Anomaly Localization for Network Data Streams with Graph Joint Sparse PCA

Ruoyi Jiang, Hongliang Fei, Jun Huan
Department of Electrical Engineering and Computer Science
University of Kansas
Lawrence, KS 66047, USA
{jiangruoyi, hfei, jhuan}@ittc.ku.edu

ABSTRACT

Determining anomalies in data streams that are collected and transformed from various types of networks has recently attracted significant research interest. Principal Component Analysis (PCA) has been extensively applied to detecting anomalies in network data streams. However, none of existing PCA based approaches addresses the problem of identifying the sources that contribute most to the observed anomaly, or anomaly localization. In this paper, we propose novel sparse PCA methods to perform anomaly detection and localization for network data streams. Our key observation is that we can localize anomalies by identifying a sparse low dimensional space that captures the abnormal events in data streams. To better capture the sources of anomalies, we incorporate the structure information of the network stream data in our anomaly localization framework. We have performed comprehensive experimental studies of the proposed methods, and have compared our methods with the state-of-the-art using three real-world data sets from different application domains. Our experimental studies demonstrate the utility of the proposed methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms, Experimentation

Keywords

Network Data Streams, Anomaly Localization, PCA, Sparse Learning

1. INTRODUCTION

Determining anomalies in data streams that are collected and transformed from various types of networks has recently attracted significant research interest in the data mining community [4, 15, 29, 32]. Applications of the work could be found in network traffic data [32], sensor network streams [4], social networks [29], and finance networks [15] among others.

The common limitation of aforementioned methods is that they are incapable of identifying the sources that contribute most to the observed anomalies, or anomaly localization. With fast-accumulating stream data, an outstanding data analysis issue is *anomaly localization*, where we aim to discover the specific sources that contribute most to the observed anomalies. Anomaly localization in network data streams is apparently critical to many applications, including monitoring the state of buildings [31], or locating the sites for flooding and forest fires [9]. In the stock market, pinpointing the change points in a set of stock price time series is critical for making intelligent trading decisions [25]. For network security, localizing the sources of the most serious threats in computer networks helps ensure security in networks [21].

Principal Component Analysis (PCA) is arguably the most widely applied unsupervised anomaly detection technique for network data streams [12, 21, 22]. However, a fundamental problem of PCA, as claimed in [28], is that the current PCA based anomaly detection methods can not be applied to anomaly localization. Our key observation is that the major obstacle for extending the PCA technique to anomaly localization lies in the high dimensionality of the abnormal space. If we manage to identify a low dimensional approximation of the high dimensional abnormal subspace using a few sources, we "localize" the abnormal sources. The starting point of our investigation hence is the recently studied sparse PCA framework [33] where PCA is formalized in a sparse regression problem where each principle component (PC) is a sparse linear combination of the original sources. However, sparse PCA does not fit directly into our problems in that sparse PCA enforces sparsity randomly in the normal and abnormal subspaces. In this paper, we explore two directions in improving sparse PCA for anomaly detection and localization.

First, we develop a new regularization scheme to simultaneously calculate the normal subspace and the sparse abnormal subspace. In the normal subspace, we do not add any regularization but use the same normal subspace as ordinary PCA for anomaly detection. In the abnormal subspace, we enforce that different PCs share the same sparse structure hence it is able to do anomaly localization. We call this method *joint space PCA* (JSPCA). Second, we observe that abnormal streams are usually correlated to each other. For example in stock market, index changes in differ-

KDD'11, August 21–24, 2011, San Diego, California, USA.

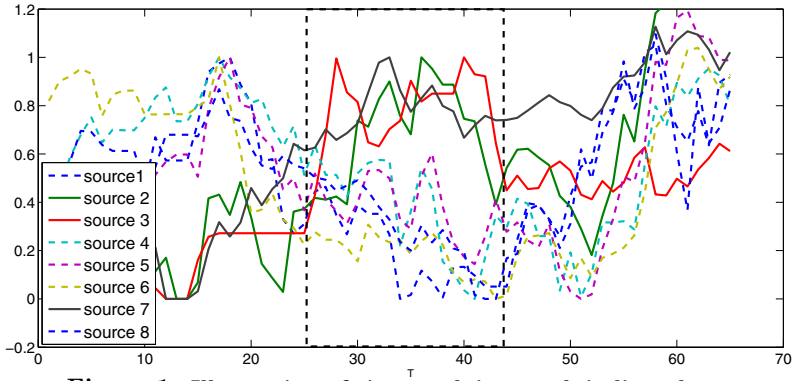


Figure 1: Illustration of time-evolving stock indices data

ent countries are often correlated. For incorporating stream correlation in anomaly localization we design a *graph guided sparse PCA* (GJSPCA) technique. Our experimental studies demonstrate the effectiveness of the proposed approaches on three real-world data sets from financial markets, wireless sensor networks, and machinery operating condition studies.

As an example of anomaly detection and anomaly localization in network data streams, we show the normalized stock index streams of eight countries over a period of three months in Figure 1. We notice an anomaly in the marked window between time stamps 25 and 42. In that window sources 1, 4, 5, 6, 8 (denoted by dotted lines) are normal sources. Sources 2, 3, 7 (denoted by solid lines) are abnormal ones since they have a different trend from that of the other sources. In the marked window, the three abnormal sources clearly share the same increasing trend while the rest share a decreasing trend.

The remainder of the paper is organized as follows. In section 2 we present related work of anomaly localization. In section 3, we discuss the challenge of applying PCA to anomaly localization. In section 4 we introduce the formulation of JSPCA and GJSPCA and the related optimization algorithm. We present our experimental study in section 5 and conclude in section 6.

2. RELATED WORK

Existing work on anomaly localization from network data streams could be roughly divided into two categories: those at the source level and those at the network level. The source level anomaly localization approaches embed detection algorithm at each stream source, resulting in a fully distributed anomaly detection system [10, 23]. The major problem of these approaches is that source level anomalies may not be indicative of network level anomalies due to the ignorance of the rest of the network [12].

To improve source level anomaly localization methods, several algorithms have been recently proposed to localize anomaly at the network level. Brauckhoff [3] applied association rule mining to network traffic data to extract abnormal flows from the large set of candidate flows. Their work is based on the assumption that anomalies often result in many flows with similar characteristics. Such an assumption holds in network traffic data streams but may not be true in other data streams such as finance data. Keogh *et al.* [20] proposed a nearest neighbor based approach to identify abnormal subsequences within univariate time series data by sliding windows. They extracted all possible subsequences and located the one with the largest Euclidean distance from its closest non-overlapping subsequences. However, the method only works for univariate time series generated from a single source. In addition, if the data is distributed on a non-Euclidean manifold, two subsequences may appear deceptively close as measured by their Euclidean distance [30]. L. Fong *et al.* developed a nonparametric change-point test based on U-statistics to detect and localize change-points in high-dimensional network traffic data [26]. The limitation is that the method is specifically designed for the Denial of Service (DOS) attack in communication networks and cannot be generalized to other types of network data streams easily.

Most related to our work, Ide *et al.* [13, 14] measured the change of neighborhood graph for each source to perform anomaly localization and developed a method called Stochastic Nearest Neighbor (SNN). Hirose *et al.* [11] designed an algorithm named Eigen Equation Compression (EEC) to localize anomalies by measuring the deviation of covariance matrix of neighborhood sources. In these two studies, we have to build a neighborhood graph for each source for each time interval, which is unlikely to scale to a large number of sources. In [18], we proposed a two step approach that first computed normal subspace from ordinary PCA and then derived a sparse abnormal subspace on the residual data subtracted from the original data.

In this paper, we design a unified approach to jointly learn normal subspace for anomaly detection and sparse abnormal subspace for anomaly localization, in which we derive a low dimensional approximation of the abnormal subspace by enforcing the loadings of normal sources to vanish across the PCs that span the abnormal subspace. Using three real world data sets across several domains, our experimental studies demonstrate the effectiveness of the proposed method over the state-of-the-art.

3. PRELIMINARIES

We introduce the notations used in this paper and background information regarding PCA and sparse PCA.

3.1 Notation

We use capital letters such as X to denote a matrix and bold lowercase letters such as $\mathbf{x}$ to denote a vector. Greek letters such as λ_1, λ_2 are Lagrangian multipliers. $\langle A, B\rangle$ represents the matrix inner product defined as $\langle A, B\rangle = tr(A^T B)$ where tr represents the matrix trace. Given a matrix X we use x_{ij} to denote the entry of X at the ith row and jth column. We use $x^{(i)}$ to represent the ith entry of a vector $\mathbf{x}$. $||\mathbf{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ denotes the l_p norm of the vector $\mathbf{x} \in \mathbb{R}^n$. Given a matrix $A \in \mathbb{R}^{p\times k}$, $\|A\|_{1,q} = \sum_{i=1}^{p} \|\tilde{\mathbf{a}}_i\|_q =$ is the l_1/l_q norm of the matrix A, where $\tilde{\mathbf{a}}_i$ is the ith row of A. Unless stated otherwise, all vectors are column vectors.

3.2 Network Data Streams

Our work focuses on data streams that are collected from multiple sources. We call the set of data stream sources together as a network since we often have information regarding the structure of the sources.

Following [7], *Network Data Streams* are multi-variate time series $\mathcal{Z}$ from p sources where $\mathcal{Z} = \{Z_i(t)\}$ and $i \in [1, p]$. p is the dimensionality of the network data streams. Each function $Z_i : \mathbb{R} \to \mathbb{R}$ is a *source*. A source is also called a "node" in the communication network community and a "feature" in the data mining and machine learning community.

Typically we focus on time series sampled at (synchronized) discrete time stamps $\{t_1, t_2, \ldots, t_n\}$. In such cases, the network data streams are represented as a matrix $X = (x_{i,j})$ where $i \in [1, n]$, $j \in [1, p]$ and $x_{i,j}$ is the reading of the stream source j at the time sample t_i.

3.3 Applying PCA for Anomaly Localization

Our goal is to explore a Principal Component Analysis (PCA) based method for performing anomaly detection and localization simultaneously. PCA based anomaly detection technique has been widely investigated in [12, 21, 22]. In applying PCA to anomaly detection, one first constructs the normal subspace $\mathbf{V}_1$ by the top k PCs and the abnormal subspace $\mathbf{V}_2$ by the remaining PCs, then projects the original data on $\mathbf{V}_1$ and $\mathbf{V}_2$ as:

$$X = X\mathbf{V}_1\mathbf{V}_1^T + X\mathbf{V}_2\mathbf{V}_2^T = X_n + X_a \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$ is the data matrix with n time stamps from p data sources, X_n and X_a are the projections of X on normal subspace and abnormal subspace respectively. The underlying assumption of PCA based anomaly detection is that X_n corresponds to the regular trends and X_a captures the abnormal behaviors in the data streams. By performing statistical testing on the squared prediction error $SPE = tr(X_a^T X_a)$, one determines whether an anomaly happens [12, 21]. The larger SPE is, the more likely an anomaly exists.

Although PCA has been widely studied for anomaly detection, it is not applicable for anomaly localization. The fundamental problem, as claimed in [28], lies in the fact that there is no direct mapping between two subspaces $\mathbf{V}_1$, $\mathbf{V}_2$ and the data sources. Specifically, let $\mathbf{V}_2 = [\mathbf{v}_{k+1}, \cdots, \mathbf{v}_p]$ be the abnormal subspace spanned by the last $p-k$ PCs, X_a is essentially an aggregated operation that performs linear combination of all the data sources, as follows:

$$\begin{aligned} X_a &= X\mathbf{V}_2\mathbf{V}_2^T \\ &= \left[\sum_{j=1}^{p} \mathbf{x}_j\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_1^T, \cdots, \sum_{j=1}^{p} \mathbf{x}_j\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_i^T, \cdots, \sum_{j=1}^{p} \mathbf{x}_j\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_{p-k}^T\right] \end{aligned} \tag{2}$$

where $\mathbf{x}_j$ is the data from the jth source and $\tilde{\mathbf{v}}_j$ is the jth row of $\mathbf{V}_2$. Considering the ith column of X_a with the value $\sum_{j=1}^{p} \mathbf{x}_j\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_i^T$, there is no correspondence between the original ith column of X and ith column of X_a. Such an aggregation makes PCA difficult to identify the particular sources that are responsible for the observed anomalies.

Although all the previous works claim PCA based anomaly detection methods *cannot* do localization, we solve the problem of anomaly localization in a reverse way. Instead of locating the anomalies directly, we filter normal sources to identify anomalies by employing the fact that normal subspace captures the general trend of data and normal sources have little or no projection on abnormal subspace. The following theorem provides a necessary condition for data sources to have no projection on abnormal subspace.

THEOREM 3.1. *Suppose $\mathbb{Z} = \{i|\tilde{v}_i = \mathbf{0}^T\}$ is the set that contains all the indices for the zero rows of $\mathbf{V}_2$, then $\forall t \in \mathbb{Z}$, $\mathbf{x}_t$ has no projection on the abnormal subspace. In other words, these sources have no contribution to the abnormal behavior.*

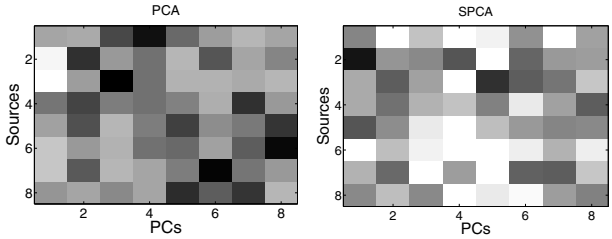
PROOF. Consider the squared prediction error $SPE = tr(X_a^T X_a)$ and plug equation 2 in:

$$\begin{aligned} tr(X_a^T X_a) &= tr(X_a X_a^T) \\ &= tr(\mathbf{V}_2^T X^T X \mathbf{V}_2) \\ &= tr((\textstyle\sum_{j=1}^{p} \mathbf{x}_j\tilde{\mathbf{v}}_j)^T(\sum_{j=1}^{p} \mathbf{x}_j\tilde{\mathbf{v}}_j)) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} tr(\tilde{\mathbf{v}}_i^T\mathbf{x}_i^T\mathbf{x}_j\tilde{\mathbf{v}}_j) \\ &= \sum_{i\notin\mathbb{Z}}\sum_{j\notin\mathbb{Z}} (\mathbf{x}_i^T\mathbf{x}_j\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_i^T) \end{aligned} \tag{3}$$

From equation (3), it is clear that $\forall i \in \mathbb{Z}$, the data $\mathbf{x}_i$ from source i has no projection on abnormal subspace and hence would be excluded from the statistics used for anomaly detection. We call such a pattern with an entire row with zeros *"joint sparsity"*. □

Unfortunately ordinary PCA does not guarantee any sparsity in PCs. Sparse PCA is a recently developed algorithms where each PC is a sparse linear combination of the original sources [33]. However existing sparse PCA method has no guarantee that different PCs share the same sparse representation and hence has no guarantee for the joint sparsity. To illustrate the point, we plotted the entries of each PC for ordinary PCA (figure 2(a)) and for sparse PCA (figure 2(b)) for the stock data set shown in figure 1. White blocks indicate zero entries and the darker color indicates a larger absolute loading. Sparse PCA produces sparse entries but that alone does not indicate sources that contribute most to the observed anomaly.

Below we present our extensions of PCA that enable us to perform anomaly localization efficiently.



(a) PCA (b) SPCA

Figure 2: Comparing PCA and Sparse PCA.

4. METHODOLOGY

In this section, we propose a novel regularization framework called joint sparse PCA (JSPCA) to enforce joint sparsity in PCs in the abnormal space while preserving the PCs in the normal subspace so that we can perform simultaneous anomaly detection and anomaly localization. Then we consider the network topology in the original data and incorporate such topology into JSPCA and develop an approach named Graph JSPCA (GJSPCA).

4.1 Joint Sparse PCA

Our objective here is to derive a set of PCs $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_3]$ such that $\mathbf{V}_1$ is exactly the ordinary normal subspace and $\mathbf{V}_3$ is a sparse approximation of the ordinary abnormal subspace with the joint sparsity. The following regularization framework guarantees the two properties simultaneously.

$$\begin{array}{ll} \min\limits_{\mathbf{V}_1,\mathbf{V}_3} & \frac{1}{2}||\mathbf{X} - \mathbf{X}\mathbf{V}_1\mathbf{V}_1^T - \mathbf{X}\mathbf{V}_3\mathbf{V}_3^T||_F^2 + \lambda||\mathbf{V}_3||_{1,2} \\ \text{s.t.} & \mathbf{V}^T\mathbf{V} = I_{p\times p} \end{array} \tag{4}$$

Equation (4) can be simplified with one variable $\mathbf{V}$:

$$\begin{array}{ll} \min\limits_{\mathbf{V}} & \frac{1}{2}||\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T||_F^2 + \lambda||\mathbf{W} \circ \mathbf{V}||_{1,2} \\ \text{s.t.} & \mathbf{V}^T\mathbf{V} = I_{p\times p} \end{array} \tag{5}$$

Here $\circ$ is the *Hadamard product* operator (entry-wise product), λ is a scalar controlling the balance between sparse and fitness, $W = [\tilde{\mathbf{w}}_1, \cdots, \tilde{\mathbf{w}}_p]^T$ with jth row vector:

$$\tilde{\mathbf{w}}_j = [\underbrace{0, \cdots, 0}_{k}, \underbrace{1, \cdots, 1}_{p-k}]^T, \quad j = 1, \cdots, p \tag{6}$$

The regularization term $\|W \circ V\|_{1,2}$ is a L_1/L_2 penalty which enforces joint sparsity for each source across in the abnormal subspace spanned by the remaining $p-k$ principal components.

The major disadvantage of equation (5) is that it poses a difficult optimization problem since the first term (the trace norm) is concave and the second term (the L_1/L_2 norm) is convex. Motivated by the formalization of sparse PCA using two variables and an alternative optimization algorithm in [33], we consider a relaxed version:

$$\begin{array}{ll} \min\limits_{A,B} & \frac{1}{2}||X - XBA^T||_F^2 + \lambda||\mathbf{W} \circ \mathbf{B}||_{1,2} \\ \text{s.t.} & A^TA = I_{p\times p} \end{array} \tag{7}$$

and with the vector form:

$$\begin{array}{ll} \min\limits_{A,B} & \frac{1}{2}\sum\limits_{i=1}^{n}||\tilde{\mathbf{x}}_i - AB^T\tilde{\mathbf{x}}_i||_2^2 + \lambda\sum\limits_{j=1}^{p}||\tilde{\mathbf{w}}_j \circ \tilde{\mathbf{b}}_j||_2 \\ \text{s.t.} & A^TA = I_{p\times p} \end{array} \tag{8}$$

where $\tilde{\mathbf{x}}_i$ is the ith row of X, $A, B \in \mathbb{R}^{p\times p}$ and $\tilde{\mathbf{b}}_j$ is the jth row of B. The advantage of the new formalization is two folds: first, equation (8) is convex to each subproblem when fixing one variable and optimizing the other. As asserted in [33] disregarding the Lasso penalty, the solution of equation (8) corresponds to exact PCA; second, we only impose penalty on the remaining $p-k$ PCs and preserve the top k PCs representing the normal subspace from ordinary PCA. Such a formalization will guarantee that we have the ordinary normal subspace for anomaly detection and the sparse abnormal subspace for anomaly localization. Note that Jenatton *et al.* recently proposed a structured sparse PCA [16], which is similar to our formalization. But their structure is defined on groups and cannot be directly applied for anomaly localization.

Figure 4(a) demonstrates the principal components generated from JSPCA for the stock market data shown in figure 1. Joint sparsity across the PCs in abnormal subspace pinpoints the abnormal sources 2,3,7 by filtering out normal sources 1, 4, 5, 6, 8. Such result matches the truth in figure 1.

4.2 Graph Guided Joint Sparse PCA

In many real-world applications, the sources generating the data streams may have structure, which may or may not change with time. As the example mentioned in figure 1, stock indices from source 2, 3 and 7 are closely correlated over a long time interval. If source 2 and 3 are anomalies as demonstrated in Figure 4(a), it is very likely that source 7 is an anomaly as well. This observation motivates us to develop a regularization framework that enforce smoothness across features. In particular, we model the structure among sources with an undirected graph, where each node represents a source and each edge encodes a possible structure relationship. We hypothesize that by incorporating structure information of sources we can build a more accurate and reliable anomaly localization model. Below, we introduce the graph guided *joint sparse* PCA, which effectively encodes the structure information in the anomaly localization framework.

To achieve the goal of smoothness of features, we add an extended l_2 (Tikhonov) regularization factor on the graph laplacian regularized matrix norm of the $p-k$ PCs. This is an extension of the l_2 norm regularized Laplacian on a single vector in [8]. With this addition, we obtain the following optimization problem:

$$\begin{array}{ll} \min\limits_{A,B} & \frac{1}{2}\sum\limits_{i=1}^{n}||\tilde{\mathbf{x}}_i - AB^T\tilde{\mathbf{x}}_i||_2^2 + \lambda_1\sum\limits_{j=1}^{p}||\tilde{\mathbf{w}}_j \circ \tilde{\mathbf{b}}_j||_2 + \\ & \frac{1}{2}\lambda_2\sum\limits_{i=k+1}^{p}\mathbf{b}_i^TL\mathbf{b}_i \\ \text{s.t.} & A^TA = I_{p\times p} \end{array} \tag{9}$$

where $\mathbf{b}_i$ is the ith column of matrix B, $\tilde{\mathbf{w}}_j$ is defined in (6) and L is the *Laplacian* of a graph that captures the correlation structure of sources [8].

By introducing W matrix defined in (6), we simplify the second regularization term with a matrix format:

$$\begin{array}{ll} \min\limits_{A,B} & \frac{1}{2}\sum\limits_{i=1}^{n}||\tilde{\mathbf{x}}_i - AB^T\tilde{\mathbf{x}}_i||_2^2 + \lambda_1\sum\limits_{j=1}^{p}||\tilde{\mathbf{w}}_j \circ \tilde{\mathbf{b}}_j||_2 + \\ & \frac{1}{2}\lambda_2 tr((W \circ B)^TL(W \circ B)) \\ \text{s.t.} & A^TA = I_{p\times p} \end{array} \tag{10}$$

In figure 4 we show the comparison of applying JSPCA and GJSPCA on the data shown in figure 1. Both JSPCA and GJSPCA correctly localize the abnormal sources 2,3,7. Comparing JSPCA and GJSPCA, we observe that in GJSPCA the entry values corresponding to the three abnormal sources 2,3,7 are closer (a.k.a. smoothness in the feature space). In the raw data, we observe that sources 2,3,7 share an increasing trend. The smoothness is the reflection of the shared trend and helps highlight the abnormal source 7. As evaluated in our experimental study, GJSPCA outperforms JSPCA. We believe that the additional structure information utilized in GJSPCA helps.

4.3 Anomaly Scoring

To quantitatively measure the degree of anomalies for each source, we define anomaly score and normalized anomaly score as following.

DEFINITION 4.1. *Given p sources, the normal subspace $V_1 = [\mathbf{v}_1, \cdots, \mathbf{v}_k]$ and the abnormal subspace $V_3 = [\mathbf{v}_{k+1}, \cdots, \mathbf{v}_n]$, the anomaly score for source i, $\quad i = 1 \cdots p$ is defined on the*

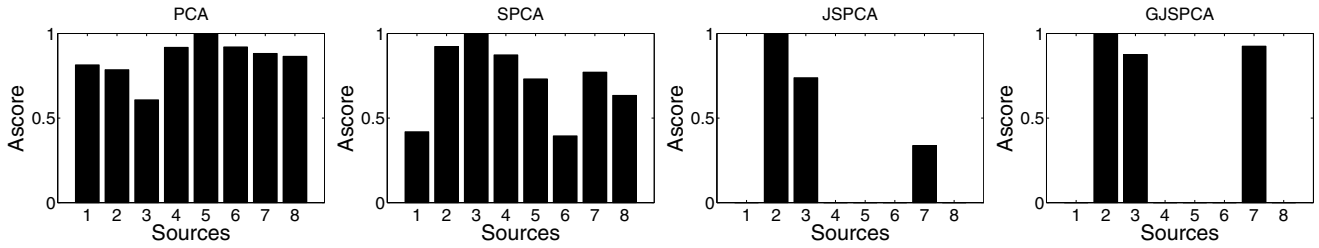


Figure 3: Comparing different anomaly localization methods. From left to right: PCA, sparse PCA, JSPCA, and GJSPCA.

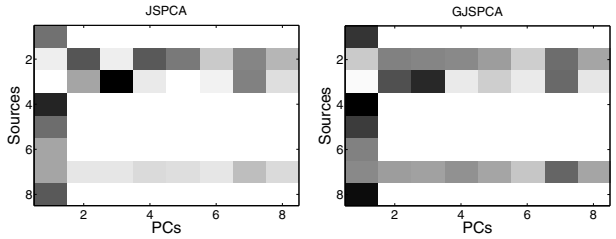
L_1 norm of the ith row of V_3, divided by the size of the row:

$$\zeta_i = \frac{||\tilde{v_i}||_1}{p-k} \tag{11}$$

where $\tilde{v_i}$ is the ith row of V_3. For each input data matrix X, (11) results in a vector $\zeta = [\zeta_1, \cdots, \zeta_p]^T$ of anomaly scores. The normalized score for source i is defined as:

$$\tilde{\zeta_i} = \zeta_i / \max\{\zeta_i, i = 1, \cdots p\}$$

A higher score indicates a higher probability that a source is abnormal. We show the anomaly scores obtained from PCA, SPCA, JSPCA and GJSPCA for the stock data in figure 3. JSPCA and GJSPCA both succeed to localize three anomalies by assigning nonzero scores to anomalous sources and zero to normal ones, while PCA and SPCA both fail. Comparing JSPCA and GJSPCA we find that JSPCA assigns higher anomaly scores to source 2 and 3 but a lower score to source 7, and GJSPCA has smooth effect on the abnormal scores. It assigns similar scores for the three sources. The similar scores demonstrate the effect of smooth regularization term induced by the graph Laplacian. The smoothness also sheds light on the reason why GJSPCA outperforms JSPCA a little in anomaly localization in our detailed experimental evaluation.



(a) JSPCA (b) GJSPCA

Figure 4: Comparing *joint sparse PCA* (JSPCA) and *graph joint sparse PCA* (GJSPCA).

4.4 Optimization Algorithms

We present our optimization technique to solve equations (8) and (10) based on accelerated gradient descent [27] and projected gradient scheme [2]. Although equations (8) and (10) are not joint convex for A and B, they are convex for A and B individually. The algorithm solves A, B iteratively and achieves a local optimum.

A **given** B: If B is fixed, we obtain the optimal A analytically. Ignoring the regularization part, equation (8) and equation (10) degenerate to

$$\begin{array}{ll} \min\limits_{A} & \frac{1}{2}||X - XBA^T||_F^2 \\ \text{s.t.} & A^TA = I_{p\times p} \end{array} \tag{12}$$

The solution is obtained by a reduced rank form of the Procrustes Rotation. We compute the SVD of GB to obtain the solution where $G = X^TX$ is the gram matrix.

$$\begin{array}{rcl} GB & = & UDV^T \\ \hat{A} & = & UV^T \end{array} \tag{13}$$

Solution in the form of Procrustes Rotation is widely discussed, see [33] for example for a detailed discussion.

B **given** A: If A is fixed, we consider equation (10) only since equation (8) is a special case of equation (10) when $\lambda_2 = 0$, Now the optimization problem becomes:

$$\begin{array}{ll} \min\limits_{A,B} & \frac{1}{2}\sum\limits_{i=1}^{n}||\tilde{\mathbf{x}}_i - AB^T\tilde{\mathbf{x}}_i||_2^2 + \lambda_1\sum\limits_{j=1}^{p}||\tilde{\mathbf{w}}_j \circ \tilde{\mathbf{b}}_j||_2 + \\ & \frac{1}{2}\lambda_2 tr((W\circ B)^T L(W\circ B)) \end{array} \tag{14}$$

Equation (14) can be rewritten as $\min\limits_{B} \; F(B) \stackrel{\text{def}}{=} f(B) + R(B)$, where $f(B)$ takes the smooth part of equation(14)

$$f(B) = \frac{1}{2}\sum_{i=1}^{n}||\tilde{\mathbf{x}}_i - AB^T\tilde{\mathbf{x}}_i||_2^2 + \frac{1}{2}\lambda_2 tr((W\circ B)^T L(W\circ B)) \tag{15}$$

and $R(B)$ takes the nonsmooth part, $R(B) = \lambda_1\sum_{j=1}^{p}||\tilde{\mathbf{w}}_j \circ \tilde{\mathbf{b}}_j||_2$. It is easy to verify that (15) is a convex and smooth function over B with Lipschitz continuous gradient and the gradient of f is: $\nabla f(B) = G(B - A) + \lambda_2 L(W \circ B)$.

Considering the minimization problem of the smooth function $f(B)$ using the first order gradient descent method, it is well known that the gradient step has the following update at step $i + 1$ with step size $1/L_i$:

$$B_{i+1} = B_i - \frac{1}{L_i}\nabla f(B_i) \tag{16}$$

In [1, 27], it has shown that the gradient step equation (16) can be reformulated as a linear approximation of the function f at point B_i regularized by a quadratic proximal term as $B_i = \underset{B}{\text{argmin}}\; f_{L_i}(B, B_i)$, where

$$f_{L_i}(B, B_i) = f(B_i) + \langle B - B_i, \nabla f(B_i)\rangle + \frac{L_i}{2}\|B - B_i\|_F^2 \tag{17}$$

Based on the relationship, we combine equations (17) and $R(B)$ together to formalize the *generalized gradient update step*:

$$\begin{array}{rcl} Q_{L_i}(B, B_i) & = & f_{L_i}(B, B_i) + \lambda_1||W \circ B||_{1,2} \\ q_{L_i}(B_i) & = & \underset{B}{\text{argmin}}\; Q_{L_i}(B, B_i) \end{array} \tag{18}$$

The insight of such a formalization is that by exploring the structure of regularization $R(.)$ we can easily solve the optimization in equation (18), then the convergence rate is the

same as that of gradient decent method. Rewriting the optimization problem in equation(18) and ignoring terms that do not depend on B, the objective can be expressed as:

$$q_{L_i}(B_i) = \underset{B \in \mathcal{M}}{\operatorname{argmin}} \left(\frac{1}{2}\|B-(B_i-\frac{1}{L_i}\nabla f(B_i))\|_F^2+\frac{\lambda_1}{L_i}||W\circ B||_{1,2}\right) \tag{19}$$

With ordinary first order gradient method for smooth problems, the convergence rate is $O(1/\sqrt{\epsilon})$ [27] where ϵ is the desired accuracy. In order to have a better convergence rate, we apply the Nestrerov accelerated gradient descent method [27] with $O(1/\sqrt{\epsilon})$ convergence rate, and solve the *generalized gradient update step* in equation (18) for each gradient update step. Such a procedure has demonstrated scalability and fast convergence in solving various sparse learning formulations [6, 17, 24]. Below we present the accelerated projected gradient algorithm. The stopping criterion is that the change of the objective values in two successive steps is less than a predefined threshold (e.g. 10^{-5}).

Algorithm 1 Accelerated Projected Gradient Descent

Input: $B_0, W \in \mathbb{R}^{p\times p}$, $L_1 > 0$, $F(.), Q_L(.,.)$ and max-iter.
Output: B.
Initialize $B_1 := B_0, t_{-1} := 0, t_0 := 1$;
for $i = 1$ to max-iter **do**
  $\alpha_i := (t_{i-2} - 1)/t_{i-1}$;
  $S := B_i + \alpha_i(B_i - B_{i-1})$;
  while (true) **do**
    Compute $q_{L_i}(S)$ in Eq. (19);
    if $F(q_{L_i}(S)) > Q_{L_i}(q_{L_i}(S), S)$ **then**
      $L_i := 2 \times L_i$;
    else
      break;
    end if
  end while
  $B_{i+1} := q_{L_i}(S), L_{i+1} := L_i$;
  $t_i := \frac{1}{2}(1 + \sqrt{1 + 4t_{i-1}^2})$;
  if (Convergence) **then**
    $B := B_{i+1}$, break;
  end if
end for
return B;

Now we focus on how to solve the generalized gradient update in equation (19). Let $C = B_i - \frac{1}{L_i}\nabla f(B_i)$ and $\tilde{\lambda} = \lambda_1/L_i$, equation (19) can be represented as:

$$\begin{aligned} q_{L_i}(B_i) &= \underset{B}{\operatorname{argmin}}(\tfrac{1}{2}||B - C||_F^2 + \tilde{\lambda}||W \circ B||_{1,2}) \\ &= \underset{\tilde{\mathbf{b}}_1,\cdots,\tilde{\mathbf{b}}_p}{\operatorname{argmin}} \textstyle\sum_{j=1}^{p}(\tfrac{1}{2}||\tilde{\mathbf{b}}_j - \tilde{\mathbf{c}}_j||_2^2 + \tilde{\lambda}||\tilde{\mathbf{w}}_j \circ \tilde{\mathbf{b}}_j||_2) \end{aligned} \tag{20}$$

where $\tilde{\mathbf{b}}_j^T, \tilde{\mathbf{c}}_j^T$ and $\tilde{\mathbf{w}}_j^T \in \mathbb{R}^p$ are row vectors denoting the jth row of matrices B, C and W. By the additivity of equation (20), we decompose equation (20) into p subproblems. For each subproblem, we ignore the row index j:

$$\min_{\mathbf{b}} \quad \frac{1}{2}||\mathbf{b} - \mathbf{c}||_2^2 + \tilde{\lambda}||\mathbf{w} \circ \mathbf{b}||_2 \tag{21}$$

For simplicity, we assume $\mathbf{b}, \mathbf{c}$ and $\mathbf{w}$ are row vectors here. The following theorem provides the analytical solution of equation (21).

THEOREM 4.1. *Given* $\tilde{\lambda}, \mathbf{w} = [\mathbf{0}_{1\times k}, \mathbf{1}_{1\times(p-k)}]$ *and* $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2]$ *where* $\mathbf{c}_1 = [c_1, \cdots, c_k]$, $\mathbf{c}_2 = [c_{k+1}, \cdots, c_p]$ *and* k *is the number of PCs representing the normal subspace, the optimal solution for (21)* $\mathbf{b}^* = [\mathbf{b}_1^*, \mathbf{b}_2^*]$ *is given by:*

$$\mathbf{b}_1^* = \mathbf{c}_1$$

and

$$\mathbf{b}_2^* = \begin{cases} (1 - \frac{\tilde{\lambda}}{||\mathbf{c}_2||_2})\mathbf{c}_2 & ||\mathbf{c}_2||_2 > \tilde{\lambda} \\ 0 & otherwise \end{cases} \tag{22}$$

PROOF. By the definition of the l_2 norm, the equation (21) can be rewritten as:

$$\min_{\mathbf{b}_1,\mathbf{b}_2} \frac{1}{2}||\mathbf{b}_1 - \mathbf{c}_1||_2^2 + \frac{1}{2}||\mathbf{b}_2 - \mathbf{c}_2||_2^2 + \tilde{\lambda}||\mathbf{b}_2||_2 \tag{23}$$

where $\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2]$. The solution can be found by decomposing (23) into two subproblems and solving one ordinary least square problem and one least square problem with l_2 norm regularization. Since there is no regularization on $\mathbf{b}_1$ and the two subproblems are independent, the optimal solution of the ordinary least square problem is $\mathbf{b}_1^* = \mathbf{c}_1$. With optimal $\mathbf{b}_1^*$, (23) degenerates to

$$\min_{\mathbf{b}_2} \frac{1}{2}||\mathbf{b}_2 - \mathbf{c}_2||_2^2 + \tilde{\lambda}||\mathbf{b}_2||_2 \tag{24}$$

The analytical solution of equation (24) is given in equation (22) and can be found by forming Lagrangian dual. A detailed proof can be found in [24]. □

We summarize what is briefly discussed previously for GJSPCA in the algorithm below. Note that JSPCA is a special case of GJSPCA, we obtain the algorithm for JSPCA by setting $\lambda_2 = 0$. Given data matrix $X \in \mathbb{R}^{n\times p}$ and the number of PCs representing normal subspace k and regularization parameters λ_1, λ_2, GJSPCA optimizes two matrix variables alternatively and returns the matrix B composed of ordinary PCs representing normal subspace and joint sparse PCs representing the abnormal subspace.

Algorithm 2 Graph Joint Sparse PCA (GJSPCA)

Input: X, k, λ_1, λ_2 and max_iter.
Output: B.
$A := I_{p\times p}, G := X^T X$;
for $iter = 1$ to max_iter **do**
  Compute B given A using Algorithm 1;
  Compute A given B via (13);
  if (Converge) **then**
    break;
  end if
end for
return B;

5. EXPERIMENTAL STUDIES

We have conducted extensive experiments with three real-world data sets to evaluate the performances of JSPCA and GJSPCA on anomaly localization. We implemented our version of two anomaly localization methods at the network level: stochastic nearest neighbor (SNN) [14] and eigen equation compression (EEC) [11] since no executables were provided by the original authors. We implemented all four methods with Matlab and performed all experiments on a desktop machine with 6 GB memory and a Intel core i7 2.66 GHz CPU.

5.1 Data Sets

We used three real-world data sets from different application domains. For each data set, we singled out several intervals with anomalies. The anomalies are either labeled by the original data provided or manually labeled by ourselves when no labeling is provided. Note that we are only interested in the intervals where anomalies really exist since we focus on localizing anomalies. We used a sliding window with fixed size L and offset $L/2$ to create multiple data windows from the given intervals. The sliding window moves forward with the offset $L/2$ until it reaches the end of the intervals. We run all four methods on each data window to evaluate and compare their performances.

To run GJSPCA we calculated the pair-wise correlation between any two sources within the window. We produced a correlation graph for the data streams with a correlation threshold δ in that if the correlation between two sources is greater than δ, we connect the two sources with an edge. This construction is meaningful because for highly correlated data, streams influence each other and such influence has been shown critical for better anomaly localization, as evaluated in our experimental studies.

Below we briefly discuss the data collection and data preprocessing procedures for the three data sets. In Table 1, we list the intervals that we selected, the dimensionality of the network data streams, the sliding window size L, and the total number of data windows W for each data set.

Table 1: Characteristics of Data Sets. D: Data sets. D1: Stock Indices, D2: Sensor, D3: MotorCurrent. T: total number of time stamps, p: dimensionality of the network data streams, I: total number of intervals, $Indices$: starting point and ending point of the intervals, W: total number of data windows, L: sliding window size.

D	T	p	I	$Indices$	W	L
D1	2396	8	4	[261-300], [361- 400] [761-800], [1631-1670]	12	20
D2	11000	7	4	[2371-2530],[3346-3550] [7191-7215], [8841-8870]	37	20
D3	1500	20	1	[1-1500]	29	50

The Stock Indices Data Set: The stock indices data set includes 8 stock market index streams from 8 countries: Brazil (Brazil Bovespa), Mexico (Bolsa IPC), Argentina (MERVAL), USA (S&P 500 Composite), Canada (S&P TSX Composite), HK (Heng Seng), China (SSE Composite), and Japan (NIKKEI 225). Each stock market index stream contains 2396 stamps recording the daily stock price indices from January 1st 2001 to March 5th 2010.

Since this data set has no ground truth, we manually labeled all the daily indices for the selected intervals. In our labeling we followed the criteria list in [5] where small turbulence and co-movements of most markets are considered as normal, dramatic price changes or significance deviation from the co-movements (e.g. one index goes up while the others in the market drop down) are considered as abnormal.

The Sun Spot Sensor Data Set: We collected a sensor data set in a car trial for transport chain security validation using seven wireless Sun Small Programmable Object Technologies (SPOTs). Each SPOT contains a 3-axis accelerometer sensor. In our data collection, seven Sun SPOTs were fixed in the separated boxes and were loaded on the back seat of a car. Each Sun SPOTs recorded the magnitude of accelerations along x, y, z axis with a sample rate of 390ms. We simulated a few abnormal events including box removal and replacement, rotation and flipping. The overall acceleration $\sqrt{(x^2+y^2+z^2)}$ was used to detect the designed anomalous events.

The Motor Current Data Set: The Motor Current Data is the current observation generated by the state space simulations available at UCR Time Series Archive [19]. The anomalies are the simulated machinery failure in different components of a machine. The current value was observed from 21 different motor operating conditions, including one healthy operating mode and 20 faulty modes. For each motor operating condition, 20 time series were recorded with a length of 1,500 samples. Therefore, there are 20 normal time series and 400 abnormal time series altogether.

In our evaluation, we randomly extracted 20 time series out of 420 with the length 1500. 10 time series are from normal series and the rest are from abnormal series.

5.2 Model Evaluation

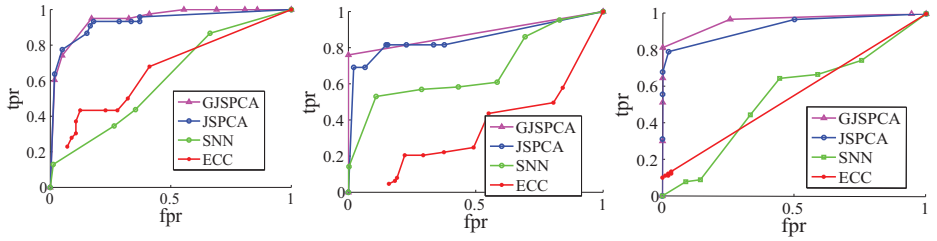
For evaluation, since our focus is anomaly localization, we did not evaluate anomaly detection although our methods are able to do both. We used the standard ROC curves and area under ROC curve (AUC) to evaluate the anomaly localization performance. Below we introduce the details regarding the construction of ROC curves.

As defined in equation 11, a higher abnormal score indicates a higher probability the source is abnormal, which is the same as that of the baseline methods [11, 14] for comparison. To have a fair comparison, we compared the normalized abnormal score among each method. The reason for normalization is that the anomaly scores generated by the baseline methods have different orders of magnitude. We used the term "anomaly score" to refer to the normalized abnormal score in the following analysis.

For each data window, the abnormal score vector $\tilde{\zeta} = [\tilde{\zeta}_1, \cdots, \tilde{\zeta}_p]^T$ was generated and compared with a cut-off threshold between $[0, 1]$ to separate abnormal sources and innocent sources. We performed the same procedure for all the data windows and finally we obtained a prediction matrix with size w by p, such that w is the number of data window and p is the number of sources. Each entry in the prediction matrix is 0 or 1 to indicate whether the source is normal or abnormal. Comparing the prediction matrix with the ground truth resulted in a pair of true positive rate (TPR) and false positive rate (FPR), where TPR is the total number of true detected abnormal sources over the total number of abnormal sources, and FPR is the total number of incorrect detected abnormal sources over the total number of normal sources in W windows. By changing the threshold, we obtained the ROC curve and the AUC value.

5.3 Anomaly Localization Performance

We have two parameters to tune in JSPCA: λ_1: controlling the sparsity, and k: the dimension of normal subspace. GJSPCA has two more parameters: λ_2: controlling the smoothness, and δ, the correlation threshold to construct the correlation graph. For the other two methods, we need to select the number of neighbors k for SSN and the number of clusters c for EEC. We first performed a grid search for each method to identify the optimal parameters and then



Dataset \ Method (AUC)	SNN	ECC	JSPCA	GJSPCA
Stock Indices	0.6119	0.6510	0.9316	0.9457
Sensor	0.6783	0.3491	0.8527	0.8798
MotorCurrent	0.5444	0.5515	0.9273	0.9601

Figure 5: ROC curves and AUC for different methods on three data sets. From left to right: ROC for the stock indices data, ROC for the sensor data, ROC curve for MotorCurrent data, AUC for the three ROC plots

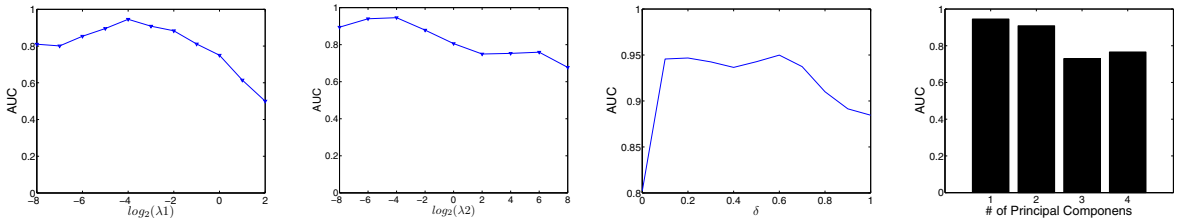


Figure 6: From left to right, sensitivity analysis on λ_1, λ_2, δ, and the dimension of the normal subspace.

compared the performances. The performances of different methods depend on the parameter selection. We evaluated the sensitivity of our results in the next selection.

For each data set, we tuned λ_1, λ_2 within $\{2^{-8}, 2^{-7}, \cdots, 2^{8}\}$ and δ from 0.1 to 0.9. k was tuned from 1 to 4 for the stock market and sensor data, and from 2 to 7 for the motor current data. All the ranges were set by empirical knowledge. Our empirical study showed that the performances did not change significantly as the parameters vary in a wide range, which reduced the parameter search space significantly.

Table 2 lists the best parameter combination for JSPCA and GJSPCA. For SNN, we tuned the number of neighbors k in the range $2 \sim 6$ (for stock index data set and sensor data) and in the range $2 \sim 10$ (for motorcurrent data) respectively. For EEC method, the number of clusters c was tuned between $2 \sim 4$.

In Figure 5, we show the performances for the four methods on three different data sets. JSPCA and GJSPCA clearly outperform the other two methods. The AUC value of JSPCA and GJSPCA are both above 0.85 on three data sets, while that of EEC and SNN are around $[0.5 \sim 0.6]$. Compared with JSPCA, GJSPCA is slightly better, which supports our hypothesis on the importance of incorporating the structure information of network data streams into anomaly localization. SNN clearly outperforms EEC on Sensor data, and is comparable with EEC for the other two data sets.

5.4 Parameter Selection

In this section, we evaluated the sensitivity of our methods to different modeling parameters. In order to do so, we selected one parameter at a time, systematically changed its value while fixing the others at their optimal values. Although our approaches have more parameters than the other two methods, the sensitivity analysis shows that performances of our methods are remarkably stable over a wide range of parameters. Next we show the sensitivity study on the stock indices data set for the parameters λ_1 and λ_2, δ, k. Similar results are observed on the other two data sets.

Table 2: Optimal parameters combinations on three data sets. J*:JSPCA, GJ*: GJSPCA.

	λ_1		k		λ_2	δ
Data set	J*	GJ*	J*	GJ*	GJ*	GJ*
Stock Indices	2^{-3}	2^{-4}	1	1	2^{-4}	0.6
Sensor	2^{-7}	2^{-5}	1	1	2^{-6}	0.6
MotorCurrent	2^{-2}	2^{-2}	5	5	2^{-8}	0.5

In Figure 6, we show the stability by changing λ_1. We observe that AUC is quite stable over a wide range of λ_1. A similar phenomenon is also observed when changing λ_2. In the middle part of figure 6, we performed sensitivity analysis on parameter δ. We observe that AUC remains stable for $\delta \in [0.15, 0.6]$. When $\delta = 0$, the graph is a complete graph and the smoothness regularization will penalize the loadings of each source across the PCs to be similar to each other. Hence very low δ leads to a worse performance. On the other hand, when $\delta = 1$, the graph is just a set of isolated sources. The structure information is missing, therefore the performance is not optimal.

An important parameter in PCA based anomaly detection is k, the number of PCs spanning the normal subspace. In [28], Ringberg *et al.* claimed that the anomaly detection performance was very sensitive to k. We demonstrate in the right part of figure 6 that our methods achieve a relatively stable performance as the dimension of normal subspace changes from 1 to 4. More specifically, the overall AUC gradually decreases from 0.96 to 0.72 as k changes from 1 to 3 and then increases to 0.77 at $k = 4$. However even in the worst case $k = 3$ it still has a good performance with AUC= 0.73. This study demonstrates that our model is not sensitive to k.

6. CONCLUSIONS AND FUTURE WORK

Previous work on PCA based anomaly detection claimed that PCA cannot be used for anomaly localization. We propose two novel approaches, *joint sparse PCA* (JSPCA) and *graph joint sparse PCA* (GJSPCA), for anomaly detection and localization in network data streams. By enforcing joint

sparsity on the PCs representing the abnormal subspace and incorporating the structure information of network via regularization, we significantly extend the applicability of PCA based technique for localization. Our experimental studies on three real world data sets demonstrate the effectiveness of our approach. Our future works will focus on three directions: (a) how to select the number of principal components that best interpret the normal subspace, and (b) how to extend our approach to kernel PCA, and (c) how to develop alternative optimization techniques to improve the scalability.

Acknowledgments

The work is partially supported by the NSF award IIS 0845951 and the Office of Naval Research N00014-07-1-1042.

7. REFERENCES

[1] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific. 2nd edition., 1999.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] D. Brauckhoff, X. Dimitropoulos, A. Wagner, and K. Salamatian. Anomaly extraction in backbone networks using association rules. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 28–34, New York, NY, USA, 2009. ACM.

[4] S. Budhaditya, D.-S. Pham, M. Lazarescu, and S. Venkatesh. Effective anomaly detection in sensor networks data streams. *IEEE International Conference on Data Mining, ICDM2009*, 0:722–727, 2009.

[5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.

[6] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. In *ICDM*, pages 746–751, 2009.

[7] P. H. dos Santos Teixeira and R. L. Milidiú. Data stream anomaly detection through principal subspace tracking. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1609–1616, New York, NY, USA, 2010. ACM.

[8] H. Fei and J. Huan. Boosting with structure information in the functional space: an application to graph classification. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2010.

[9] C. Franke and M. Gertz. Orden: outlier region detection and exploration in sensor networks. In *SIGMOD Conference*, pages 1075–1078, 2009.

[10] X. Gu and H. Wang. Online anomaly prediction for robust cluster systems. In *ICDE*, pages 1000–1011, 2009.

[11] S. Hirose, K. Yamanishi, T. Nakata, and R. Fujimaki. Network anomaly detection based on eigen equation compression. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1185–1194, New York, NY, USA, 2009. ACM.

[12] L. Huang, M. I. Jordan, A. Joseph, M. Garofalakis, and N. Taft. In-network pca and anomaly detection. In *In NIPS*, pages 617–624, 2006.

[13] T. Idé, A. C. Lozano, N. Abe, and Y. Liu. Proximity-based anomaly detection using sparse structure learning. In *SDM*, pages 97–108, 2009.

[14] T. Idé, S. Papadimitriou, and M. Vlachos. Computing correlation anomaly scores using stochastic nearest neighbors. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 523–528, Washington, DC, USA, 2007. IEEE Computer Society.

[15] M. Isaac, B. Raul, E. Gerard, and G. Moisès. On-line fault diagnosis based on the identification of transient stages. In *in Proc. of 20th European Symposium on Computer Aided Process Engineering ĺC ESCAPE20*. Elsevier B.V., 2010.

[16] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, 2010.

[17] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 457–464, New York, NY, USA, 2009. ACM.

[18] R. Jiang, H. Fei, and J. Huan. Anomaly localization by joint sparse pca in wireless sensor networks. In *Proceedings of the The 4th International Workshop on Knowledge Discovery from Sensor Data (SensorKDD-2010)*, 2010.

[19] E. Keogh and T. Folias. The ucr time series data mining archive. Website, 2002. `http://www.cs.ucr.edu/eamonn/TSDMA/index.html`.

[20] E. Keogh, J. Lin, and A. Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 226–233, Washington, DC, USA, 2005.

[21] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *In ACM SIGCOMM*, pages 219–230, 2004.

[22] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *In ACM SIGCOMM*, pages 217–228, 2005.

[23] J.-G. Lee, J. Han, and X. Li. Trajectory outlier detection: A partition-and-detect framework. In *ICDE*, pages 140–149, 2008.

[24] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2,1-norm minimization. In *Conference on Uncertainty in Artificial Intelligence (UAI) 2009*, 2009.

[25] X. Liu, X. Wu, H. Wang, R. Z. 0003, J. Bailey, and K. Ramamohanarao. Mining distribution change in stock order streams. In *ICDE*, pages 105–108, 2010.

[26] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Distributed detection/localization of change-points in high-dimensional network traffic data. *CoRR*, abs/0909.5524, 2009.

[27] Y. Nesterov. Introductory lectures on convex optimization: A basic course. 2003.

[28] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of pca for traffic anomaly detection. In *SIGMETRICS '07: Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 109–120, New York, NY, USA, 2007. ACM.

[29] J. Silva and R. Willett. Detection of anomalous meetings in a social network. In *42nd Annual Conference on Information Sciences and Systems, 2008. CISS 2008.*, pages 636 –641, 2008.

[30] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319 – 2323, 2000.

[31] N. Xu, S. Rangwala, and *et al.* . A wireless sensor network for structural monitoring. In *IN SENSYS*, pages 13–24, 2004.

[32] J. Zhang, Q. Gao, and H. Wang. Anomaly detection in high-dimensional network data streams: A case study. In *IEEE International Conference on Intelligence and Security Informatics, 2008. ISI 2008.*, pages 251 –253, June 2008.

[33] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.