCSVM with RBF kernel function is $\boldsymbol{O(dN^2)}$. Similarly, the computational complexity of one class support tensor machine with RBF kernel is $\boldsymbol{O(N^2d_1^2d_2)}$ [14] whereas $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ denotes the second-order tensor such that $\boldsymbol{d = d_1 \times d_2 \approx d}$. The computational complexity of one class support tensor machine with bounded loss function is the complexity for dual optimization with RBF kernel $\boldsymbol{O(N^2d_1^2d_2)}$ and one-class support tensor machines with randomized projection is $\boldsymbol{O(kN)}$, where N is the size of the training dataset. The computational complexity of the Lagrange multiplier is $\boldsymbol{\alpha}$ in each iteration. The complexity of the auxiliary variable and complexity of the margin parameter $\boldsymbol{p}$ is $\boldsymbol{N}$. Thus, the computational complexity of OCSTM-BH is $\boldsymbol{O(H_{BH}((N^2+N+N)d_1^2d_2)}$ and $\boldsymbol{O(H_{BH}((kN+N^2+N)d_1^2d_2)}$ with RBF and randomized kernel respectively. Neglecting the lower order terms, we get $\boldsymbol{O(H_{BH}((kN^2))}$ and $\boldsymbol{O(H_{BH}((kN))}$ for RBF and randomized kernel respectively, where $\boldsymbol{H_{BH}}$ is the complexity of the half quadratic optimization.

To further observe the run time complexity, compare the performance in terms of training, testing time on breast cancer dataset as shown in table 9.8. We can observe that computational and space complexity (both train and test) are much better as compare to the state of the art methods. Furthermore, it requires much less number of iterations to converge.

9.4 Summary

In this work, we presented a novel approach for anomaly detection for large scale datasets with the aim to not only improve the robustness of anomaly detection but also overcome the complexity challenge. We replaced the hinge loss function with bounded the loss function and utilized randomized features projection rather than finding the optimal support vectors. We showed empirically that providing randomized features to a one-class support tensor machine with bounded loss function produces much better results in comparison to state-of-the-art methods. Furthermore, the computational an space complexity is very attractive not only

for large datasets but also for small that validate the scalability of the proposed approach.

Table 9.7: Performance evaluation (Accuracy, AUC and number of iteration) of R1STM-BH with different methods on different training sample size

		1STM				R1STM				1STM-BH				R1STM-BH			
Sample Size	Target Class	Accuracy		AUC		Accuracy		AUC		Accuracy		AUC		Accuracy		AUC	
2	Class 1	63.60	±	99.22	±	73.43	±	99.32	±	81.54	±	99.33	±	82.86	±	99.80	±
		15.32		0.16		7.87		0.12		12.21		0.10		14.85		0.15	
	Class 2	70.21	±	84.76	±	76.64	±	88.65	±	82.34	±	90.21	±	84.06	±	90.11	±
		6.21		20.54		12.54		4.65		8.43		66.51		6.77		16.78	
4	Class 1	75.71	±	98.76	±	81.43	±	99.12	±	83.43	±	99.64	±	84.87	±	99.82	±
		11.43		1.86		12.98		4.67		8.89		5.66		14.34		14.27	
	Class 2	79.91	±	92.43	±	84.54	±	96.75	±	85.58	±	97.54	±	86.43	±	97.86	±
		6.95		10.43		6.75		4.65		12.12		10.76		10.11		10.45	
6	Class 1	82.47	±	98.31	±	87.65	±	98.87	±	89.54	±	99.22	±	91.65	±	99.42	±
		12.12		6.66		7.65		2.43		4.55		6.76		9.81		4.65	
	Class 2	84.54	±	93.21	±	87.53	±	95.32	±	88.77	±	97.54	±	89.91	±	97.86	±
		4.55		3.65		5.43		3.45		6.66		7.54		7.22		6.75	
8	Class 1	84.32	±	98.50	±	88.45	±	93.54	±	89.32	±	93.67	±	91.05	±	94.45	±
		10.37		6.78		11.32		5.78		5.59		4.65		11.24		7.54	
	Class 2	83.65	±	93.43	±	87.54	±	97.65	±	90.32	±	98.65	±	92.11	±	98.76	±
		6.87		8.86		24.54		4.76		5.55		0.19		4.95		3.45	
10	Class 1	84.54±		98.76	±	88.33	±	99.21	±	89.54	±	99.43	±	92.11	±	99.89	±
		8.76		4.56		4.45		2.43		5.56		0.76		8.87		2.65	
	Class 2	88.43	±	94.61	±	91.32	±	96.43		92.22	±	97.21	±	94.67	±	98.95	±
		3.45		8.65		3.45		±2.43		8.86		1.59		4.75		4.5	

Table 9.8: Comparative evaluation of training time (sec), test time (sec) and number of iterations on Breast Cancer dataset

Sample size	Training Time			Test Time			Number of Iterations		
	OCSVM	OCSTM	R1STM-BH	OCSVM	OCSTM	R1STM-BH	OCSVM	OCSTM	R1STM-BH
2	0.0963 ± 0.082	0.0483 ± 0.0674	0.0372 ± 0.047	0.0632 ± 3.32	0.024 ± 2.22	0.014± 1.35	16.43 ± 5.27	11.54 ± 3.96	9.65 ± 4.21
4	0.1759 ±0.098	0.0968 ± 0.087	0.0502 ± 0.065	0.067 ± 4.11	0.054 ± 1.89	0.019 ± 1.43	14.53 ± 3.45	10.32 ± 4.11	8.71 ± 2.76
6	0.2154 ±0.076	0.1043 ± 0.089	0.0614 ± 0.092	0.081 ± 3.76	0.065 ± 2.02	0.021 ± 1.39	14.59 ± 3.46	8.93 ± 3.65	6.43 ± 3.27
8	0.2334 ±0.065	0.1232 ± 0.0932	0.0698 ± 0.099	0.087 ± 2.68	0.084 ± 2.32	0.026 ± 1.33	12.43 ± 4.79	7.43 ± 3.76	4.87 ± 3.29
10	0.2782 ± 0.104	0.1365 ± 0.108	0.0783 ± 0.078	0.096 ± 3.11	0.089 ± 2.43	0.037 ± 1.74	12.40 ± 3.78	5.87 ± 2.76	4.11 ± 2.56

CHAPTER 10

CONCLUSIONS AND FUTURE DIRECTION

Once we know something, we find it hard to imagine what it was like not to know it.

(Chip Dan Heath)

It is no surprise that most of the real-world data have such a high sparsity, i.e., only a small number of features are important. An ad-hoc approach to deal with such problems is achieving the sparsity artificially by considering only those loadings that are greater than threshold however, in general, it is an inefficient approach espicially for small and high dimensional data. Classification of such a small number of noisy data samples that are high in dimensional is a challenging task that require selection of robust features to capture the intrinsic and structural properties. Thus, in this case, sparse models (SSVM), low rank (BSVM, SMM) or low rank plus sparse methods are not sufficient to capture the underlying structural and intrinsic property of the data entirely. ℓ_1 regularizer term has some limitations due to the fact that the selected features are upper bounded by the data sample size. Hence, it provides structural sparsity and does not discover the intrinsic group structure, resulting in the selection of features without considering all the classes. Furthermore, there could be outliers in the data that could affect the classification performance. To deal with aforementioned challenges, we simultaneously explored dimensionality reduction, matrix recovery, feature extraction, and classification.

In chapter 4 and chapter 5, we preformed joint dimensionality reduction and

feature extraction. Chapter 4 presents a robust dimensionality reduction method that by relaxing the orthogonal constraints of the transformation matrix and imposing a penalty function on the regularization term [74]. We presented outliers robust two-dimensional principal component analysis by efficiently integrating the robustness of traditional 2DPCA and the regularization term $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\mathbf{2}}$ that relaxes the orthogonal constraint. The regularization term $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\mathbf{2}}$ reduces the constraints and enables the objective function to select features jointly. Furthermore, the regularization parameter $\|\boldsymbol{Q}\|_{\boldsymbol{F}}^{\mathbf{2}}$ is convex and can be easily optimized. Penalty term penalizes all regression coefficients corresponding to the single feature as a whole to make PCA possible to select features jointly. Hence, ORPCA approximates high-dimensional representation in a flexible manner. As such, ORPCA has more freedom to select low-dimensional features efficiently. The one major drawback of F-norm is its sensitivity against outliers as outlying measurement arbitrarily skew the solution from desired due to squared objective function. As a result, F-norm is not able to utilize the underlying geometric structure in a real sense. To cope with the sensitivity due to squared F-norm, recently, non-square F-norm has been used. Although, ORPCA and 2D-JSPCA have more freedom to select robust features jointly for low dimensional representation that helps to minimize the effect of outliers as well as redundancy. However, it does not guarantee fully sparse solution but it (joint feature selection and alternative derivation of the objective function) makes the objective function robust against outliers. To deal with data redundancy explicitly, in chapter 5, we present an additional penalty term $\|\boldsymbol{Q}\|_{\mathbf{2,1}}$ reduces the constraints and enables objective function to select features jointly and discard the features that already exist in other principal components. Furthermore, both the regularization terms are convex and can be easily optimized. In contrast to previous works on robustness in PCA, we jointly select the important features. The introduction of penalty terms results in the spare and robust solution against outliers by reducing their impact in projection matrix. Compared with state-of-the-art methods, our evaluation results show the improvement in effectiveness proposed approach for the task of data reconstruction and classification. Eventually, 2D-JSPCA has poor reconstruction error because it suffers from loss of information. However, it provides a better reconstruction error with respect to SPCA and JSPCA. It might be due to the selection of important features that helps to reproduce the image. We have noticed that discriminant features selected by 2D-JSPCA are those important and contributive features such as nose, eyes, lips in case of the face

image, while contours of different objects in non-facial datasets. In conclusion, the numerical results suggest that our methods (ORPCA and 2D-JSPCA) are superior to previous approaches. In comparison to ORPCA, 2D-JSPCA provided better performance however, it has a high reconstruction error. However, this calls for further analysis and variations of the ORPCA and 2D-JSPCA. For example, having more than one $\boldsymbol{P}$ and one $\boldsymbol{Q}$, offers more flexibility in accommodating the discriminant features.

It turns out that the nuclear norm can also be used as a convex relaxation of this optimization problem, which greatly simplifies the problem and allows further room for interesting applications such as accelerated algorithms for matrix completion (compressed sensing). Recently, classifier based on combination of hinge loss, nuclear norm and Frobenius norm [1, 46, 142], $\boldsymbol{\ell_1}$ [140, 141] has been presented. Although these methods showed excellent performance by taking advantage of the correlation between rows and columns of the regression matrix under the low-rank assumptions. But, they simply consider entities in the matrix as explanatory factors and do not consider the intrinsic group structure of data and are sensitive to outliers. Furthermore, they also tend to select the features without considering all classes. We propose a novel classifier RSMM works by effectively combining the hinge loss function for model fitting and the elastic net penalty for regularization on the regression matrix. RSMM can achieve the goal stated above, by employing the regularizer term which promotes structural sparsity. The regularization term helps to avoid the inevitable upper bound for the number of selected features occurring in $\boldsymbol{\ell_{2,1}}$-norm SVM. The linear combination of the nuclear norm, $\boldsymbol{\ell_{2,1}}$ inherits the property of low-rank and sparsity together which not only helps to deal with outliers but also selects features across all data points with joint sparsity. Since the optimization is convex but non-smooth and one of the major challenges is, how to efficiently solve non-smooth optimization, we devised an efficient algorithm to solve the proposed objective function based on the Generalized Forward-Backward (GFB) splitting framework. RSMM modeled the group intrinsic structure. The regularization term helps to select the features across all data points with joint sparsity i.e. each feature either has small scores or large scores over all data points. The results on contaminated data show that RSMM provided better results as compared to state of the art methods, which validate our claim that RSMM is robust against outliers and able to model the intrinsic property of the data entirely. RSMM works well for small data corruption, however,

fragile to the presence of outliers: even a single corrupted data point can arbitrarily alter the quality of the approximation, what if a fraction of columns are corrupted then the quality may be poor. To deal with data having extensive corruption, we simultaneously performing matrix recovery, feature selection, and classification through joint minimization of $\ell_{2,1}$ and nuclear norm. We assume that the data consists of a low-rank clean matrix plus a sparse noise matrix by effectively combining the hinge loss function for model fitting, low-rank matrix recovery and an elastic net penalty for regularization on the regression matrix. We performed a simultaneous matrix recovery and classification, which first performs matrix recovery followed by clean feature extraction and classification. Since the convex optimization cannot perform an exact recovery of the corrupted matrix, thus, we used an Oracle Problem for matrix recovery. As a result, convex optimization-based SSMRe performs correct matrix recovery as well as the identification of outliers, which improves the classification performance. We achieve the goal stated above, by employing the regularizer term (a combination of low rank and $\ell_{2,1)}$) which promotes structural sparsity and matrix recovery as well as selects features across all data points with joint sparsity. The low-rank matrix recovery helps to recover the unobserved entities as well as to avoid the inevitable upper bound for the number of selected features occurring in $\ell_{2,1}$-norm SVM. Since the optimization is convex but non-smooth and one of the major challenges is, how to efficiently solve non-smooth optimization, we devised an efficient algorithm to solve the proposed objective function. A comprehensive experimental study on publicly available datasets of image classification and EEG classification was carried out to validate the proposed approach. The experiment results showed the effectiveness of RSSM and SMMRe approach for solving classification problems even fraction of columns are corrupted while keeping a reasonable number of support vectors. Although, RSSM and SMMRe showed amazing performance, however, are binary classifier and could be used for multiclass classification by breaking multiclass problem into series of binary class classification problem such as one-vs-rest (OvR) or one-vs-one (OvO) strategies (e.g. In OvsR, the multi-class problem is solved by splitting it into n binary class classification problems, whereas OvsO approach splits the problem into $\frac{c(c-1)}{2}$ binary classification problems.) but are computationally expensive and may result in the unbalanced distribution of input samples. We extend RSSM to the multiclass Support Matrix Machine (M-SMM) approach by utilizing the maximization of the inter-class margins (i.e. margins between pairs of classes).

The proposed model is a combination of binary hinge loss for models fitting, and elastic net penalty as a regularization on the regression matrix. The binary hinge loss uses C matrices to simulate a one-vs-one classifier of all classes rather than $\frac{c(c-1)}{2}$ models. The regularization term which promotes the structural sparsity and shares similar sparsity patterns across multiple predictors is a combination of Frobenius and nuclear norm. Thus, the proposed objective function not only maximizes the inter-class margins but is a spectral extension of the conventional elastic net that combines the property of low rank and joint sparsity together, to deal with complex high dimensional noisy data. MSMM works by effectively combining the binary hinge loss function (to maximize the inter-class hyperplane margin for model fitting) and elastic net penalty (to promote low-rank plus sparsity), as a regularization on regression matrix. Unlike one vs one classification strategy, we have used C matrices to simulate the binary classification that not only helps to overcome the complexity issue but also maximizes the inter-class margin. Since the optimization is convex and one of the major challenges is how to efficiently solve nonsmooth optimization?, thus, we devised an efficient algorithm for solving the proposed objective functions. In future, multiclass support matrix machine based on $||.||_{2,1}$ and $||.||_*$ as well as privilege information-based support matrix machine can explored. Furthermore, we will be exploring cooperative evolution based support matrix machines that will help to minimizing the training time too by breaking down the original m-Class problem into sub-problems and solving them in cooperative fashion.

Recently, several non-convex and bounded loss functions have been presented to substitute the hinge loss function in order to suppress the effect of outliers and improve the robustness of support vector machines. However, there is no work done for the improvement of one-class tensor machines [1]. Furthermore, the computational complexity of traditional support tensor machines is high and increases with the increase of training samples. Thus, it limits the applicability of OCSTM for large datasets. To overcome the aforementioned challenges, we replaced the hinge loss with bounded loss function and used randomized features [61, 70, 71] rather than finding the optimized support tensors which result in not only improving the robustness against outliers as well as significantly reduces the training time. The proposed support tensor machine with bounded hinge loss is monotonic, bounded and nonconvex, thus robust to outliers by limiting the loss due to outliers. We further extends the approach using randomized non-linear set

of features rather than finding the support vectors, thus, eliminating the need to deal with large kernel matrices for large datasets resulting that results reduction in both time and space complexity. To solve the non-convex objective function, we devised an iterative approach using the half quadratic optimization. Extensive experimental analysis shows that proposed bounded one-class support tensor machines with randomized kernel considerably improves the robustness against outliers and significantly reduces the computational complexity as compared to state of the art anomaly detection methods. We can observe that computational and space complexity (both train and test) are much better as compare to the state of the art methods. Furthermore, it requires much less number of iterations to converge. In future, regularize terms $||.||_*$ and $||.||_{\mathbf{2,1}}$ can be considered with bounded hinge loss function.

BIBLIOGRAPHY

[1]

[2] A. S. AGHAEI, M. S. MAHANTA, AND K. N. PLATANIOTIS, *Separable common spatio-spectral patterns for motor imagery bci systems*, IEEE Transactions on Biomedical Engineering, 63 (2016), pp. 15–29.

[3] A. ANAISSI, Y. LEE, AND M. NAJI, *Regularized tensor learning with adaptive one-class support vector machines*, in International Conference on Neural Information Processing, Springer, 2018, pp. 612–624.

[4] K. K. ANG, Z. Y. CHIN, C. WANG, C. GUAN, AND H. ZHANG, *Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b*, Frontiers in Neuroscience, 6 (2012), p. 39.

[5] P. N. BELHUMEUR, J. P. HESPANHA, AND D. J. KRIEGMAN, *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection*, tech. rep., Yale University New Haven United States, 1997.

[6] S. BHATTACHARYYA, A. KHASNOBISH, S. CHATTERJEE, A. KONAR, AND D. TIBAREWALA, *Performance analysis of lda, qda and knn algorithms in left-right limb movement classification from eeg data*, in Systems in Medicine and Biology (ICSMB), 2010 International Conference on, IEEE, 2010, pp. 126–131.

[7] C. L. BYRNE, *Iterative Optimization in Inverse Problems*, Chapman and Hall/CRC, 2014.

[8] D. CAI, X. HE, AND J. HAN, *Learning with tensor representation*, tech. rep., 2006.

[9] X. CAI, F. NIE, H. HUANG, AND C. DING, *Multi-class l2, 1-norm support vector machine*, in Data Mining (ICDM), 2011 IEEE 11th International Conference on, IEEE, 2011, pp. 91–100.

[10] A. P. CASTAÑO, *Support vector machines*, in Practical Artificial Intelligence, Springer, 2018, pp. 315–365.

[11] C.-C. CHANG AND C.-J. LIN, *Libsvm: a library for support vector machines*, ACM transactions on intelligent systems and technology (TIST), 2 (2011), p. 27.

[12] Y. CHEN, Z. LAI, J. WEN, AND C. GAO, *Nuclear norm based two-dimensional sparse principal component analysis*, International Journal of Wavelets, Multiresolution and Information Processing, 16 (2018), p. 1840002.

[13] Y. CHEN, L. LU, AND P. ZHONG, *One-class support higher order tensor machine classifier*, Applied Intelligence, 47 (2017), pp. 1022–1030.

[14] Y. CHEN, K. WANG, AND P. ZHONG, *One-class support tensor machine*, Knowledge-Based Systems, 96 (2016), pp. 14–28.

[15] ——, *A linear support higher order tensor domain description for one-class classification*, Journal of Intelligent & Fuzzy Systems, (2018), pp. 1–11.

[16] A. D'ASPREMONT, L. E. GHAOUI, M. I. JORDAN, AND G. R. LANCKRIET, *A direct formulation for sparse pca using semidefinite programming*, in Advances in neural information processing systems, 2005, pp. 41–48.

[17] A. DATTA AND R. CHATTERJEE, *Comparative study of different ensemble compositions in eeg signal classification problem*, in Emerging Technologies in Data Mining and Information Security, Springer, 2019, pp. 145–154.

[18] D. S. DE LUCENA, S. R. MORENO, V. C. MARIANI, AND L. DOS SANTOS COELHO, *Support vector machine optimized by artificial bee colony applied to eeg pattern recognition*, in Brain Function Assessment in Learning: First International Conference, BFAL 2017, Patras, Greece, September 24-25, 2017, Proceedings, vol. 10512, Springer, 2017, p. 213.

[19] S. ERFANI, M. BAKTASHMOTLAGH, S. RAJASEGARAR, S. KARUNASEKERA, AND C. LECKIE, *R1svm: a randomised nonlinear approach to large-scale anomaly detection*, (2015).

[20] S. M. ERFANI, M. BAKTASHMOTLAGH, S. RAJASEGARAD, V. NGUYEN, C. LECKIE, J. BAILEY, AND K. RAMAMOHANARAO, *R1stm: One-class support tensor machine with randomised kernel*, in Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 198–206.

[21] J. FENG, H. XU, AND S. YAN, *Online robust pca via stochastic optimization*, in Advances in Neural Information Processing Systems, 2013, pp. 404–412.

[22] Q. GAO, L. MA, Y. LIU, X. GAO, AND F. NIE, *Angle 2dpca: a new formulation for 2dpca*, IEEE transactions on cybernetics, (2017).

[23] Q. GAO, S. XU, F. CHEN, C. DING, X. GAO, AND Y. LI, *R,ÇÅ-2-dpca and face recognition*, IEEE Transactions on Cybernetics, (2018), pp. 1–12.

[24] R. GROSS, I. MATTHEWS, J. COHN, T. KANADE, AND S. BAKER, *Multi-pie*, Image and Vision Computing, 28 (2010), pp. 807–813.

[25] Y. GUERMEUR, *Combining discriminant models with new multi-class svms*, Pattern Analysis & Applications, 5 (2002), pp. 168–179.

[26] R. HAMID, Y. XIAO, A. GITTENS, AND D. DECOSTE, *Compact random feature maps*, in International Conference on Machine Learning, 2014, pp. 19–27.

[27] L. HE, X. KONG, P. S. YU, X. YANG, A. B. RAGIN, AND Z. HAO, *Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages*, in Proceedings of the 2014 SIAM International Conference on Data Mining, SIAM, 2014, pp. 127–135.

[28] X. HE AND P. NIYOGI, *Locality preserving projections*, in Advances in neural information processing systems, 2004, pp. 153–160.

[29] M. A. HEARST, S. T. DUMAIS, E. OSUNA, J. PLATT, AND B. SCHOLKOPF, *Support vector machines*, IEEE Intelligent Systems and their applications, 13 (1998), pp. 18–28.

[30] H. HOTELLING, *Analysis of a complex of statistical variables into principal components.*, Journal of educational psychology, 24 (1933), p. 417.

[31] W.-C. HSU, L.-F. LIN, C.-W. CHOU, Y.-T. HSIAO, AND Y.-H. LIU, *Eeg classification of imaginary lower limb stepping movements based on fuzzy support vector machine with kernel-induced membership function*, International Journal of Fuzzy Systems, 19 (2017), pp. 566–579.

[32] N. T. T. IAN T. JOLLIFFE AND M. UDDIN, *A modied principal component technique based on the lasso*, Journal of Computational and Graphical Statistics, 3 (2003), pp. 531–547.

[33] M. I. M. B. IMRAN RAZZAK, MUHAMMAD KHURRAM AND G. XU, *Randomized nonlinear one-class support vector machines with bounded loss function for outliers detection, future generation computer systems*, Future Generation Computer System.

[34] T. JOACHIMS, T. FINLEY, AND C.-N. J. YU, *Cutting-plane training of structural svms*, Machine Learning, 77 (2009), pp. 27–59.

[35] T.-E. KAM, H.-I. SUK, AND S.-W. LEE, *Non-homogeneous spatial filter optimization for electroencephalogram (eeg)-based motor imagery classification*, Neurocomputing, 108 (2013), pp. 58–68.

[36] Z. KHAN, F. SHAFAIT, AND A. MIAN, *Joint group sparse pca for compressed hyperspectral imaging*, IEEE Transactions on Image Processing, 24 (2015), pp. 4934–4942.

[37] T. KOBAYASHI AND N. OTSU, *Efficient optimization for low-rank integrated bilinear classifiers*, in Computer Vision–ECCV 2012, Springer, 2012, pp. 474–487.

[38] Y. LECUN, L. BOTTOU, Y. BENGIO, P. HAFFNER, ET AL., *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[39] S. LEMM, B. BLANKERTZ, G. CURIO, AND K.-R. MULLER, *Spatio-spectral filters for improving the classification of single trial eeg*, IEEE transactions on biomedical engineering, 52 (2005), pp. 1541–1548.

[40] K.-L. LI, H.-K. HUANG, S.-F. TIAN, AND W. XU, *Improving one-class svm for anomaly detection*, in Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693), vol. 5, IEEE, 2003, pp. 3077–3081.

[41] M. LI, H. XU, X. LIU, AND S. LU, *Emotion recognition from multichannel eeg signals using k-nearest neighbor classification*, Technology and Health Care, (2018), pp. 1–11.

[42] T. LI, M. LI, Q. GAO, AND D. XIE, *F-norm distance metric based robust 2dpca and face recognition*, Neural Networks, 94 (2017), pp. 204–211.

[43] Y. LI, X. WU, AND J. KITTLER, *L1-(2d) 2pcanet: A deep learning network for face recognition*, arXiv preprint arXiv:1805.10476, (2018).

[44] H. LIAN AND Z. FAN, *Divide-and-conquer for debiased l 1-norm support vector machine in ultra-high dimensions*, The Journal of Machine Learning Research, 18 (2017), pp. 6691–6716.

[45] X. LIAO, D. YAO, D. WU, AND C. LI, *Combining spatial filters for the classification of single-trial eeg in a finger movement task*, IEEE Transactions on Biomedical Engineering, 54 (2007), pp. 821–831.

[46] L. LUO, Y. XIE, Z. ZHANG, AND W.-J. LI, *Support matrix machines*, in Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, pp. 938–947.

[47] M. LUO, F. NIE, X. CHANG, Y. YANG, A. G. HAUPTMANN, AND Q. ZHENG, *Avoiding optimal mean ,Ñì2, 1-norm maximization-based robust pca for reconstruction*, Neural computation, 29 (2017), pp. 1124–1150.

[48] O. L. MANGASARIAN AND D. R. MUSICANT, *Lagrangian support vector machines*, Journal of Machine Learning Research, 1 (2010), pp. 161–177.

[49] A. M. MARTINEZ, *The ar face database*, CVC technical report, (1998).

[50] A. M. MARTÍNEZ AND A. C. KAK, *Pca versus lda*, IEEE transactions on pattern analysis and machine intelligence, 23 (2001), pp. 228–233.

[51] A. F. MARTINS, N. A. SMITH, P. M. AGUIAR, AND M. A. FIGUEIREDO, *Structured sparsity in structured prediction*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1500–1511.

[52] H. N. S. H. I. MI RAZZAK, U NASEEM, *Nuclear norm minimization in frequency domain for complex noise*, in International Conference on Neural Information Processing, ICONIP 19, 2019, pp. 1–6.

[53] A. NASEER, M. RANI, S. NAZ, M. I. RAZZAK, M. IMRAN, AND G. XU, *Refining parkinson,Äôs neurological disorder identification through deep transfer learning*, Neural Computing and Applications, (2019), pp. 1–16.

[54] S. NENE, S. NAYAR, AND H. MURASE, *Technical report cucs-006-96*, Columbia Object Image Library (COIL-100), (1996).

[55] P. NETRAPALLI, U. NIRANJAN, S. SANGHAVI, A. ANANDKUMAR, AND P. JAIN, *Non-convex robust pca*, in Advances in Neural Information Processing Systems, 2014, pp. 1107–1115.

[56] T. NEZAM, R. BOOSTANI, V. ABOOTALEBI, AND K. RASTEGAR, *A novel classification strategy to distinguish five levels of pain using the eeg signal features*, IEEE Transactions on Affective Computing, (2018).

[57] T. NEZAM, R. BOOSTANI, V. ABOOTALEBI, AND K. RASTEGAR, *A novel classification strategy to distinguish five levels of pain using the eeg signal features*, IEEE Transactions on Affective Computing, (2018), pp. 1–1.

[58] F. NIE, X. WANG, AND H. HUANG, *Multiclass capped lp-norm svm for robust classifications*, in Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017), 2017.

[59] J. OH AND N. KWAK, *Generalized mean for robust principal component analysis*, Pattern Recognition, 54 (2016), pp. 116–127.

[60] T. PANG, F. NIE, J. HAN, AND X. LI, *Efficient feature selection via $\boldsymbol{l}_{\mathbf{2,0}}$-norm constrained sparse regression*, IEEE Transactions on Knowledge and Data Engineering, (2018).

[61] S. PAUL, C. BOUTSIDIS, M. MAGDON-ISMAIL, AND P. DRINEAS, *Random projections for support vector machines*, in Artificial intelligence and statistics, 2013, pp. 498–506.

[62] K. PEARSON, *Liii. on lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2 (1901), pp. 559–572.

[63] G. PFURTSCHELLER, C. NEUPER, D. FLOTZINGER, AND M. PREGENZER, *Eeg-based discrimination between imagination of right and left hand movement*, Electroencephalography and clinical Neurophysiology, 103 (1997), pp. 642–651.

[64] P. J. PHILLIPS, H. WECHSLER, J. HUANG, AND P. J. RAUSS, *The feret database and evaluation procedure for face-recognition algorithms*, Image and vision computing, 16 (1998), pp. 295–306.

[65] H. PIRSIAVASH, D. RAMANAN, AND C. C. FOWLKES, *Bilinear classifiers for visual recognition*, in Advances in neural information processing systems, 2009, pp. 1482–1490.

[66] M. PONTIL AND A. MAURER, *Excess risk bounds for multitask learning with trace norm regularization*, in Conference on Learning Theory, 2013, pp. 55–76.

[67] F. QI, Y. LI, AND W. WU, *Rstfc: A novel algorithm for spatio-temporal filtering and classification of single-trial eeg*, a, a, 1 (2015), p. 1.

[68] L. QIAO, S. CHEN, AND X. TAN, *Sparsity preserving projections with applications to face recognition*, Pattern Recognition, 43 (2010), pp. 331–341.

[69] H. RAGUET, J. FADILI, AND G. PEYRÉ, *A generalized forward-backward splitting*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1199–1226.

[70] A. RAHIMI AND B. RECHT, *Random features for large-scale kernel machines*, in Advances in neural information processing systems, 2008, pp. 1177–1184.

[71] ——, *Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning*, in Advances in neural information processing systems, 2009, pp. 1313–1320.

[72] I. RAZZAK, M. BLUMENSTEIN, AND G. XU, *Robust support matrix machine*, Pattern Recognition.

[73] ——, *Multiclass support matrix machines by maximizing the inter-class margin for single trial eeg classification*, IEEE Transactions on Neural Systems and Rehabilitation Engineering, (2019).

[74] I. RAZZAK, I. HAMEED, AND G. XU, *Robust sparse representation and multiclass support matrix machines for the classification of motor imagery eeg signals*, IEEE Journal of Translational Engineering in Health and Medicine, (2019).

[75] I. RAZZAK, M. IMRAN, B. MICHAEL, AND X. GUANDONG, *Robust sparse support matrix machines for single trial eeg classification*, Artificial Intelligence in Medicine.

[76] I. RAZZAK, M. IMRAN, AND G. XU, *Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks*, IEEE journal of biomedical and health informatics, (2018).

[77] I. RAZZAK, Z. KHURRAM, M. IMRAN, B. MICHAEL, AND X. GUANDONG, *One-class support tensor machines with bounded hinge loss function for classification of high-dimensional data*, Future Generation Computer System.

[78] I. RAZZAK, B. MICHAEL, AND X. GUANDONG, *Robust support matrix machine for classification of corrupted data*, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–12.

[79] ——, *Support matrix machine via joint l21 and nuclear norm minimization under matrix completion framework for classification of corrupted data*, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–12.

[80] I. RAZZAK, R. A. SARIS, M. BLUMENSTEIN, AND G. XU, *Integrating joint feature selection into subspace learning: A formulation of 2dpca for outliers robust feature selection*, Neural Networks, 121 (2020), pp. 441–451.

[81] I. RAZZAK, R. A. SARIS, B. MICHAEL, AND X. GUANDONG, *Robust two dimensional joint sparse pca with f-norm minimization*, IEEE Transactions on Image Processing.

[82] M. I. RAZZAK, M. IMRAN, AND G. XU, *Big data analytics for preventive medicine*, Neural Computing and Applications, (2019), pp. 1–35.

[83] M. I. RAZZAK, M. K. KHAN, K. ALGHATHBAR, AND R. YOUSAF, *Face recognition using layered linear discriminant analysis and small subspace*, in 2010 10th IEEE International Conference on Computer and Information Technology, IEEE, 2010, pp. 1407–1412.

[84] M. I. RAZZAK, S. NAZ, AND A. ZAIB, *Deep learning for medical image processing: Overview, challenges and the future*, in Classification in BioApps, Springer, 2018, pp. 323–350.

[85] M. I. RAZZAK, R. A. SARIS, M. BLUMENSTEIN, AND G. XU, *Robust 2d joint sparse principal component analysis with f-norm minimization for sparse modelling: 2d-rjspca*, in 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–7.

[86] B. RECHT, M. FAZEL, AND P. A. PARRILO, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM review, 52 (2010), pp. 471–501.

[87] A. ROSALES-PEREZ, S. GARCIA, H. TERASHIMA-MARIN, C. A. C. COELLO, AND F. HERRERA, *Mc2esvm: Multiclass classification based on cooperative evolution of support vector machines*, IEEE Computational Intelligence Magazine, 13 (2018), pp. 18–29.

[88] A. ROSALES-PÉREZ, A. E. GUTIERREZ-RODRÍGUEZ, S. GARCÍA, H. TERASHIMA-MARÍN, C. A. C. COELLO, AND F. HERRERA, *Cooperative multi-objective evolutionary support vector machines for multiclass problems*, in Proceedings of the Genetic and Evolutionary Computation Conference, ACM, 2018, pp. 513–520.

[89] L. RUFF, N. GÖRNITZ, L. DEECKE, S. A. SIDDIQUI, R. VANDERMEULEN, A. BINDER, E. MÜLLER, AND M. KLOFT, *Deep one-class classification*, in International Conference on Machine Learning, 2018, pp. 4390–4399.

[90] Z. SAEED, R. A. ABBASI, O. MAQBOOL, A. SADAF, I. RAZZAK, A. DAUD, N. R. ALJOHANI, AND G. XU, *What‚Äôs happening around the world? a survey and framework on event detection techniques on twitter*, Journal of Grid Computing, (2019), pp. 1–34.

[91] Z. SAEED, R. A. ABBASI, I. RAZZAK, O. MAQBOOL, A. SADAF, AND G. XU, *Enhanced heartbeat graph for emerging event detection on twitter using time series networks*, Expert Systems with Applications, (2019).

[92] Z. SAEED, R. A. ABBASI, M. I. RAZZAK, AND G. XU, *Event detection in twitter stream using weighted dynamic heartbeat graph approach*, arXiv preprint arXiv:1902.08522, (2019).

[93] Z. SAEED, R. A. ABBASI, A. SADAF, M. I. RAZZAK, AND G. XU, *Text stream to temporal network-a dynamic heartbeat graph to detect emerging events on twitter*, in Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 534–545.

[94] F. S. SAMARIA AND A. C. HARTER, *Parameterisation of a stochastic model for human face identification*, in Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on, IEEE, 1994, pp. 138–142.

[95] Y.-H. SHAO, C.-N. LI, M.-Z. LIU, Z. WANG, AND N.-Y. DENG, *Sparse lq-norm least squares support vector machine with feature selection*, Pattern Recognition, 78 (2018), pp. 167–181.

[96] M. SIGNORETTO, L. DE LATHAUWER, AND J. A. SUYKENS, *A kernel-based framework to tensorial data analysis*, Neural networks, 24 (2011), pp. 861–874.

[97] M. SIGNORETTO, E. OLIVETTI, L. DE LATHAUWER, AND J. A. SUYKENS, *Classification of multichannel signals with cumulant-based kernels*, IEEE Transactions on Signal Processing, 60 (2012), pp. 2304–2314.

[98] 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. 53–58.

[99] ——, *The cmu pose, illumination, and expression (pie) database*, in Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, IEEE, 2002, pp. 53–58.

[100] L. SMALLMAN, A. ARTEMIOU, AND J. MORGAN, *Sparse generalised principal component analysis*, Pattern Recognition, (2018).

[101] Z. SONG, D. P. WOODRUFF, AND P. ZHONG, *Low rank approximation with entrywise l 1-norm error*, in Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2017, pp. 688–701.

[102] J. SPILKA, J. FRECON, R. LEONARDUZZI, N. PUSTELNIK, P. ABRY, AND M. DORET, *Sparse support vector machine for intrapartum fetal heart rate classification*, IEEE journal of biomedical and health informatics, 21 (2017), pp. 664–671.

[103] W. N. STREET, W. H. WOLBERG, AND O. L. MANGASARIAN, *Nuclear feature extraction for breast tumor diagnosis*, in Biomedical image processing and biomedical visualization, vol. 1905, International Society for Optics and Photonics, 1993, pp. 861–871.

[104] A. SUBASI AND M. I. GURSOY, *Eeg signal classification using pca, ica, lda and support vector machines*, Expert systems with applications, 37 (2010), pp. 8659–8666.

[105] C. TIAN, Q. ZHANG, J. ZHANG, G. SUN, AND Y. SUN, *2d-pca representation and sparse representation for image recognition*, Journal of Computational and Theoretical Nanoscience, 14 (2017), pp. 829–834.

[106] Y. TIAN, M. MIRZABAGHERI, S. M. H. BAMAKAN, H. WANG, AND Q. QU, *Ramp loss one-class support vector machine; a robust and effective approach to anomaly detection problems*, Neurocomputing, 310 (2018), pp. 223–235.

[107] R. TOMIOKA, G. DORNHEGE, G. NOLTE, B. BLANKERTZ, K. AIHARA, AND K.-R. MÜLLER, *Spectrally weighted common spatial pattern algorithm for single trial eeg classification*, Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep, 40 (2006).

[108] M. TURK AND A. PENTLAND, *Eigenfaces for recognition*, Journal of cognitive neuroscience, 3 (1991), pp. 71–86.

[109] N. VASWANI, T. BOUWMANS, S. JAVED, AND P. NARAYANAMURTHY, *Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery*, IEEE Signal Processing Magazine, 35 (2018), pp. 32–55.

[110] C. VIDAURRE, N. KRÄMER, B. BLANKERTZ, AND A. SCHLÖGL, *Time domain parameters as a feature for eeg-based brain–computer interfaces*, Neural Networks, 22 (2009), pp. 1313–1319.

[111] H. WANG, Q. TANG, AND W. ZHENG, *L1-norm-based common spatial patterns*, IEEE Transactions on Biomedical Engineering, 59 (2012), pp. 653–662.

[112] H. WANG AND J. WANG, *2dpca with l1-norm for simultaneously robust and sparse modelling*, Neural Networks, 46 (2013), pp. 190–198.

[113] L. WANG, T. TAN, H. NING, AND W. HU, *Silhouette analysis-based gait recognition for human identification*, IEEE transactions on pattern analysis and machine intelligence, 25 (2003), pp. 1505–1518.

[114] L. WANG, B. WANG, Z. ZHANG, Q. YE, L. FU, G. LIU, AND M. WANG, *Robust auto-weighted projective low-rank and sparse recovery for visual representation*, Neural Networks, 117 (2019), pp. 201–215.

[115] Q. WANG AND Q. GAO, *Robust 2dpca and its application*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 79–85.

[116] Q. WANG, Q. GAO, X. GAO, AND F. NIE, *Optimal mean two-dimensional principal component analysis with f-norm minimization*, Pattern Recognition, 68 (2017), pp. 286–294.

[117] P. S. WEERASINGHE, T. ALPCAN, S. M. ERFANI, AND C. LECKIE, *Unsupervised adversarial anomaly detection using one-class support vector machines*, (2018).

[118] F. WEN, R. YING, P. LIU, AND R. C. QIU, *Robust pca using generalized nonconvex regularization*, IEEE Transactions on Circuits and Systems for Video Technology, (2019).

[119] L. WOLF, H. JHUANG, AND T. HAZAN, *Modeling appearances with low-rank svm*, in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–6.

[120] Y. XIAO, H. WANG, AND W. XU, *Ramp loss based robust one-class svm*, Pattern Recognition Letters, 85 (2017), pp. 15–20.

[121] H.-J. XING AND M. JI, *Robust one-class support vector machine with rescaled hinge loss function*, Pattern Recognition, 84 (2018), pp. 152–164.

[122] H. XU, C. CARAMANIS, AND S. SANGHAVI, *Robust pca via outlier pursuit*, in Advances in Neural Information Processing Systems, 2010, pp. 2496–2504.

[123] J. XU, X. LIU, Z. HUO, C. DENG, F. NIE, AND H. HUANG, *Multi-class support vector machine via maximizing multi-class margins*, in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 3154–3160.

[124] J. YANG, D. ZHANG, A. F. FRANGI, AND J.-Y. YANG, *Two-dimensional pca: a new approach to appearance-based face representation and recognition*, IEEE transactions on pattern analysis and machine intelligence, 26 (2004), pp. 131–137.

[125] J. YANG, D. ZHANG, X. YONG, AND J.-Y. YANG, *Two-dimensional discriminant transform for face recognition*, Pattern recognition, 38 (2005), pp. 1125–1129.

[126] Q. YE, L. FU, Z. ZHANG, H. ZHAO, AND M. NAIEM, *Lp-and ls-norm distance based robust linear discriminant analysis*, Neural Networks, 105 (2018), pp. 393–404.

[127] S. YI, Z. LAI, Z. HE, Y.-M. CHEUNG, AND Y. LIU, *Joint sparse principal component analysis*, Pattern Recognition, 61 (2017), pp. 524–536.

[128] H. ZHANG, H. YANG, AND C. GUAN, *Bayesian learning for spatial filtering in an eeg-based brain–computer interface*, IEEE transactions on neural networks and learning systems, 24 (2013), pp. 1049–1060.

[129] K. ZHANG, L. LAN, Z. WANG, AND F. MOERCHEN, *Scaling up kernel svm on limited resources: A low-rank linearization approach*, in Artificial Intelligence and Statistics, 2012, pp. 1425–1434.

[130] Y. ZHANG, Y. WANG, J. JIN, AND X. WANG, *Sparse bayesian learning for obtaining sparsity of eeg frequency bands based feature vectors in motor imagery classification*, International journal of neural systems, 27 (2017), p. 1650032.

[131] Y. ZHANG, Y. WANG, G. ZHOU, J. JIN, B. WANG, X. WANG, AND A. CICHOCKI, *Multi-kernel extreme learning machine for eeg classification in brain-computer interfaces*, Expert Systems with Applications, 96 (2018), pp. 302–310.

[132] Y. ZHANG, Z. ZHANG, J. QIN, L. ZHANG, B. LI, AND F. LI, *Semi-supervised local multi-manifold isomap by linear embedding for feature extraction*, Pattern Recognition, 76 (2018), pp. 662–678.

[133] Z. ZHANG, F. LI, M. ZHAO, L. ZHANG, AND S. YAN, *Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification*, IEEE Transactions on Image Processing, 25 (2016), pp. 2429–2443.

[134] ——, *Robust neighborhood preserving projection by nuclear/l2, 1-norm regularization for image feature extraction*, IEEE Transactions on Image Processing, 26 (2017), pp. 1607–1622.

[135] Z. ZHANG, J. REN, W. JIANG, Z. ZHANG, R. HONG, S. YAN, AND M. WANG, *Joint subspace recovery and enhanced locality driven robust flexible discriminative dictionary learning*, arXiv preprint arXiv:1906.04598, (2019).

[136] Z. ZHANG, S. YAN, AND M. ZHAO, *Pairwise sparsity preserving embedding for unsupervised subspace learning and classification*, IEEE Transactions on Image Processing, 22 (2013), pp. 4640–4651.

[137] M. ZHAO, T. W. CHOW, Z. WU, Z. ZHANG, AND B. LI, *Learning from normalized local and global discriminative information for semi-supervised regression and dimensionality reduction*, Information Sciences, 324 (2015), pp. 286–309.

[138] Q. ZHAO, G. ZHOU, T. ADALI, L. ZHANG, AND A. CICHOCKI, *Kernel-based tensor partial least squares for reconstruction of limb movements*, in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 3577–3581.

[139] M. ZHENG, J. BU, C. CHEN, C. WANG, L. ZHANG, G. QIU, AND D. CAI, *Graph regularized sparse coding for image representation*, IEEE Transactions on Image Processing, 20 (2011), pp. 1327–1336.

[140] Q. ZHENG, F. ZHU, AND P.-A. HENG, *Robust support matrix machine for single trial eeg classification*, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26 (2018), pp. 551–562.

[141] Q. ZHENG, F. ZHU, J. QIN, B. CHEN, AND P.-A. HENG, *Sparse support matrix machine*, Pattern Recognition, 76 (2018), pp. 715–726.

[142] Q. ZHENG, F. ZHU, J. QIN, AND P.-A. HENG, *Multiclass support matrix machine for single trial eeg classification*, Neurocomputing, 275 (2018), pp. 869–880.

[143] H. ZHOU AND L. LI, *Regularized matrix regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76 (2014), pp. 463–483.

[144] J. ZHU, S. ROSSET, R. TIBSHIRANI, AND T. J. HASTIE, *1-norm support vector machines*, in Advances in neural information processing systems, 2004, pp. 49–56.

[145] H. ZOU, T. HASTIE, AND R. TIBSHIRANI, *Sparse principal component analysis*, Journal of computational and graphical statistics, 15 (2006), pp. 265–286.