基于ADMM的PCA异常空间稀疏化方法

（申请清华大学工学硕士学位论文）

培养单位	：	计算机科学与技术系
学　　科	：	计算机科学与技术
研 究 生	：	安奈威
指导教师	：	陈文光 教授

二〇一六年五月

基于ADMM的PCA异常空间稀疏化方法

安奈威

An ADMM approach to estimate Sparse Abnormal Subspace for Anomaly Detection and Interpretation

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Master of Science

in

Computer Science & Technology

by

Kazi Abir Adnan

Thesis Supervisor： Professor Wenguang CHEN

May, 2016

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。本人保证遵守上述规定。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：________________ 导师签名：________________

日　　期：________________ 日　　期：________________

摘　要

异常检测在当前的诸多研究领域都是一个重要的问题。在如今的“大数据”时代，大部分已有的异常数据检测技术都是面向特定领域且难以扩展的。此外，大部分数据集都未经标注，需要非监督的检测技术来识别其中的异常值。一般地，数据集中的绝大部分实例都是正常的，而剩下的那些最不符合模型的实例最终被判定为异常。异常解释是另一个用于检测异常原因的关键技术，我们需要一个模型用于解释所探测到的异常的原因。这样的一个模型应能识别出不同类型的异常，并且保留原始的有意义的特征（输入）变量的子集。

主成分分析（Principal Component Analysis, PCA）是用于异常检测的一个开创性方法，但它的不足之处在于可解释性较差。而稀疏 PCA 仅关注于头个主成分，更加剧了这一问题。为克服这一困难，紧缩（deflation）技术令最重要的几个主成分去解释其中最大的差异，然而这一方法低效且无法解决模糊主成分的问题。此外，获得稀疏主成分是一个最优化问题，很难用紧缩方法计算它的全局解。

本文提出了一种 d 维稀疏子空间估计问题（这是一个半定规划问题）的解决方案。在 PCA 的基础上，结合有噪声输入数据矩阵的奇异值分解（Singular Value Decomposition, SVD）以及凸松弛（Convex Relaxation），使得异常实例在异常子空间中具有更长的投影长度。其中的凸松弛问题可以由 Alternating Direction Method of Multipliers (ADMM) 有效而高效地解决，ADMM 计算出最不重要的主成分的 d 维异常子空间，而这些主成分是非常易于解释的。张成异常空间的稀疏正交负载向量（基）能够很好地解释每个单独的异常。在此基础上，本文提出了另一种可选方案，即使用串行的 ADMM 来计算这些负载向量以获得更优的解释。我们的方法在一个合成的数据集和两个实际数据集上表现出比传统方法更高的准确度，并且能够对不同类型的异常作出合理的解释。

关键词：异常子空间，异常检测，ADMM，稀疏 PCA，异常解释, 非监督算法

Abstract

Anomaly detection is an essential phenomena in diverse research areas and application domains now a days. In this era of Big Data many anomaly data detection techniques are being proposed which is mostly supervised, domain specific and not scalable. Furthermore, most of the datasets are unlabeled and we need unsupervised detection techniques to recognize the anomalies. In general, utmost of the instances in the datasets are normal while rest fit least to the model eventually identified as anomaly. Interpretation is also another key concern to reveal the explanations of being anomaly. So, we need a model to interpret the reasons of its detection as well. The model should also detect different kind of outliers with original meaningful subset of feature variables.

Principal Component Analysis (PCA) is a pioneering technique for anomaly detection which has a major weakness of poor interpretability. To improve, Sparse PCA has been proposed which introduces new issue of focusing only on most significant PC explaining maximum variance. To overcome, deflation technique is used to get top PCs explaining maximum variance which is inefficient, sequential and doesn’t deal the problem of indistinct eigenvectors. Above all, obtaining sparse PCs is an optimization problem whose global solution is intractable to compute using deflation method.

In this paper, we propose a solution to ***d***-dimensional sparse abnormal subspace estimation problem as a semidefinite program with convex relaxation and using the connection of PCA with singular value decomposition (SVD) incorporating noisy input matrix. The solution gives us a subspace where anomalies have higher projection length. The convex problem is solved efficiently and effectively using Alternating Direction Method of Multipliers (ADMM) computing the ‘abnormal’, d least significant PC’s covering less variance of dataset and very much interpretable. The sparse and orthogonal loading vectors (basis) representing the abnormal subspace can explain each individual anomalies eloquently. Afterwards, we proposed another alternate solution to compute this subspace sequentially using ADMM for better interpretation performance. Our methods achieved even better accuracy than traditional detection techniques on one synthetic and two real world dataset and can provide valid explanations for different types of anomalies.

Keywords: Abnormal Subspace, Anomaly Detection, ADMM, Sparse PCA, Anomaly Interpretation, Unsupervised Algorithm

Contents

List of Illustrations

List of Tables

List of Abbreviations

PCA	Principal Component Analysis
PC	Principal Component
sPCA	Sparse PCA
ADMM	Alternating Direction Method of Multipliers
DSPCA	Direct Sparse PCA
FPS	Fantope Projection and Selection
LFPCA	Localized Functional PCA
ASPCA-FG	Abnormal Subspace Sparse PCA - Forward
ASPCA-BG	Abnormal Subspace Sparse PCA - Backward
AUC	Area under Curve
ROC	Receiver Operating Characteristic
SDP	Semi-definite Programming
s.t.	Subject to

Chapter 1 INTRODUCTION

1.1 Background

Variety, Velocity and Volume, the 3V’s define Big Data. Data is forever and present everywhere. As organization grows, the data associated with it also grows and today there are lots of complexity and inter-dependency in individual datasets. Moreover, Big Data is not just about tons of data, it is a concept providing an opportunity to find hidden insights of existing data and provide guidelines as well to predict the future representation of it.

Data Mining is a process of exploring these large sets of data in order to explore the potential knowledge and significant information that it contains. So, it is basically all about investigating data for different unique requirements. Likewise, anomalous data detection is an example of Data Mining and very much needed in this era of Big Data.

An Anomaly is an instance of dataset which is significantly exceptional from the majority or usual trend of data. Anomalies are also referred as outliers, abnormal data, discordant, noise in the field Data Mining and Data Science. In most applications, data is generated by one or more specific internal processes, which could either reflect activity in the system or observations collected from different entities. Eventually when the process behaves in an unusual way, it results in the formation of abnormal data. Therefore, an outlier often contains useful information about abnormal characteristics of the systems or entities, which comes out of the data generation process. The recognition of such unusual characteristics provides valuable domain specific insights. Some examples are as follows: network intrusion detection systems, credit card fraudulent, sequential sensor events, health science, medical diagnosis, law enforcement, earth science and etc.

Three comprehensive categories of anomaly detection techniques exist ranging from supervised to unsupervised. Unsupervised anomaly detection technique finds anomalies in an unlabeled dataset under the assumption that the majority of the instances in the dataset are of usual trend and looks for instances that seem to fit least to the remainder of the data set. On the other hand, supervised anomaly detection technique requires a training dataset that has been labeled as "normal" and "abnormal" data and involves learning a classifier using those labeled data. Semi-supervised anomaly detection technique defines a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test data instance to be generated by the trained model.

Another aspect of Anomaly Detection is the interpretation. The interpretability of an outlier detection model is extremely important from the perspective of the analyst as well. It is often desirable to determine why a particular data point is an outlier in terms of its behavior with respect to the majority of data. This provides the analyst further suggestions about the diagnosis required in an application specific scenario. For example in the case of different sets of data where one would like to see few significant variables, making it easy to interpret by human.

Thus, the necessity of detecting the abnormal data and the hidden reason behind it is conspicuous these days. In most common scenarios it is inexpedient to expect labeled data. Therefore, people need an unsupervised method which identifies anomaly independently and clearly as well.

1.2 Motivation

PCA is a revolutionary technique for abnormal data detection. For many years, anomaly detection based on PCA has emerged as a powerful method for detecting anomalies. The main reason behind this is that PCA is comprehensible and very easy to implement. Due to few limitations of interpretation, PCA now a days is not the best choice where anomaly interpretation is needed. So, the main motivation of our work is to find a method which acts like PCA and the result of PCA can also be achieved under some configuration for fair comparison. To sum up, we need an easy and reasonable method for anomaly detection and interpretation that eradicates the downside of PCA.

In most of the domains, pinpointing the sources of anomalies from log files is very imperative such as diagnosing failures or recovering systems/networks. Ranging from Computer Science to Health Science and other sectors as well, we need a system that exposes the different reasons of diverse abnormal reality. However, for each detected anomaly, an ideal model should also be able to interpret the motives of its detection. We know that, different datasets have diversified level of interpretability. Typically, models which work with the original attributes, and use fewer transforms on data such as sparse as possible have higher interpretability. But in reality, most of the methods fail to interpret the reasons of anomaly. At the same time it is very much domain dependent and unscalable. Henceforward, our method should be domain independent and sparse enough to explain reasons easily and clearly for different settings of environment.

Another concerned aspect is that in today's modern Big Data era, most of the datasets are unlabeled. Labeling dataset is expensive and at the same time mostly unfeasible. It needs a lot of human effort to classify each data and also need cross validation of each labeling. So, to detect outliers we need an efficient unsupervised method that can easily detect abnormality without any prior knowledge or training.

The last but not the least reason of our motivation is the computational complexity of existing methods in literature. Our study concentrated to find a way to reduce the current computation overhead to detect anomalies with interpretation using PCA based models. Most of the studies have solution using deflation techniques which we are going to introduce in later chapters. We want to shape our model incorporating this aspect of reducing the overhead of computation.

As computer science is ameliorating day by day, people already focused all the mentioned motivations. But, till now to the best of our knowledge performing anomaly detection and at the same the interpretation is still a hard computational job in the perspective of PCA. Therefore, our one of the prime focuses is to reduce this hard job to a simple and understandable one that can be easily implementable.

1.3 Thesis Objective

The main objective of this thesis is to forget Big Data and find the Right Data. In this paper we attempt to detect anomalies in an unsupervised manner and interpret individual anomalies with original important feature variables at the same time. Given a ($\boldsymbol{n} \times \boldsymbol{p}$)-dimensional dataset, where $\boldsymbol{n}$ is the sample size of dataset and $\boldsymbol{p}$ is the feature space size, a model can be constructed by forming an abnormal subspace with $\boldsymbol{d}$ least significant sparse and orthogonal loading vectors (eventually basis) of the subspace. Our prime focus is to find a $\boldsymbol{d}$-dimensional ($\boldsymbol{d} \lll \boldsymbol{p}$) subspace which represents the set of abnormal data. Outliers should have higher projection values on this subspace so that we can easily represent the individual outliers from the basis of this estimated subspace. Afterward, our objective is to identify individual anomalies and its reasons from the loading vectors. To conclude, we need a method to find out the reason or behavior of different abnormalities which eventually means to find combinations of the original attributes and the specific criteria for data points being interpreted as outliers.

Eventually, we propose a sparse ***d***-dimensional abnormal subspace estimation problem as a semidefinite program (SDP) with convex relaxation. The convex problem can be solved efficiently using Alternating Direction Method of Multipliers (ADMM) which computes the abnormal subspace with least significant PC's covering least variance of dataset and referred as "Abnormal PC" which is sparse and orthogonal. These loading vectors represent the abnormal subspace and can interpret each individual anomalies and can achieve accuracy at least as good as traditional anomaly detection techniques. We also proposed another alternate method to get the least significant PCs sequentially rather than simultaneously for better interpretation and compared with the result of simultaneous extraction of PCs from abnormal subspace. We have used the ADMM algorithm again for the second method to compute abnormal PCs one by one for anomaly detection and interpretation which is ultimately a special case of first method.

1.4 Summary of Contribution

We have contributed two efficient and effective method to estimate ***d***-dimensional abnormal subspace which is actually computing least significant ***d*** sparse principal components for anomaly detection and interpretation. The main contribution of this paper can be summarized as follows:

1. We formulate the sparse ***d***-dimensional abnormal subspace estimation problem eventually revealing least significant sparse and orthogonal loading vectors that identifies anomalies, especially of different kinds. In other words, the model computes the least ***d*** sparse orthogonal loading vectors (basis of the subspace) of a covariance matrix, ***S*** based on a noisy input data matrix, ***D***.
2. An efficient model for individual anomaly detection and direct way of interpretation with the subset of the original feature space (input) variables which is sparse in nature.
3. One (ASPCA-ADMM) of the two efficient and effective parametric ADMM algorithm to solve a novel convex semidefinite programming (SDP)

formulation with a Fantope constraint to get the ***d***-dimensional abnormal subspace simultaneously (span of least orthogonal sparse PC's).

4. Another ADMM based algorithm (ASPCA-ADMMSEQ) to get the abnormal PC's sequentially rather computing the whole subspace at a time for better anomaly interpretation performance using deflation technique.
5. A method that can also formulate PCA result in a special parametric input combination circumstance.
6. A comprehensive and comparative analysis with other well-known and traditional unsupervised methods (PCA, ASPCA-BG) in the field of outlier detection with respect to Accuracy, Explained variance, Sparsity and Interpretation.

1.5 Thesis Structure

The rest of this thesis is organized as follows. Chapter 2 discusses about the related work done in the field of Computer Science especially in outlier detection using PCA. In Chapter 3 the precursors or background knowledge required to understand the methodology of our proposed methods has been discussed. Chapter 4 describes the method of estimating the abnormal subspace of anomalous data. It contains both of our proposed simultaneous and sequential method to estimate the abnormal subspace. Chapter 5 introduces the technique to detect anomaly and at the same time the procedure to interpret exposed individual anomalies. Chapter 6 reveals the comprehensive experimental details of well-known datasets including synthetic and real world with all possible evaluations. We compared our proposed methods with PCA (as baseline) and another outstanding method ASPCA-BG. An example of clustering performance is also analyzed in this chapter. Finally, Chapter 7 ends with some concluding brief remarks with possible extensions as future work.

Chapter 2 RELATED WORKS

Unsupervised Anomaly Detection is of great demand for extensive period in different domain perspective. It is a technique of Data Mining which detects outliers in an unlabeled dataset under the assumption that some of the instances in the dataset are outliers. There are different unsupervised methods to detect anomalies in dataset. But, in our study we mainly focused on PCA-based methods. Because, PCA is a well-known popular method which is easy to implement and at the same time comprehensible enough to understand.

Principal Component Analysis (PCA) is a powerful and efficient unsupervised method for detecting a wide variety of anomalies [10, 20]. Indeed, PCA is mostly known as a dimension reduction tool [1]. It usually helps us to visualize data in lower dimensions. People have used PCA for anomaly detection in different domains. KDD’99 Network Intrusion is one of the well-known benchmark dataset for anomaly detection. We can find many works where people took help of PCA to detect anomalies of this dataset [10, 19]. Usually in most of the datasets, we have very small set of abnormal data and the main advantage of PCA is that it doesn’t need to accommodate a trained set to identify these anomalies. But, PCA suffers from a major weakness of interpretability in higher dimensions [10]. PCA involves most of the feature variables in the computed eigenvectors or principal components.

Various proposals have been introduced later in the literature to improve PCA for interpretation. Jolliffe (1995) described several rotation techniques that are helpful for interpreting PCs [14]. To address the drawbacks of PCA new technique called Sparse PCA (sPCA) has been introduced with the essence of PCA with the assumption that most of the principal components depend mostly on few variables imposing the lasso penalty on the regression coefficients [14]. As a result, the objective becomes finding principal component that explain maximum amount of variance like PCA but includes less feature space variables. So, the number of variables created by linear combinations in PCs is much lower than the number of variables in original feature space (input variables). It introduces new cardinality constraint which is widely known was sparsity constraint. But, sparse PCA is a hard combinatorial problem [6]. Many techniques were proposed to solve the problem. It has been proposed that sparse PCA can be approximated by semidefinite programming (SDP) [8]. Since, cardinality constraint makes the problem numerically intractable, to overcome the difficulty, convex relaxation approach has been proposed to solve it approximately. In [8] cardinality constraint is represented by a 1-norm convex constraint and we get a semidefinite programming relaxation, which can be solved efficiently in polynomial time. But, this results only computing significant sparse PC which explains most variance. But, in reality this PC might explain less variance than what is expected. Therefore, in most of the cases we need more than one PC which explains maximum variance of data instances.

Dominant multiple sparse PCs are important to explain the discrepancies present in a dataset. Previous method of sPCA mainly focused only on top PC that explains the maximum variance. To get multiple sparse PCs, sPCA is solved with an additional outer layer of iteration. In other words, a sequence of sparse PCA problems are solved via the deflation technique for each sparse PC [15, 21]. This means to find sparse PC's one by one until maximum variance has been covered by computed PC's cumulatively. This deflation method solves a SDP problem in each iteration and this leads to a new problem of computation inefficiency. Because, in each iteration of deflation technique solving a SDP problem is a time consuming task.

Fundamentally, the principal components are directions of variation corresponding to eigenvectors of the population or covariance matrix. It is also not guaranteed that largest eigenvalues are distinct. Individual eigenvectors can be unreliable if the gap between their eigenvalues is small [3]. In [3] one method has been proposed which emphasizes on the span of eigenvectors called the principal subspace of maximum variation. This subspace captures the highest variance of dataset and normal data mostly projects on this subspace of most significant sparse PC's. This reduces the need of using deflation technique which saves a lot of computation by solving only one SDP problem with convex relaxation. In [3], the author proposed a novel convex optimization problem to estimate the ***d***-dimensional principal subspace of a population matrix ***S*** based on an input matrix ***D***.

Simultaneous subspace estimation introduces another issue of interpretation performance. [4] Proposed a method which focused on sequential estimation of subspace rather a simultaneous way and exposed that interpretation performance improves if the PCs are computed sequentially. This method estimates the ***d***-dimensional principal subspace of a population matrix ***S*** by sequential computation of most significant ***d*** principal components enforcing orthogonality constraint to be represented as eigenvectors.

Later, in [2] an idea is proposed to estimate an abnormal subspace to represent the anomalies and find the reasons of their abnormality. This abnormal subspace is mainly composed of least significant PCs which covers very less variance of input data and assumed this should represent the abnormal trend present in dataset. This idea to analyze least significant PCs is somehow novel to all other mentioned related works in literature. Almost all methods just emphasized on important principal components with maximum variance, [2]'s method just concerns about the PCs with lowest variance. The advantage of this method is clearly visible in anomaly detection. Abnormal data will have almost zero length projection on normal subspace and they will form an "abnormal subspace" defined by removing the normal subspace. The main assumption of this method is, dataset must need to have at least a small set of abnormal data.

What is found so far, we can now detect anomalies and at the same time interpret the reason behind it using least significant PCs which cannot be done analyzing most significant PCs. But, the method of [2] uses deflation technique and eventually solves a SDP problem in each iteration. The author proposed methods named ASPCA-FG and ASPCA-BG to estimate the subspace in two direction. ASPCA-FG finds the sparse PCs in same direction like PCA. One of the drawbacks of the ASPCA-FG framework is that the last abnormal PCs tend to have poor sparsity. To solve this problem, another method, Backward ASPCA framework (shorted as ASPCA-B) is proposed that extracts the least significant orthogonal PCs first. This prioritizes the optimization of the sparsity of the abnormal subspace. But, the sparsity level hasn't improved much which eventually ended including another step of optimization problem to get sparse loading vectors and named it as ASPCA-BG method.

In this thesis, we thought of an efficient process to overcome the flaws of ASPCA-BG method. We want a method that estimates abnormal subspace of ***d*** dimension with least significant orthogonal and sparse PCs which is close to the idea mentioned in [3] but with completely different objective function and motivation. Here we want to define a sparse abnormal subspace estimation problem as a semidefinite program with convex relaxation. We want to use the connection of PCA with singular value decomposition (SVD) or eigenvalue decomposition of the noisy input data matrix and extract the least significant PCs (basis of abnormal subspace) through solving a low rank matrix approximation problem using SDP formulation with convex relaxation and orthogonal guarantee on computed PCs. As a result, regularization penalties will also be added to encourage the corresponding minimization problem to increase sparsity in loading vectors. But, incorporating sparsity and low rank approximation constraint at the same time is not an easy problem to solve.

ADMM is a well-known algorithm to optimize convex problems [7] and it has been used in wide range of problems ranging from optimization to distribution [24]. In our case, we have a convex optimization problem and that problem can be solved efficiently and effectively using ADMM. This problem computes the sparse abnormal subspace with least ***d***-significant sparse PC's which is very much interpretable. The sparse and orthogonal loading vectors represented as eigenvectors and as a basis of the estimated abnormal subspace. These PCs can interpret each individual anomalies eloquently present in the dataset. This proposed method achieved even better accuracies than traditional anomaly detection techniques on some real world and synthetic dataset.

We also propose another method to estimate abnormal subspace by computing least significant PCs one by one (sequentially) using ADMM rather approximating whole abnormal subspace. The main motivation of this method is to improve the interpretation of abnormal PCs. From [4] we know that individual eigenvectors has better interpretability performance. So, our second method estimates the abnormal subspace sequentially rather than estimating the whole subspace at a time eventually improving the interpretation performance. This scheme is mainly deflation technique and it doesn't require same period of time as deflated sPCA due to the blessing of ADMM.

Chapter 3 BACKGROUND

3.1 Notations

Input/data matrix, ***D*** is a ***p*** dimensional data (feature space) with ***n*** samples forming (***n*** × ***p***) dimensional matrix where each row vector corresponds to a ***p***-dimensional data instance, and each column vector corresponds to a single feature variable. ***S*** is defined to be the population or covariance matrix of ***D***. *Tr(**S**)* represents the trace of covariance matrix, ***S***. *Card(**S**)* or $\| \boldsymbol{S} \|_0$ denotes the cardinality (number of non-zero elements) of matrix ***S***. $\| \boldsymbol{S} \|_q$ is the usual ℓ_q norm ***S*** and ***I*** is defined as the identity matrix.

For matrices *A*, *B* of compatible dimension $\langle A, B \rangle = Tr(A^T B)$ is the Frobenius inner product, and $||A||_2^2 := \langle A, A \rangle$ is the squared Frobenius norm. Eigen decomposition of covariance matrix, $\boldsymbol{S} = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \lambda_3 v_3 v_3^T + \cdots + \lambda_p v_p v_p^T$ where, $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$ (eigenvalues) and $v_i^T v_j = \delta_{ij}$ are eigenvectors. ***d*** is the size of abnormal subspace which represents the outliers and ***d***-dimensional projection matrix of the data is, $D \rightarrow \prod_d D$ where $\prod_d = V_d V_d^T$ is the ***d***-dimensional projection matrix and $V_d = (v_1, v_2, v_3, .., v_d)$ are the eigen vectors. $V_{abnormal} = (v_1, .., v_d)$ is referred as the abnormal subspace with least significant ***d*** principal components explaining less variance of dataset or represents the outliers. $\boldsymbol{\lambda}$ is the sparsity parameter and $\boldsymbol{\rho}$ is the step size or penalty parameter of Augmented Lagrangian.

3.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a well-known method for dimension reduction, with various applications in Engineering, Market Research, Biology, and Social Science etc. It focuses mainly to capture variations present in data and finds the strong patterns with in it. Now a days, it's mostly used as a visualization model in Data Science. The main assumption of PCA model is that the data can be embedded in a lower dimensional subspace. In PCA, using the covariance matrix of the original dataset, one can calculate the matrix of the principal components. As a result, we can say PCA is a dimensionality reduction technique that captures the highest variance of a multi-dimensional dataset in a lower dimensional subspace defined by a set of orthogonal eigenvectors. It is a procedure to identify smaller number of uncorrelated variables mainly known as principal components from high dimensional input feature space. Reader can refer to [1] to get more idea on PCA.

We know PCA finds a linear combinations of the original feature variables in a new direction of orthogonal vectors which captures maximum variance of data. This can be computed through the Singular Value Decomposition (SVD) or eigenvalue decomposition of the data matrix. Let an input data matrix, $\boldsymbol{D}$ is ($\boldsymbol{n} \times \boldsymbol{p}$) dimensional, where $\boldsymbol{n}$ and $\boldsymbol{p}$ are the number of observations and the number of input feature space variables, respectively. Let, $\boldsymbol{S}$ be the covariance matrix of input data matrix, $\boldsymbol{D}$. So, the SVD of $\boldsymbol{S}$ is,

$$S = UXU^T \tag{3-1}$$

Where, $Z = UX$ are the principal components (PCs), and the columns of U are the respective loadings of the PCs. Usually the leading, significant PCs represent the data as these can exploit most of the variance present in dataset and achieves a great reduction of dimension. PCA has two major benefits which made this method very widespread. PCs can sequentially capture the variance of data matrix in the directions of columns U and ensures minimum information loss. Another major advantage of PCs are the independence among themselves. So, one can concentrate only on a single PC without refereeing or thinking of another PC. It is believed that the different dimensions in real data sets are highly correlated with one another and PCA can exploit this. This is because the different attributes are usually generated by the same underlying process in closely related ways.

However, PCA also has a major drawback that each PC is a linear combination of all $\boldsymbol{p}$ feature variables and the loading vectors of PCs are usually nonzero. This makes it often difficult to interpret the derived new computed PCs for better understanding. As a result, it cannot identify important feature variables present in the PCs and at the same time very difficult to interpret anything out of this non-sparse PCs.

3.3 PCA for Anomaly Detection

Principal Component Analysis is much more stable to the presence of few outliers, than the dependent variable analysis methods [5]. For example, if someone needs to detect fraudulent dealings, enough data instances of fraud examples to train the model might not be present in the dataset, but there might have many examples of normal data transactions. However, using the PCA-based anomaly detection method, the model can be trained using the available data features to determine what is being established as a normal data. On the other hand using different available distance measurement evaluation metrics to identify cases which are different than normal data eventually defined as outliers.

Given a $\boldsymbol{p}$-dimensional feature space, a PCA based anomaly detection model can be constructed by forming a normal subspace defined by the most significant PCs which captures most variance of dataset and an abnormal subspace with the remaining least significant ones by removing the normal subspace [2, 5]. The normal subspace captures the maximum variance present among the dataset, it is worth assuming that this subspace resembles to the usual trends of the dataset. All normal data tends to have almost zero length projection on the abnormal subspace. Therefore, the model can detect whether it is an anomaly or not based on whether it is primarily expressed by the normal or abnormal subspace [18].

Let, $V_{normal} = (v_1, v_2, v_3, .., v_k\ )$ be the normal subspace defined by the first or leading significant k principal components with $(v_1, v_2, v_3, .., v_k\ )$ being the orthogonal loading vectors. Given the $\boldsymbol{p}$-dimensional single data instance y, its residual of projection length on normal subspace, $\hat{y}$ is defined as:

$$\hat{y} = y - \ V_{normal} V_{normal}^{T} y \qquad (3\text{-}2)$$

The squared length of $\hat{y}$, entitled as the Squared Prediction Error (SPE), is the metric to indicate whether y is an anomaly or not [2]. The higher the SPE value is, it is more likely that y belongs to the set of anomaly. It is important to note that most of the contribution to the SPE score is computed by summing up the projection length along each of the PCs present in the subspace; see [5] for more elaborate discussion.

3.4 Anomaly Interpretation

The interpretation of the anomalies is very much intuitive in different domains with respect to different scenarios. A data instance may have several measured quantities, and significantly abnormal behavior of this instance may be represented only with a small subset of these quantities [5]. If we consider the example of [5], an airplane mechanical fault uncovering situation, the results of thousands of different airframe tests on the plane may mostly be normal, with some noisy variations, which are not significant. On the other hand, some deviations in a small subset of tests may be significant enough to be symbolic of anomalous behavior. So, the aim is to reconnoiter PCA based method to identify the vital reasons of outliers. However, an essential issue of PCA as discussed earlier and also claimed in [18], is that the classical PCA based anomaly detection methods cannot be applied directly to anomaly interpretation.

In our anomaly detection step, if the SPE score of a given data instance y is over a predefined threshold, y is identified as an anomaly. Now, at next step we need to understand and pinpoint the source of abnormality for data instance, y. In other words, we need to find anomalous feature behaviors of input feature space which are responsible for isolating y from normal trend. We denote this problem as Anomaly Interpretation. But it is worth to be noted again that the anomaly interpretation using PCA based model is pretty difficult due to the absence of direct mapping from original feature space to PCA's reduced dimension for individual anomalies [18].

Now, given the normal subspace V_{normal} and abnormal $V_{abnormal}$, we can rewrite $\hat{y}$ using the orthogonal property as:

$$\hat{y} = y - V_{normal}V_{normal}^{T}y = V_{abnormal}V_{abnormal}^{T}y \tag{3-3}$$

Using orthogonal property of loading vectors mentioned and proved in [2], we can redefine SPE as,

$$SPE = \hat{y}^{T}\hat{y} = \sum_{i=1}^{d}(v_i^T y)^2 \,, where\ V_{abnormal=}(v_1, v_2, v_3, .., v_d) \tag{3-4}$$

In other words, SPE is equal to the squared sum of y's scalar projection on each abnormal PCs present and we can identify the set of responsible PCs for the abnormality indicated by projection lengths. But, at this point, these PCs are not easy to interpret, because of its linear combination of all feature variables present in feature space.

Hereafter, if we manage to extract PCs not only orthogonal loading vectors but also sparse to represent abnormal subspace, these can be useful to interpret anomalies as well. Precisely, orthogonality constraint ensures that anomaly can be translated to high projection values on a set of abnormal components and the sparsity safeguards the interpretation of abnormality.

Nonetheless, the major weakness of our anomaly interpretation is the non-sparse high dimensionality of the input feature space involved in components. If we can estimate a low cardinal representation of the high non-sparse components using few input variables, we can easily pinpoint the important abnormal behavior or source input variables. Unfortunately, PCA does not ensure sparsity in PCs. In consequence, Sparse PCA is an advanced developed algorithm where each PC is a sparse linear combination of the original sources. As a result, sparsity can help here to identify the important feature variables from the PCs.

3.5 Sparse Principal Component Analysis (sPCA)

The sparse Principal Component Analysis (Sparse PCA) problem is a variant of the PCA problem, which realizes a trade-off between the explained variance along a normalized vector, and the number of non-zero components of that vector [14]. We already know the key shortcoming of PCA is that the factors are linear combinations of PCs involve all original feature variables. Most of the loadings are non-zero in PCs. We also know that PCA facilitates model visualization by concentrating the information in a few principal components, but the principal components are still constructed using all present input variables and hard to interpret.

In many applications, the coordinate axes involved in the PCs have a direct interpretation. In financial or biological applications, each axis might correspond to a specific asset or gene [9]. In problems such as these, it is natural to seek a trade-off between the two goals of explaining most of the variance in the data and making sure that the factors involve only a few which might facilitate to interpret each PCs. The aim here is to efficiently derive sparse principal components, a set of sparse loading vectors that explain maximum amount of variance like PCA. The motivation is that in many applications, the decrease in statistical fidelity required to obtain sparse factors is small and relatively benign [6].

In what follows, [14] focused on the problem of finding sparse PC which explain a maximum amount of variance of the original data and written as

$$\max_{||X||\leq 1} X^T SX - \lambda Card(X) \tag{3-5}$$

In the variable $X\epsilon\boldsymbol{R}^n$, where $S\epsilon\, \boldsymbol{S_n}$ is the (Symmetric positive semi-definite) sample covariance matrix, λ is a parameter regulating sparsity, and $Card(X)$ denotes the cardinality or the number of non-zero coefficients of X.

While PCA is analytically easy to solve and each factor requires computing a leading eigenvector, sparse PCA is a hard combinatorial problem [6]. A problem that is directly related to it is that of computing a cardinality constrained maximum eigenvalue, by solving optimization problem mentioned in [14]

$$\begin{gathered}\max_{X} X^T S X \\ s.t.\,\mathrm{Card(X)} \leq \mathrm{k}\,, \|X\| = 1\end{gathered} \tag{3-6}$$

And also rewritten using semidefinite relaxation as

$$\begin{gathered}\max_{X} Tr(SX) \\ s.t. Tr(X) = 1 \\ Card\ (X) \leq\ k^2 \\ X \succcurlyeq 0\ , Rank(X) = 1\end{gathered} \tag{3-7}$$

Where X is a positive semi-definitive matrix with the constraint $Rank(X) = 1$, which can be uniquely decomposed as xx^T. $Tr(X) = 1$ is equivalent to $x^T x = 1$, $Card\ (X) \leq\ k^2$ is equivalent to $Card\ (X) \leq\ k$ [14]. So, an evolution is introduced by turning the convex maximization objective and the nonconvex constraint into a linear constraint and linear objective. But, this solution only returns the most significant PC and often we need more than one PC to explain required variance captured by these.

As a result, various other methods [15, 21] later has been proposed to get more than one sparse PC solving a sequence of sparse PCA problems using the deflation technique. This leads to a new problem of solving SDP problem iteratively and costs a great computation. Thus, to get rid of this drawback we need a method that computes all the PCs at a time. Eventually this leads to estimate a subspace of PCs where PCs can be represented as a basis of the subspace by enforcing orthogonality among them to treat as eigenvectors as well. Sparsity constraint is also added to get a subspace to make it easy enough to interpret.

3.6 Subspace Estimation

Now we need to formulate the Sparse Subspace estimation using PCA problem to compute loading vectors simultaneously rather than sequentially. The recently studied sPCA framework [5] added a sparsity constraint on the PCs to get sparse result. However, this framework cannot be used directly to solve abnormal PC estimation problem in our case. The main reason is that the sPCA framework usually does not enforce orthogonality on the resultant sparse PCs what is essential for our study. Consequently, the resultant sparse PCs cannot be used to define the normal and abnormal subspaces, as the abnormal PCs are not the orthogonal complement of the normal PCs. By enforcing orthogonality, sparse PCA is used in [2] to solve abnormal subspace estimation problem, which [2] denoted as forward ASPCA (shorted as ASPCA-F). Given a covariance matrix ***S*** and a sparsity constraint constant k, for each $i = 1, \ldots, p$, ASPCA-F tries to solve:

$$\begin{gathered}\max_{v_i} v_i^T S v_i \\ s.t.\, v_i^T v_i = 1, v_i^T v_j = 0\,, \forall\, 1 \leq j < i, Card(v_i) \leq k\end{gathered} \tag{3-8}$$

The last ***d*** loading vectors obtained by solving the equation are used for detecting and interpreting anomalies [2]. One of the drawbacks of the ASPCA-F framework is that the last abnormal PCs tend to have poor sparsity. To solve this problem, they proposed another Backward ASPCA framework (shorted as ASPCA-B) that extracts the least significant orthogonal PCs first, which prioritizes the optimization of the sparsity of the abnormal subspace.

$$\begin{gathered}\min_{v_i} v_i^T S v_i \\ s.t.\, v_i^T v_i = 1, v_i^T v_j = 0\,, \forall\, 1 \leq j < i, Card(v_i) \leq k\end{gathered} \tag{3-9}$$

Let $V_i = (v_1, v_2, v_3, .., v_i\ )$ be the subspace and $\mathrm{R_i} = \mathrm{V_i V_i^T}$, the orthogonality constraint $v_j^T v_i = 0\,;\, \forall\, 1 \leq j < i$ is equivalent to $||V_{i-1}^T V_i||_2^2 = 0$ and $||V_{i-1}^T V_i||_2^2 = V_i^T V_{i-1} V_{i-1}^T V_i = Tr(R_{i-1} X_i) = 0$. Similarly as in [14], [2] relaxed $Card\ (X_i) < k^2$ to $||X_i||_{1,1} < k$ and added it to the objective function with a sparsity coefficient λ. Later, the nonconvex constraint $rank\ (X_i) = 1$ is dropped, and formulated an objective function that can be solved by SDP.

$$\begin{gathered}\min_{X_i \in S^p} Tr(SX_i) +\ \lambda ||X_i||_{1,1} \\ s.t.\, X_i \succcurlyeq 0\,, Tr(X_i) = 1, Tr(R_{i-1} X_i) = 0\end{gathered} \tag{3-10}$$

But still, to estimate the subspace, [2] uses a sequential deflation process of computing PCs and an extra step of global optimization for better sparseness. And an improvement is expected by computing the PCs at a time simultaneously and reducing the computation. Given a symmetric input matrix ***S***, we want a sparse principal subspace estimator $\hat{X}$ that is defined to be a solution of the semidefinite program and projects in a convex body, and uses a sparsity constraint or regularization parameter $\lambda\ \geq\ 0$ like (3-10) to encourage sparsity. So, the objective function becomes,

$$\min_X Tr(SX) + \lambda\ ||X||_{1,1} \tag{3-11}$$

$$s.t. X \epsilon\ Convhull(d - dimension\ projection\ matrices)$$

Later we will explain that the objective function is not easy to find a tractable global solution. To ease the computation, we are going to introduce a popular fashion to solve convex minimization function in next section.

3.7 Alternating Direction Method of Multipliers (ADMM)

We noted that solving the objective function of our problem is not an easy task. We need an efficient method which can deal the constraints and difficulty involved in computation at the same time. In our case, we will have a model which involves convex functions. Many problems in machine learning and big data can be modeled in the framework of convex optimization. The Alternating Direction Method of Multipliers (ADMM) is well suitable to convex optimization problems, and in particular to large-scale problems arising in statistics, machine learning, and related areas [7]. The method was developed in the 1970s, with roots in the 1950s, and is equivalent or closely related to many other algorithms [7]. To understand the steps involved in ADMM, we can start from the basics of Dual Ascent method which has significant relevance with it. We will try to present the basics of ADMM in this section. Afterwards, if someone needs to have an elaborate idea on ADMM, reader is referred to [7] which is an excellent starting point to investigate the working procedure of ADMM.

3.7.1 Dual Ascent

We will focus on Dual Ascent method to minimize an objective function here in this section which has resemblance with ADMM. Most of the equality constrained optimization problem has the form like,

$$\begin{gathered}\min_{x} f(x) \\ s.t. Ax = b\end{gathered} \tag{3-12}$$

With variable $x \in \mathbf{R}^n$, where $A \in \mathbf{R}^{p \times n}$ and $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is convex. The Lagrangian $L(x, y)$ of the mentioned problem is defined as,

$$L(x, y) = f(x) + y^T (Ax - b) \tag{3-13}$$

And the dual function [see 7 for background] of the Lagrangian is

$$g(y) = inf\ L(x, y) = -f^*(-A^T y) - b^T y \tag{3-14}$$

Where y is a dual variable or Lagrange multiplier, and f^* is the convex conjugate of f; Reader is referred to [7] for background. Now, the dual problem can be defined as,

$$\max_{y} g(y) \tag{3-15}$$

With variable $y \in \mathbf{R}^m$ and we assume that strong duality property holds here. Eventually it seems that the optimal values of the primal and dual problems are the same [7]. Afterwards, we can compute the primal optimal point of the objective problem, x^* using the dual optimal point, $y^*$from the dual problem [7],

$$x^* = argmin\ L(x, y^*) \tag{3-16}$$

Providing there is only one minimizer of $L(x, y^*)$ which means we need to have a saddle point for a global solution and it is only possible if f is a strictly convex function. If we refer to dual ascent method again, the dual problem of our primal one is solved using well-known gradient ascent method. We again assume that g is differentiable here for our convenience, and the gradient can be evaluated effortlessly. We first find $x^+ = argmin\, L(x, y)$; then we have gradient, $\nabla g(y) = Ax^+ - b$, which is the residual to uphold the equality constraint. The dual ascent method mentioned in [7] consists of iterating the following updates

$$x^{k+1} := argmin\, L(x, y^k) \tag{3-17}$$

$$y^{k+1} := y^k + \alpha^k(Ax^{k+1} - b) \tag{3-18}$$

Where $\alpha^k > 0$ a step size, and the superscript is the iteration number of algorithm needed to converge. The first step is called x-minimization step, and the second step is dual variable update. This algorithm is called Dual Ascent and with appropriate choice of α^k, the dual function increases in each step to converge for a global optimal solution; see [7] for background

If α^k is chosen appropriately and several other assumptions hold, x^k converges to optimal point and y^k converges to optimal dual point. However, these assumptions do not hold in many applications, so dual ascent often cannot be used directly [7].

3.7.2 Augmented Lagrangian and the Method of Multipliers

Method of multipliers is introduced here to robustify the method of Dual Ascent. Let us assume, we have to optimize following function.

$$\begin{gathered} \min_{x,z} f(x) + g(z) \\ s.t. Ax + Bz = c \end{gathered} \tag{3-19}$$

With variables $x \in \mathbf{R}^n$ and $z \in \mathbf{R}^m$, where $A \in \mathbf{R}^{p\times n}$, $B \in \mathbf{R}^{p\times m}$, and $c \in \mathbf{R}^p$. We will also assume that f and g are convex as mentioned in [7]. The only difference from the general linear equality constrained problem mentioned earlier in the previous section, is that the variable, called *x* there, has been divided into two variables, called *x* and *z*, with the objective function separated across this splitting. The optimal value of the problem is then denoted by; [see [7] for background]

$$p^* = inf\{f(x) + g(z) \mid Ax + Bz = c\}. \tag{3-20}$$

As like the method of multipliers, the augmented Lagrangian can be formed as,

$$\begin{aligned} L_\rho(x,z,y) = f(x) + g(z) + y^T (Ax + Bz - c) \\ + \left(\frac{\rho}{2}\right) \left|\left|Ax + Bz - c\right|\right|_2^2 \end{aligned} \tag{3-21}$$

And the method of multipliers has the following steps to form a solution

$$(x^{k+1}, z^{k+1}) := argmin\, L\rho(x,z,y^k) \tag{3-22}$$

$$y^{k+1} := y^k + \rho(x^{k+1} + B\, z^{k+1} - c) \tag{3-23}$$

The good news of method of multipliers is that it converges under much more relaxed conditions. But, introducing the quadratic penalty destroys one of the major advantages of decomposition. It destroys the decomposition of *x*-update step. Alternating Direction Method of Multipliers (ADMM) can exploit the fact and support decomposition and converges to optimal global solution.

3.7.3 ADMM Algorithm

ADMM is an algorithm that is intended to split the first *x*-update step of Dual Ascent method and at the same time uses the convergence properties of the method of Multipliers. The algorithm solves optimization problems in the form,

$$\begin{aligned} \min_{x,z} f(x) + g(z) \\ s.t.\ Ax + Bz = c \end{aligned} \tag{3-24}$$

And the augmented Lagrangian with quadratic penalty with two sets of variables with separable objective is, [see 7 for background]

$$L_\rho(x,z,y) = f(x) + g(z) + y^T (Ax + Bz - c) + \left(\frac{\rho}{2}\right) \left|\left|Ax + Bz - c\right|\right|_2^2$$

According to the definition of [7], ADMM algorithm consists of the following iterations of *x*-minimization, *z*-minimization and a dual update respectively.

$$x^{k+1} := argmin\, L_\rho(x,z^k,y^k) \tag{3-25}$$

$$z^{k+1} := \; argmin\, L_\rho(x^{k+1}, z, y^k) \tag{3-26}$$

$$y^{k+1} := \; y^k \; + \; \rho(x^{k+1} \; + \; B\, z^{k+1} \; - \; c), \tag{3-27}$$

Where step size or penalty parameter, $\rho > 0$. The algorithm is very similar to Dual Ascent and the method of Multipliers. It basically consists of *x*-minimization step, a *z*-minimization step, and a dual variable update in a sequence. The method of multipliers uses dual variable update as step size equal to the augmented Lagrangian parameter ρ.

Now, we want to find the similarities with method of multipliers. We know that the method of multipliers has the steps mentioned in (3-22, 3-23).

Here the augmented Lagrangian is minimized together with respect to the two primal variables [7]. In ADMM, on the other hand, *x* and *z* are updated in an alternating or sequential fashion, which explains the term alternating direction. ADMM can be considered as a special version of the method of multipliers where a single *Gauss-Seidel* pass over *x* and *z* is used instead of the usual joint minimization [7]. Unravelling the minimization over *x* and *z* into two steps is precisely what allows for a decomposition if *f* or *g* functions are separable.

Now, *x*-update or *z*-update step often requires minimizing common patterns. Suppose for *x*-update it requires to minimize the following function which we can find after some analytical mathematical steps in scaled form where $v = Bz^k - c + u^k$.

$$x^{k+1} := \; \min_x f(x) + \left(\rho/2\right) \|Ax - v\|_2^2 \tag{3-28}$$

$$z^{k+1} := \; \min_z g(z) + \left(\rho/2\right) \|Ax^{k+1} - v\|_2^2 \tag{3-29}$$

$$y^{k+1} := \; y^k \; + \; \rho(x^{k+1} \; + \; B\, z^{k+1} \; - \; c), \tag{3-30}$$

In reality, several similar types of special cases often shows up while optimizing such kinds of convex functions. As a result, we can simplify the update steps exploiting the structure in similar classes.

3.7.4 Proximity Operator

Consider an *x*-update step where $A = I$ and let *f* is a convex function. Now, we want to find an approximate solution of *x*-update step and define it as

$$x^{+} = \min_x (x) + \left(\rho/2\right) \|x - v\|_2^2 = \; prox_{f,\rho}\,(v) \tag{3-31}$$

The proximal operator is used to approximate value, while making a negotiation between the accuracy of the estimate and a cost associated to the new value. Suppose we have the following minimization problem:

$$\min_{x} \frac{1}{2} \parallel u - x \parallel_2^2 \tag{3-32}$$

It is evident that the best value to x is the x itself. Now, adding a cost to the choice what we made of u to be x.

$$prox_h(x) = \min_{x} \frac{1}{2} \parallel u - x \parallel_2^2 + h(u) \tag{3-33}$$

Now the best approximation depends how expensive it is to make $u = x$, given $h(u)$, and a concession must be made. If we consider the case that $h(u) = 0$. Then, the proximal operator $prox_h(x)$ is just the analytic solution to the minimization problem.

There are some special cases often comes around in convex optimization and we suggest reader to remember these mostly used convex forms. For example, two popular proximal operators are mentioned below [7].

$f = I_C(Indicator\ function\ of\ set\ C)$ $\qquad$ $x^+ = \Pi_C(v)\ (projection\ onto\ C)$

$f = \lambda \parallel . \parallel_1 (l_1\ norm)$ $\qquad$ $x^+ = S_{\lambda/\rho}(v)\ (soft\ thresholding)$

We can refer a good and sufficient reading material [26] for better understanding on Proximal Algorithms. We will mention two proximal algorithms which will be used extensively to approximate our objective of estimating abnormal subspace.

3.7.4.1 Fantope Projection

A Fantope is the convex hull of that set comprising outer product of all orthonormal matrices of particular dimension [16]. The column dimension forms a corresponding Fantope. The Fantope has vital involvement in the implementation of low rank constraints of semidefinite programs. And in our case we want to use this constraint to solve our low rank approximation problem.

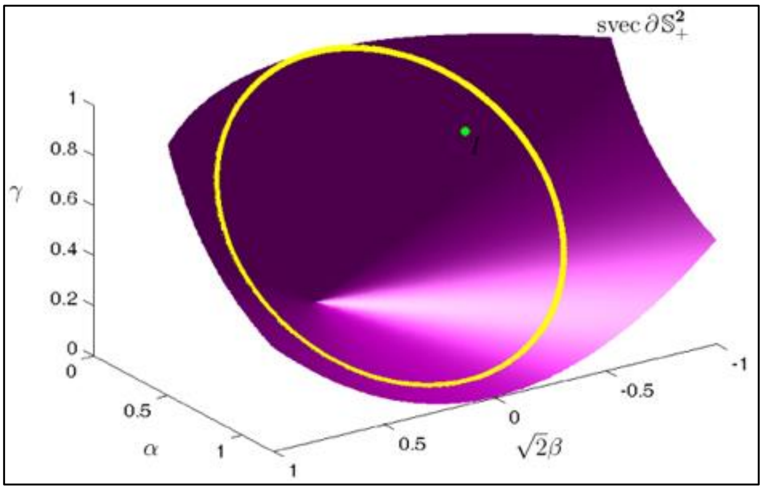


Figure 3-1 Fantope structure

In the example illustrated above and referring [16], the circular Fantope represents outer product of all *2×2* rank-1 orthonormal matrices. The identity matrix is the Fantope comprising outer product of all *2×2* rank-2 orthonormal (orthogonal, in this case) matrices. Now we can define a body, $F_{p_d} := \{ X : 0 \preccurlyeq X \preccurlyeq I \, and \, Tr(X) = d\}$ which is a convex body and called the Fantope [3]. We can use the Fantope body, defined earlier to represent this ***d***-dimensional projection matrix and we define it as,

$$Convhull(d - \dim matrice) = \begin{cases} X : X = X^T ; 0 \preccurlyeq X \preccurlyeq I_p; \, Trace(X) = d \\ \qquad\qquad F_{p,d} \\ \qquad The \, Fantope \end{cases}$$

At this stage, we need an approximation value of this Fantope projection using proximal algorithm. We will use this approximation in our *x*-update step of both proposed ADMM algorithms extensively.

Suppose, P_{F^d} is the Euclidean projection onto $F_{p,d}$and is given in closed form in the following lemma from [3].

If, $X = \sum_i \gamma_i u_i u_i^T$is a spectral decomposition of X, then

$$P_{F^d}(X) = \sum\nolimits_i \gamma_i^+(\theta) u_i u_i^T \tag{3-34}$$

Where $\gamma_i^+(\theta) = min(max(\gamma_i - \theta, 0), 1)$ and θ satisfies the equation $\sum_i \gamma_i^+(\theta) = d.$

This is the Fantope Projection and computed using a spectral decomposition of X. P_{F^d} is the ***d***-dimensional proximal operator of Fantope projection.

For our ASPCA-ADMMSEQ method which will be introduced later, we have used deflated Fantope projection for sequential estimation of eigenvectors constraining the subspace rank to 1. In this case, we also need to guarantee orthogonality of estimated eigenvectors. Let $\widehat{U}_{j-1} \in R^{p \times p-j+1}$be an orthonormal complement basis of $\widehat{V}_{i-1}$, then the proximity operator of deflated Fantope is; see [4] for background,

$$P_{D_j}(X) = \widehat{U}_{j-1} \left[\sum_{i=1}^{p-j+1} \gamma_i^+(\theta) \eta_i \eta_i^T \right] \widehat{U}_{j-1}^T \tag{3-35}$$

Where,

I. (γ_i, η_i) are the eigen components of $\widehat{U}_{j-1}^T X \widehat{U}_{j-1}$

II. $\gamma_i^+(\theta) = min(max(\gamma_i - \theta, 0), 1)$,

III. And θ is chosen such that $\sum_{i=1}^{p-j+1} \gamma_i^+(\theta) = 1$

3.7.4.2 Soft Thresholding

Soft thresholding is a very popular tool in computer vision and machine learning. Essentially it permits to add l_1 penalties to take into account the subsequent objective:

$$\min_X ||X - b||_2^2 + \lambda||X||_1 \tag{3-36}$$

It can be solved in a very efficient manner using an approach known as Soft-thresholding. Since l_1 penalties are being used nearly everywhere, the property of soft thresholding to efficiently find the solution becomes very useful. Considering, $h(u) = \lambda||X||_1$, the minimization function of (3-38) becomes

$$\frac{1}{2}||u - X||_2^2 + \lambda||u||_1 \tag{3-37}$$

We want to find the minimum of the objective and for that we must solve the gradient of this function. It is worth to be noted that both terms in (3-39) are divisible in *u* and we can minimize each element of *u* individually. This simplifies the expression to:

$$\frac{1}{2}||u - X||_2^2 + \lambda||u_i||, \forall i \tag{3-38}$$

To find the derivative of $|u_i|$, we have two circumstances need to consider. Either $u_i > 0$ or $u_i < 0$ and for the first case the derivative is:

$$\begin{gathered} u_i - X_i + \lambda = 0 \\ u_i = X_i - \lambda \end{gathered} \tag{3-39}$$

Since, $u_i > 0$, this is only valid for $X_i > \lambda$. The same logic applies for other circumstance of $u_i < 0$, with which we get:

$$u_i = X_i + \lambda \tag{3-40}$$

And it is only valid for $X_i < -\lambda$. Now, for $-\lambda < X_i < \lambda$, $u_i = 0$ to make the derivative somewhere between −λ and λ.

Summing all up:

$$\begin{cases} X_i - \lambda & X_i > \lambda \\ X_i + \lambda & X_i < -\lambda \\ 0 & -\lambda < X_i < \lambda \end{cases}$$

This is equivalent to the expression:

$$u_i = prox_h(X) = max(|X_i| - \lambda, 0)sign(X_i) \tag{3-41}$$

This is known as the shrinkage operator and is used in iterative soft thresholding algorithms. We are also going to use this proximal operator in second update step of ADMM algorithm.

Chapter 4 ABNORMAL SUBSPACE ESTIMATION

4.1 Introduction

Now we need to formulate the objective function of the sparse abnormal subspace estimation problem. In literature, most of the existing studies not only compute significant sparse PCs sequentially only but also without orthogonal constraints [8]. Therefore, the resultant sparse PCs cannot be used to define the normal and abnormal subspaces, as the abnormal PCs are not the orthogonal complement of the normal PCs [2]. By enforcing orthogonality, sparse PCA is used to solve subspace estimation problem, which is denoted as backward ASPCA (shorted as ASPCA-BG) method [2]. To ensure sparsity in the abnormal loading vectors [2] introduced a global optimization step in the end. This study of ASPCA-BG framework finds the abnormal subspace by computing the abnormal PCs one by one (sequentially). But, to improve the sequential computation, our idea is to compute the whole abnormal subspace at a time reducing computation complexity proportional to dimension size.

4.2 Problem definition

Our motivation is to find a sparse abnormal subspace where outliers have high projection length. At the same time, we want to use the idea of PCA with eigenvalue decomposition of input's covariance matrix and extract the least significant PCs (basis of abnormal subspace). As a result it becomes solving a low rank matrix approximation problem using SDP formulation with convex relaxation for a tractable global solution. As we want behavior like eigenvectors, we also need orthogonal guarantee on computed PCs. It is worth to mention that, we want to get all the sparse PCs at the same time which actually never done before emphasizing least significant PCs covering less variance especially for outliers detection.

Given a covariance matrix $\boldsymbol{S}$ and a sparsity constraint constant $\boldsymbol{k}$, the ASPCA-BG [2] method tries to solve a problem to obtain last $\boldsymbol{d}$ abnormal loading vectors sequentially where $i \in 1, \dots,$ d and from (3-9) we know that:

$$\min_{v_i} v_i^T S v_i$$

$$s.t.\ v_i^T v_i = 1, v_j^T v_j = 0, \forall\, 1 \leq j < i, Card\,(v_i) \leq k$$

To solve the above equation, in (3-10) we already have demonstrated the SDP problem with convex relaxation

$$\min_{X_i \in S^P} Tr(SX_i) + \lambda\, ||X_i||_{1,1}$$

$$s.t. X_i \succcurlyeq 0\,, Tr(X_i) = 1, Tr(R_{i-1}X_i) = 0$$

These $\boldsymbol{d}$ loading vectors obtained by solving the above equation are used for detecting and interpreting anomalies. Afterwards, a global optimization step is introduced for better sparsity of loading vectors in the subspace. But, our goal is to compute the subspace simultaneously and we also want to skip the global optimization step of the method. Therefore, the target is to estimate a sparse abnormal subspace simultaneously rather than sequentially which is a solution of a semidefinite program in the variable X.

To illustrate the need in our case, the PCA problem with objective to capture the lowest variance, what we have in our hand is:

$$\min_{U \epsilon\, R^{p \times d}} Tr(U^T S U)$$

$$s.t. U^T U \;=\; I_d$$

Afterwards, we have converted the quadratic problem to a linear problem by considering, $X = \;UU^T$ making it positive semidefinite.

$$\min_{X,U}\; Tr(SX)$$

$$s.t. X \;=\; UU^T\,; U^T U \;=\; I_d$$

Or equivalently it can be formulated as,

$$\min_{X}\; Tr(SX)$$

$$s.t. X\ is\ a\ d - dimensional\ projection\ matrix$$

Because $Tr(SX)$ is linear in X, it is equivalent to

$$\min_{X}\; Tr(SX)$$

$$s.t. X\ \in\ Convhull(all\ d - dim\ projection\ matrices)$$

Incorporating the sparse condition in the objective function we get.

$$\min_{X} Tr(SX) + \lambda\, ||X||_{1,1}$$

$$s.t. X \epsilon\ Convhull(d - dimesnional\ projection\ matrices)$$

4.2.1 Simultaneous approach (ASPCA-ADMM)

Our proposed ASPCA-ADMM method estimates the abnormal subspace computing it simultaneously. Given a symmetric input matrix ***S***, we propose a sparse principal subspace estimator $\hat{X}$ that is defined to be a solution of the semidefinite program incorporating the facts mentioned in [3] and (3-11).

$$\min_{X} Tr(SX) + \lambda\, ||X||_{1,1}$$
$$s.t. X\epsilon\, F_{p_d}$$

In the variable X, where

$$F_{p_d} := \{\, X : \; 0 \preccurlyeq \; X \; \preccurlyeq \; I \; and \; Tr(X) = \; d\}$$

Is a convex body and it is the Fantope; see section 3.7.4.1 on pervious chapter for background. $\lambda \; \geq 0$ is a regularization parameter that ensures sparsity. The difficulty in directly solving the problem is the direct interaction between the l_1 norm and the Fantope constraint. Without either of the constraints, the optimization problem would be much easier. We are going to use ADMM to exploit the fact of convex functions and find tractable global solution to estimate ***d***-dimensional subspace in next section.

4.2.2 Sequential approach (ASPCA-ADMMSEQ)

This method estimates the abnormal subspace by computing the least significant principal components sequentially one by one. The computed loading vectors are sparse and orthogonal to each other eventually guarantees to be eigenvectors. First, we try to get the first least significant component by forcing rank 1 constraint on Fantope projection. So, the first abnormal PC is easy to compute and it is extracted solving the following optimization problem,

$$\min_{X_1} Tr(SX_1) + \lambda\, ||X_1||_{1,1}$$
$$s.t. X_1\epsilon\, F_{P_1} \tag{4-1}$$

Afterwards, we extract the leading eigenvector of this rank 1 estimator to compute the first abnormal PC or eigenvector. Now, we have the least significant PC or first abnormal PC from (4-1) and to ensure next ***d***-1 sequential least PCs we use the idea of deflated Fantope projection introduced in previous chapter. This process computes one dimensional subspace eventually revealing the least significant PC in each iterations which is orthogonal to previous computed PCs. To understand more clearly, let’s compare it with ASPCA-ADMM method where ***d*** = 1. ASPCA-ADMM computes ***d***-dimensional subspace at a time. But, our ASPCA-ADMMSEQ method computes 1-

dimensional subspace (rank 1). So, to get ***d***-dimensional subspace we need an iterative process which can compute all the required ***d*** least significant loading vectors one by one. But, here is the tricky part. It doesn't guarantee orthogonality constraint in computed PCs. While computing the ***d*** abnormal PCs, it is needed to assure the orthogonality constraint to be considered as eigenvectors. So, each time while computing new PC we should also guarantee it is orthogonal to all other previously computed PCs. So, to get next ***d***-1 Eigen vectors we used deflated Fantope projection with orthogonal guarantee in each iteration. If $j \in 2, \dots,$ d and $\hat{V}_{j-1}$ contains all previously computed eigenvectors then Deflated Fantope projection $DF_{p_j}$for each j is,

$$DF_{p_j} \ = \{ X_j \in F_{P_1};\ X_j\hat{V}_{j-1} \ = \ 0\} \qquad (4\text{-}2)$$

Eventually the objective problem for the iterations of $j \in 2, \dots,$ d is,

$$\min_{X_j} Tr(SX_j) + \lambda \ ||X_j||_{1,1}$$
$$s.t. X_j \in DF_{p_j} \qquad (4\text{-}3)$$

In this sub-process we will get rest of the abnormal PCs which will later form the abnormal subspace. So, this sequential method has two major steps. In first step, we compute the first abnormal PC using Fantope projection. In second step, we use the deflation technique to compute rest of the abnormal PCs. In other words, we are using the deflated Fantope projection by forcing rank 1 projection and compute PCs sequentially one by one. We will use ADMM in each iteration of computing PCs sequentially. This version of ADMM implementation is almost same as the simultaneous version except some minor changes. The algorithm of ASPCA-ADMM and ASPCA-ADMMSEQ are mentioned in following section.

4.3 ADMM implementation of Simultaneous approach

As we mentioned earlier the principal hurdle in directly solving (3-11) is the direct presence of the penalty and the Fantope constraint at the same time. Without either of these requirements, the optimization problem would be easier to solve. Interestingly, ADMM can exploit this fact easily if we first rewrite the objective function as the equivalent equality constrained problem divided in two variables.

$$\min_{X} \infty. 1_{F_{p_d}}(X) + Tr(SX) + \ \lambda \left||Y|\right|_{1,1}$$
$$s.t. X - Y = 0 \qquad (4\text{-}4)$$

4.3.1 ASCPA-ADMM Algorithm

We have rewritten the equivalent equality constrained problem in the variables X and Y, where $1_{F_{p_d}}$ is the 0-1 Indicator Function for F_{p_d} and we assumed the convention $\infty . 0 = 0$. The augmented Lagrangian associated with it has the form

$$\begin{aligned} L\rho(X,Y,U) := \infty \cdot 1_{F_{p_d}}(X) + Tr(SX) + \lambda ||Y||_{1,1} \\ + \frac{\rho}{2} (||X - Y - U||_2^2 + ||U||_2^2) \end{aligned} \tag{4-5}$$

Where $U = (1/\rho)Z$ is the scaled ADMM dual variable and ρ is the ADMM penalty parameter; see [7] for background. ADMM consists of iteratively minimizing $L\rho$ with respect to X in the x-update step, minimizing Lρ with respect to Y, and then updating the dual variable.

Here we are trying to minimize the linear objective function of X over the intersection of two convex sets. First, the Fantope and secondly the l_1 ball. First step can be defined as Fantope Projection,

$$P_{F^d}\left(Y - U - {}^{S}/_{\rho}\right) = \min_{X \in F^d} \frac{1}{2} \left\|X - (Y - U - {}^{S}/_{\rho})\right\|_2^2 \tag{4-6}$$

The second step is called the shrinkage step and can be defined as Soft Thresholding using the idea of (3-41),

$$\begin{aligned} S_{\lambda/\rho}(X + U) = \min_{Y} \frac{\lambda}{\rho} \left||Y|\right|_{1,1} + \frac{1}{2} \|(X + U) - Y)\|_2^2 \\ = sign\,(X + U)\ \max(|X + U| - \lambda/\rho\ , 0) \end{aligned} \tag{4-7}$$

Our method solves two sub-problems in each iteration. One sub-problem has a closed-form solution that corresponds to projecting a given matrix onto the Fantope. This projection requires spectral decomposition of input variable with abnormal dimension space constraint. The other problem has a closed form solution that corresponds to a vector shrinkage operation onto the l_1 ball. Thus, our method produces two iterative points at each iteration. One iterative point is a semidefinite matrix with trace equal to the abnormal space dimension and the other one ensures a sparse matrix. Eventually these two points will converge to some point, and thus we get an optimal solution which is a sparse and semidefinite.

Algorithm 1 ASPCA-ADMM

```
Input: D, d , λ, ρ , ϵ
S ←covariance(D)                                                        ◃ Covariance matrix
$Y^{(0)} \leftarrow 0$, $U^{(0)} \leftarrow 0$                          ◃ Initialization
repeat t = 0, 1, 2, 3, . . .
        $X^{(t+1)} \leftarrow P_{F^d}(Y^{(t)} - U^{(t)} - S/\rho)$      ◃ Fantope projection
        $Y^{(t+1)} \leftarrow S_{\lambda/\rho}(X^{(t+1)} + U^{(t)})$    ◃ Soft-thresholding
        $U^{(t+1)} \leftarrow U^{(t)} + X^{(t+1)} - Y^{(t+1)}$          ◃ Dual update
until $\max(|||X^{(t)} - Y^{(t)}|||_2^2, \rho^2|||Y^{(t)} - Y^{(t-1)}|||_2^2) \leq d\epsilon^2$      ◃Stopping criterion
return $Y^{(t)}$
```

In light of the separation of X and Y in (4-4) and some analytical manipulation, the X and Y updates are reduced to computing the proximal operators; see 3.7.4 for background.

If, $X = \sum_i \gamma_i u_i u_i^T$is a spectral decomposition of X, then Fantope Projection can be computed by proximal operator and using the knowledge from (3-34),

$$P_{F^d} = \sum_i \gamma_i^+(\theta) u_i u_i^T$$

Where $\gamma_i^+(\theta) = min(max(\gamma_i - \theta, 0), 1)$ and θ satisfies the equation $\sum_i \gamma_i^+(\theta) = d$.

Thus, P_{F^d}(X) involves computing an Eigen decomposition of X, and then modifying the eigenvalues by solving a monotone, piecewise linear equation.

We try to keep the primal and dual residual norms (the two sum of squares in the stopping criterion of Algorithm 1) within a constant factor, ϵ of each other. For the stopping criteria of ADMM, we consider both the primal and dual residuals. Note that in our problem, the primal residual at iteration k is measured by

$$r^t := X^t - Y^t$$

And we chose to terminate our ADMM when

$$max(|||X^{(t)} - Y^{(t)}|||_2^2, \rho^2||| Y^{(t)} - Y^{(t-1)}|||_2^2) \leq d\epsilon^2$$

4.3.2 Abnormal Subspace representation

The constraint of our abnormal subspace estimator $\hat{X} \in F_{p_d}$ guarantees that its rank is at least ***d***. However, the solution need not be an extremal point of F_{p_d} (rank-d projection matrix). As a result to obtain a proper ***d***-dimensional abnormal subspace, we can extract the ***d*** leading eigenvectors of the estimator and denote it as $\hat{V}$, and form the projection matrix $\hat{\Pi} = \hat{V}\hat{V}^T$. The projection is unique, but the choice of basis is subjective. Therefore, we can then pick the top ***d*** eigenvectors of the projection matrix and represent those loading vectors as the basis as the subspace.

When we extract ***d*** loading vectors $(v_1, .., v_d)$ to span a subspace $V_{abnormal}$, we make sure that the orthogonal complement of $V_{abnormal}$ has major variance for describing the normal patterns in the dataset, so that $V_{abnormal}$ is the abnormal subspace and the resultant ***d*** sparse principal components can be used to detect and interpret anomalies.

4.4 ADMM implementation of Sequential approach

ASCPA-ADMMSEQ method mainly consists of two steps. First, we compute the least significant component by rank 1 constraint on Fantope projection and the objective function is,

$$\begin{gathered}\min_X \infty . 1_{F_{p_1}}(X) + Tr(SX) + \lambda \left||Y|\right|_{1,1} \\ s.t. X - Y = 0\end{gathered} \tag{4-8}$$

Now, to estimate next sequential ***d***-1 Eigen vectors we used deflated Fantope projection as a proximal operator in each iteration for j ϵ 2,…,d and the objective functions we have to optimize in each iteration is,

$$\begin{gathered}\min_{X_j} \infty . 1_{DF_{p_j}}\left(X_j\right) + Tr(SX_j) + \lambda \left||Y|\right|_{1,1} \\ s.t. X_j - Y = 0\end{gathered} \tag{4-9}$$

4.4.1 ASPCA-ADMMSEQ Algorithm

The augmented Lagrangian associated with (4-8) has the form to get first abnormal PC is,

$$\begin{aligned}L\rho(X, Y, U) := \infty \cdot\ & 1_{F_{p_1}}(X) + Tr(SX) + \lambda \parallel Y \parallel_{1,1} \\ & + \frac{\rho}{2}\ (|X - Y - U|||_2^2 + |||\ U|||_2^2)\end{aligned} \tag{4-10}$$

We can compute the Fantope projection like (3-36) by setting the value of ***d*** to 1.

$$P_{F^1} = \sum_i \gamma_i^+(\theta) u_i u_i^T$$

Where $\gamma_i^+(\theta) = min(max(\gamma_i - \theta, 0), 1)$ and θ satisfies the equation $\sum_i \gamma_i^+(\theta) = 1$.

To get next ***d***-1 Eigen vectors we compute the deflated Fantope projection in each iteration where, $j \in 2, \dots, \mathrm{d}$.

$$L\rho(X,Y,U) := \infty \cdot \; 1DF_{p_j}(X_j) + Tr(SX_j) + \; \lambda \parallel Y \parallel_{1,1} \\ + \frac{\rho}{2} \; (|X_j - Y - U|||_2^2 + \; ||| \, U|||_2^2) \tag{4-11}$$

Using the idea of (3-40), Let $\widehat{U}_{j-1} \in \; R^{p \times p-j+1}$be an orthonormal complement basis of $\widehat{V}_{i-1}$, then

$$P_{D_j}(X) = \; \widehat{U}_{j-1} \left[\sum_{i=1}^{p-j+1} \gamma_i^+(\theta) \eta_i \eta_i^T \right] \widehat{U}_{j-1}^T$$

Where, (γ_i, η_i) are the Eigen components of $\widehat{U}_{j-1}^T X \widehat{U}_{j-1}$

$$\gamma_i^+(\theta) = min(max(\gamma_i - \; \theta, 0), 1) \;,$$

And θ is chosen such that $\sum_{i=1}^{p-j+1} \gamma_i^+(\theta) = 1$

And the second update step is defined as Soft Thresholding like first method,

$$S_{\lambda/\rho}(X + U) = \; sign \; (X + U) \; \max(|X + U| - \frac{\lambda}{\rho} \;, 0)$$

Thus, $P_{D_j}(X)$ involves computing an eigenvector decomposition of X, and then modifying the eigenvalues by solving a monotone, piecewise linear equation by setting the dimension size to 1 (rank 1).

We try to keep the primal and dual residual norms (the two sum of squares in the stopping criterion of Algorithm 2) within a constant factor of each other like Algorithm 1. For the stopping criteria of ADMM, we again considered both the primal and dual residuals and the primal residual at iteration k is measured by in each outer iteration,

$$r^t := \; X^t - Y^t$$

And we chose to terminate our ADMM in each iteration when

$$max(|||X^{(t)} - \; Y^{(t)}|||_2^2, \rho^2 ||| \, Y^{(t)} - \; Y^{(t-1)}|||_2^2) \; \leq \; d\epsilon^2$$

Our Sequential method has another outer iteration. One iteration is for sequential computation of PCs and the inner one is for ADMM convergence. So, the outer iteration terminates when we finish computing ***d*** least significant abnormal PCs to represent the abnormal subspace.

Algorithm 2 ASPCA-ADMMSEQ

```
Input: D, d, $\lambda$, $\rho$, $\epsilon$ , j = d
S = covariance (D)                                                        ◃ Compute covariance matrix
$P = 0$                                                                   ◃ Rank-d Projection Matrix
$Y_1^{(0)} \leftarrow 0, U_1^{(0)} \leftarrow 0$                          ◃ Initialization
repeat $t = 0, 1, 2, \ldots .$                                            ◃ First Eigen Vector
    $X_1^{(t+1)} \leftarrow P_{F_1}(Y_1^{(t)} - U_1^{(t)} - S/\rho)$      ◃ Rank-1 Fantope projection
    $Y_1^{(t+1)} \leftarrow S_{\lambda/\rho}(X_1^{(t+1)} + U_1^{(t)})$    ◃ Soft-thresholding
    $U_1^{(t+1)} \leftarrow U_1^{(t)} + X_1^{(t+1)} - Y_1^{(t+1)}$        ◃ Dual update
until $\max(|||X_1^{(t)} - Y_1^{(t)}|||_2^2, \rho^2 |||Y_1^{(t)} - Y_1^{(t-1)}|||_2^2) \leq d\epsilon^2$    ◃Stopping criterion
$P = P \cup Y_1^{(t)}$                                                    ◃ Add the first eigenvector

for $j = 2, \ldots, d$                                                    ◃ compute d-1 eigenvectors
    $Y_j^{(0)} \leftarrow 0, U_j^{(0)} \leftarrow 0$                      ◃ Initialization
    repeat $t = 0, 1, 2, \ldots$
        $X_j^{(t+1)} \leftarrow P_{DF_j}(Y_j^{(t)} - U_j^{(t)} - S/\rho)$ ◃ Deflated Fantope projection
        $Y_j^{(t+1)} \leftarrow S_{\lambda/\rho}(X_j^{(t+1)} + U_j^{(t)})$ ◃ Soft-thresholding
        $U_j^{(t+1)} \leftarrow U_j^{(t)} + X_j^{(t+1)} - Y_j^{(t+1)}$    ◃ Dual update
    until $\max(|||X_j^{(t)} - Y_j^{(t)}|||_2^2, \rho^2 |||Y_j^{(t)} - Y_j^{(t-1)}|||_2^2) \leq d\epsilon^2$    ◃Stopping criterion
    $P = P \cup Y_j^{(t)}$                                                ◃ Add the j^th eigenvector
end
return $P$
```

4.4.2 Abnormal Subspace representation

The rank 1 constraint guarantees that in each iteration we one dimensional subspace. Now, in order to obtain the abnormal PC, we can extract the leading eigenvector of our estimated projection for each iteration of the algorithm.

In the ASPCA-ADMMSEQ framework we have enforced orthogonality on the resultant each sparse abnormal PCs. As a result, the resultant sparse PCs can be used as a basis to define the abnormal subspace, as the abnormal PCs are orthogonal and unique to each other. Let, $V_{abnormal} = (v_1, .., v_d)$ is the ***d***-dimensional abnormal subspace with least significant ***d*** principal components. ASPCA-ADMMSEQ method computes the least significant eigenvectors which captures least variance PC in first iteration using Fantope proximity operator and finds all the rest $\mathrm{d} - 1$ abnormal PCs sequentially in increasing direction of variance which covered less using deflated Fantope proximity operator. Later, these sparse loading vectors can be used directly to identify the outliers and interpret the reasons of being abnormal.

4.5 Discussion

Our ADMM method solves SDP relaxation of the sparse PCA problem to estimate the subspace which represents abnormal data. This method eventually incorporates variable splitting technique to separate the l_1 norm constraint, which controls the sparsity of the abnormal loading vectors, and the Fantope projection which constraints close form projection. This iterative method resulted in two sub-problems that have closed form solutions in each iteration.

We also proposed another method using ADMM where sequentially obtained eigenvectors have guaranteed orthogonality and are allowed to form a subspace which represents the abnormal data. Our intuition behind this method is that, sequential computation of loading vectors might have better interpretation than simultaneous abnormal loading vectors extraction. We are going to have a comprehensive analysis of both of this methods in experimental analysis chapter with proper real world and synthetic examples.

Now, we have the abnormal subspace which represents the anomalies in the dataset. Both of our methods give us the subspace and the basis of it. The next step is to find the explanation of this detection and there we will use the extracted abnormal loading vectors to interpret each of the individual anomalies present in our dataset.

Chapter 5 ANOMALY DETECTION & INTERPRETATION

5.1 Introduction

At this moment, we have already estimated the abnormal subspace from the input data points using one of our two proposed methods. We also have the abnormal PCs in our hand which we assume till now represents the sparse and orthogonal loading vectors. To be specific, abnormal PCs are the least significant eigenvectors which should explain the outliers. These abnormal PCs are none other than the basis of the abnormal subspace. This basis set is orthonormal in its nature. Now, it is the time to demonstrate whether our estimated abnormal subspace can detect anomaly and interpret it. In other sense whether this subspace can represent the abnormal data of our input dataset.

5.2 Anomaly Detection

In this step, we want to detect the outliers from the dataset and in our hand we have normal and abnormal subspace. Again we assume that, $V_{abnormal} = (v_1, .., v_d)$ is the ***d***-dimensional abnormal subspace with least significant ***d*** principal components and $V_{normal} = (v_{d+1}, v_2, v_3, .., v_p)$ is the normal subspace defined by the significant ***p*** - ***d*** principal components which covers maximum variance.

Now, our goal to identify anomalies using the abnormal principal components present in $V_{abnormal}$ subspace. The plan is to project all data into these subspaces and measure the SPE value. Our heuristic is that, normal data would have higher projection on normal subspace and abnormal data would have lower projection. In other words, abnormal data should have higher projection on abnormal subspace. We know that the normal subspace explains the maximum variance of the dataset. PCA based detection methods adopt that this subspace represents the major trends of the dataset, and all normal data likely to have almost zero length projection on the abnormal subspace. Therefore, given a ***p***-dimensional feature space, our model detects outliers using the projection length of data points in subspace.

So, for a ***p***-dimensional features of a point y, its residual $\hat{y}$ is defined in (3-2) as:

$$\hat{y} = y - V_{normal}V_{normal}^{T}y$$

This is the projection residual length on normal subspace. The squared length of projection residual length $\hat{y}$, called as the Squared Prediction Error (SPE), is the metric to indicate whether y is an anomaly or not. The higher the SPE, it is more likely that data instance is an outlier. Subsequently, we also need a perfect threshold value to ensure the most detection of true abnormal data.

5.3 Anomaly Interpretation

At this moment, we have identified the abnormal data from the dataset. Finally, the next crucial and important step remaining is to interpret the reason why they are outliers. When the SPE score of a given instance y is over a predefined threshold, y is considered as an anomaly. Afterwards, it is important to understand where the abnormality of y comes from, i.e., what anomalous feature behaviors of y are more responsible for distinguishing y from normal data and how we define it as Anomaly Interpretation. The anomaly interpretation for PCA-based model is difficult, as there is no direct mapping between PCA's dimension reduced subspace and the original feature space for anomaly [10]. Therefore, the length of $\hat{y}$ can be used to detect anomaly. But still, interpreting $\hat{y}$ directly is meaningless because it only involves the subspace of normal data.

Now, given the normal subspace V_{normal} and abnormal $V_{abnormal}$, from (3-3) we know that

$$\hat{y} = y - V_{normal}V_{normal}^{T}y = V_{abnormal}V_{abnormal}^{T}y$$

And from (3-4) we know that,

$$SPE = \hat{y}^{T}\hat{y} = \sum_{i=1}^{d}(v_i^{T}y)^2$$

In other words, SPE is equal to the square sum of y's scalar projection on each abnormal PCs, so that we can identify the set of PCs that are responsible for the abnormality indicated by high projection values. At this point, these PCs are easy to interpret, since each abnormal PC is as simple as a linear combination of important feature variables responsible for abnormality. The reason behind it is the sparse nature of loading vectors which is achieved by estimating a sparse subspace.

Hence, we managed to extract PCs with sparse and orthogonal loading vectors to represent abnormal subspace, these loading vectors can be used directly to detect and interpret anomalies. Here the orthogonality promises the abnormality can be translated to projection values on a set of abnormal PCs eventually interpreated as eigenvectors and the sparsity ensures the interpretation ability of outliers.

Summing up, the interpretation of a detected anomaly is conducted in two steps. First, we identify the set of projections on abnormal PCs that contribute the most for a high SPE score. Then, we interpret these projections one by one, by identifying which original feature variables are responsible for each projection, and how each projection triggers a high SPE score.

5.4 Discussion

Interpretation of anomalies solely depends on sparsity of loading vectors and explained variance. We are going to plot heatmap of data points on computed abnormal PCs in next chapter and we can visualize the individual anomalies that exists in the dataset. It is important to understand the interpretation method and how we are interpreting using this PCs. One can think of it as observing the heatmap of data points on abnormal PCs. Normal data usually should not have any projection on this heatmap. And anomalies should have a clear heatmap on respective PCs depending on their type of anomaly. In chapter 6 reader can have a clear idea of methodology with further example.

Abnormal subspace dimension size, ***d*** is a vital parameter for detecting anomaly using this method. We should have a perfect ***d*** to get a pure subspace representation of anomalies. Otherwise, the estimated subspace would be polluted by normal data and eventually it might represent false positive anomalies. On the other hand, the ratio of normal and abnormal data is also important. We usually assume that anomalies are very less in number compared to majority trend of data. In other words, these anomalies are small subset of original input dataset. So, the domain space we are analyzing need to have a perfect ratio of anomalies to detect them.

It is also important to be noted that we need to have a clear understanding of correlation between input variables or feature space to interpret anomalies. Most of the time, the reason of anomaly is dependent on more than one feature variables. So, we need to represent this correlation of variables with perfect weight vectors respectively.

But, to make the loading vectors sparser, sometimes this dependency gets disappeared. Due to this reason, we need to keep the explained variance as low as possible and sparsity at a maximum level. In our empirical observations, we identified that sparsity increases variance and hence confirming the trade-off relationship between sparsity and variance.

Nonetheless, PCA plays a vital role as a baseline method to keep track of this trade-off. We need to keep the variance as close as possible to PCA. Otherwise, our model might represent false positive anomalies represented by the abnormal PCs of our estimated abnormal subspace and moreover, we cannot make any fair comparison.

Chapter 6 EXPERIMENTAL RESULTS

6.1 Evaluation metrics

There are different benchmark criterion to evaluate computation performance and sparsity result of an anomaly detection model (ex. Accuracy, F1, Recall, ROC curve etc.). In our model, we evaluated our Anomaly Detection performance mainly by (ROC & AUC curve) and Sparsity is evaluated by entry wise norm 1($||V||_{1, 1}$) of loading vectors, Card$_{0.1}$ (number of entries with absolute value higher than 0.1), and Card$_{0.01}$ (number of entries with absolute values higher than 0.01). Variance is also another metric to identify how much data the loading vectors explains. Our main concern is to identify outliers and therefore we tried different SPE threshold values for different dataset to ensure higher accuracy. We also considered Consumed Time as a metric to compute the abnormal subspace to evaluate the performance of our two proposed methods. Another assessment on clustering is also performed, where tried to demonstrate that our proposed method can also be used as a clustering algorithm.

This comprehensive comparative analysis includes four methods of anomaly detection taking into account two of our proposed ASPCA-ADMM and ASPCA-ADMMSEQ. PCA is used as a baseline method to evaluate performance and we have also included the ASPCA-BG method in comparative analysis as it is closely related to our proposed method. Our methods eventually demonstrate extensive improved performance in terms of Accuracy and Computation complexity. The performance of our proposed two methods are almost same. But, ASPCA-ADMMSEQ method has better interpretation result and finally demonstrated its result to explain working procedure. We used Matlab programming and performed all experiments on a laptop Computer with 8GB memory with an Intel (R) Core(TM) i5-4200U CPU 2.30 GHz.

6.2 Datasets

We tried to evaluate our method ranging from small to mediocre dataset which includes both synthetic and well-known real world for anomaly detection. We have used the synthetic dataset mentioned in [2] to test and compare the performance of our proposed methods with other methods in outlier detection field. To evaluate the performance on real word dataset, we have used Breast Cancer and dataset of KDD'99. And last of all the performance of clustering is assessed using IRIS dataset.

6.2.1 Synthetic Dataset

We have used the synthesized dataset of [2], which contains 500 normal records and 15 anomalies. So, the dimension of the sample is 515. The feature dimension size is 7 named from A to G, and the normal records were generated using four patterns, A ≈ B, D ≈ C + A, F ≈ 0, and G ≈ 0. The anomalies were generated by breaking the first three patterns. Therefore, there are three kinds of anomaly and each type contains five abnormal data. The code to generate the dataset can be found in the Appendix.

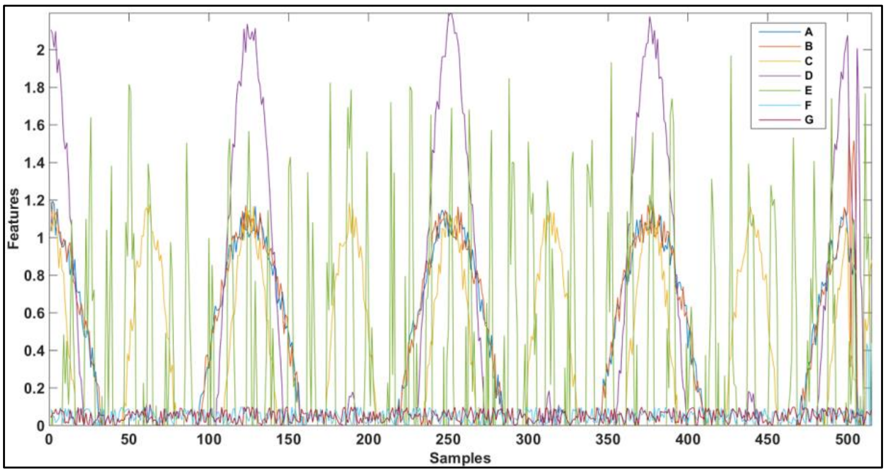


Figure 6-1 Instances and Features of Synthetic Data

Projecting the data on first two principal components, we can see there is no meaningful way to distinguish abnormal one from the normal one. Moreover, first two PCs almost cover 80% variance of the dataset.

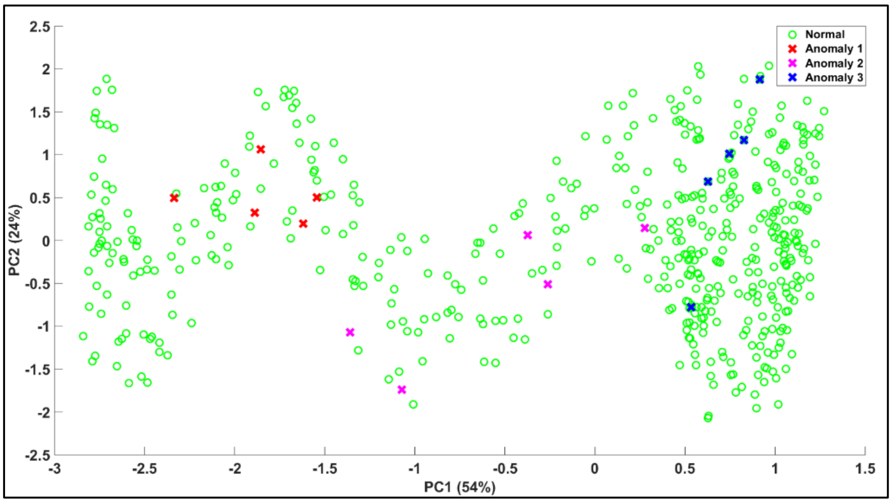


Figure 6-2 Projection of Synthetic Instances on leading 2 PCs

6.2.2 Breast Cancer Dataset

Breast Cancer Wisconsin (Diagnostic) Data Set provides features to distinguish between malignant and benign tumors [11]. The features demonstrate characteristics of the cell nuclei present in a digitized image of a fine needle aspirate (FNA) of a breast mass [2]. We have used the dataset from [2], which is prepared by keeping 357 benign records, and included the rest 10 malignant records to transform the classification task to an anomaly detection task. Projecting data on first two PCs we can see that about 60% variance of data is covered by the PCs.

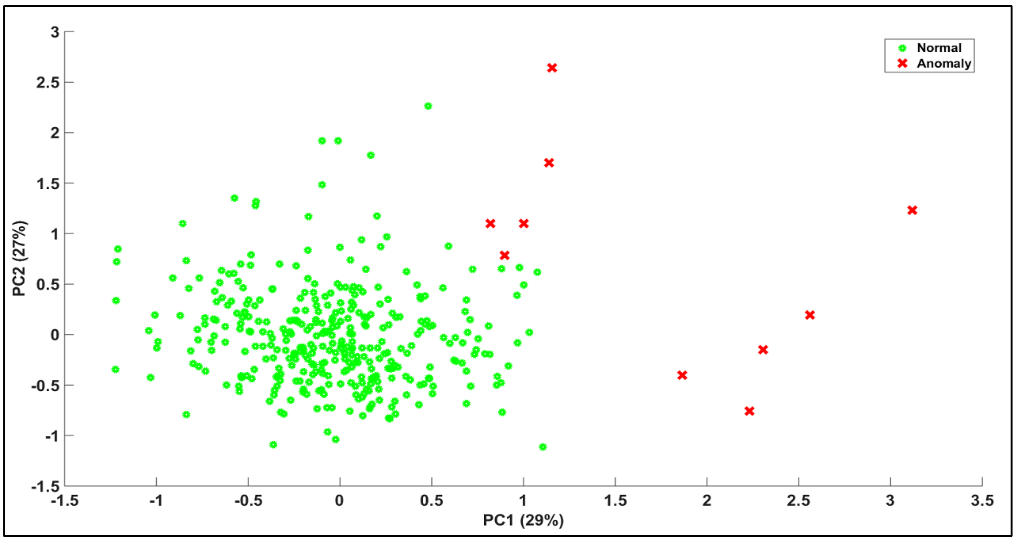


Figure 6-3 Projection of Breast Cancer Instances on leading 2 PCs

There are many cells of a breast mass and the features are three important measures (mean, standard error, and worst value) on 10 features of each cell: radius, texture, perimeter, area, smoothness, compactness, concavity, concave-points, symmetry and fractal-dimension. All 30 real-valued features were subtracted by the mean values and normalized to [-1, 1].

6.2.3 KDD’99 Network Intrusion Dataset

KDD’99 Network Intrusion Dataset is a widely used data for anomaly and intrusion detection [12]. It’s a very popular real-world dataset which is mostly considered as benchmark for outlier detection. We have only used the 10% KDD’99 dataset which contains about 103,000 data instance. Among these connection records, about 5% are attacks or abnormal records. There are different kinds of attacks which are also further grouped together depending on their behavior. We want to detect each of these anomalies individually and interpret their respective reason of detection.

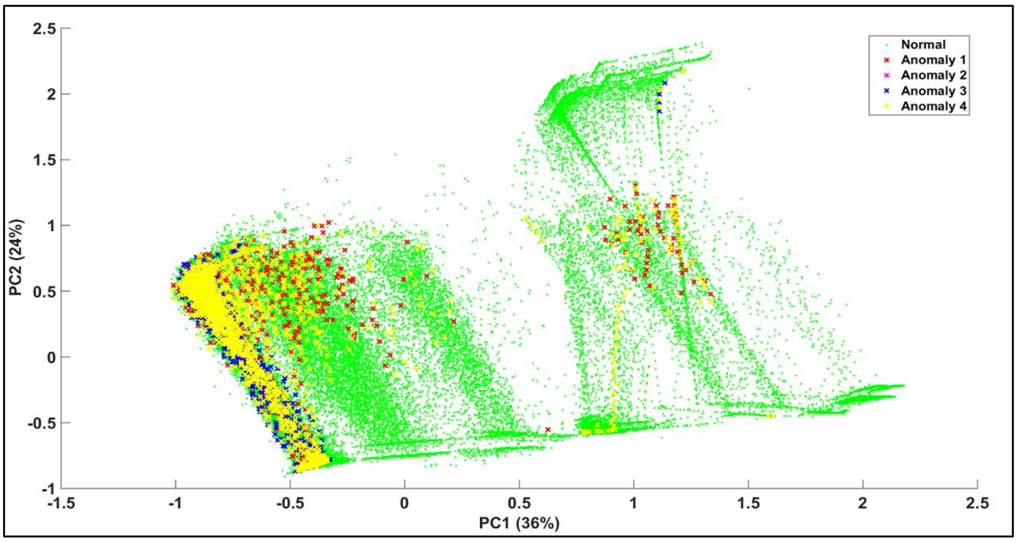


Figure 6-4 Projection of KDD'99 Instances on leading 2 PCs

We followed a similar preprocessing procedure as described in [2] to prepare the dataset for outlier analysis. In this dataset, each data instance is a connection record classified as normal or one of 22 classes of attacks (abnormal). Attacks can be classified into four main groups: DoS, Remote- to-local (R2L), User-to-root (U2R), and Probe. As mentioned in [2], all normal and part of the abnormal records in 10% KDD'99 datasets were picked from the rest 500 records on smurf, neptune, back, teardrop, satan, ipsweep, and portsweep and all records on other attacking types. There are in total 41 feature variables including seven categorical and those were mapped into distinct positive integers ranging from 0 to $state\ -1$ where $state$ is the number of states involved in the categorical feature. For example, *0* to *2* in protocol_type feature defined for TCP, UDP, and ICMP. Logarithmic scaling was also introduced on duration, src_bytes, and dst_bytes feature. All feature values were subtracted by the mean values and normalized to [-1, 1].

6.2.4 IRIS Dataset

This dataset consists of 3 different types of irises' flower (Setosa, Versicolour, and Virginica) and the feature space represents the petal and sepal length of the flower [28]. The size of the data matrix is $150\ \times\ 4$. The rows being the 150 samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width respectively. We again subtracted the mean value of the features from the original one and normalized to [-1, 1] to have a fair comparison with PCA method.

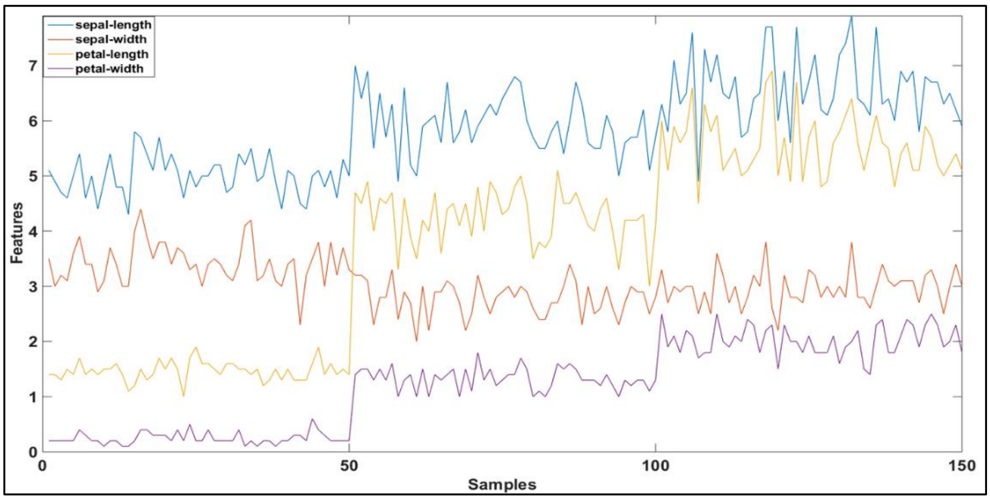


Figure 6-5 Instances of IRIS data with Features

We will try to demonstrate clustering performance of our model using this dataset. In these dataset there are three kinds of flower and we will try to generate cluster of similar kind of flower on each individual clusters. So, we converted our anomaly detection model to clustering algorithm demonstrating this dataset.

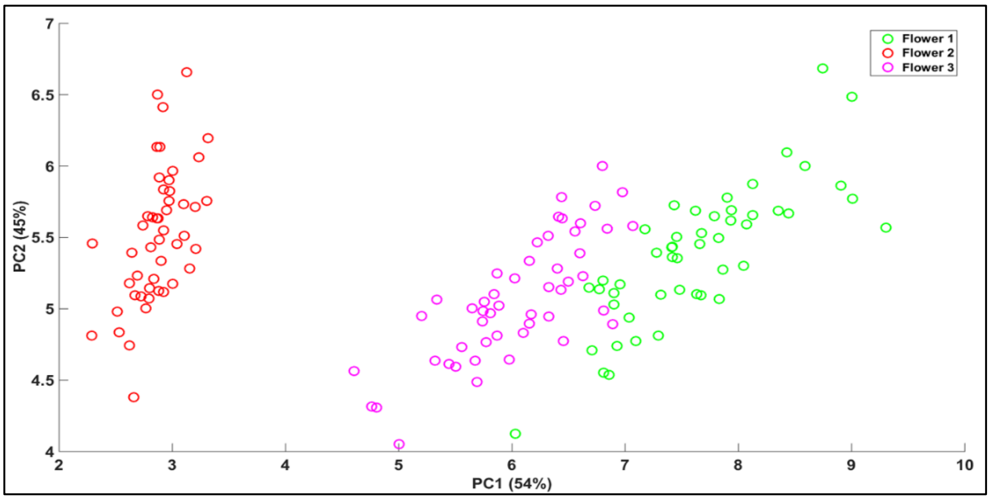


Figure 6-6 Projection of IRIS Instances on leading 2 PCs

6.3 Detection evaluation

We have used Receiver Operating Characteristic (ROC) curve to represent the anomaly detection performance for different dataset. It is a plot of the true positive rate against the false positive rate for the different threshold settings. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test result is. We also plotted the Squared Prediction Error (SPE) of each data points and visually demonstrated the projection length on abnormal subspace to get a clear visualization of anomaly subspace.

6.3.1 Synthetic Dataset

For the synthetic dataset, all of the methods obtained perfect result to detect abnormal data. So, our proposed two methods ADMM and ADMMSEQ obtains perfect ROC curve like all other methods shown below. The abnormal subspace dimension size used here is 4. That means, we have used 4 abnormal PCs to represent the abnormal subspace estimated by our proposed two methods as like ASPCA-BG method.

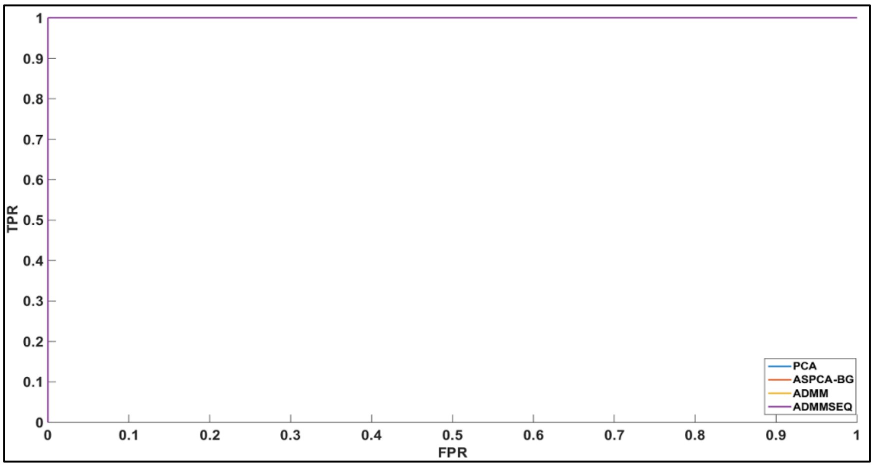


Figure 6-7 ROC Curve for Synthetic Dataset

We projected all the data on abnormal subspace for better understanding of abnormal subspace. From (Figure 6-8) we can see that, normal records (green bars (1:500)) has very low projection on abnormal subspace. Whereas all abnormal data has higher projection value and we can easily identify the outliers from the dataset. The three outliers in our synthetic dataset have more or less different projection length on abnormal subspace. We have used the estimated subspace of ASPCA-ADMMSEQ method to represent the SPE values of each of the records. ASPCA-ADMM also has very similar projection length on abnormal subspace for this synthetic dataset.

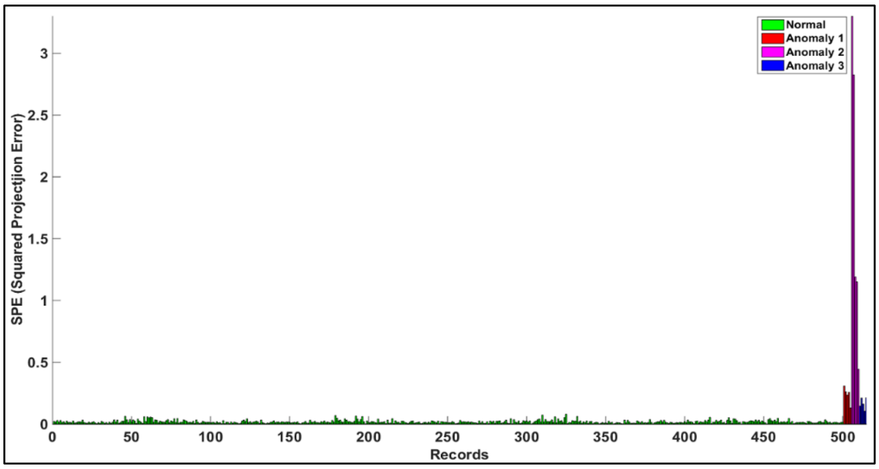


Figure 6-8 SPE values of Synthetic Data instances on Abnormal Subspace

6.3.2 Breast Cancer Dataset

Our proposed ASPCA-ADMMSEQ method achieved best ROC curve among all other methods for breast cancer dataset to detect malignant tumors. It just outperforms all other concurrent methods. It even achieved better performance than classical PCA. The feature dimension size of this dataset is 30. We have used 10 abnormal PCs to represent our estimated abnormal subspace. That means rest 20 PCs represents the normal subspace as these are complement of abnormal PCs from the abnormal subspace.

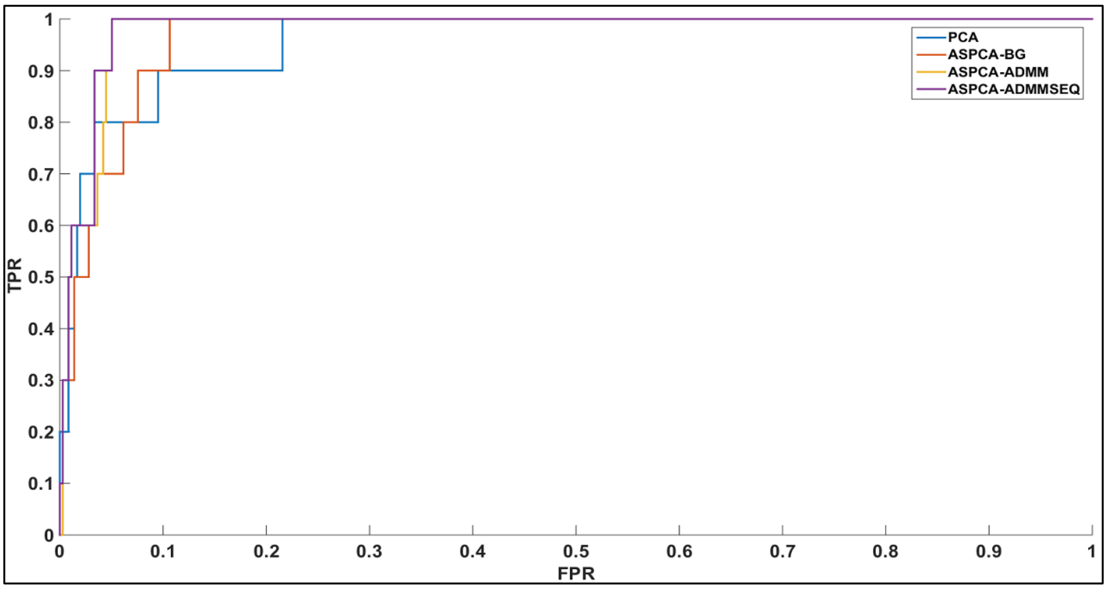


Figure 6-9 ROC Curve for Breast Cancer Dataset

The AUC score of the methods mentioned below in the table. All of the methods have better detection performance than PCA method but ASPCA-ADMM has the best AUC. ASPCA-ADMM has also good detection rate for Breast Cancer Dataset. But, it should be worth to be noted that AUC is dependent on the dimension size of abnormal subspace.

Method	AUC
PCA	0.958
ASPCA-BG	0.965
ASPCA-ADMM	0.978
ASPCA-ADMMSEQ	0.981

Table 6-1 AUC values for Breast Cancer Dataset

We again projected all of the malignant and benign breast cancer data on abnormal subspace for better visualization of projection length. From (Figure 6-10) we can see that, almost all of the normal (benign) records (green bars (1:357)) has very low projection length on abnormal subspace. Whereas all abnormal data has higher projection value and we can easily identify the outliers from the dataset projecting it on our estimated abnormal subspace. ASPCA-ADMM method also shows almost similar behavior if we project SPE of each data point.

We can easily see that our ASPCA-ADMMSEQ model has detected some false positive anomalies. These false positive anomaly data instance has higher projection length on abnormal subspace. There might be several reasons for this detection. These data might be noise or misinterpreted as positive in labeled dataset. Another reason might be, we should include better feature in input feature space to identify the dissimilarities and detect properly. Again, the size of the abnormal subspace also matters to avoid false positive data instances. As we know, an ideal abnormal subspace should not represent any normal trend of data, we should minimize it by choosing a proper value of ***d***.

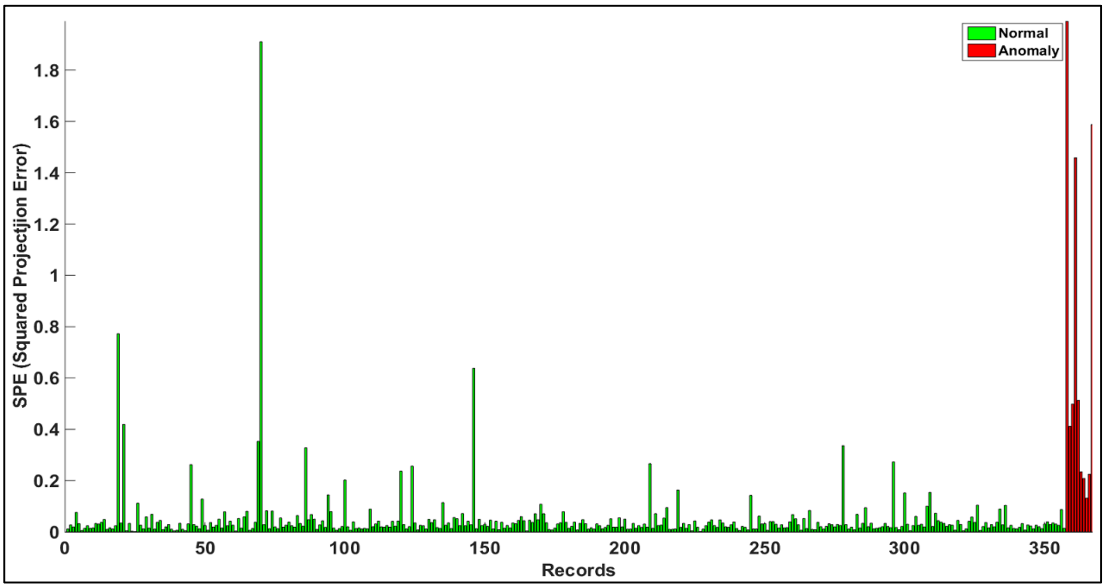


Figure 6-10 SPE values of Breast Cancer Data instances on Abnormal Subspace

6.3.3 KDD'99 Dataset

Our proposed ASPCA-ADMMSEQ method again achieved best ROC curve for KDD'99 Network Intrusion dataset. It can be seen that the following figure that ROC curve of ASPCA-ADMMDSEQ follows closer to the left-hand border and the top border of the ROC space. The detection performance of ASPCA-ADMM is slightly lower than previous datasets. If we do not require a sparse subspace, the AUC is close to PCA method. But, due to achieve it, detection performance slightly dropped.

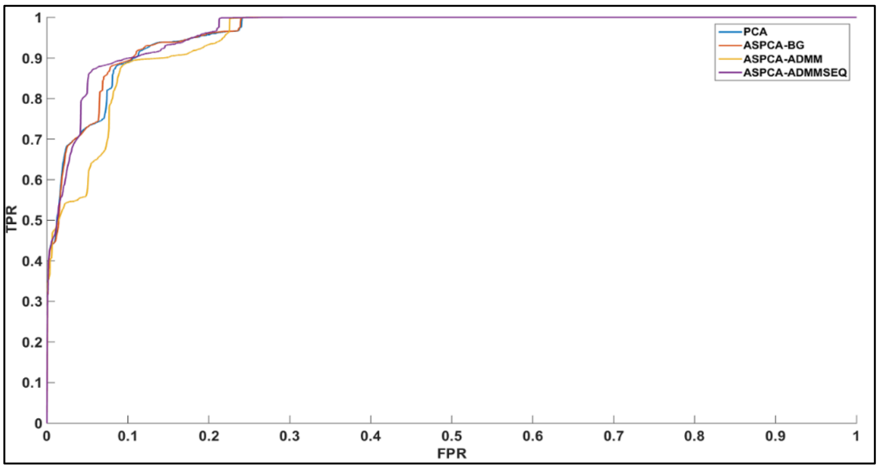


Figure 6-11 ROC Curve for KDD'99 Dataset

The AUC score of each of the methods mentioned below in the table. The abnormal subspace dimension is 35. That means, only 6 significant PCs represent the normal subspace and projects the major trends of the dataset. The detection performance sequence is different as the performance of ASPCA-ADMM falls down.

Method	AUC
PCA	0.962
ASPCA-BG	0.963
ASPCA-ADMM	0.952
ASPCA-ADMMSEQ	0.967

Table 6-2 AUC values for KDD'99 Dataset

If we project the squared prediction error of data points on ASPCA-ADMMSEQ abnormal subspace, outliers can be easily visualized. But, in this case we cannot distinguish the individual anomalies visually from the projection length. Almost all of the four major kind of outliers have similar SPE score on abnormal subspace.

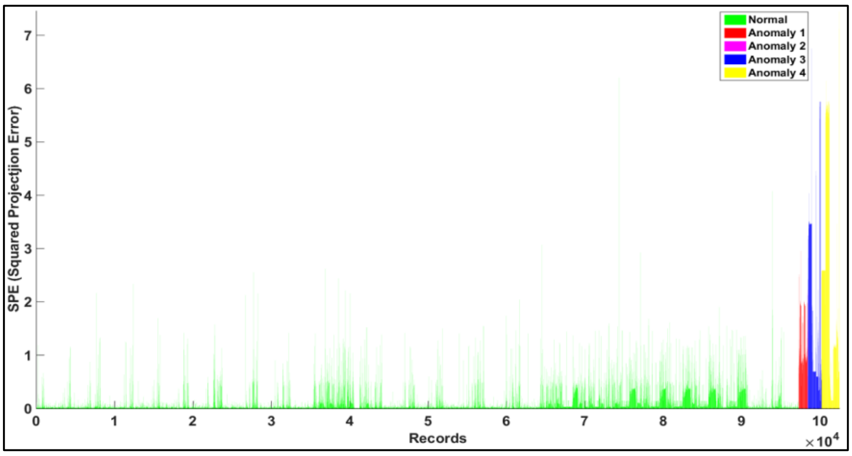


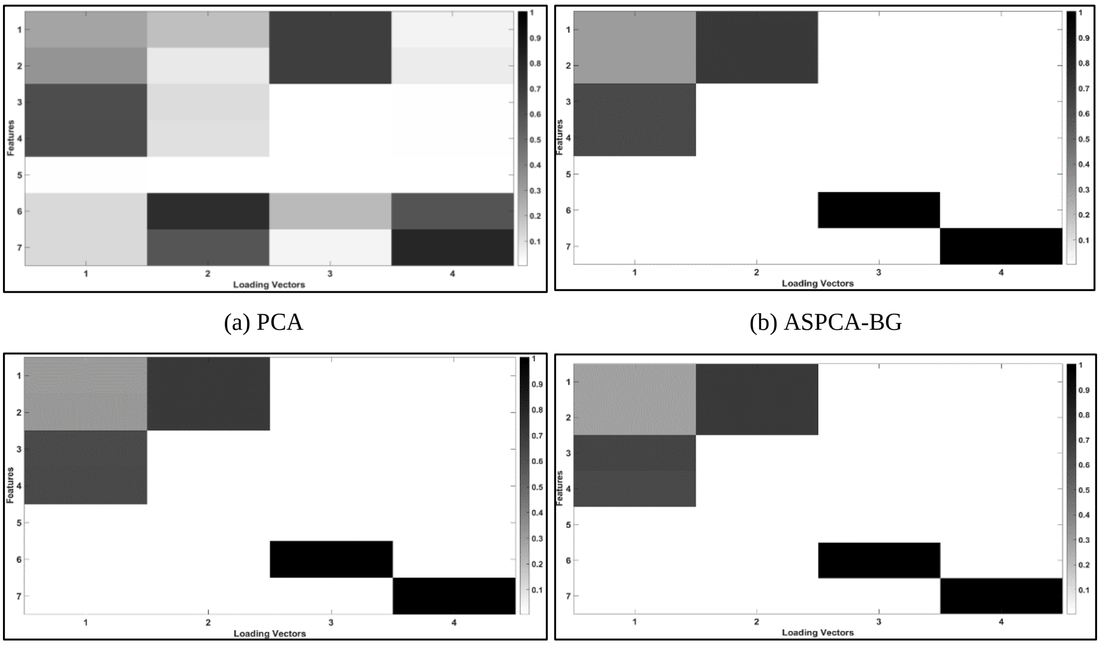
Figure 6-12 SPE values of KDD'99 Data instances on Abnormal Subspace

6.4 Sparsity evaluation

The next stage of experiment is to compare the sparsity or cardinality performance of the loading vectors generated by our proposed methods and again we used the result of PCA method as a baseline. We used three metrics to evaluate the sparsity of the abnormal subspace of the abnormal PCs, namely, norm 1($\|V\|_{1,\,1}$) of loading vectors, $\text{Card}_{0.1}$ (number of entries with absolute value higher than 0.1), and $\text{Card}_{0.01}$ (number of entries with absolute values higher than 0.01). We will see that our proposed methods demonstrate sparser result than PCA. For all datasets overall, the ASPCA-ADMMSEQ model achieved better sparsity performance than all other methods in terms of $\|V\|_{1,\,1}$ and cardinality.

6.4.1 Synthetic Dataset

For Synthetic dataset, both of our proposed model performs as good as ASPCA-BG method in terms of Sparsity. We can compare the involved variables in abnormal PCs with the PCA version of abnormal PCs. Cardinality values has been reduced in a great extent than PCA. ASPCA-BG method uses a global optimization step in the end to improve the sparsity result. But, our proposed two methods returns sparser result without needing any further improvement. From the Table we can compare the performance of sparsity using the entry wise norm and cardinality metric.



(a) PCA

(b) ASPCA-BG

(c) ASPCA-ADMM

(d) ASPCA-ADMMSEQ

Figure 6-13 Loading matrix of PCs for Synthetic Data

The main advantage of ASPCA-ADMMSEQ method is that it can formulate the result of PCA under if we don't need sparser abnormal PCs. But, our proposed ASPCA-ADMM method cannot formulate the result of PCA even setting the cardinality constraint to zero. We will see details on this topic later in parameter selection section in this chapter.

Method	**$\|\|V\|\|_{1,1}$**	**$Card_{0.1}$**	**$Card_{0.001}$**
PCA	7.17	16	23
ASPCA-BG	5.30	8	8
ASPCA-ADMM	5.31	8	8
ASPCA-ADMMSEQ	5.28	8	8

Table 6-3 Sparsity performance of Abnormal Subspace for Synthetic Dataset

The sparsity evaluation metrics is presented in bars below for easy visual comparison.

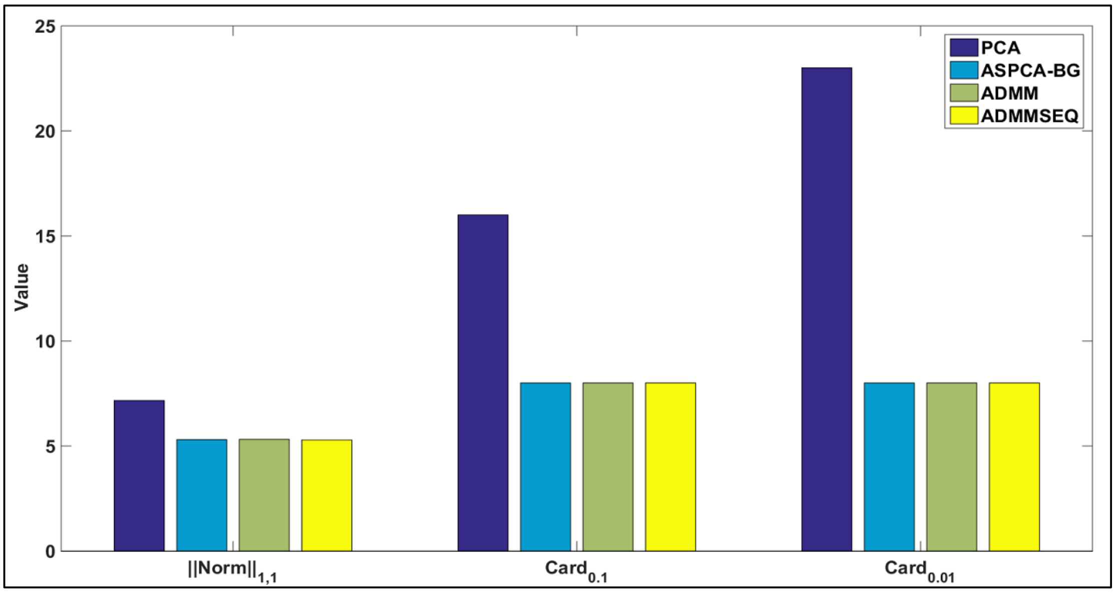
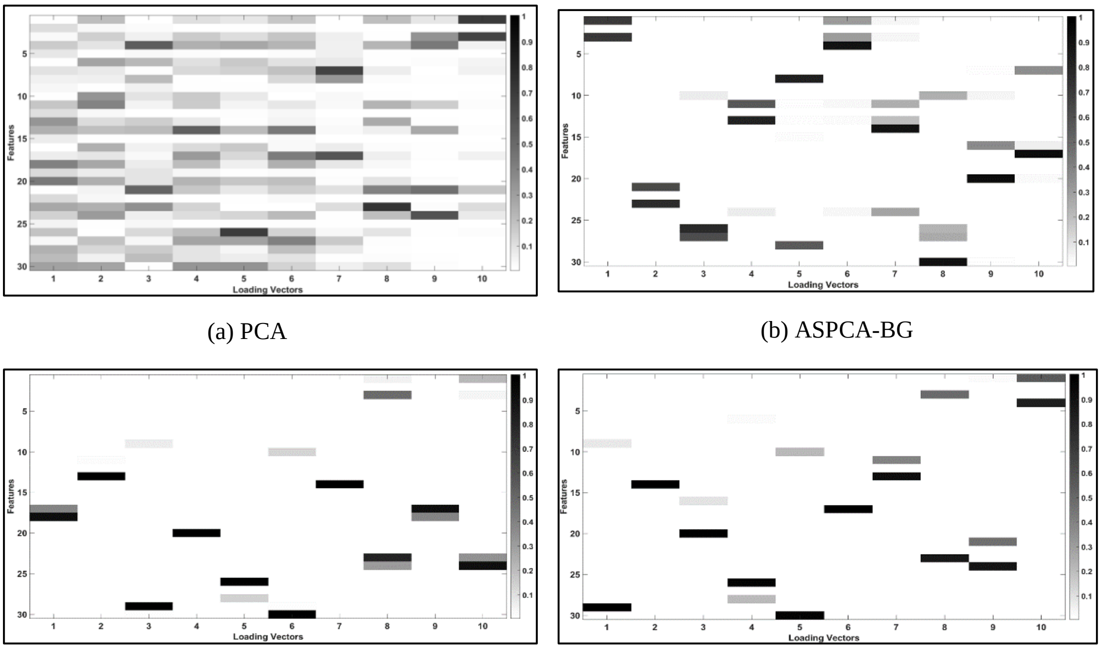


Figure 6-14 Sparsity Evaluation metrics in Bar Graph for Synthetic dataset

6.4.2 Breast Cancer Dataset

For the Breast Cancer dataset we have chosen abnormal subspace dimension size as 10. We chose a convenient dimension size so that it can achieve highest AUC values for our proposed to get a decent and fair comparison with PCA and ASPCA-BG. If we compare the sparsity of the PCs returned by PCA and our proposed ASPCA-ADMMSEQ, it can easily be seen that our methods gives us a sparse subspace and eventually sparse PCs explaining almost same variance.



(a) PCA

(b) ASPCA-BG

(c) ASPCA-ADMM

(d) ASPCA-ADMMSEQ

Figure 6-15 Loading matrix of PCs for Breast Cancer Data

Norm 1, $||V||_{1,1}$ values can be compared to get better idea on sparsity. Card$_{0.1}$ and Card$_{0.001}$ also reduced. If we look at the result of ASPCA-BG method, it seems that this method has best sparsity result. But, as we know that sparsity increases variance and our aim is to keep the variance as low as possible. Our proposed method can also achieve the result of ASPCA-BG method. But, in that case, variance increases and get similar reading to ASPCA-BG method which we don’t want. We should also remember another important aspect regarding sparsity is that sparsity introduces another disadvantage. Interpretability performance and correlation of feature variables loses, if we make the PCs sparser more than it is needed.

Method	**$\|\|V\|\|_{1,1}$**	**Card$_{0.1}$**	**Card$_{0.001}$**
PCA	34.22	111	237
ASPCA-BG	16.51	25	40
ASPCA-ADMM	12.19	18	24
ASPCA-ADMMSEQ	12.01	16	20

Table 6-4 Sparsity performance of Abnormal Subspace of Breast Cancer Dataset

The sparsity evaluation metrics are presented in a bar graph below for better visual comparison for this Breast Cancer Dataset. It easily demonstrates that, ASPCA-ADMMSEQ method has best sparsity result for Breast Cancer dataset.

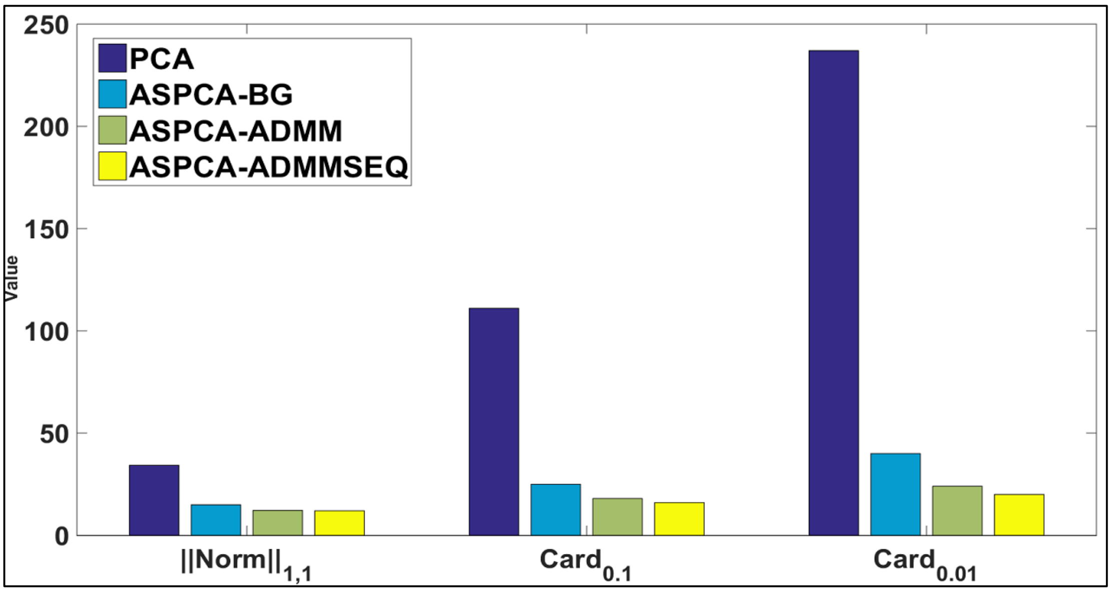
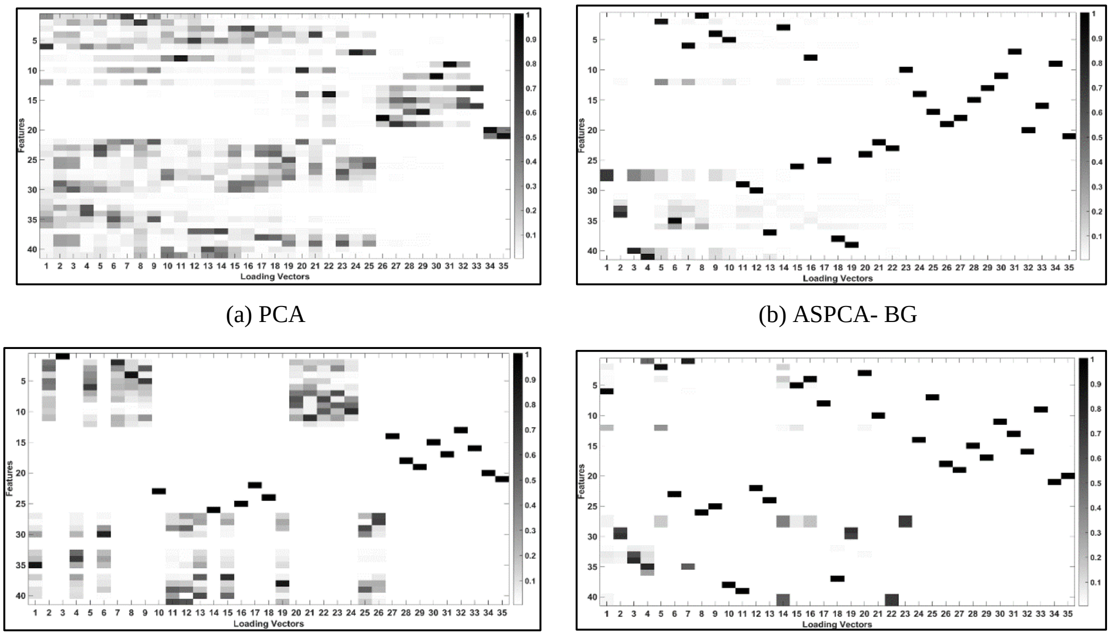


Figure 6-16 Sparsity Evaluation metrics in Bar Graph for Breast Cancer dataset

6.4.3 KDD'99 Dataset

For KDD'99 dataset, we have chosen a convenient abnormal dimension size of 35 to achieve better AUC values for our proposed algorithms and also to get a decent and fair sparsity comparison. If we compare the sparsity of the PCs returned by PCA and our proposed methods, it can easily be seen that ASPCA-ADMMSEQ method gives us a sparse subspace and eventually sparse PCs. But, the sparsity performance of ASPCA-ADMM is not as good as ASPCA-ADMMSEQ or ASPCA-BG.



(a) PCA

(b) ASPCA- BG

(c) ASPCA-ADMM

(d) ASPCA-ADMMSEQ

Figure 6-17 Loading matrix of PCs for KDD'99 Data

From the following table we can see that ASPCA-ADMMSEQ algorithm returns the sparsest subspace. But, simultaneous estimation of subspace this time doesn’t give as sparse as sequential method. Our observation is that, simultaneous method focuses mainly on subspace rather than eigenvectors. But, ASPCA-ADMMSEQ captures most of the correlations and also returns a sparse subspace maintaining a low variance.

Method	$\|V\|_{1,1}$	Card$_{0.1}$	Card$_{0.001}$	Variance
PCA	97.17	248	691	21518
ASPCA-BG	44.00	57	192	21564
ASPCA-ADMM	62.00	146	229	35000
ASPCA-ADMMSEQ	41.77	55	82	23480

Table 6-5 Sparsity performance of Abnormal Subspace for KDD’99 dataset

The sparsity evaluation metrics are presented in a bar graph below for better visual comparison for this KDD’99 Dataset.

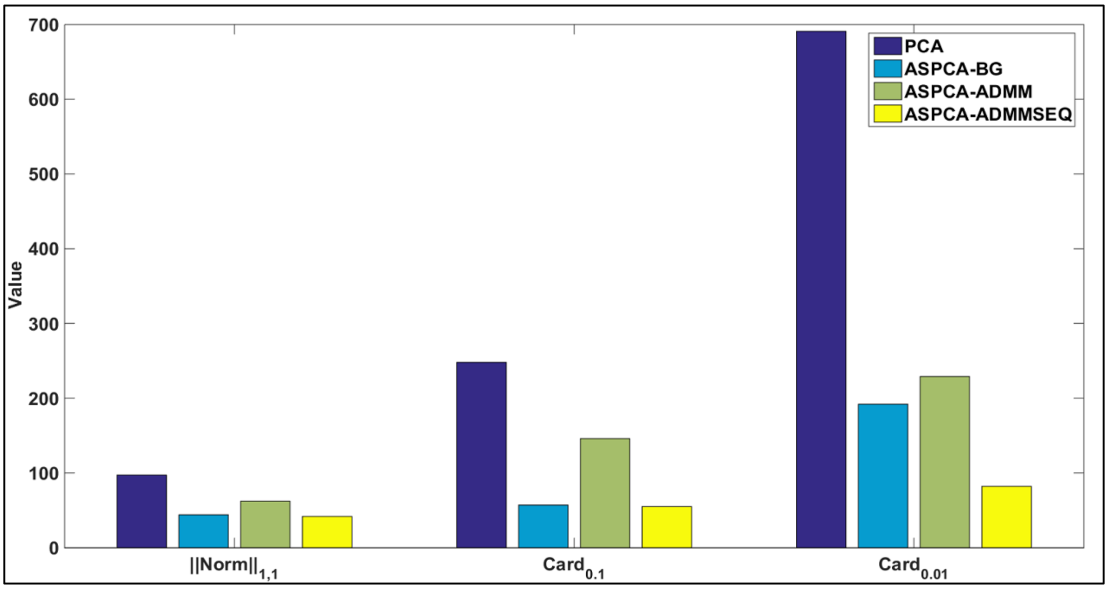


Figure 6-18 Sparsity Evaluation metrics in Bar Graph for KDD’99 dataset

6.5 Interpretation evaluation

Now we are in the last but most interesting stage of our proposed models. Here we will evaluate the interpretation performance of the ASPCA-ADDMMSEQ model. Because, it has better performance in terms of interpretation. Since we want to see how anomalies are interpreted by our model, we tried to choose a feasible threshold value on SPE to ensure that most of the true anomalies are detected. We show the threshold values, false positive rates (FPR), and true positive rates (TPR) for all three datasets in Table 6-6.

Dataset	Threshold	TPR	FPR
Synthetic	0.08	1	0
Breast Cancer	0.12	1	.05
KDD’99	0.5341	0.8735	.0600

Table 6-6 SPE Threshold values for 3 datasets

6.5.1 Synthetic Dataset

The four abnormal PCs returned by abnormal subspace and the projection values of 15 anomalies on these PCs are shown in Table 6-7 and Figure 6-19, respectively. As shown in Table 5, the first three PCs correspond to the rules of D ≈ C + A, A ≈ B, and F ≈ 0, respectively. The anomalies breaking these rules indeed have large projection values on the corresponding PCs. Thus, our model not only identifies the set of features that are responsible for an anomaly, but also tell the cause of the anomaly breaking the rules indicated by the abnormal PCs. It is worth to be noted that ASPCA-ADMM and ASPCA-ADMMSEQ both of the methods returns same PCs for this synthetic dataset.

Index	Components
1	-0.29148 A – 0.291663 B – 0.654865 C + 0.633345 D
2	-0.70733 A + 0.706883B
3	1 F
4	1 G

Table 6-7 Components on Synthetic Data

The heatmap of abnormal records on components returned by the abnormal subspace also shows which PC is responsible for what anomaly. We know that in our dataset, there are three kind of anomalies and first three PC represents those anomalies. And the last 4th PC doesn’t contain any heatmap of abnormal records. The first PC shows the breaking of D ≈ C + A rule. The second PC interprets breaking of A ≈ B rule and the 3rd one represents the breaking of 3rd rule of F ≈ 0. We achieved the similar result like ASPCA-BG method. That method takes a lot of time to compute these PCs. But, here by computing the subspace simultaneously reduced a lot of computation time by reducing SDP problem solving.

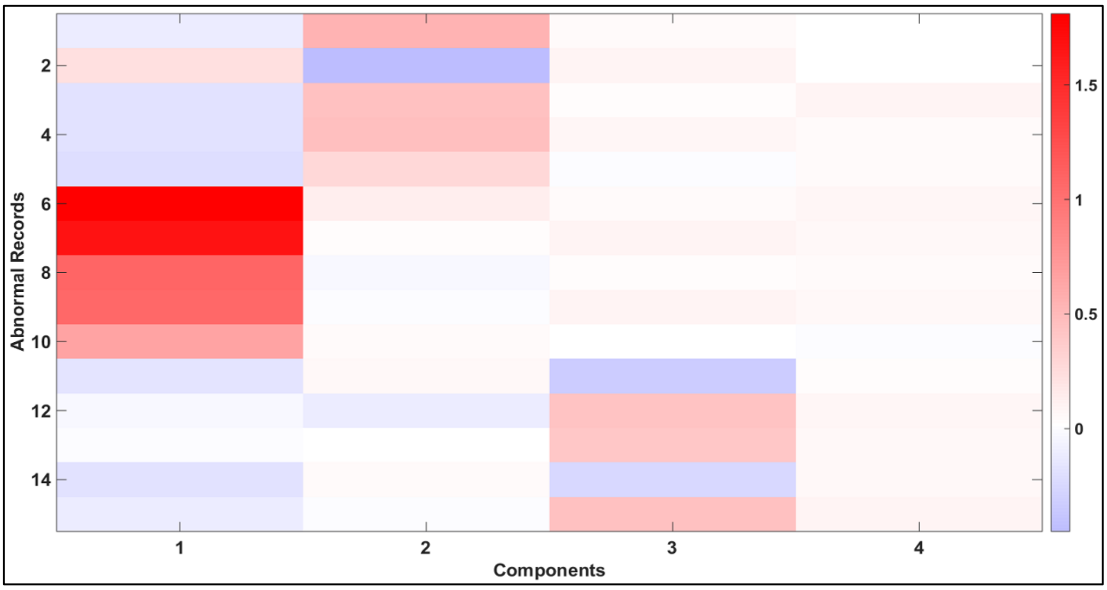


Figure 6-19 Heatmap of Abnormal Synthetic Records on Components

6.5.2 Breast Cancer Dataset

We will demonstrate interpretation performance of ASPCA-ADMMSEQ method using the real world well-known Breast Cancer dataset. We want to remind the reader that this dataset has 357 benign (normal) records and 10 malignant (anomalies) records. The heatmap projection values of these anomalies on the abnormal PCs obtained are shown in Figure 6-20.

We also found similar result like [2] that there are two kinds of malignant records in this dataset. Records 1, 2, 3, 5, and 7 has large projection on 2, 7, 8, and 9 PCs shown in the following table. Rest of the anomalies have mainly on 1, 4, and 5 PC. The features appearing in the first kinds of malignant are related to area, perimeter, radius and especially area-se, area-worst, radius-worst, perimeter-se, perimeter-worst, perimeter-mean which is also reported by [2, 26] as being effective for classifying malignant records.

In our study we found similar results where area-worst actually has a linear relation with radius-worst like ASPCA-BG method. On the other hand other kind of malignant records are detected by PCs related to symmetry, fractal dimension, compactness and concave features, which clearly indicates another type of malignant records. ASPCA-ADMM also shows almost similar interpretation result. But, it doesn’t cover all the correlation of variables and sparsity of individual PCs is not as promising as ASPCA-ADMMSEQ. Because, simultaneous estimation doesn’t consider the sparsity of individual eigenvectors rather focuses on the whole abnormal subspace.

Index	Components
1	- 0.08symmetry-mean + 0.996symmetry-worst
2	1area-se
4	- 0.02 compactness-mean + 0.97 compactness-worst - 0.20 concave-points-worst
5	- 0.21 fractal dimension-mean + 0.97 fractal dimension-worst
7	- 0.39 radius-se + 0.91 perimeter-se
8	- 0.49 perimeter-mean + 0.86 perimeter-worst
9	- 0.02 radius-mean - 0.47 radius-worst + 0.87 area-worst

Table 6-8 Components on Breast Cancer Data

The loading matrix of the abnormal PCs obtained by ASPCA-BG on Breast-Cancer is also similar; see [2] for comparison. The relevant features for Breast Cancer dataset are radius, concavity, area, fractal-dimension, perimeter, and compactness. For SPE, the threshold value of 0.12 our ASPCA-ADMMSEQ method has perfect TPR and tiny FPR of 0.05.

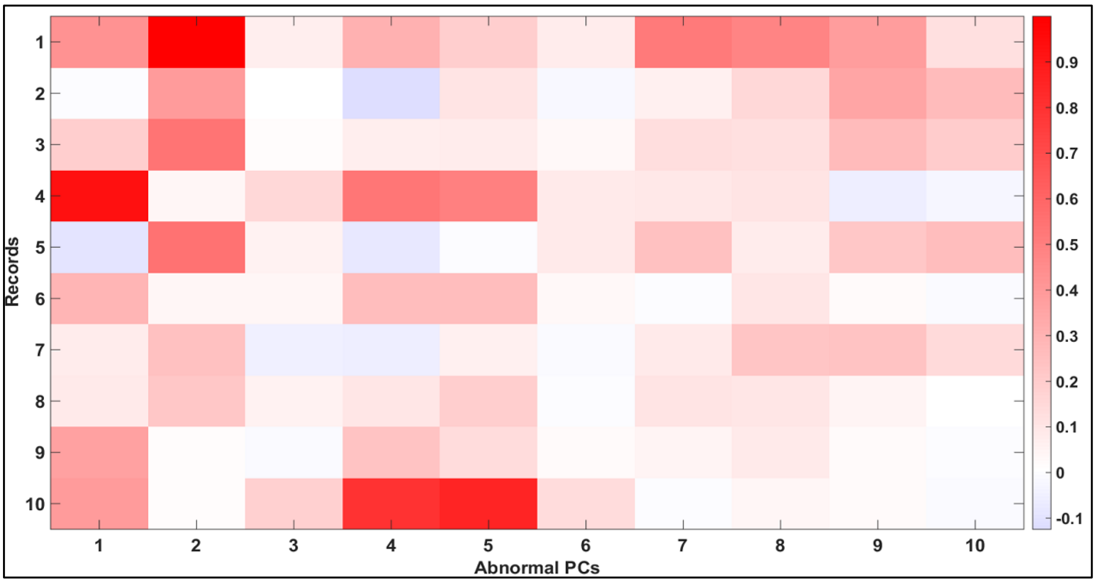


Figure 6-20 Heatmap of Abnormal Breast Cancer Records on Components

6.5.3 KDD'99 Network Intrusion Dataset

Our ASPCA-ADMMSEQ method dominated other methods in terms of detection accuracy with AUC value of 0.97 for KDD'99 dataset. The dataset we are using has 102, 472 connection data instance including 5194 outliers. With a given SPE threshold of 0.37, this model detected 4530 true anomalies which is better than ASPCA-BG (4397).

Index	Components
1	- 0.96 dst_bytes + 0.21 logged_in
2	0.66 same_srv_rate - 0.74 diff_srv_rate - 0.05 dst_host_srv_count
3	-0.06dst_host_count+0.61dst_host_srv_count-0.78 dst_host_same_srv_rate
4	-0.51duration - 0.10dst_host_srv_count - 0.78dst_host_diff_srv_rate + 0.30dst_host_same_src_port_rate
5	- 0.90protocol_type - 0.34logged_in - 0.16rerror_rate - 0.16srv_rerror_rate
6	1count
7	-0.83duration + 0.01dst_bytes + 0.54dst_host_diff_srv_rate
8	1srv_serror_rate
9	1serror_rate
10	1dst_host_serror_rate
11	1dst_host_srv_serror_rate
12	1is_guest_login
13	1srv_count
14	- 0.41rerror_rate - 0.41srv_rerror_rate + 0.55dst_host_rerror_rate + 0.55dst_host_srv_rerror_rate
15	- 0.99src_bytes + 0.08logged_in - 0.07rerror_rate - 0.07srv_rerror_rate
16	- 0.964flag + 0.18rerror_rate + 0.18srv_rerror_rate
17	1wrong_fragment
18	- 0.028dst_host_count - 0.99dst_host_srv_diff_host_rate
19	- 0.74same_srv_rate - 0.66diff_srv_rate
20	- 0.998service - 0.037logged_in - 0.029dst_host_srv_count
21	1hot
22	- 0.70dst_host_rerror_rate + 0.71dst_host_srv_rerror_rate
23	- 0.70rerror_rate + 0.70srv_rerror_rate
24	1root_shell
25	1land
26	1num_shells
27	1num_access_files
29	1num_file_creations
30	1num_failed_logins

Table 6-9 Components on KDD'99 Network Intrusion Data

Our model should distinguish individual anomalies present in dataset and at the same time we should also summarize interpretations of similar anomalies. It is not always the case that one PC represents a single anomaly. We can interpret anomalies where anomalies have high projection on several PCs at the same time. We are going to introduce this fact in this experiment. From the dataset definition of KDD'99, in low level there are 23 kinds of anomalies present in the dataset. But, those anomalies are further grouped into 4 kinds which consists of R2L (1126), U2R (52), Probe (1731), and DoS (2285).

Anomaly Group	Anomaly	Active Components
R2L	'guess_passwd'	3, 18, 30
	'ftp_write'	3, 12
	'imap'	20
	'phf'	24
	'multihop'	14
	'warezmaster'	12
	'warezclient'	1, 3, 12
	'spy'	7, 26
U2R	'buffer_overflow'	3,18,24
	'loadmodule'	3, 24, 26
	'perl'	24
	'rootkit'	15
Probe	'satan'	2
	'ipsweep'	4, 7
	'portsweep'	2
	'nmap'	5
DoS	'smurf'	5, 6, 13
	'neptune'	3, 8,9,10,11,16
	‘back’	3
	'pod'	5
	'teardrop'	17
	'land'	3, 8, 9, 10, 16, 25

Table 6-10 Active Components of individual anomalies for KDD'99 dataset

From Table 7, we can see that the projections on abnormal PCs for different anomaly types varied a lot, from which we can identify the characteristics of each of the anomalies. In the table we have demonstrated the important components to represent the anomalies. Active components means that, anomalies have high projection heatmap on those PCs. Some of those anomalies might have lower projection on other PCs. But, we ignored those PCs to understand our interpretation model in a better way. Next, we will demonstrate interpretation performance at least two kinds of anomalies from each of the groups.

In R2L anomalies group if we want to talk about 'guess_passwd' anomaly, it has higher projection values on 3, 18, 30 PCs. If we look the variables involved in these PCs includes 'num_failed_logins', 'dst_host_count', 'dst_host_srv_count', 'dst_host_same_srv_rate', 'logged_in'. And eventually it makes sense that these variables should be responsible for the anomaly type 'guess_passwd'. Another anomaly of this group is Warezclient and attackers try to download files from forbidden directories from the FTP servers. Our interpretation is consistent with [2] as this anomaly contains the features 'logged_in' with 'low_dst_bytes' (not shown as low projection), 'is_guest_login', 'dst_host_count', 'dst_host_srv_count' and 'dst_host_same_srv_rate'.

From the U2R anomaly group, 'buffer_overflow' attackers try to gain the root authority of the host, and high projection on 24th component indicates that the user has logged into the server with root shell. Other variables related to destination host and server is also involved.

For the next group 'probe', we want to establish similar trends. If we consider for example 'ipsweep' attacks, it involves sweeping of different host IPs and tries to hack it and eventually 4th PC interprets it clearly incorporating 'dst_host_diff_srv_rate' variable. Another anomaly named 'portsweep' attacks try to visit different service ('diff_srv_rate'), 'same_srv_rate' and is presented by 2nd PC.

From DoS group, for example, smurf attacks are common form of DoS packet floods. Our model shows this anomaly must have high 'srv_count', 'srv_rerror_rate', 'protocol_type' and 'count' values and which is presented by 5th, 6th , and 13th PC. Another anomaly named 'teardrop' attack in smurf group where attackers try to break the host by sending mangled IP fragments, which led to high 'wrong_fragment' and which is presented by 17th PC.

All of our observations are consistent with the result of ASPCA-BG method and even includes more meaningful variables are involved to represent and interpret the anomalies. As we know, KDD’99 dataset has several types of abnormal data, our model almost detected each of the anomalies with its reasons. Next, we presented the heatmap projection of anomaly records in computed abnormal PCs below.

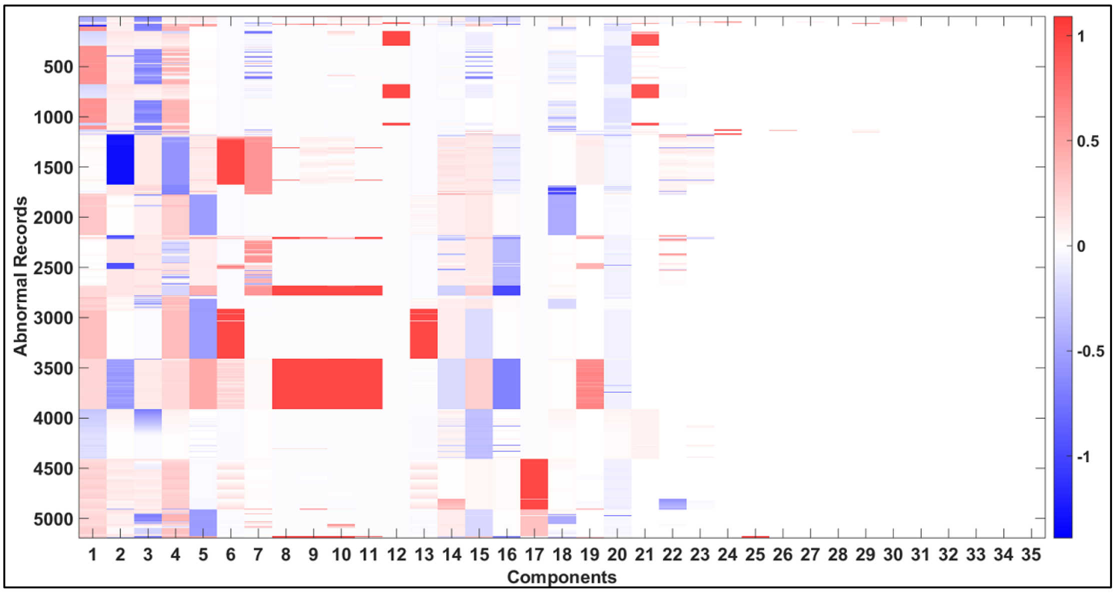


Figure 6-21 Heatmap of Abnormal KDD’99 records on Components

6.6 Computation Time

Our proposed methods have the most impressive performance on this metric. Both of the methods just outperform all other concurrent methods (ASPCA-BG) in terms of consumed time. As, ADMM converges much faster than ASPCA-BG method it takes much lesser time to return the estimated abnormal subspace. Moreover, ASPCA-BG method computes the abnormal PCs sequentially and ASPCA-ADMM method estimates the subspace simultaneously. In other words, this method doesn’t use any deflation method to compute the abnormal PCs rather just returns all the PCs on a single iteration. For this reason, this method is very much faster than all other methods including ASPCA-ADMMSEQ and mostly close to PCA. On the other hand ASPCA-ADMMSEQ method returns abnormal PCs sequentially or one by one. So, it takes more time than ASPCA-ADMM method to estimate the whole subspace. If we compare the result of ASPCA-ADMMSEQ with ASPCA-BG method, ADMM algorithm saves a lot of time for computation. Both of the methods use deflation technique to compute the subspace. But, due the blessing of ADMM and better tuning parameter even than

ASPCA-ADMM, it converges faster. For example if we compare the execution time of ASPCA-ADMMSEQ with ASPCA-BG method for synthetic dataset, it is 20 times faster than ASPCA-BG method. The result of other datasets can be found in the following table in details.

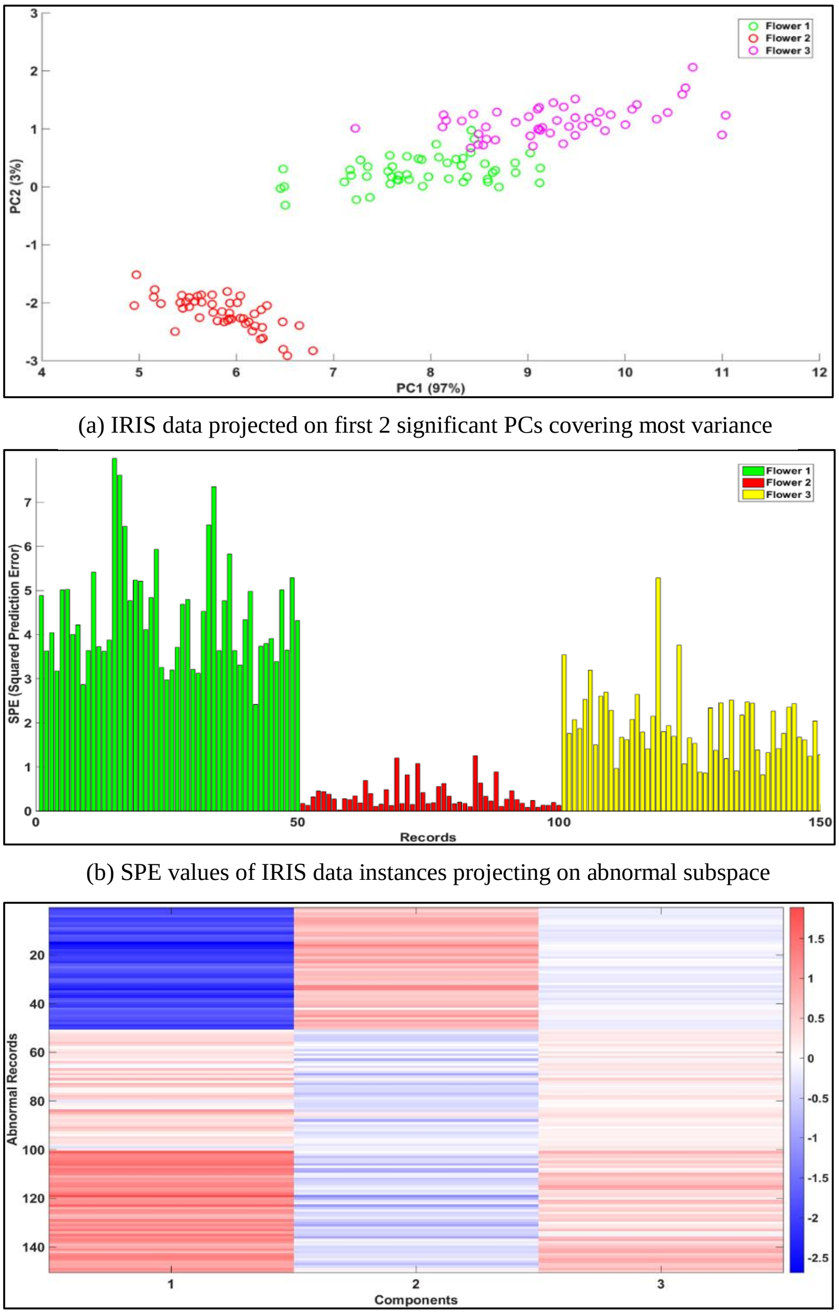
Dataset	Method	Time consumed (seconds)
Synthetic Dataset	PCA	3.87e-04
	ASPCA-BG	2.30
	ASPCA-ADMM	0.15
	ASPCA-ADMMSEQ	0.11
Breast Cancer Dataset	PCA	6.70e-04
	ASPCA-BG	14.02
	ASPCA-ADMM	0.275
	ASPCA-ADMMSEQ	2.31
KDD’99 Dataset	PCA	0.0016
	ASPCA-BG	75
	ASPCA-ADMM	14.49
	ASPCA-ADMMSEQ	20.79

Table 6-11 Computation time to estimate Abnormal Subspace

6.7 Clustering Performance

We converted the anomaly detection task to perform clustering task using the IRIS dataset. This dataset has three flowers and each flower contains 50 data instances. So, the sample size of the dataset is 150. And we will try to estimate the clusters of single flower using the ASPCA-ADMMSEQ method. It is worth to be noted that, ASPCA-ADMM has similar properties and replicates the same performance as ASPCA-ADMMSEQ method. That’s why we have only shown the clustering performance of ASPCA-ADMMSEQ method in this section.

If we project the data points on first two principal components, it can be easily seen that three clusters cannot be easily formed from this point of view. Now, we will use our method to explore the clustering performance. As, first PC almost covers all the data, we set the abnormal subspace dimension size to 3. Our ASPCA-ADMMSEQ returns an abnormal subspace and we will project the data points on this subspace. If we plot the SPE of each data point, we can see that there are more or less three types of SPE value in the graph. The mid 50 (Flower 2) has lowest prediction as it was assumed as the normal data while estimating the abnormal subspace. And all other two types of flower has highest projection length on this subspace but Flower 1 has the maximum one. Even from the heatmap of records on 3 abnormal PCs we can distinguish the individual cluster of individual flowers.



(a) IRIS data projected on first 2 significant PCs covering most variance

(b) SPE values of IRIS data instances projecting on abnormal subspace

(c) Heatmap of IRIS data records on components

Figure 6-22 Clustering performance on IRIS Dataset

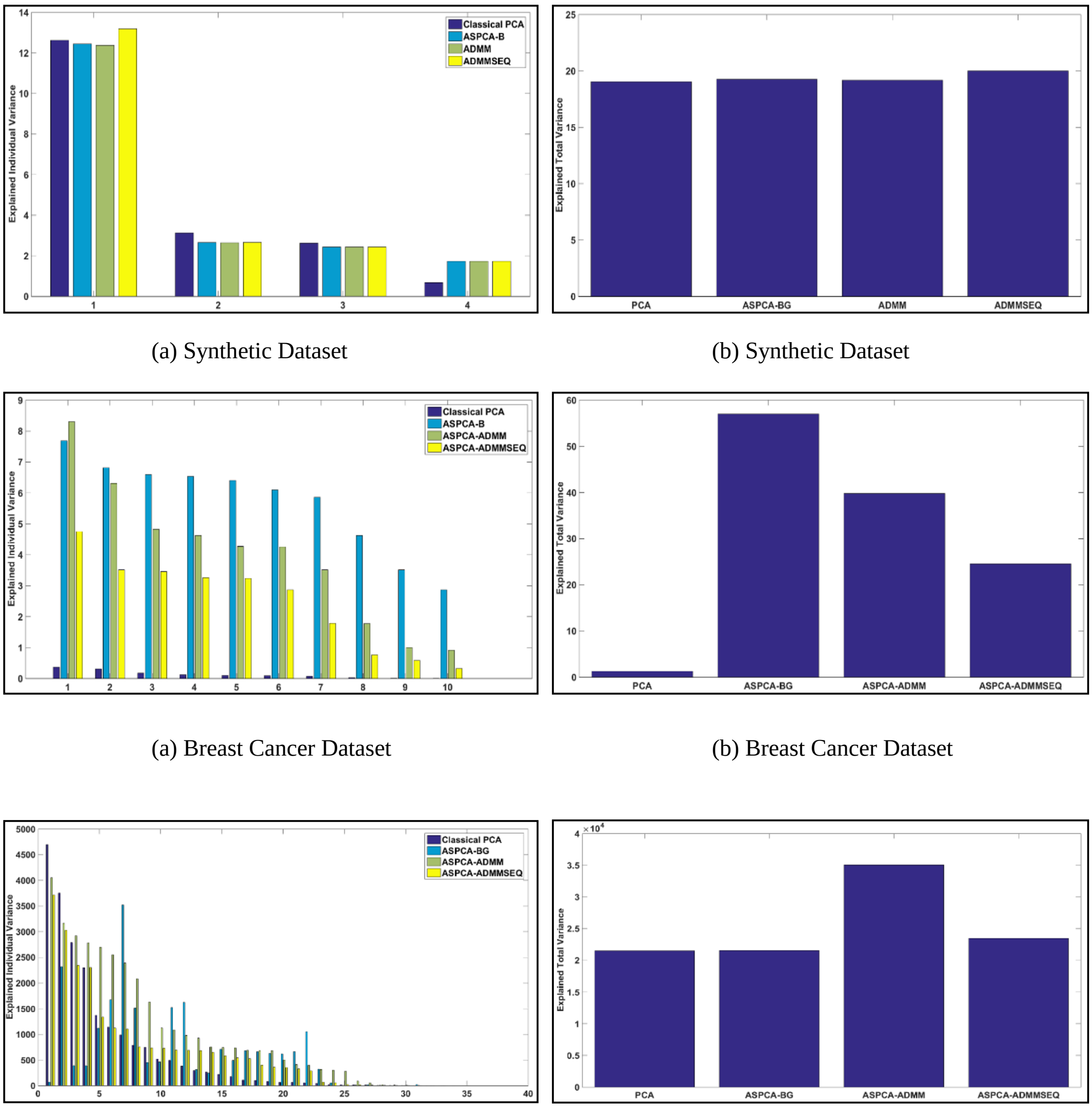
6.8 Explained Variance

Cardinality parameter, λ introduces a trade-off between the sparsity and the variance explained by the abnormal components. With less sparsity, the variances on the whole abnormal subspace for our proposed two methods are almost closer to PCA. The ASPCA-ADMMSEQ even can replicate the exact result of PCA if the sparsity parameter is set to zero. Moreover, ASPCA-ADMMSEQ explains less variance than ASPCA-ADMM method and that's why we prefer sequential estimation of PCs.

Dataset	*d*	Method	Variance Explained
Synthetic	4	PCA	19.04
		ASPCA-BG	19.26
		ASPCA-ADMM	19.17
		ASPCA-ADMMSEQ	19.50
Breast Cancer	20	PCA	1.27
		ASPCA-BG	20.29
		ASPCA-ADMM	39.79
		ASPCA-ADMMSEQ	24.53
KDD'99	35	PCA	21518
		ASPCA-BG	21564
		ASPCA-ADMM	35000
		ASPCA-ADMMSEQ	23480

Table 6-12 Variance explained by the Abnormal Subspace

Our proposed methods explains almost same variance for the synthetic dataset. For other two datasets we tried to keep the variance as low as possible to be close to PCA method. But, at the same time we need sparser result for better interpretation. So, we cannot keep variance as like as PCA and get sparse principal components. If we compare the performance in this regard, ASPCA-ADMMSEQ method has better result as shown in figures. For Breast Cancer and KDD'99 dataset ASPCA-ADMMSEQ method shows close result to PCA. The dimension size of abnormal subspace, ***d*** also plays a vital role to measure the explained variance. If the dimension size is not chosen properly FPR would be increased and estimated abnormal subspace would get polluted by normal data. We have presented the total variance covered by the subspace and also by the abnormal principal components for better visualization and understanding.



(a) Synthetic Dataset

(b) Synthetic Dataset

(a) Breast Cancer Dataset

(b) Breast Cancer Dataset

(a) KDD’99 Dataset

(b) KDD’99 Dataset

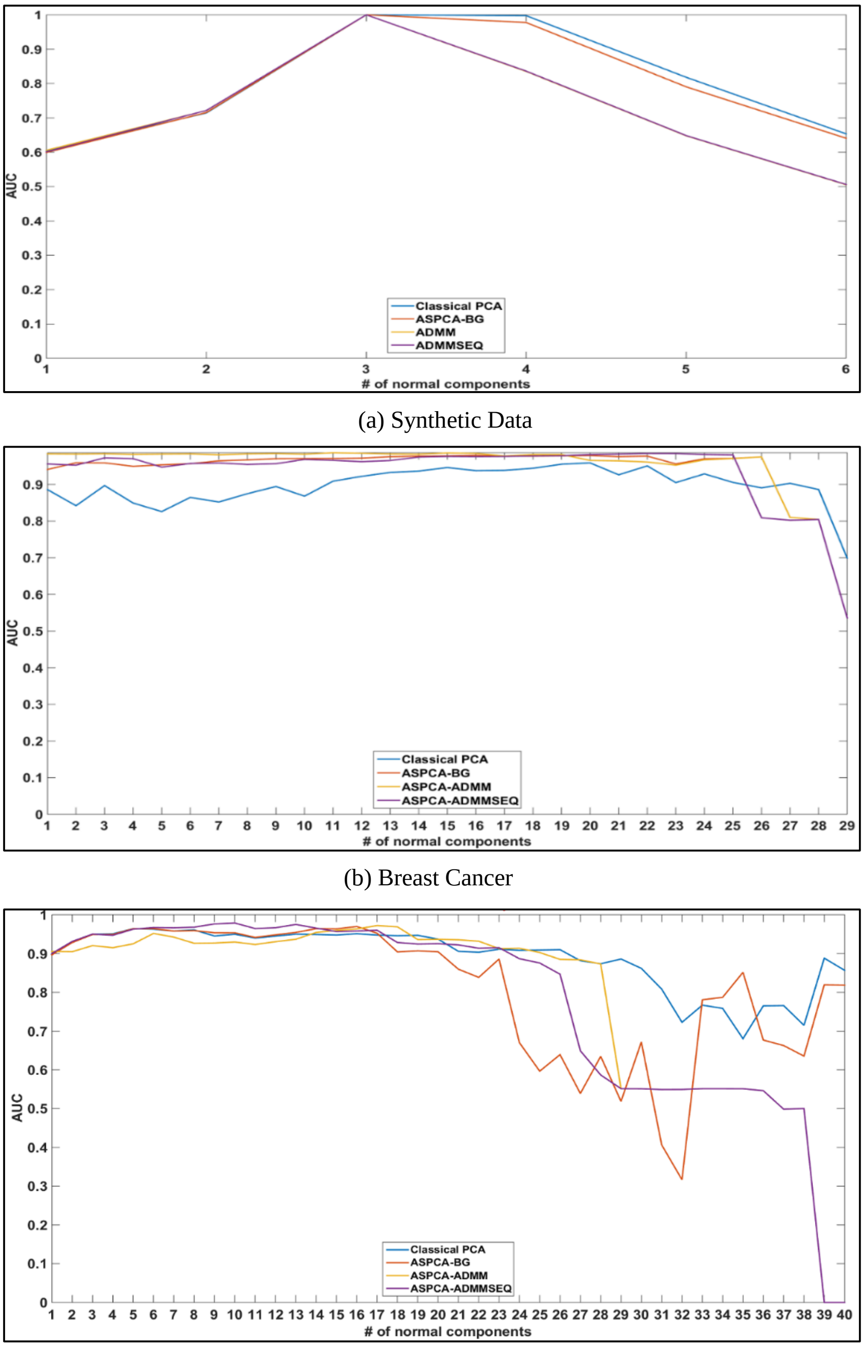
Figure 6-23 Captured Variance (a) Individual Components (b) Abnormal Subspace

6.9 Parameter selection

Both of the ASCPA-ADMM and ASPCA-ADMMSEQ methods are sensitive to the input parameters. All of the parameters involved in our algorithms influences on the performance of detection accuracy, AUC but also on sparsity. We will show a comprehensive analysis of parameters influence on our proposed methods.

6.9.1 Abnormal Subspace dimension- (*d*)

The perfect abnormal subspace dimension is the key issue of better anomaly detection accuracy and interpretation as well. We considered two factors which influenced highly to choose the *d*. It is chosen such that the AUC value is high and we get a subspace which can detect most of the anomalies. On the other hand captured variance by the significant principal components are also taken into consideration. Leading significant principal components mostly explains the variance on dataset. So, we have chosen the residuals of principal components where the variance is less covered. According to the following demonstrations, we have chosen ***d*** as 4 for Synthetic dataset, 10 for Breast Cancer Dataset and 32 for KDD’99 dataset.



(a) Synthetic Data

(b) Breast Cancer

(c) KDD’99

Figure 6-24 Selection on number of Abnormal PCs

The choice to select a perfect dimension size, ***d*** is of great interest. There are several factors are being considered while choosing it. Better AUC and FPR value played vital role to choose ***d*** in our study**.** Other key factors which are also monitored are the ratio of normal and abnormal data, captured variance by significant principal components.

For, Synthetic dataset, we get perfect AUC value for ***d*** **=** 4. Moreover, the explained variance by first 3 significant PCs is around 98% which is roughly equal to anomaly ratio in the dataset. For, Breast Cancer dataset ASPCA-ADMMSEQ method got AUC of 0.981 which was highest and we chose ***d*** **=** 10. It is worth to mention that in [2] ASPCA-BG method also found similar result and selected 10 abnormal PC to represent the abnormal subspace. For, KDD'99 dataset the abnormal subspace dimension size is set to 35 considering best AUC value and also captured variance by the significant principal components of normal subspace.

6.9.2 Sparsity parameter- *(λ)*

We know that sparsity parameter is introduced to increase the number of non-zero loadings of PCs. But, this parameter has very sensitive impact on ADMM convergence too. Our empirical study suggests that the ratio of λ and ρ has great influence on ADMM convergence as we used the ratio (λ/ρ) as the shrinkage parameter of soft thresholding. In some cases, the algorithm didn't even converge to give a global tractable solution with in a short period of time.

In essence, sparsity parameter introduces additional variance on the final solution. The more you make the PCs sparser, the variance gets higher. All of our experiments suggests the fact. It is true for both of our proposed methods and also for other sparse PCA methods like ASPCA-BG. To make a fair comparison with other methods, our goal was to keep the variance as close as to the result of PCA. The proposed ASPCA-ADMMSEQ method formulated the result of PCA by setting the sparsity parameter, λ to zero. It is usual that sparsity lowers the value of norm $||V||_{1,1}$ and it is also evident in our study. For, $\lambda = 0$, the norm is exactly equal to the norm calculated by PCA. And the other sparsity parameters like cardinality values are also equal to the PCA.

The impact of changing λ has also influence on AUC values. For Breast Cancer and Synthetic dataset, in Table 6-14 and Table 6-15, AUC reaches to a maximum value and then again starts to reduce. For example for breast cancer dataset, (λ = .015 and ρ = .001) both of our proposed method has highest detection performance.

			ASPCA-ADMM				ASPCA-ADMMSEQ			
λ	ρ	λ/ρ	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)
0	0.0005	0	7.81	19.04	1	0.002	7.16	19.04	1	0.002
0.05		100	5.31	19.17	1	0.04	6.69	19.25	1	3.68
0.1		200	5.71	19.17	1	0.09	6.56	19.14	1	5.94
0.15		300	5.54	19.17	1	0.13	6.50	19.14	1	6.96
1.5		3000	5.31	19.17	1	1.39	5.30	19.23	1	28.68
15		30000	5.31	19.27	1	13.45	N/A	N/A	N/A	N/A
0	0.5	0	7.19	19.04	1	0.003	7.16	19.04	1	0.003
0.05		0.1	5.78	19.17	1	0.031	6.68	19.23	1	0.02
0.1		0.2	5.31	19.17	1	0.004	6.57	19.15	1	0.009
0.15		0.3	5.31	19.17	1	0.002	6.50	19.14	1	0.008
1.5		3	5.31	19.17	1	0.003	5.30	19.22	1	0.022
15		30	5.31	19.27	1	0.015	5.28	20.00	1	0.11

Table 6-13 Selection on λ parameter for Synthetic Dataset

			ASPCA-ADMM				ASPCA-ADMMSEQ			
λ	ρ	λ/ρ	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)
0	.001	0	41.51	2.06	0.944	0.005	34.22	1.27	0.958	0.023
.01		10	20.46	11.37	0.978	0.172	13.32	12.96	0.966	6.30
.015		15	22.50	22.38	0.982	0.236	12.00	24.42	0.981	9.03
.018		18	12.19	39.79	0.978	0.332	11.08	38.48	0.978	11.19
.05		50	10	57.00	0.965	0.362	10	57.00	0.965	12.11
.5		500	10	160.39	0.982	0.436	10	201.43	0.652	40.10
0	.004	0	41.37	5.01	0.930	0.003	38.01	2.06	0.944	0.020
.01		2.5	22.41	11.40	0.977	0.043	13.43	12.79	0.968	1.61
.015		3.75	21.77	23.97	0.981	0.057	12.01	24.53	0.981	2.41
.018		4.5	23.50	44.53	0.977	0.078	10.83	42.94	0.975	2.64
.05		12.5	10	57.00	0.965	0.091	10	57.00	0.965	3.12
.5		125	10	160.39	0.982	0.085	10	201.43	0.652	9.93

Table 6-14 Selection on λ parameter for Breast Cancer Dataset

Sparsity parameter depends on the type of covariance matrix used as an input in our proposed methods. For KDD'99 dataset, we didn't normalize the sample covariance matrix with number of entries for better comparison with ASPCA-BG method. That's why the sparsity parameter, λ is high in the study case of KDD'99 dataset.

			ASPCA-ADMM				ASPCA-ADMMSEQ			
λ	ρ	λ/ρ	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)
0	4	0	168.21	21518	0.962	0.08	97.21	21518	0.962	0.25
100		25	156.66	22142	0.964	0.984	54.49	21839	0.965	82
700		175	140.49	23773	0.967	4.17	41.76	235075	0.967	261
1000		250	149.57	24561	0.965	5.84	41.59	24229	0.965	344
5000		1250	60.16	35072	0.951	17.53	N/A	N/A	N/A	N/A
100	50	2	160.60	22110	0.963	0.14	54.60	21814	0.965	7.82
700		14	142.98	23774	0.967	0.38	41.77	23480	0.967	23.50
1000		20	152.47	24562	0.965	0.49	41.74	24076	0.966	32.03
5000		100	62.23	35076	0.951	1.46	36.76	44343	0.943	80.05

Table 6-15 Selection on λ parameter for KDD'99 Dataset

6.9.3 Penalty parameter - *(ρ)*

Rather than updating the ADMM penalty parameter ρ, we have used some constant value depending on the dataset which are being used. We should dynamically update ρ to converge the algorithm in a faster fashion and also to keep the norms within a constant factor as recommended by [7]. This might exclude hardens of tuning another parameter and might lead to a faster convergence. We continue discussing the sensitivity of the parameter on our subspace. As, we said earlier the impact of ρ cannot be judged alone without λ. But, still if we keep the sparsity parameter to constant and change the penalty parameter, our proposed model shows almost similar AUC which is really nice. The impact on variance is considerable. If we increase the value of step size variance also remains almost constant.

			ASPCA-ADMM				ASPCA-ADMMSEQ			
ρ	λ	λ/ρ	$\|\|V\|\|_{1,1}$	Var	AUC	Time	$\|\|V\|\|_{1,1}$	Var	AUC	Time(s)
0.0005	0.175	350	5.31	19.17	1	0.16	6.48	19.14	1	8.45
0.005		35	5.31	19.17	1	0.01	6.48	19.14	1	0.92
0.05		3.5	5.74	19.17	1	0.004	6.48	19.14	1	0.08
0.5		.35	5.31	19.17	1	0.002	6.48	19.14	1	0.01
0.0005	15	30000	5.31	19.27	1	13.45	N/A	N/A	N/A	N/A
0.005		3000	5.53	19.27	1	1.3	5.28	20.00	1	12.16
0.05		300	5.31	19.27	1	0.13	5.28	20.00	1	1.18
0.5		30	5.31	19.27	1	0.01	5.28	20.00	1	0.11

Table 6-16 Selection on ρ parameter for Synthetic Dataset

			ASPCA-ADMM				ASPCA-ADMMSEQ			
ρ	λ	λ/ρ	$\|V\|_{1,1}$	Var	AUC	Time(s)	$\|V\|_{1,1}$	Var	AUC	Time(s)
0.001	0.015	15	22.50	22.38	0.982	0.246	12.00	24.42	0.981	9.37
0.004		3.75	21.77	23.97	0.981	0.058	12.01	24.53	0.981	2.44
0.005		3	23.23	24.33	0.981	0.044	12.02	24.85	0.981	1.96
0.001	.018	18	12.19	39.79	0.978	0.302	11.08	38.48	0.978	11.17
0.004		4.5	23.50	44.53	0.977	0.098	10.83	42.94	0.975	2.83
0.005		0.36	12.41	44.16	0.977	0.062	10.86	42.55	0.975	2.15
0.5		.036	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 6-17 Selection on ρ parameter for Breast Cancer Dataset

			ASPCA-ADMM				ASPCA-ADMMSEQ			
ρ	λ	λ/ρ	$\|V\|_{1,1}$	Var	AUC	Time	$\|V\|_{1,1}$	Var	AUC	Time(s)
100	700	7	148.54	23594	0.966	0.24	41.80	23458	0.967	15.26
50		14	142.98	23774	0.967	0.40	41.77	23480	0.967	23.70
10		70	155.73	23773	0.967	2	41.77	23510	0.967	103.81
5		140	1.3947	23773	0.967	4.15	41.76	23507	0.967	206.78
4		175	140.49	23773	0.967	4.17	41.76	23507	0.967	261
100	5000	50	64.23	35064	0.951	0.71	36.76	44334	0.943	43.14
50		100	62.23	35076	0.951	1.33	36.76	44343	0.943	81.48
10		500	62.89	35072	0.951	6.70	36.76	44350	0.943	383.21
5		1000	62.16	35072	0.951	12.87	N/A	N/A	N/A	N/A
4		1250	60.16	35072	0.951	17.53	N/A	N/A	N/A	N/A

Table 6-18 Selection on ρ parameter for KDD'99 Dataset

But, we don't want to make a rigid comment whether it is solely depends on sparsity parameter or ADMM penalty parameter. Expectedly, this penalty parameter controls the convergence time for both of our proposed methods which is evident on tables. If we increase the step size ADMM algorithm converges faster for all of our datasets. Choosing a perfect step size depends on the covariance matrix we are using. We have used sample covariance matrix (normalizing by the number of entries) for Synthetic and Breast Cancer dataset and covariance matrix without normalizing for KDD'99 dataset for better comparison with ASPCA-BG method. In a nutshell, step size parameter doesn't effect AUC or variance directly except the convergence time of ADMM process.

6.10 Discussion

According to the experiment results, we can realize the trade-off of sparsity, variance and accuracy. Sparsity introduces additional variance for each principal components. But, we can go back to PCA by setting the sparsity parameter to zero which is really nice. But, increase in variance also introduces another issue of lower accuracy. If explained variance and sparsity increases the accuracy performance of the model goes down. So, AUC score is also kept in consideration while improving the interpretation performance.

Initially our motivation was to find the PCs simultaneously estimating the abnormal subspace and which is done nicely. ASPCA-ADMM and ASPCA-ADMMSEQ has more or less similar performance with respect to anomaly detection. But, the later method has better performance in term of interpretation. The main reason behind is that estimating the whole subspace at a time doesn't emphasize the loading vectors interpretation performance. Rather if we find the PCs one by one, it gives us a better sparse result with better interpretation performance. It doesn't mean that our simultaneous estimation is not sparse. Nevertheless, it is sparse and only drawback is, it cannot capture the most of the correlation of input feature space.

The sequential computation of abnormal PCs by ASPCA-ASDMMDEQ method might mislead readers in terms of computation performance. But, we found that this method performs much faster than ASPCA-BG method of [2]. It even converges faster than ASPCA-ADMM method with better tuning parameter. Say for comparative example, estimating the whole subspace simultaneously ASPCA-ADMM takes thousands of iterations to converge and ASPCA-ADMMSEQ takes may be hundreds of iterations for each PC. Eventually empirical result suggests that somehow the computation time is very close to ASPCA-ADMM method and many times faster than ASPCA-BG method.

The performance of anomaly interpretation is very much meaningful for synthetic dataset which is relatively easy to understand. But, before interpreting we need to have application domain entity knowledge to understand how one abnormal PC can explain anomaly behavior in the dataset. Observation on input feature variables is really important here.

Tuning the perfect parameter for a better result is really a hard task. It is true for both of our proposed methods. But, the initial method of our proposed method is much sensitive to parameters than the second one. We need to keep eyes open on several issues while choosing perfect parameter for the model. To remind you again, our model has three parameters which consists of sparsity, step size and abnormal dimension. So, it is not hard to understand the difficulty of tuning three parameters at the same time. We will discuss about the way to improve its performance regarding tuning in the next chapter and leave it as a future extension.

Chapter 7 CONCLUSIONS

7.1 Summary

We have seen that PCA based anomaly detection models are very easy to implement for detecting anomaly but not suitable for interpretation at all. Therefore, we used the connection of PCA with singular value decomposition (SVD) or eigenvalue decomposition, allowing a noisy input data matrix and extracted the least significant PCs (basis of abnormal subspace). It is solved through a low rank matrix approximation problem using SDP formulation with convex relaxation incorporating orthogonal guarantee on computed PCs. We also introduced sparsity constraint to get a sparse result. Two efficient and effective ADMM algorithms have been proposed to estimate abnormal subspace. The main contribution of this paper is a SDP based abnormal subspace approximation, two efficient algorithms for noisy input data instances is a significant support towards estimating abnormal subspace. We estimated the abnormal subspace in two ways. One method is discovering the abnormal loading vectors sequentially (one by one) and other one is finding the whole abnormal subspace simultaneously (at a time) and represents its basis as sparse, orthogonal loading vectors named as abnormal PCs.

Three tuning parameters and its sensitivity on our methods are discussed elaborately. We designed and evaluated them on two real world datasets and one famous synthetic dataset for our primal purpose. One extra study is done for analyzing clustering performance of our model on a well-known flower dataset.

We have considered three criterion to investigate the performance of our models. We tried to analyze the result with respect to AUC, Variance and Cardinality or Sparsity. Simple thresholding approach is also used in our study to investigate accuracy of true anomaly detection. However, the solemn concern with simple thresholding is that it can misidentify the real important variables. This is still considered as benchmark for any possibly superior method.

Nonetheless, it has been a traditional research topic for years to derive principal components with sparse loadings. But, our methods solely focused on least significant PCs to represent anomalies which is an extraordinary idea. Estimating sparse abnormal subspace in high dimensions poses both computational and statistical challenges. From a practical point of view, a good method to achieve the accuracy and sparseness at the same time keeping the variance at minimal should consider other fundamental properties of incorporating both small and big feature space relative to sample size of the dataset. We know, the ratio of feature space size and sample data instance size is a big issue for many PCA based algorithms. But, in our case it doesn’t pose any significant issue. Furthermore, a perfect model should also avoid misidentifying the important variables. Our methods consider all these significant factors and throws strong challenge to other existing well-known methods.

The two parameters- *sparsity (**λ**)* and *step-size (**ρ**)* which we consider empirically as coupled allows flexible control on the sparse structure of the resulting loading vectors of abnormal subspace. Two efficient iterative ADMM algorithm proposed to compute SDP based problem and outperforms all other existing available solutions in literature. As a principled procedure, our method also enjoys advantages in several aspects, including computational efficiency, explained variance and an ability in identifying important features to interpret anomaly evidently.

In a nutshell, we have presented a solution to convex relaxation of SDP to estimate sparse abnormal subspace, and derived tractable sufficient conditions for convergence and found a solution of the optimization problem. It is worth to be noted that we extensively use the convex relaxation of Fantope projection (ASPCA-ADMM), deflated Fantope projection (ASPCA-ADMMSEQ) and provide idea on sparse extremal eigenvalues.

7.2 Research Contribution and Limitation

In this paper, our interpretable PCA based anomaly detection model using ADMM algorithm achieved the perfect sparse cardinality and orthogonality of the least significant loading vectors. Our model achieved even better outlier detection performance than the traditional PCA model and existing promising models, and provided meaningful interpretation for individual anomalies. The overall method worked especially well for the synthetic and real world dataset, which is often the case in applications where interpretability by a human is significant. Major contributions of our works are mentioned below.

Firstly, very few methods handle the problem of simultaneous finding several sparse and orthogonal loading vectors especially the least ones at a time. According to my knowledge, this is the first work which tried to emphasize on finding all least significant sparse loading vectors concurrently eventually termed as abnormal subspace for outlier detection. The algorithms we proposed has far better computational and time consumption performance than other well-known promising competitive deflated methods and in practice and converges much faster. Even for better interpretation performance, we proposed sequential method assuming that individual loading vectors might have better interpretation result. Thus, we have also investigated sequential computation of abnormal loading vectors. Our experiments on datasets just outperforms that compared to subsequent competitive models.

We have evaluated the performance of our proposed two methods with traditional PCA and ASPCA-BG model on different criterions. We compared norm, variance, cardinality, interpretation performance of each algorithms and presented in a meaningful and clear technique.

In this paper we also tried to use our developed method to analyze IRIS dataset clustering problem. When this model is used as a clustering tool, sparse constraint allowed us to identify the different flowers with the action of only a few variables. The heatmap of all data points on the components demonstrated the individual dependencies on respective PCs. The clustering structure remained visually clear. But, we limited our clustering analysis only using IRIS dataset. Moreover, choosing the perfect dimension size has a significant impact on clustering performance using our proposed model. We need to extend this model exploring other big datasets to make more robust, scalable.

Whether sparse loading vectors are easier to interpret than other methods or not, depends on the dataset we are investigating. Sometimes one is more interested in the low dimensional representation of the data or in other words, in the principal components. It is only in the last case that sparse PCA can have main benefits for the interpretation. This study recommends that using only PCA based models for anomaly interpretation is harder than it appears due to non-sparse structure. In our ongoing work, we are also investigating other techniques that may be able to interpret anomaly in a more robust and effective manner. Furthermore, dataset deficiency is a big limitation in our study. In some cases, we do lack proper understanding of the behavior of data which is eventually is a big factor to interpret individual anomaly types.

7.3 Future Work

Experimenting with various real and synthetic data sets and exploring a range of parameter tuning, our empirical observation is that selecting the appropriate parameter value for abnormal subspace dimension-***d*** is unexpectedly pretty difficult task. Very small change in dimension, ***d*** can have a significant influence on the variance, accuracy and obviously on interpretation. In addition, techniques for automatically estimating the dimension ***d*** are inadequate. In fact, explained variance is a way to choose the parameter but not the sole perfect measure to estimate the dimension parameter, ***d***. In our study, we have chosen ***d*** such that we our FPR is low and AUC is high. So, there is a big room of improvement of tuning ***d*** automatically which solely depends on data instances and its features.

Furthermore, the normal subspace also may in fact become polluted by large anomalies comparing to normal data, which deteriorates the explanations of interpretation. Eventually, for automated, unsupervised detection of anomaly, we need more effective techniques for determining the dimensionality of the abnormal subspace, preventing its contamination for detection.

There are few other open problems and many possible extensions related to this work. For instance, it would be interesting to inspect the performance projecting to other tight convex hulls rather than Fantope and deflated Fantope.

Finally, the choices of penalty parameter ρ and regularization parameter λ are of great empirical interest. Our observation is that regularization parameter can be chosen internally from values of covariance matrix. We are using the soft-thresholding in the second update of ADMM algorithm and the value of λ can be chosen from highest values present in each column of covariance matrix. Moreover, ADMM algorithm can easily be modified to handle the step size parameter also known as Lagrange multiplier. It can be dynamically updated easily in each iteration to converge in an effective way.

Our future extension works can focus mainly in these directions: 1) how to improve interpretation and make it sparser on high dimensional datasets; 2) prove the convergence of ADMM algorithm 3) how to make the model more robust and scalable for different big data problems. And last but not the least 4) make non-parametric algorithm instead of depending on input parameters which is parametric with respect to dimension, sparsity and step size.

References

[1] I. Jolliffe. Principal component analysis. Wiley Online Library, 2002.

[2] X.Bin, Y. Zhao, and B. Shen. Abnormal Subspace Sparse PCA for Anomaly Detection and Interpretation. ACM SIGKDD Workshop on ODDx3: Outlier Definition, Detection, and Description, 2015.

[3] V. Q. Vu, J. Cho, J. Lei, and K. Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In Advances in Neural Information Processing Systems, pp. 2670-2678, 2013.

[4] K. Chen, and J. Lei. Localized Functional Principal Component Analysis. Journal of the American Statistical Association, 110(511), pp.1266-1275, 2015.

[5] C. C. Aggarwal. Outlier analysis. In Data Mining, pp. 237-263. Springer International Publishing, 2015.

[6] Y. Zhang, A. d’Aspremont, and L. E. Ghaoui, L. Sparse PCA: Convex relaxations, algorithms and applications. In Handbook on Semidefinite, Conic and Polynomial Optimization (pp. 915-940). Springer US, 2012.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning, 3(1), pp.1-122, 2011.

[8] A. d'Aspremont, L. E. Ghaoui, Jordan, M.I. jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. SIAM review, 49(3), pp.434-448, 2007.

[9] S. Ma. Alternating direction method of multipliers for sparse principal component analysis. Journal of the Operations Research Society of China, 1(2), pp.253-274, 2013.

[10] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. In ACM SIGMETRICS Performance Evaluation Review (Vol. 35, No. 1, pp. 109-120). ACM, 2007.

[11] Breast Cancer Wisconsin (Diagnostic) Data Set, 1995.

[12] KDD Cup 1999 Data, 1999.

[13] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), p.15, 2009.

[14] H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), pp.265-286, 2006.

[15] L.W.Mackey. Deflation methods for sparse PCA. In Advances in neural information processing systems (pp. 1017-1024), 2009

[16] J. Dattorro. Convex optimization & Euclidean distance geometry. Lulu. com, 2010

[17] R. Luss, and A. d'Aspremont. Clustering and feature selection using sparse principal component analysis. Optimization and Engineering, 11(1), pp.145-157, 2010.

[18] C.M. Bishop. Pattern Recognition. Machine Learning, 12(12.1), pp.561-569, 2006.

[19] X. Xu, and X. Wang. An adaptive network intrusion detection method based on PCA and support vector machines. In Advanced Data Mining and Applications (pp. 696-703). Springer Berlin Heidelberg, 2005.

[20] F.R. Jolliffe. Survey design and analysis. Ellis Horwood, 1986.

[21] Z. Lu, and Y. Zhang. An augmented Lagrangian approach for sparse principal component analysis. Mathematical Programming, 135(1-2), pp.149-193, 2012.

[22] R. Jiang, H. Fei, and J. Huan. Anomaly Localization for Network Data Streams with Graph Joint Sparse PCA. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 886-894). ACM, 2011.

[23] R. Jiang, H. Fei, and J. Huan. A family of joint sparse PCA algorithms for anomaly localization in network data streams. Knowledge and Data Engineering, IEEE Transactions on, 25(11), pp.2421-2433, 2013.

[24] D. Hallac, J. Leskovec, and S. Boyd. Network lasso: Clustering and optimization in large graphs. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 387-396). ACM, 2015, August.

[25] R. Jiang, H. Fei, and J. Huan. Anomaly localization for network data streams with graph joint sparse PCA. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 886-894). ACM, 2011.

[26] W. H. Wolberg, W. Street, D. M. Heisey, and O. L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26(7), pp.792 -796, 1995.

[27] N. Parikh, and S.P. Boyd. Proximal Algorithms. Foundations and Trends in optimization, 1(3), pp.127-239, 2014.

[28] R Fisher, Iris flower dataset.

Acknowledgements

I would like to express my gratitude to my supervisor Professor Wenguang Chen for his support and assistance during the whole period of my Master's program. I would also like to express my thankfulness to my co-supervisor Professor Zhao Ying for her valuable time, kind assistance and guideline throughout the period of this thesis work. And lastly, I appreciate Xingyan Bin for helping me to understand the project from the beginning and providing valuable comments on my work.

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他人和集体，均已在文中以明确方式表明。

签 名：__________ 日 期：______________

Appendix

The code to generate synthetic data using Matlab

```
ee = (1:normal_num)';

a = cos(ee/20);

b = a;

c = cos(ee/10);

d = a+c;

standA = [a b c d];

randA = standA + 0.2*rand(normal_num,4);

e = rand(normal_num,1)*2+cos(ee/2)-1;

f = rand(normal_num,1)*0.1;

g = rand(normal_num,1)*0.1;

randA = [randA e f g];


an1 = randA(11:10+anNum,:);

an1(:,2) = an1(:,2)+(0.6+rand(anNum,1)).*((rand(anNum,1)>0.5)-0.5);

an2 = randA(31:30+anNum,:);

an2(:,4) = an1(:,4)+(0.6+rand(anNum,1)).*((rand(anNum,1)>0.5)-0.5);

an3 = randA(51:50+anNum,:);

an3(:,6) = an1(:,6)+(0.6+rand(anNum,1)).*((rand(anNum,1)>0.5)-0.5);
```

Proposition proof

Given the normal subspace V_{normal} and abnormal $V_{abnormal}$ subspace with orthogonal loading vectors, SPE can be expressed by

$$SPE = \hat{y}^T\hat{y} = \sum_{i=1}^{d}(v_i^T y)^2\ , where\ V_{abnormal=}(v_1, v_2, v_3, .., v_d)$$

$$\begin{aligned} SPE = \hat{y}^T\hat{y} &= y^T V_{abnormal} V_{abnormal}^T V_{abnormal} V_{abnormal}^T y \\ &= (y^T V_{abnormal})(V_{abnormal}^T V_{abnormal})(V_{abnormal}^T y) \\ &= (V_{abnormal}^T y)^T (V_{abnormal}^T y) \\ &= \sum_{i=1}^{d}(v_i^T y)^2 \end{aligned}$$

Soft Thresholding proof

Take the following problem

$$\min_x ||x-b||_2^2 + \lambda||x||_1$$

We know this has 3 unique solutions

I. $\min_x ||x-b||^2 + \lambda x$→ assuming $x \geq 0$

II. $\min_x ||x-b||^2 - \lambda x$→ assuming $x < 0$

III. $\min_x ||x-b||^2$ → assuming $x = 0$

For (i),

$$\frac{\delta}{\delta x}\left(\min_x \left|\left|x-b\right|\right|^2 + \lambda x\right) = 0$$

$$\frac{\delta}{\delta x}(x^2 - 2xb + b^2 + \lambda x) = 0$$

$$2x - 2b + \lambda = 0$$

So,$x = b - \frac{\lambda}{2}$

Similarly for (ii),

$$x = b + \frac{\lambda}{2}$$

By deducting (iii),

$$x = 0$$

(i) holds if $x \geq 0$

$$b - \frac{\lambda}{2} > 0$$

$$b > \frac{\lambda}{2}$$

(ii) only holds if $x < 0$

$$b + \frac{\lambda}{2} < 0$$

$$b < -\frac{\lambda}{2}$$

(iii) only holds if

$$\left|\left|b\right|\right| < \frac{\lambda}{2}\ due\ to\ (i) + (ii)$$

Then we can get the

Resume

Kazi Abir Adnan – Bangladeshi nationality
Address: Flat# 12/A, 33/1 Mirpur Road, New Market, Dhaka – 1205, Bangladesh
Email: i_know1988@yahoo.com

EDUCATION:

- Master in Advanced Computing [2014 – 2016]:-
  Tsinghua University, Haidian District, Beijing – 100084, P.R. China
- Bachelor of Science in Computer Science and Engineering[2008 - 2013]:-
  Bangladesh University of Engineering & Technology, Dhaka, Bangladesh

WORK EXPERIENCE:

- Research Assistant at Pacman Group, Tsinghua University, P.R. China [June, 2015 – July, 2016]
- Teaching Assistant, Department of Computer Science & Technology, Tsinghua University, P.R. China
  Introduction to Big Data Systems (80240693) 2016-Spring
  [March 2016 – July, 2016]
  Combinatorics & Algorithm Design (70240384) 2015-Autumn
  [September, 2015 – January, 2016]
- Software Engineer at Enosis Solutions, Dhaka, Bangladesh
  [February, 2013 – July, 2014]
- Junior Software Engineer (Part Time) at Evatix, Dhaka, Bangladesh
  [March, 2012 – December, 2012]