Turkish Journal of Science & Technology
19(1), 107-112, 2024

Research Paper
https://doi.org/10.55525/tjst.1293057

An Application of Robust Principal Component Analysis Methods for Anomaly Detection

Kübra BAĞCI GENEL[1*], Halit Eray ÇELİK[2,3]
[1,2] Department of Econometrics, Faculty of Economics and Administrative Sciences, Van Yüzüncü Yıl University, Van, Türkiye
[3] Department of Computer Engineering, Engineering Faculty, Khoja Akhmet Yassawi International Kazakh- Turkish University, Kazakhstan
[*1] kubrabagci@yyu.edu.tr, [2] ecelik@yyu.edu.tr

(Geliş/Received: 05/05/2023; Kabul/ Accepted: 01/03/2024)

Abstract: Ensuring a secure network environment is crucial, especially with the increasing number of threats and attacks on digital systems. Implementing effective security measures, such as anomaly detection can help detect any abnormal traffic patterns. Several statistical and machine learning approaches are used to detect network anomalies including robust statistical methods. Robust methods can help identify abnormal traffic patterns and distinguish them from normal traffic accurately. In this study, a robust Principal Component Analysis (PCA) method called ROBPCA which is known for its extensive use in the literature of chemometrics and genetics is utilized for detecting network anomalies and compared with another robust PCA method called PCAGRID. The anomaly detection performances of these methods are evaluated by injecting synthetic traffic volume into a well-known traffic matrix. According to the application results, when the normal subspace is contaminated with large anomalies the ROBPCA method provides much better performance in detecting anomalies.

Key words: Anomaly detection, outlier, Principal Component Analysis, robust statistics.

Dayanıklı Temel Bileşenler Analizi ile Anomali Tespiti Üzerine Bir Uygulama

Öz: Dijital sistemlere yönelik artan sayıda tehdit ve saldırılar sebebi ile güvenli bir ağ ortamı sağlamak önemli bir problemdir. Anomali tespiti gibi yöntemlerin uygulanması, herhangi bir anomal trafik hacminin tespit edilmesine yardımcı olabilmektedir. Dayanıklı istatistiksel yöntemler de dahil olmak üzere ağ anomalilerini tespit etmek için çeşitli istatistiksel ve makine öğrenmesi yaklaşımları kullanılmaktadır. Dayanıklı yöntemler, anormal trafik modellerini belirlemeye ve bunları normal trafikten doğru bir şekilde ayırmaya yardımcı iyi bir araçtır. Bu çalışmada, ağ anomalilerini tespit etmek için kemometri ve genetik literatüründe yaygın kullanımıyla bilinen ROBPCA adlı dayanıklı bir Temel Bileşen Analizi (PCA) yöntemi kullanılmış ve PCAGRID adlı başka bir dayanıklı PCA yöntemi ile karşılaştırılmıştır. Bu yöntemlerin anomali tespit performansları, iyi bilinen bir trafik matrisine sentetik trafik hacmi enjekte edilerek değerlendirilmiştir. Uygulama sonuçlarına göre anomali tespitinde ROBPCA yöntemi daha iyi performans sağladığı görülmüştür.

Anahtar kelimeler: Anomali tespiti, aykırı değer, dayanıklı istatistik, Temel Bileşenler Analizi.

1. Introduction

Network security has become more important due to the rapid development of network technologies. Providing a stable network is a critical task considering even some basic services use network technologies. To fulfill this task, anomaly detection is one of the approaches to detecting abnormal behaviors in the traffic volume. Anomaly detection is also effective with unknown attack patterns, unlike the signature-based approaches.

Although machine learning methods are promising and generally used in detecting network anomalies, statistical approaches are also frequently utilized in related fields successfully. Robust versions of the classical statistical methods can be useful in solving many statistically related computer network issues, and their potential as a tool for detecting anomalies [1]. Principal Component Analysis (PCA) known as a classical data reduction technique has been widely used in network anomaly detection over the last 20 years. To use the dimension reduction method for detecting outliers i.e., anomalies, the PCA is regarded as a classification procedure and utilized to detect outlying and normal observations. Since outliers can seriously bias or influence estimates that may be of substantive interest [2] it is not desirable for a statistical method to be sensitive to outliers even though the objective is to detect the outliers. The outcomes caused by this sensitivity are extensively discussed in previous studies [3–5] on network anomaly detection. Consequently, these works suggested that using robust methods may be beneficial to cope with such problems as subspace contamination and sensitive false positive rates. Due to the potential costs of an incorrect classification, robust methods are considered in critical cases, such as the detection of network anomalies.

* Corresponding author: kubrabagci@yyu.edu.tr. ORCID Number of authors: [1] 0000-0002-6679-9738, [2] 0000-0001-7490-8124

In this study, two robust PCA methods called the ROBPCA [6] and PCAGRID [7] are considered. The reason for considering these methods is based on a study by [8] that compares the outlier detection performances of many robust PCA methods throughout a simulation study. According to their study, amongst the many robust PCA algorithms, the ROBPCA and PCAGRID algorithms stand out in detecting anomalies. Motivated by this, performances of the ROBPCA and PCAGRID methods for detecting network anomalies are compared through a well-known traffic matrix. In addition, PCAGRID is recognized for its ability to yield precise estimates in cases of uncontaminated datasets while also demonstrating robustness when dealing with contaminated data [8]. The PCAGRID and ROBPCA, have versatile applications across different fields. The efficiency of these methods in outlier detection is compared in the studies from various fields (see [9, 10]) including chemometrics, and time series analysis, recently. Chen et al. [9] showed that the PCAGRID method had performed better in terms of false positives over the ROBPCA algorithm for both simulated and real biological data with varying outlierness levels and Kazemi et al. [10] pointed out the PCAGRID is computationally faster than the ROBPCA method. Although the PCAGRID method has already been used in network anomaly detection previously [1], the ROBPCA method is utilized by [11] in the field of cybersecurity for detecting outliers most recently. Consequently, these two methods continue to find applications across various fields for detecting outliers. Considering the methods have advantageous features, the studies to detect network anomalies using these methods, are quite limited. This study contributes to the literature by evaluating the performance of two established robust PCA methods, ROBPCA and PCAGRID, for the purpose of network anomaly detection. Building upon prior research that highlighted their effectiveness in various research areas, our work emphasizes their applicability to network security by comparing their recalls of varying sizes of injections.

The rest of the study is organized as follows. In section 2, related literature is reviewed. Then, the methods used in the study are briefly described and the application scheme is explained. In section 4, the anomaly detection performances of the robust PCA methods are evaluated. In section 5, the study is finalized with a discussion of the results in light of previous studies.

2. Literature

In this section, some of the studies that addressed the advantages of using robust PCA methods in anomaly detection are reviewed. Lakhina et al. [12] pioneered the use of a PCA-based approach for network anomaly detection. Subsequently, many studies have contributed to this field. After the study of Lakhine at al., [12] there have been studies [3] and [13] concerning the sensitivity of the PCA for traffic anomaly detection and stated that using robust methods may overcome this problem. In this manner, Pascoal et al. [1], proposed a new detection scheme that utilizes a robust PCA method, the PCAGRID for detecting outliers after a robust feature selection step. They showed that the robust PCA successfully detected the anomalies in different traffic conditions. Wang et al. [14] proposed the Relaxed Principal Component Pursuit method as a new decomposition model and discussed the limitations of the classical PCA when the traffic matrix is contaminated with large anomalies. Kudo et al. [15] stated that considering a robust PCA method may perform better even though their PCA-based anomaly detection scheme is able to decrease the false alarm rate. Also, they proposed a detection scheme using the daily or weekly periodicity of the traffic matrix to avoid normal subspace contamination problems. Subspace contamination problem is explained as, problems encountered such as a high false alarm rate, due to contamination of normal subspace with the presence of very large anomalies [3]. By this means, Matsuda et al. [16] utilized an existing robust PCA method for the network anomaly detection scheme in [15] which is based on the periodicity of the traffic volume. They considered the Minimum Covariance Determinant (MCD) estimator MCD to be a robust covariance estimation method also included in the ROBPCA algorithm. Also applied the mentioned scheme to the Abilene data set. Hadri et al. [17] proposed a nonlinear feature extraction method called Nonlinear Fuzzy Robust PCA for anomaly detection is effective with noisy data as well. They showed that the Nonlinear Fuzzy Robust PCA method resulted in low false positive alarms using well-known KDDcup99 and NSL-KDD data. There are several studies in network intrusion detection literature employing PCA-based methods. Fernandes et al. [5] reviewed some of these studies on network anomaly detection including a detailed discussion of PCA-based methods as well. In recent years, the PCA has continued to be used in many network anomaly detection systems mostly in combination with various other approaches. For example, Vilaca et al., [18] introduce a semi-supervised machine learning model called RPCA-MD, which utilizes robust PCA and Mahalanobis Distance to automatically identify anomalies in network traffic and detect potential attacks, offering a promising solution for enhanced network security addressing the issue of botnet attacks. Wang et al., [19] combined PCA and Single-Stage Headless Face Detector algorithms offering superior detection speed and accuracy in experiments conducted using IDS2017 and IDS2012 datasets, making it a promising solution for efficient traffic attack detection in network security. Lu [20] proposes an anomaly detection system based on PCA to enhance the security of networked medical devices,

and the evaluation using real-world data demonstrates its effectiveness in detecting malicious attacks with a high detection rate and a low false alarm rate.

3. PCA-Based Outlier Detection

In this section, the PCA-based robust ROBPCA, and PCAGRID methods are described briefly. Since the methods introduced by [6, 7] before, a brief description of the methods is only given.

As mentioned the classical PCA is a statistical method generally used for dimension reduction. It is based on transforming relatively large numbers of variables into fewer unrelated variables by finding orthogonal linear combinations of original variables with the greatest variance [1]. The PCAGRID is introduced by Croux et al. [7] as a robust PCA method that uses the projection pursuit algorithm. It is based on finding projections of the data that have maximal dispersion with the grid search algorithm. Instead of using the variance, the algorithm employs a robust scale estimator called the Median Absolute Deviation. Similarly, the ROBPCA method includes both projection pursuit and robust covariance estimation. In the robust estimation of covariance, the MCD method is employed in the algorithm. In both methods, each observation is classified using different outlying scores based on distance, as normal observations and anomalies. See also [6] for more information on the ROBPCA method. The ROBPCA and PCAGRID methods are implemented by using the functions of LIBRA, which is a library of MATLAB published by [21], and the "rrcov" package featured in the R application, respectively.

4. Data and Application

4.1. Abilene data set and synthetic outliers

The traffic matrix used in this study is based on the traffic volumes measured on the origin-destination (OD) flows in the Abilene network on 1-8 March 2004. Abilene is a backbone network that connects various US campuses consisting of 12 nodes and 30 links. The Abilene dataset is a well-known data that continues to be used in various intrusion detection studies such as, [14–16]. It is available in [22]. The topology of the Abilene network is given in Figure 1. 12 nodes (n1-n12) and 30 links (e1-e30) of the Abilene network are visualized in the network topology.

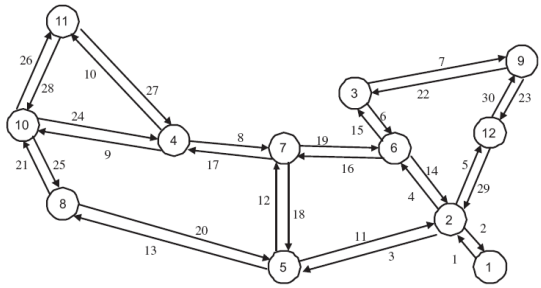

Figure 1. Abilene Network Topology [15].

For evaluating the anomaly detection performance of the PCAGRID and ROBPCA methods, synthetic anomalies are injected randomly into the Abilene traffic matrix which consists of 2016 observations and 144 variables. Anomalies are injected into traffic volumes measured in generated random places between the links e1-e7. The data is contaminated with different sizes of anomalies. As suggested in [23] the large anomalies injected as 10 times larger than the given traffic volume and small anomalies are injected as 2 times larger than the given traffic volume. Moreover, 1.2, 1.5, 8, and 12 times larger injections from background traffic volume are injected to examine both methods' recall (detection) performances in more detail.

4.2. Application scheme

The proposed application scheme is given as follows

1. Create random numbers using discrete uniform distribution (I_{1xn}) to determine where to inject n synthetic anomalies. n=100.

For, i=1…n,

2. set i=1

3. Add synthetic traffic into the original traffic volume ($X_{2016x144}$) for the corresponding element given in I_{1x1} and store injected data as a new data matrix (Xi).

4. Implement the ROBPCA for Xi using the "robpca" function in the LIBRA, MATLAB library.

5. Extract outlier score from $flag_{2016x1}$ matrix obtained by implementing the "robpca" function. If the corresponding element of $flag$ matrix (injection) is labeled as 0, classify observation as an outlier, and if it is 1, classify it as regular observation.

6. Set i=i+1 and repeat 3-5 steps.

Consequently, a detection rate as given in equation 1, is obtained by considering outcomes in each loop. A similar scheme is followed for the PCAGRID method by using the PcaGrid function of the "rrcov" package featured in the R software. The ROBPCA method initializes a parameter denoted as 'alpha,' which serves as a control parameter to regulate the algorithm's sensitivity to outliers. Alpha is a continuous variable within the range of 0.5 to 1, with a default value set at 0.75. Here default value of this parameter is chosen since the injections are not more than %25 of the data (around %1). The PCAGRID is also implemented using default parameter settings as well.

$$\text{Detection rate} = \text{True positives}/\text{ False negatives} + \text{True positives} \quad (1)$$

4.3. Application

In this section, anomaly detection performances of the ROBPCA and PCAGRID methods are compared by injecting synthetic anomalies into the Abilene dataset. The number of principal components to be used in the ROBPCA and PCAGRID algorithms is determined with the help of the scree plots. A scree plot is a practical tool used in deciding the number of PCs. The anomaly detection rate is considered a measure of detection performance.

Scree plots of eigenvalues for the ROBPCA algorithms are given in Figure 2.

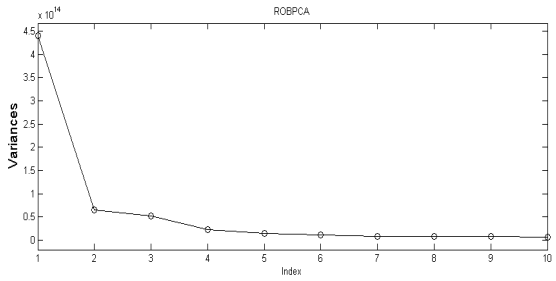

Figure 2. Scree plot for the ROBPCA method.

Scree plots of eigenvalues for the PCAGRID algorithm is given in Figure 3.

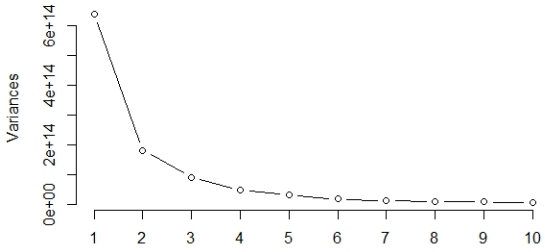

Figure 3. Scree plot for the PCAGRID method.

According to Figure 2 and Figure 3, it is decided to consider the first 4 PCs by taking into account the level-off point on the line.

Hair et al. [24] stated that the percentage of total explained variance of at least 60% is a satisfactory rate. A high variance explanation rate (86%) is reached by considering 4 PCs.

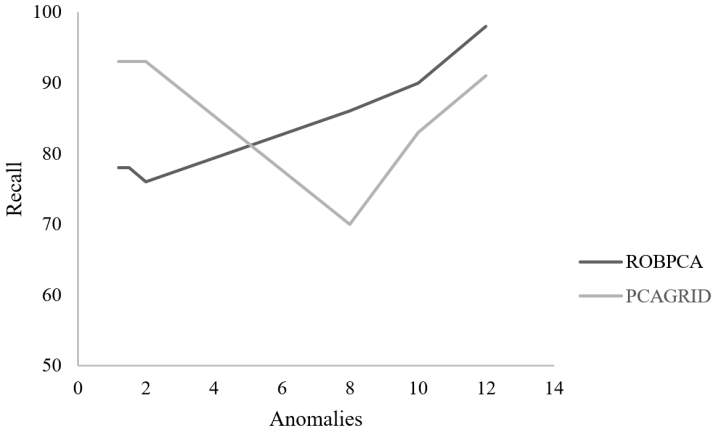

Figure 4. Detection rates for different sizes of injected anomalies.

Detection rates for various injections are given for both methods in Figure 4. With the ROBPCA algorithm, the %98, %90, %86, %76, %78, and %78 of injections are detected for 12, 10, 8, 2, 1.5 and 1.2 sized injections, respectively. With the PCAGRID algorithm, the %91, %83, %70 %94, %93, and %93 detection rates are obtained for 12, 10, 8, 2, 1.5 and 1.2 sized injections, respectively. Consequently, it can be said that the PCAGRID method is effective in detecting small anomalies better than ROBPCA. Moreover, although in larger anomalies the ROBPCA performed better, the PCAGRID reached higher detection rates overall. According to the results, it is seen that the ROBPCA provides better performance in detecting larger network anomalies and PCAGRID provides better performance in detecting smaller network anomalies. Also, anomaly detection rates are consistent with previous studies, such as [12] and [25] which reached around a % 90 detection rate. The false alarm rate is not given since it would require a data cleaning process. Also, it is seen that the ROBPCA method, which has been successfully used in chemometrics and genetics [9, 26, 27] can be adapted to network anomaly detection as well.

5. Conclusion

In this study, the anomaly detection performance of two robust PCA methods, ROBPCA and PCAGRID, across various injection scenarios are investigated. First, the background traffic volume is contaminated with synthetic anomalies by injecting them into random links of Abilene data. Then, it is examined whether these two methods can detect these outliers. Results show that the ROBPCA method performed better when the data are contaminated with large anomalies and the PCAGRID method is superior in recall of small anomalies. Consequently, the two methods that are highly popular in different fields, such as chemometrics, biology, and engineering, have been applied to network anomaly detection, providing significant performance in the subjected area, which should be investigated in more detail in future studies.

References

[1] Pascoal C, Oliveira MR de, Valadas R, et al. Robust feature selection and robust PCA for internet traffic anomaly detection. 2012 Proceedings IEEE INFOCOM 2012[Online] 2012.

[2] Zimmerman DW. A Note on the Influence of Outliers on Parametric and Nonparametric Tests. J Gen Psychol Routledge 1994; 121(4):391–401.

[3] Ringberg H, Soule A, Rexford J, et al. Sensitivity of PCA for Traffic Anomaly Detection. SIGMETRICS Perform. Eval. Rev. Association for Computing Machinery: New York, NY, USA 2007; 35(1):109–20.

[4] Brauckhoff D, Salamatian K, May M. Applying PCA for Traffic Anomaly Detection: Problems and Solutions. IEEE INFOCOM 2009 2009[Online] 2009.

[5] Fernandes G, Rodrigues JJPC, Carvalho LF, et al. A comprehensive survey on network anomaly detection. Telecommun Syst 2019; 70(3):447–89.

[6] Hubert M, Rousseeuw PJ, Branden K Vanden. ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics Taylor & Francis 2005; 47(1):64–79.

[7] Croux C, Filzmoser P, Oliveira MR. Algorithms for Projection–Pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems 2007; 87(2):218–25.

[8] Pascoal C. and Oliveira MR and PA and VR. Detection of Outliers Using Robust Principal Component Analysis: A Simulation Study. Combining Soft Computing and Statistical Methods in Data Analysis 2010[Online] Springer Berlin Heidelberg: Berlin, Heidelberg 2010.

[9] Chen X, Zhang B, Wang T, et al. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinformatics 2020; 21(1):269.
[10] Kazemi M, Rodrigues PC. Robust singular spectrum analysis: comparison between classical and robust approaches for model fit and forecasting. Comput Stat 2023;
[11] Burr B. Intruder Alert: Dimension Reduction and Density-Based Clustering for a Cybersecurity Application. 2021[Online] Ottawa 2021.
[12] Lakhina A, Crovella M, Diot C. Diagnosing Network-Wide Traffic Anomalies. Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications 2004[Online] Association for Computing Machinery: New York, NY, USA 2004.
[13] Abdelkefi A, Jiang Y, Wang W, et al. Robust Traffic Anomaly Detection with Principal Component Pursuit. Proceedings of the ACM CoNEXT Student Workshop 2010[Online] Association for Computing Machinery: New York, NY, USA 2010.
[14] Wang Z, Hu K, Xu K, et al. Structural analysis of network traffic matrix via relaxed principal component pursuit. Computer Networks 2012; 56(7):2049–67.
[15] Kudo T, Morita T, Matsuda T, et al. PCA-based robust anomaly detection using periodic traffic behavior. 2013 IEEE International Conference on Communications Workshops (ICC) 2013[Online] 2013.
[16] Matsuda T, Morita T, Kudo T, et al. Traffic anomaly detection based on robust principal component analysis using periodic traffic behavior. IEICE Transactions on Communications The Institute of Electronics, Information and Communication Engineers 2017; 100(5):749–61.
[17] Hadri A, Chougdali K, Touahni R. A Network Intrusion Detection Based on Improved Nonlinear Fuzzy Robust PCA. 2018 IEEE 5th International Congress on Information Science and Technology (CiSt) 2018[Online] 2018.
[18] Vilaça ESC, Vieira TPB, Sousa RT de, et al. Botnet traffic detection using RPCA and Mahalanobis Distance. 2019 Workshop on Communication Networks and Power Systems (WCNPS) 2019[Online] 2019.
[19] Wang Z, Han D, Li M, et al. The abnormal traffic detection scheme based on PCA and SSH. Conn Sci Taylor & Francis 2022; 34(1):1201–20.
[20] Lu W. Detecting Malicious Attacks Using Principal Component Analysis in Medical Cyber-Physical Systems. In: Traore I, Woungang I, Saad S, Eds. Artificial Intelligence for Cyber-Physical Systems Hardening Springer International Publishing: Cham 2023; pp. 203–15.
[21] Verboven S, Hubert M. LIBRA: a MATLAB library for robust analysis. Chemometrics and Intelligent Laboratory Systems 2005; 75(2):127–36.
[22] Zhang Y. Abilene Data. https://WwwCsUtexasEdu/~yzhang/Research/AbileneTM/ [Online].
[23] Nagaraja S, Jalaparti V, Caesar M, et al. P3CA: Private Anomaly Detection Across ISP Networks. Privacy Enhancing Technologies 2011[Online] Springer Berlin Heidelberg: Berlin, Heidelberg 2011.
[24] Hair JF, Black WC, Babin BJ, et al. Multivariate data analysis: Pearson new international edition. Essex: Pearson Education Limited 2014; 1(2).
[25] Rubinstein BIP, Nelson B, Huang L, et al. ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors. Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement 2009[Online] Association for Computing Machinery: New York, NY, USA 2009.
[26] Shieh AD, Hung YS. Detecting Outlier Samples in Microarray Data 2009; 8(1).
[27] Granzotto C, Sutherland K, Arslanoglu J, et al. Discrimination of Acacia gums by MALDI-TOF MS: applications to micro-samples from works of art. Microchemical Journal 2019; 144:229–41.