

# Robust Methods for Unsupervised PCA-based Anomaly Detection

Roland Kwitt  
Advanced Networking Center  
Salzburg Research  
Austria, Salzburg 5020  
Email: rk Witt@salzburgresearch.at

Ulrich Hofmann  
Advanced Networking Center  
Salzburg Research  
Austria, Salzburg 5020  
Email: uhofmann.salzburgresearch.at

**Abstract**—The paper discusses the need for robust unsupervised anomaly detection. We focus on an approach that employs robust principal component analysis (PCA) to detect malicious behaviour. By using robust PCA, we can overcome the problem that we have to have enough anomaly-free data in the training phase of a detection system.

## I. INTRODUCTION

According to CERT/CC statistics, the number of security incidents caused by malicious internet traffic is dramatically increasing. Network Intrusion Detection Systems (NIDS) that rely on pattern (signature) detection algorithms are commonly used in production systems. However, an attack will stay unrecognized in case the system lacks the according signature.

Anomaly detection is a promising approach to tackle this problem. In the most general case, an anomaly detector can detect deviations from an established baseline profile, that characterizes normal behavior. If we consider malicious traffic to be inherently abnormal, then an ideal anomaly detector can detect even novel and modified attacks. However, our definition of anomaly detection implies, that we have to have completely anomaly free data, so that the detector can *learn* what is actually normal. Besides, the usually high false positive rates, this is an important issue, why anomaly detectors are still rarely used in production systems. Network data has to be cleaned from anomalous activities in a preprocessing step, in order to be used for training purposes. Furthermore, since network traffic is everything but static, training will have to be repeated in certain time intervals. As a matter of fact, this *cleaning* step is a rather time consuming activity.

A possible way to approach this problem is to use robust methods in the training phase of an anomaly detector. In this context, the term *robust* means, that we allow the contamination of training data and still accurately estimate normal behavior. It is clear, that contamination can only be tolerated to a certain extend, depending on the employed method. In this paper, we focus on the use of robust estimation methods for covariance and correlation matrix in PCA-based anomaly detection, which is closely related to the field of outlier detection in multivariate data.

The paper is organized as follows. In section II will we briefly review some related work on this topic. Section III introduces PCA for anomaly detection, followed by some

preliminary results in section IV. Section V concludes the paper with a brief summary and an outlook on further research.

## II. RELATED WORK

Principal Component Analysis (PCA) has already been used in recent research work on anomaly detection [1], [2]. Our work here is mainly based on the work done in [1]. There, the authors proposed and successfully employed a PCA-based classifier, to filter out anomalies in a 34-dimensional connection record dataset, used in the KDD Cup 1999 [3], a classifier learning contest. Furthermore, the authors took a first step forward towards robustifying their detector. They used an iterative approach called *multivariate trimming*, which we will discuss later in section III, to clean the training dataset from possible anomalies. The only drawback that can be argued is, that the compiled feature set for KDD Cup 1999 contains features from several domains (host and network) and is hardly available in that form in real-life, without considerable effort.

## III. ROBUST ANOMALY DETECTION WITH PCA

This section briefly introduces the theoretical background of PCA and how it can be used in anomaly detection. Then we go further and discuss how to robustify PCA. The actual detector is discussed in III-C.

### A. Principal Component Analysis

PCA is variable-oriented method, with transforms a set of correlated original variables into a set of uncorrelated variables, called principal components (PC). These principal components are linear combinations of the original variables. By carrying out PCA we hope that a few PCs can explain most of the variation in the original data. Thus, dimensionality can be reduced with almost no loss of information. It is obvious, that in case of a priori uncorrelated data, PCA makes no sense.

If  $\mathbf{x}^T = (x_1, \dots, x_p)$  denotes a  $p$ -dimensional vector of random variables with expected value  $\mu$  and covariance matrix  $\Sigma$ , then we try to find a set of new, uncorrelated random variables, whose variance decreases with increasing  $j = 1, \dots, p$ . (Notation: the transpose of a vector is denoted by  $^T$ ). Hence, for the first principal component we look for a linear function  $\alpha_1^T \mathbf{x}$  having maximum variance. Next, we look for  $\alpha_2^T \mathbf{x}$ , uncorrelated with  $\alpha_1^T \mathbf{x}$  and maximum variance, etc.

Furthermore,  $\alpha_j, j = 1, \dots, p$  is scaled to meet the constraint  $\alpha_j^T \alpha_j = 1$ . The deviation of the PCs leads to the result, that the the vectors of coefficients  $\alpha_1, \dots, \alpha_p$  for each PC are the eigenvectors of  $\Sigma$  corresponding to  $\lambda_1, \dots, \lambda_p$  eigenvalues, with  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p$ . As mentioned before, we hope that  $m \ll p$  PCs will account for most of the variance in  $\mathbf{x}$ . If the  $p \times p$  matrix of eigenvectors is denoted by  $\mathbf{A} = (\alpha_1, \dots, \alpha_p)$ , the vector  $\mathbf{z}$  of principal components can be written as  $\mathbf{z} = \mathbf{A}^T \mathbf{x}$ .

Now that we have briefly covered the theoretical background of PCA (for greater detail, see [4]), we go on to explain, how to obtain a sample version of the PCs from a number of observations. Let  $\mathbf{x}_1, \dots, \mathbf{x}_n$  denote  $n$  independent observations of the random vector  $\mathbf{x}$  and let  $\mathbf{X}$  be the  $n \times p$  matrix of these observations. Next, let  $\mathbf{S}$  denote the sample covariance matrix of  $\mathbf{X}$ , where the  $(j, k)$ -th element is given by

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k),$$

and  $\bar{x}_j$  and  $\bar{x}_k$  are the mean of the  $j$ -th and the  $k$ -th variable. By taking the  $p \times p$  matrix  $\mathbf{A}$  ( $k$ -th column is the eigenvector corresponding to the  $k$ -th largest eigenvalue  $\lambda_k$ ) and the matrix of observations  $\mathbf{X}$ , we can obtain the  $n \times p$  matrix of sample PC scores  $\mathbf{Z}$  from  $\mathbf{Z} = \mathbf{X}\mathbf{A}$ . Accordingly,  $z_{ik}$  is the score of the  $i$ -th observation on the  $k$ -th PC with zero mean.

It is important to note, that PCA based on a covariance matrix, is quite sensitive regarding the units of measurement. That is, if there are large differences between the variances of variables, those variables with high variance will dominate the first PCs [4]. In such a case, it is sensible to use the correlation matrix instead of the covariance matrix to obtain the PC scores. Especially, when we observe network traffic variables with different units and ranges, such as byte counts or connection times, then the correlation matrix is the first choice.

## B. Robustness

Robustness is a central point when trying to employ PCA for unsupervised anomaly detection. We have seen in III-A, that in order to obtain PC scores based on a number of observations, we have to determine the sample covariance/correlation matrix. Furthermore, our aim is to obtain reasonable results for sample covariance/correlation matrix estimation, even if the data is contaminated. Statistically speaking, the term contaminated signifies the presence of outliers. To completely understand the problem, we introduce the so called *breakdown point*. We consider the simple case, where we want to estimate the center of a point cloud, with  $n$  points in  $p$  dimensions. Now, let  $T$  denote a location estimator and  $C$  denotes a covariance estimator. The standard estimator of multivariate location  $T(\mathbf{X})$  is the arithmetic mean  $T(\mathbf{X}) = \bar{\mathbf{x}}$ .

The breakdown point is defined as the smallest fraction of contamination that can cause  $T$  to take on arbitrarily values far away [5]. Formally, if  $\mathbf{X}'$  denotes the corrupted samples, obtained by replacing  $m$  original points by arbitrary values, the so called *maximal bias* is defined by

$$bias(m; T, \mathbf{X}) = \sup_{\mathbf{X}'} \left\| T(\mathbf{X}') - T(\mathbf{X}) \right\|$$

and the *breakdown point* can be formulated as

$$\epsilon_n^*(T, \mathbf{X}) = \min\{m/n; bias(m; T, \mathbf{X}) \text{ is infinite}\}.$$

It is obvious that the multivariate arithmetic mean has 0% breakdown, if  $n \rightarrow \infty$ . Now, since the arithmetic mean is used in the standard estimators for covariance and correlation matrix, breakdown is 0% as well. Hence, data has to be anomaly free (outlier free) in order to obtain reasonable PC scores with the standard estimators. Even one bad connection record for instance can destroy our estimator.

In [1], *multivariate trimming* (iterative trimming) [6] is used to obtain reasonable estimates. To understand the method, we introduce the squared Mahalanobis distance

$$MD^2(\mathbf{x}_i, \mathbf{X}) = (\mathbf{x}_i - T(\mathbf{X}))^T C(\mathbf{X})^{-1} (\mathbf{x}_i - T(\mathbf{X})).$$

The procedure works as follows: Start with  $\mathbf{X}^{(1)} = \mathbf{X}$  and define  $\mathbf{X}^{(k+1)}$  recursively as the set of observations with the  $(1 - \alpha)n$  smallest values of  $\{MD^2(\mathbf{x}_i, \mathbf{X}^{(k)}), \mathbf{x}_i \in \mathbf{X}\}$ . The procedure stops if  $T(\mathbf{X}^{(k)})$  and  $C(\mathbf{X}^{(k)})$  are stable. Here,  $C$  and  $T$  are the classic estimates. According to [5], it was shown that the breakdown point is at most  $1/p$ .

Here, we propose to take estimators with an even higher breakdown point to estimate location and covariance, such as the *Minimum Volume Ellipsoid (MVE)* or the *Minimum Covariance Determinant (MCD)* [7] estimator. Usually, MCD is used more often, since there exists a fast calculation algorithm [8]. If we consider the MVE estimator,  $T(\mathbf{X})$  is the center of the minimal volume ellipsoid covering at least  $h = \lfloor n/2 \rfloor + 1$  data points.  $C(\mathbf{X})$  is given by the ellipsoid itself, multiplied by a suitable consistency factor. The breakdown point is 50% as  $n \rightarrow \infty$ . In case of the MCD estimator,  $T(\mathbf{X})$  is the mean of the  $h$  points  $\in \mathbf{X}$  for which the determinant of the covariance matrix is minimal.  $C(\mathbf{X})$  is again, the ellipsoid itself. (if  $h = n/2$ , breakdown is 50%)

In [5], Rousseeuw points out that if one is certain that the fraction of outliers is at most  $0 < \alpha \leq 1/2$ , then  $h$  can be replaced by  $k(\alpha) = \lfloor n(1 - \alpha) \rfloor + 1$ , with a resulting breakdown point of  $\alpha$  if  $n \rightarrow \infty$ .

## C. Outlier (Anomaly) Detection

A first intuitive approach to outlier detection would be to use a robustified version of the Mahalanobis distance, to directly identify outliers in the data. By replacing the standard estimates by MCD or MVE estimates, a robust distance measure can be obtained. However, the method presented in [1] can identify a broader range of outliers at a higher speed, which is important with regards to online anomaly detection. The basic idea is, that outliers violating the correlation structure of the main bulk of data, are often not detectable by looking at the major PCs, but the minor ones. In contrast, outliers that cause an increase in variance and covariance in the original

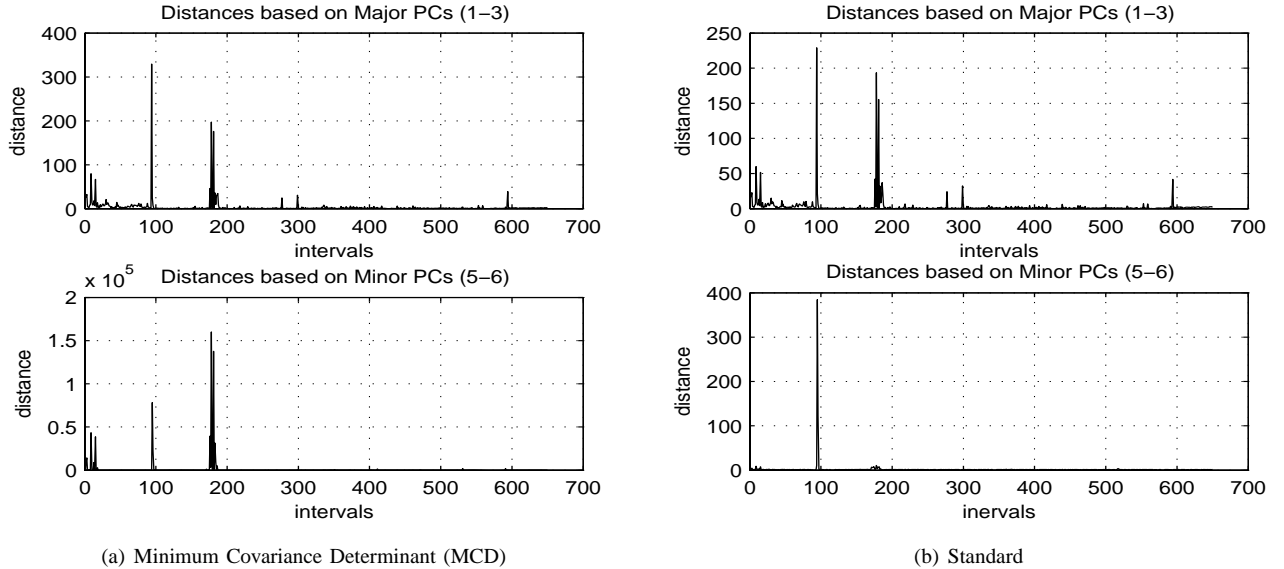


Fig. 1. Comparison of anomaly detection results, based on MCD and standard correlation matrix estimation

variables must be extreme on those variables as well and thus detectable by the major PCs. It now is reasonable to introduce two functions of the PCs [1], with the PCs obtained from robust PCA. It can be shown that the sum of the squared PC scores divided by the corresponding eigenvalues

$$d_i^2 = \sum_{k=1}^p \frac{z_{ik}^2}{l_k} \quad (1)$$

is equivalent to  $MD^2$  of observation  $\mathbf{x}_i$ . By splitting up Eq. (1) into

$$v_1^2 = \sum_{k=1}^q \frac{z_{ik}^2}{l_k} \quad \text{and} \quad v_2^2 = \sum_{k=p-r+1}^p \frac{z_{ik}^2}{l_k} \quad (2)$$

we obtain two different distance measures. On the basis of some (possibly contaminated) data sample, the anomaly detection thresholds for  $v_1^2$  and  $v_2^2$  can be determined as follows: First, we calculate (2) and determine the corresponding empirical CDFs for  $v_1^2, d_2^2$ . Then, we fix the false positive rate at a preferred level and determine the corresponding quantiles to reach this level. Based on the robust estimation methods in III-B, no cleaning step is required anymore.

#### IV. PRELIMINARY RESULTS

This section presents some preliminary results based on the proposed combination of PCA and robust estimation. We used the first 12 hours of thursday in week 2 (labeled attack data) of the DARPA 1999 ID data sets (see [9]) as an input to `tcpstat` to obtain the following features over consecutive 60 second time intervals: `#bytes/s`, `load`  $\in [0, 1]$ , `#packets`, `#TCP packets`, `#packets/s`, `#UDP packets`. Based on this feature set, we detect the *satan* (at  $\approx$  start +100 minutes), *portsweep* and *neptune* (at  $\approx$  start +170 minutes) attacks. If we take a closer look at the

distances (see Fig. 1(b)) based on the minor PCs obtained by standard PCA, we can see a clear *masking* effect caused by the aforementioned attacks.

#### V. CONCLUSION

In this short paper, we have pointed out the need for unsupervised anomaly detection and especially for robust estimation methods in PCA-based unsupervised anomaly detection. Furthermore, we have shown that the use of the MCD estimator permits the use of contaminated training data and still produces promising detection results. Further work on this topic includes research on the applicability this approach with regards to online anomaly detection and an enhancement of the feature set to detect a broader range of attacks.

#### REFERENCES

- [1] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A Novel Anomaly Detection Scheme Based on Principal Component Classifier," in *Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, Florida, USA, November 2003, pp. 172–179.
- [2] K. Ramah, H. Ayari, and F. Kamoun, "Traffic Anomaly Detection and Characterization in the Tunisian National University Network," in *Networking 2006*, Coimbra, Portugal, May 2006, pp. 136–147.
- [3] Knowledge Discovery and Data Mining Cup 1999 Data. [Online]. Available: <http://www.ics.uci.edu/kdd/databases/kddcup99/kddcup99.html>
- [4] I. Jolliffe, *Principal Component Analysis*. Springer, 2002.
- [5] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. John Wiley & Sons, 2003.
- [6] R. Gnanadesikan and J. Kettenring, "Robust Estimates, Residuals and Outlier Detection with Multiresponse Data," *Biometrics*, vol. 28, pp. 81–124, 1972.
- [7] P. Rousseeuw, "Multivariate Estimation with High Breakdown Point," *Mathematical Statistics and Applications*, vol. B, pp. 283–297, 1985.
- [8] P. Rousseeuw and K. Van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, vol. 41, pp. 212–223, 1997.
- [9] D. Marchette, *Computer Intrusion Detection and Network Monitoring - A Statistical Viewpoint*. Springer, 2001.