

**T² Control Chart based on PCA with KDE Control Limit for Monitoring Intrusion**

A I Jaya, T R Aulia, F D Putri, and T Rakhmawati

1Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia

* Email: andir721@gmail.com

**Abstract.** In monitoring network anomaly, the traditional $T^2$ chart can be an alternative owing to its ability to capture the network anomaly. However, the new problem emerges in consequence of the hardship of the network traffic data to satisfy the multivariate normal distribution assumption in Hotelling’s $T^2$ chart. As a result, many false alarms will be found during the monitoring process. In this work, the combination between Hotelling’s $T^2$ control chart and the Principal Component Analysis (PCA) is utilized to observe the network traffic data. The PCA is used to minimize the data dimension which can reduce computational time. Meanwhile, the Kernel Density approach is employed in estimating the control limit of the non-normal process. The proposed method is applied to the famous KDD99 dataset, and its performance is compared with the other methods. Compared to the other charts, the proposed control chart yields a higher detection accuracy with a lower false alarm rate. Moreover, the proposed control chart also produces a faster computational time.

**Keywords:** Hotelling’s $T^2$; Principal Component Analysis; Kernel Density Estimation; Intrusion

1. Introduction

Monitoring quality using the control chart is commonly conducted using Shewhart [1–5], EWMA [6,7], and CUSUM [8,9] charts. Not only these charts but several types of control charts are also used in monitoring product quality. The Principal Component Analysis (PCA) has been widely employed in observing quality using control charts approach. Several previous works have adopted PCA in control charts [10–13]. By using this method, the drawbacks of control chart caused by a large number and highly correlated quality characteristics have been resolved [14]. The large numbers of dimensions often happen in modern production processes and services. When this happens, the calculation $T^2$ statistic becomes challenging due to the singularity of the covariance matrix [13,15]. The PCA approach can cope with these troubles by reducing quality characteristics involved. The lower number of quality characteristics will also make a swifter execution time [16].

The multivariate control charts can also be utilized for monitoring anomalies in the network. Several studies employ this method for Intrusion Detection System (IDS). IDS based on robust PCA is developed by [17]. PCA combination with Support Vector Machine (SVM) is developed by [18,19]. Integration between PCA dan genetic algorithm is proposed by [20] Also, neural network-based PCA for network intrusion detection is proposed by [21]. However, there are two main drawbacks when PCA based Multivariate chart is applied for network anomaly detection. First, the assumption of normality distribution cannot be fulfilled due to the existence of the anomaly which results in extreme values events [16,22]. Consequently, there will be many false alarms occurred [23]. Second, the computational time will be longer owing to the large number of features used.
In order to resolve the high false alarm that occurred, the nonparametric method, such as kernel density estimation (KDE) is used. This KDE method has been proven to possess good results for non-normal distribution [24,25]. The combination between PCA and Hotelling’s $T^2$ statistic can be a good alternative in reducing execution time while also reducing the probability of the singularity of the covariance matrix. Therefore, this paper proposes the combination of Hotelling’s $T^2$ chart and the PCA with KDE control limit. This combination is expected to have higher accuracy detection with a faster running time. The remaining sections of this paper are composed as follows: Section 2 presents the PCA-based $T^2$ chart. Section 3 discusses the methodology and algorithms of the proposed IDS. Section 4 contains the result and discussion. Summary and research opportunities are displayed in Section 5.

2. PCA based hotelling’s $T^2$ control chart
In this section, the proposed $T^2$ based on PCA using KDE control limit is presented. First, statistics of PCA-based Hotelling’s $T^2$ are presented. Further, the calculation procedure of the KDE control limit is also displayed.

2.1. Hotelling’s $T^2$ based on PCA
The initial step in this method is calculating the covariance matrix of the quality characteristics $C$ using the following expression:

$$ C = \frac{1}{n} X^T X, $$

Calculating the eigenvalue and eigenvector of $C$ is conducted using eigenvalue-decomposition as follows:

$$ C = V \Lambda V^T, $$

where $V = (v_1, ..., v_p)$ is the eigenvector matrix and $\Lambda = diag(\lambda_1, ..., \lambda_p)$ is the eigenvalues diagonal matrix.

To calculate the Principal Component Scores (PCs) the eigenvector is multiplied by the original data as presented in the following equation:

$$ Y = XV = (y_1, ..., y_p)^T. $$

The PCA based $T^2$ chart uses the first $k$ PCs to create the statistic of a control chart as:

$$ T^2_{PCA} = \sum_{i=1}^{k} \frac{(y_i - \mu)^2}{\lambda_i}, $$

where $k$ is the number of PCs involved, $\lambda_i$ is the eigenvalue corresponding to the $l$-th PC, and $\mu$ is the mean vector target. The traditional control limit can be calculated using the following expression:

$$ CL = \frac{k(n+1)(n-1)}{n^2-nk} F_{\alpha,k,n-k}, $$

where $n$ is the number of samples and $\alpha$ is a false alarm rate. As a note, this control limit is developed under the assumption of the multivariate normal distribution.

2.2. Control limit calculation with kernel density estimation (KDE)
Let $T^2_{PCA}$ is estimated from in-control data. The empirical distribution of the proposed $T^2_{PCA}$ statistic can be estimated as follows:

$$ \hat{f}_h(t) = \frac{1}{n} \sum_{i=1}^{n} K \left( \frac{t-T^2_{PCA,i}}{h} \right), $$
where \( \hat{h} \) and \( K \) represent the estimated bandwidth and the kernel function, respectively. The control limit of \( T^2_{PCA} \) based on KDE is computed by picking the quantile of kernel distribution as follows:

\[
CL_{Kernel} = \hat{h}^{-1}(T)(1 - \alpha),
\]

with \( \alpha \) is a false alarm rate. The distribution function \( F_h(t) \) is estimated using a numerical approach with the trapezoid rule method as follows:

\[
\hat{F}_h(T) = \int_{\eta_{\min}}^{\eta_{\max}} \hat{f}_h(t) dt \approx \frac{\eta_{\max} - \eta_{\min}}{2n} \sum_{i=1}^{n} \left( \hat{f}_h(t_i) + \hat{f}_h(t_{i+1}) \right),
\]

where \( \eta_{\min} \) and \( \eta_{\max} \) are the minimum and maximum values. Furthermore, the Gaussian Kernel employed in this paper is defined as:

\[
K(u) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} u^2 \right), \quad -\infty < u < \infty.
\]

### 3. Intrusion detection system (IDS) algorithms

The algorithms of the proposed IDS are characterized into three stages. The normal profile such as PC score, mean vector, and covariance matrix obtained from the normal connection are calculated in stage 1. In stage 2, the KDE control limit is calculated using the normal profile captured from stage 1. Stage 3 performs the monitoring process from new data.

#### Stage 1: Determining the in-control parameter

**Step 1** Crate principal component matrix \( Y_{normal} \) calculated normal connection data \( X_{normal} \).

**Step 2** Calculate mean vector of \( Y_{normal} \), denoted as \( \mu, l = 1, 2, ... , p \).

**Step 3** Calculate the variance of \( Y_{normal} \) and create a diagonal matrix of eigenvalues denoted as \( S \).

#### Stage 2: Calculating the control limit

**Step 1** Calculate the \( T^2 \) based PCA statistic, denoted as \( T^2_{PCA} \).

**Step 2** Calculate the empirical density using KDE.

**Step 3** Calculate the \( F_h(t) \).

**Step 4** Determine \( \alpha \) and estimate the control limit of \( T^2_{PCA} \).

#### Stage 3: Monitoring Anomalies Stage

**Step 1** Create matrix of new connection data \( X_{test} \).

**Step 2** Estimate \( Y_{test} = X_{test}S \).

**Step 3** Determine \( \alpha \) and \( k \) used in the monitoring.

**Step 4** Calculate statistics.

**Step 5** If \( T^2_{PCA} > CL_{KDE} \) then declare the connection is an intrusion.
4. Results and discussions

4.1. Dataset and performance metric

This paper uses the KDD99 dataset as the benchmark dataset. Furthermore, this study uses 32 out of 34 quantitative variables. Table 1 shows the intrusion detection confusion matrix.

| Prediction | Intrusion | Normal |
|------------|-----------|--------|
| Intrusion  | True Positives (TP) | False Negatives (FN) |
| Normal     | False Positives (FP) | True Negatives (TN) |

The level of accuracy is estimated using the following equation:

\[
\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total observations}}.
\]

On the other hand, False Positive Rate (FPR) and False Negative Rate (FNR) are calculated as

\[
\text{FPR} = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}}.
\]

\[
\text{FNR} = \frac{\text{False Negatives}}{\text{True Positives} + \text{False Negatives}}.
\]

4.2. Monitoring results

The monitoring results are presented in this subsection. Table 2 shows the performance of the proposed IDS for the several numbers of quality characteristic \( k \). From the table, it can be seen that according to the highest accuracy criterion, the optimum number of principal component scores \( k \) is nine. The proposed IDS has an accuracy of 0.9809, FPR of 0.0539, and FNR of 0.0106.

| \( k \) | Accuracy | FP | FN | FPR | FNR |
|--------|----------|----|----|-----|-----|
| 3      | 0.8031   | 97,278 | 0  | 1.0000 | 0.0000 |
| 5      | 0.9625   | 16,925 | 1602 | 0.1740 | 0.0040 |
| 7      | 0.9782   | 1,071 | 9684 | 0.0110 | 0.0244 |
| 8      | 0.9783   | 936 | 9793 | 0.0096 | 0.0247 |
| 9      | **0.9809** | **5,241** | **4209** | **0.0539** | **0.0106** |
| 10     | 0.9794   | 2,614 | 7565 | 0.0269 | 0.0191 |
| 11     | 0.9788   | 3,010 | 7477 | 0.0309 | 0.0188 |
| 13     | 0.9789   | 2,917 | 7529 | 0.0300 | 0.0190 |
| 15     | 0.9790   | 3,239 | 7157 | 0.0333 | 0.0180 |
| 20     | 0.9774   | 3,233 | 7908 | 0.0332 | 0.0199 |

These results indicate that the proposed IDS can detect more intrusion which can be seen from the lower number of FNR. It also can be seen that the proposed IDS still produces more false alarms which can be seen from the higher value of FPR than FNR.
4.3. Performance comparison

In presenting the advantage of the proposed IDS, some IDS based on control charts are compared to then proposed IDS based on PCA-based $T^2$ with KDE control limit. Table 3 shows the performance comparison. It can be seen that the traditional IDS-based $T^2$ and PCA-based $T^2$ charts have identical accuracy detection of 0.97. According to the results, it can be known that according to the level of accuracy, the proposed IDS performs better than the performance of IDS-based $T^2$ and PCA-based $T^2$ charts with an $F$ distribution control limit. According to the FPR value, the proposed IDS also yields a lower false alarm compared to the conventional Hotelling’s $T^2$ chart. Meanwhile, according to the FNR value, the proposed IDS finds intrusions more accurately than IDS based on PCA-based $T^2$ charts with an $F$ distribution control limit.

Table 3. Performance of the proposed IDS compare to the existing control charts

| Control Charts | Accuracy | FP  | FN  | FPR  | FNR  |
|----------------|----------|-----|-----|------|------|
| $T^2$          | 0.9799   | 6542| 3384| 0.0673| 0.0085|
| PCA-based $T^2$| 0.9779   | 2472| 8441| 0.0254| 0.0213|
| Proposed Method| **0.9809**| **5241**| **4209**| **0.0539**| **0.0106**|

Figure 1. Execution time (sec) comparison

The execution time (in the second) comparisons of the proposed IDS and the existing charts are presented in Figure 1. From the figure, it can be known that the PCA-based $T^2$ chart has the fastest computational time compared to the conventional $T^2$ chart and the proposed IDS. However, the PCA-based $T^2$ chart has a lower accuracy than the proposed chart. Also, the computational time difference is not more than 0.1 seconds. Therefore, it can be concluded that the proposed IDS demonstrates higher performance in terms of the faster computational time and high level of accuracy.

5. Conclusion

In this paper, the integration of the Hotelling’s $T^2$ chart and PCA is proposed to observe the intrusion. The KDE method is used to comply with the non-normality concern that happens in the network traffic. Using nine principal component scores, the proposed IDS has an accuracy of 0.9809, FPR of 0.0539, and FNR of 0.0106. By comparing the proposed IDS to the other control charts-based IDS, it can be concluded that the proposed IDS yields the highest accuracy. Compared to the conventional $T^2$-
based IDS, the proposed IDS has a lower false alarm. Meanwhile, compared to the IDS based on PCA-$T^2$ with $F$ control limit, the proposed IDS can discover intrusions more accurately which can be known from the lower FNR. The proposed IDS also demonstrates the faster computational time with a better detection rate compared to the benchmark methods.

6. References

[1] Haddad F, Alsmadi M K, Badawi U, Farag T, Alkhasawneh R, Almarashdeh I and Hassan W 2019 Bivariate modified hotelling’s T2 charts using bootstrap data Int. J. Electr. Comput. Eng. 9 4721–7

[2] Ahsan M, Mashuri M and Khusna H 2017 Evaluation of Laney p’ Chart Performance Int. J. Appl. Eng. Res. 12 14208–17

[3] Ahmad M R and Ahmed S E 2021 On the distribution of the T2 statistic, used in statistical process monitoring, for high-dimensional data Stat. Probab. Lett. 168 108919

[4] Ahsan M, Mashuri M and Khusna H 2018 Hybrid James-Stein and successive difference covariance matrix estimators based hotelling’s T2 chart for network anomaly detection using bootstrap J. Theor. Appl. Inf. Technol. 96 6828–41

[5] Mashuri M, Haryono H, Aksioma D F, Wibawati W, Ahsan M and Khusna H 2019 Tr (R2) control charts based on kernel density estimation for monitoring multivariate variability process Cogent Eng. 6 1665949

[6] Khusna H, Mashuri M, Suhartono, Prastyo D D and Ahsan M 2018 Multioutput Least Square SVR Based Multivariate EWMA Control Chart J. Phys. Conf. Ser. 1028 12221

[7] Khusna H, Mashuri M, Suhartono S, Prastyo D D and Ahsan M 2018 Multioutput least square SVR based multivariate control chart: The performance evaluation and application Cogent Eng.

[8] Khusna H, Mashuri M, Ahsan M, Suhartono S and Prastyo D D 2020 Bootstrap-based maximum multivariate CUSUM control chart Qual. Technol. Quant. Manag. 17 52–74

[9] Mahmoud M A and Maravelakis P E 2013 The performance of multivariate CUSUM control charts with estimated parameters J. Stat. Comput. Simul. 83 719–36

[10] Ahsan M, Mashuri M, Khusna H and Lee M H 2020 Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics Symmetry (Basel). 12 1838

[11] Ahsan M, Mashuri M, Kuswanto H, Prastyo D D and Khusna H 2019 Outlier detection using PCA mix based T2 control chart for continuous and categorical data Commun. Stat. - Simul. Comput. 1–28

[12] Ahsan M, Mashuri M, Kuswanto H, Prastyo D D and Khusna H 2018 Multivariate Control Chart based on PCA Mix for Variable and Attribute Quality Characteristics Prod. Manuf. Res. 6 364–84

[13] Kouri T 2005 Application of latent variable methods to process control and multivariate statistical process control in industry Int. J. Adapt. Control Signal Process. 19 213–46

[14] Montgomery D 2009 Introduction to statistical quality control (New York)

[15] Mason R L and Young J C 2002 Multivariate Statistical Process Control with Industrial Applications (Society for Industrial and Applied Mathematics)

[16] Mashuri M, Ahsan M, Lee M H and Prastyo D D 2021 PCA-based Hotelling’s T2 chart with Fast Minimum Covariance Determinant (FMCD) Estimator and Kernel Density Estimation (KDE) for Network Intrusion Detection Comput. Ind. Eng. 107447

[17] Pascoal C, Oliveira M R de, Valadas R, Filzmoser P, Salvador P and Pacheco A 2012 Robust feature selection and robust PCA for internet traffic anomaly detection 2012 Proceedings IEEE INFOCOM pp 1755–63

[18] Kuang F, Xu W and Zhang S 2014 A novel hybrid KPCA and SVM with GA model for intrusion detection Appl. Soft Comput. J. 18 178–84

[19] Xu X and Wang X 2005 An adaptive network intrusion detection method based on PCA and support vector machines International Conference on Advanced Data Mining and Applications (Springer) pp 696–703
[20] Chabathula K J, Jaidhar C D and Kumara M A A 2015 Comparative study of Principal Component Analysis based Intrusion Detection approach using machine learning algorithms *Signal Processing, Communication and Networking (ICSCN), 2015 3rd International Conference on* pp 1–6

[21] Liu G, Yi Z and Yang S 2007 A hierarchical intrusion detection model based on the PCA neural networks *Neurocomputing* **70** 1561–8

[22] Zhu X 2006 *Anomaly Detection Through Statistics-Based Machine Learning For Computer Networks* (The University of Arizona)

[23] Ahsan M, Mashuri M, Kuswanto H and Prastyo D D 2018 Intrusion Detection System using Multivariate Control Chart Hotelling’s T2 based on PCA *Int. J. Adv. Sci. Eng. Inf. Technol.* **8** 1905–11

[24] Phaladiganon P, Kim S B, Chen V C P and Jiang W 2013 Principal component analysis-based control charts for multivariate nonnormal distributions *Expert Syst. Appl.* **40** 3044–54

[25] Phaladiganon P, Kim S B, Chen V C P, Baek J-G and Park S-K 2011 Bootstrap-Based T 2 Multivariate Control Charts *Commun. Stat. - Simul. Comput.* **40** 645–62