This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2009 proceedings.

Applying PCA for Traffic Anomaly Detection: Problems and Solutions

Daniela Brauckhoff
ETH Zurich
Zurich, Switzerland
brauckhoff@tik.ee.ethz.ch

Kave Salamatian
Lancaster University
Lancaster, UK
kave@lancaster.ac.uk

Martin May
ETH Zurich
Zurich, Switzerland
may@tik.ee.ethz.ch

***Abstract*—Spatial Principal Component Analysis (PCA) has been proposed for network-wide anomaly detection. A recent work has shown that PCA is very sensitive to calibration settings. Unfortunately, the authors did not provide further explanations for this observation. In this paper, we fill this gap and provide the reasoning behind the found discrepancies.**

We revisit PCA for anomaly detection and evaluate its performance on our data. We develop a slightly modified version of PCA that uses only data from a single router. Instead of correlating data across different spatial measurement points, we correlate the data across different metrics. With the help of the analyzed data, we explain the pitfalls of PCA and underline our argumentation with measurement results. We show that the main problem is that PCA fails to capture temporal correlation. We propose a solution to deal with this problem by replacing PCA with the Karhunen-Loeve Transform. We find that when we consider temporal correlation, anomaly detection results are significantly improved.

I. Introduction

Principal Component Analysis (PCA) has been first proposed as a method for traffic anomaly detection in [1]. While being known in other domains before, Lakhina et al made its application very popular in the networking community. Subsequent publications [2] proposed extensions to the initial method. Recently, it has been shown by Ringberg et al [3] that PCA is very sensitive to its parameter settings. The authors have reported about instability problems encountered when using PCA. However, they did not provide a precise explanation for their observation.

In this paper, we will provide the missing explanations for the encountered problems. In particular, we revisit PCA-based approaches for anomaly detection from a signal processing point of view. When we applied the PCA method to our dataset, we found similar inconsistencies as those reported in [3]. Further investigating the results, we found that the main problem of PCA, as used today, is that it does not consider the temporal correlation of the data. The fact that the data is temporally correlated breaks the underlying hypothesis of the PCA-based anomaly detection method. The main contributions of this paper are that we (i) show what kind of problems arise when PCA is not carefully applied to anomaly detection; and (ii) provide a profound theoretical explanation for the encountered problem; and (iii) provide correction to the previously published method and alter it to an efficient anomaly detection mechanism. We validate our improved method by applying it to real network traffic with well known and identified anomalies.

The paper is structured as follows. In section II, we revisit the PCA theory and develop its basic properties. We thereafter extend the analysis to stochastic processes and explain why the simple PCA is not applicable and should be replaced by Karhunen-Loeve (KL) expansion in section III. We describe this expansion and develop a Galerkin-based approach to calculate the KL expansion from a finite number of samples of the process. We then use the KL expansion to develop the predictive model that could be used for anomaly detection. Section IV applies the KL extension method developed in section III to three weeks of backbone data and examines therewith the reasons for the bad performance of classic PCA. We are able to validate that the source of the poor performance really is the temporal correlation. Applying KL extension instead of classic PCA to our dataset we find the detection results to be improved significantly. We further show that the non-stationarity is a critical issue and recalibration of the model is mandatory for good anomaly detection performance.

II. A signal processing view on PCA

A browse in the literature shows two closely related but different interpretations of PCA:

- As an efficient representation that transforms the data to a new coordinate system such that the projection on the first coordinate contains the greatest variance, the projection on second coordinate has the second greatest variance, and so on.
- As a modeling technique using a finite number of terms of an orthogonal serie expansion of the signal with uncorrelated coefficients.

Interestingly, the literature mainly motivates the application of PCA to network anomaly detection by the first interpretation although it is indeed following the second interpretation. This has resulted in some erroneous interpretation and practices that have widely spread among the community. We are devoting this paper to describing these erroneous practices and to present a way of correcting them.

Let's suppose that we have a column vector of correlated random variables $\mathbf{X} = (X_1, \ldots, X_K)^T \in \mathbb{R}^k$. One observes these random variables through N independent realization vectors $\mathbf{x}^i = (x_1^i, \ldots, x_K^i)^T,\ i = 1, \ldots, N$ and arranges them

in a $N \times K$ observation matrix $\mathbf{x}$ with each row containing an observation vector $\mathbf{x}^i$. We are searching for the most "**suitable**" non-canonical basis $(\mathbf{e}_1, \ldots, \mathbf{e}_K)$ for the vector space $\mathbb{R}^K$ to represent the random variables $\mathbf{X}$.

For the class of random variables that are linear (*i*.e., they can be decomposed to a combination of independent linear random variables) and have as sufficient statistics their means and covariances (*i*.e., means and covariances entirely describe their joint probability distributions), the most suitable basis is the one that maximizes the variance of each projected component. One very popular case where these two assumptions hold is when $(X_1, \ldots, X_K)$ are jointly gaussian.

Nonetheless, the literature is full of examples where using such an orthonormal basis results in erroneous interpretation because the linearity or the sufficiency of mean and covariance is not valid. In particular, the lack of linearity can be a serious problem that could be dealt with by applying a non-linear transform to the dataset before PCA. This approach is called Kernel PCA and is widely used. Another approach is to replace PCA by Independent Component Analysis (ICA) that could result in a non-orthogonal basis [4]. However, the involved complexity is much higher than with classic PCA.

Assuming linearity and sufficiency of mean and variance, the most suitable basis is $(\phi_1, \ldots, \phi_K)$, where ϕ_i is an eigenvector of the covariance matrix of $\mathbf{X}$ defined as $\mathbf{\Sigma} = \mathbb{E}\{(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T\}$ where μ is a column vector containing the means of X_i). We derive these eigenvectors by solving the following linear equation:

$$\Sigma\phi_i = \lambda_i\phi_i \tag{1}$$

where λ_i are the eigenvalues of the covariance matrix. As the covariance matrix is positive definite, this equation has at most K positive eigenvalues and K different orthonormal eigenvectors. The basis change matrix $U = [\phi_1, \ldots, \phi_K]$ contains in its columns the eigenvectors ϕ_i. Solving the above problem is called in matrix theory the Singular Value Decomposition (SVD) of the covariance matrix.

It is noteworthy that U is a basis change matrix only when $\mathbf{X}$ is zero mean, and in general one has to work with $\tilde{\mathbf{X}} = \mathbf{X} - \mu$ in place of $\mathbf{X}$, *i*.e., the coordinate change is $\tilde{\mathbf{y}} = U\tilde{\mathbf{x}}$. This last point is frequently overlooked in the literature, and not taking care of it could lead to large errors when using PCA[1]. In the forthcoming we will assume that we have taken care of this obvious precaution so we can drop the $\sim$.

After applying PCA one can rewrite the initial vector of random variables $\mathbf{X}$ in the new coordinate system as:

$$\mathbf{X} = \sum_{i=1}^{K} Y_i\phi_i \tag{2}$$

where Y_i are jointly independent random variables with $\mathbb{E}\{Y_i\} = 0$ and $\mathbb{V}\text{ar}\{Y_i\} = \lambda_i$. PCA replaces the correlated random variables $\mathbf{X}$ by a vector of independent random variables $\mathbf{Y}$ that are linearly equivalent. The independence of Y_i is therefore an ***essential*** property as this is the main reason that the PCA representation is "suitable".

The above discussion remains theoretical, and in practice one has a set of observations and has to pick the suitable basis. Whenever the dataset under study is not flagrantly in contradiction with the conditions of mean and variance sufficiency and linearity we can apply PCA and find a convenient representation of the data.

First one has to estimate the covariance matrix using the popular sum of product formula $\hat{\Sigma} = \frac{1}{N-1}\mathbf{x}\mathbf{x}^T$. This gives a reliable estimation because of independence between observations. Thereafter applying the SVD factorization is just a straightforward and mechanical step that provides the needed basis as well as the basis transform matrix.

III. Extension of PCA to stochastic processes

The extension of PCA to stochastic processes is mandatory as the signals used for anomaly detection are samples of stochastic processes that have temporal as well as spatial correlations. Let's assume we have a K-vector of zero mean stationary stochastic processes $\mathbf{X}(t) = (X_1(t), \ldots, X_K(t))^T$ with a covariance function $\sigma_{i,j}(\tau) = \mathbb{E}\{X_i(t)X_j(t-\tau)\}$ defined over an interval $[a, b]$. The multi-dimension Karhunen-Loeve theorem [2] [5] states that one can rewrite this vector as a serie expansion (named KL expansion):

$$X_l(t) = \sum_{i=1}^{K}\sum_{j=1}^{\infty} Y_{i,j}^l\Phi_{i,j}(t) \tag{3}$$

where $Y_{i,j}^l$ are pairwise independent random variables and $\Phi_{i,j}(t)$ are pairwise orthogonal **deterministic** (non-random) functions defined on $[a, b]$, *i*.e., $\int_a^b \Phi_{i,j}(t)\Phi_{m,n}^*(t)dt = 0$ for $i \neq m$ or $j \neq n$. Generally, the basis functions $\Phi_{i,j}(t)$ are rescaled such that $\int_a^b |\Phi_{i,j}|^2(t)dt = 1$.

This theorem extends PCA to a vector of stochastic processes as Eq. 3 is the equivalent of Eq. 2. The family of deterministic functions $\Phi_{i,j}(t)$ is an orthonormal basis for the space of linear stochastic processes and the random variables $Y_{i,j}^l$ are coordinates of the stochastic process $X_l(t)$ in this new space. We can formally derive the basis functions $\Phi_{i,j}(t)$ by solving the following set of linear integral equations:

$$\sum_{i=1}^{K}\int_a^b \sigma_{i,l}(s)\Phi_{i,j}(s-t)ds = \lambda_{l,j}\Phi_{l,j}(t),\ j > 0.$$

This set of equations is the equivalent of Eq. 1. The random variables $Y_{i,j}^l$ are obtained by projecting each stochastic process over an eigenfunction:

$$Y_{i,j}^l = \int_a^b X_l(s)\Phi_{i,j}(s)ds$$

[1]It is noteworthy that even if [1], [2] did not state clearly the necessity of removing the mean, they have used zero mean signals in their implementation code.

[2]The KL theorem was initially defined for one-dimensional stochastic processes.

The KL expansion considers the temporal correlation between time t and $t+\tau$ as well as the spatial correlation between process $X_i(.)$ and $X_j(.)$. This results in a more complex analysis than the simple PCA described earlier. However, this higher complexity is unavoidable because of the temporal correlation. Not taking it into account leads to the errors described in [3].

In practice, we have only access to a finite set of samples observed each T time unit from the vector of stochastic processes. In the forthcoming, we will use the notation $[k]$ to represent the discrete version of a time continuous process sampled at times kT. Let's assume that we have n samples of the multidimensional stochastic process and the covariance values $\sigma_{i,j}(\tau)$ can be assumed as negligible for $\tau > NT$. We can therefore truncate the KL expansion to N terms. The Galerkin method [6] transforms the above integral equations to a matrix problem that can be solved by applying the SVD technique. This makes it possible to derive the KL expansion using only a finite number of samples. The Galerkin method generates a set of eigenvectors in a $K \times N$ dimensional vector space, that are time-sampled versions $\Phi_{i,j}[k] = \Phi_{i,j}(kT)$ of the originally continuous function $\Phi_{i,j}(t)$. Finally, we obtain a discrete version of the KL expansion as:

$$X_l[k] = \sum_{i=1}^{K} \sum_{j=1}^{N} Y_{i,j}^{l} \Phi_{i,j}[k]. \tag{4}$$

We first have to estimate the spatio-temporal correlation matrix. Let's construct a $KN \times (n-N)$ observation matrix:

$$\mathbf{x} = \begin{pmatrix} x_1(1) & \dots & x_1(n-N) \\ x_1(2) & \dots & x_1(n-N+1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \dots & x_1(n) \\ x_2(1) & \dots & x_2(n-N) \\ \vdots & \ddots & \vdots \\ x_2(N) & \dots & x_2(n) \\ \vdots & \ddots & \vdots \\ x_K(1) & \dots & x_K(n-N) \\ \vdots & \ddots & \vdots \\ x_K(N) & \dots & x_k(n) \end{pmatrix}$$

The matrix $\hat{\Sigma} = \frac{1}{n-N-1}\mathbf{x}\mathbf{x}^T$ is a $KN \times KN$ matrix that contains all the needed spatio-temporal covariance estimates. It is noteworthy that because of temporal correlation one needs more data to estimate correctly the covariance here than for the independent case we had in section II.

The Galerkin method consists of applying PCA as described in section II to this large matrix. This results in KN eigenvectors $\Phi_{i,j}[.]$ of length KN that are used to construct a basis transform matrix U. The coefficients $Y_{i,j}^{l}$ are obtained by applying the basis change transform $\mathbf{y} = U\mathbf{x}$. Applying KL expansion to K stochastic processes entails diagonalizing a $KN \times KN$ matrix (in place of a $K \times K$ matrix in section II). However, this added complexity is unavoidable when one has to deal with correlated observations.

A. PCA as a modeling method

Up to now, we described KL expansion as a tool for creating an equivalent (in probability) and suitable representation of stochastic processes. If we neglect some of the smaller terms of the expansion (terms with small values of $\mathbb{V}\mathrm{ar}\{Y_{i,j}^{l}\}$), we obtain a linear approximation of the initial process in a smaller dimension vector space. The discrete expansion in Eq. 4 is therefore approximated as:

$$\hat{X}_l[k] = \sum_{i=1}^{L} \sum_{j=1}^{M} Y_{i,j}^{l} \Phi_{i,j}[k]. \tag{5}$$

where $M < N$ and $L < K$. This approximation has a noteworthy optimality property. Among all approximations defined over a linear space of dimension LM, this is the linear approximation with the smallest approximation error variance ($\mathbb{V}\mathrm{ar}\{X(t) - \hat{X}(t)\}$. The basis change transform becomes a $KN \times LM$ matrix U_{LM} that contains the LM eigenfunctions $\Phi_{i,j}[.]$ in its columns. This is the theoretical basis to use the KL expansion as a non-parametric and generic technique for modeling a large class of processes where we cannot reject the linearity and sufficiency of mean and variance (see section II).

The non-parametric nature of the above modeling technique is simultaneously its strength and Achilles heel; a non-parametric method is not based on any precise form of the distribution (out of the linearity and the mean and variance sufficiency) meaning it is more robust. At the same time being non-parametric means that no prior knowledge can be incorporated into the model.

Before going further, let's first give some details about the obtained model. The expansion in Eq. 5 provides a synthesis method for generating an approximated process $\hat{X}_l[k]$ by a bank of ML filters with Finite Impulse Response equal to $\Phi_{i,j}[k], k = 0; \dots, KN$; each filter being excited by the random variable input $Y_{i,j}^{l}$. Predicting the values of the realization of the KN random variables $Y_{i,j}^{l}$ by applying the basis change matrix to $\mathbf{X}[.]$, we can use this synthesis filter as a predictive filter. This is the approach followed in PCA and KL expansion based anomaly detectors.

IV. Validation

In this section, we introduce our data set and present the results of applying PCA and KL expansion to this data set.

A. Data Set and Metrics

We use for our experiments three weeks of Netflow data coming from one of the peering links of a medium-sized ISP (SWITCH, AS559). These data were recorded in August 2007 and comprise a variety of traffic anomalies happening in daily operation such as network scans, denial of service attacks, alpha flows, *etc*. For computing the detection metrics we distinguish between incoming and outgoing traffic, as well as UDP and TCP flows. For each of these four categories, we computed seven commonly used traffic features: byte, packet, and flow counts, source and destination IP address entropy, as well as unique source and destination IP address counts. All

metrics were obtained by aggregating the traffic at 15 minute intervals resulting in a 28×96 data matrix per measurement day. Anomalies in the data were identified using available manual labeling methods: visual inspection of time series and top-n queries on the flow data. This resulted in 28 detected anomalous events in UDP and 73 detected in TCP traffic.

It is noteworthy that our dataset is different in nature from the one used in [1], [2]. There, the data were collected from different network border routers. Also, [1] only analyzes the traffic volume. In the follow-up publication [2], the authors extended their dataset to include the 4 entropy values of the source and destination IP addresses as well as the source and destination port numbers. In this work, we collected the data at a single link. The observed metrics are correlated in time and in space. Our observations and statements are valid whenever both spatial and temporal correlation exist; they can be extended to the case studied by [1], [2].

We use the vector of metrics $\mathfrak{X}[1 : 192]$ containing the first two days of metrics for building the model. For this purpose, we derive a spatio-temporal correlation matrix with the temporal correlation range set to $N = 1, .., 5$ as described in III. Note that setting $N = 1$ gives the standard PCA approach. Then, we apply SVD decomposition to the data, resulting in a basis change matrix. We choose the component numbers to be included in the model according to the methodology proposed in [1], [3] to englobe more than 95% of the variance in the initial metric. Finally, we compute a prediction $\hat{\mathfrak{X}}[k]$ based on the model, and the squared prediction error as $Q[k] = \mathbf{e}[k]^T \mathbf{e}[k]$ where $\mathbf{e}[k] = \mathfrak{X}[k] - \hat{\mathfrak{X}}[k]$.

B. Bias and correlation in the decision variable

For PCA based anomaly detectors, [7] proposed to use a non-linear function of the squared prediction error $Q[k]$ as decision variable $D[k]$

$$D[k] = \frac{Q[k]}{\theta}^h \tag{6}$$

This non-linear function has been tailored to make its distribution converging as closely as possible to a gaussian distribution. In [7], the authors refer to the work of Jensen [8] showing that, under the hypothesis that elements of $\mathbf{e}[k]$ are statistically ***independent*** and follow a gaussian distribution, the distribution of $D[k]$ converges to a normal distribution with known mean and variance. [8] gives formulas for deriving from PCA characteristics the mean and variance of $D[k]$, as well as the parameters θ and h. It is therefore, possible to normalize $D[k]$ to a gaussian random variable with zero mean and a variance of one. Because of space restrictions, we refer the reader to [8], [7], and [1] for details.

To verify if the problems reported in [3] are really coming from the temporal correlation, we checked the bias of $\mathbf{D}[k]$ after application of classical PCA to the dataset. Actually, we observed that the mean of the decision variable over the 192 values used for calibration was 7.8 in place of zero as predicted by theory. The first possible explanation for this discrepancy was a well known argument, that was not stated in previous papers on anomaly detection: PCA should be applied to zero mean random variables. Meaning that one needs to first get the mean as close as possible to zero before applying PCA. We verified if ignoring the non-zero mean could have led to the observations. We checked the bias in the decision variable after removing the mean in all input variables and found that it is reduced but is still important as the mean remains equal to 1.6 in place of zero. We show the distribution of the decision variable and the normal plot for this setting in Fig. 1.

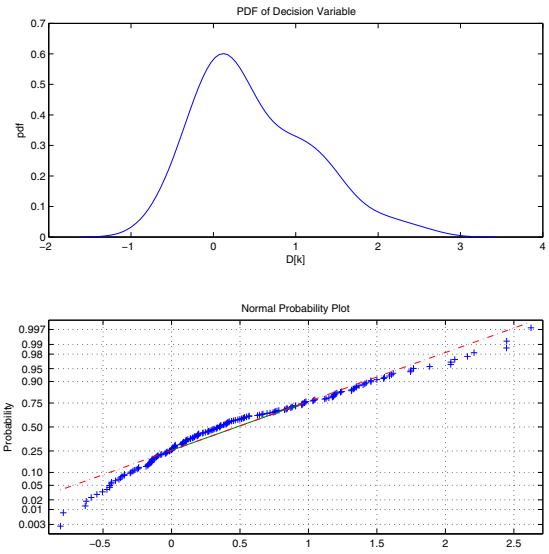


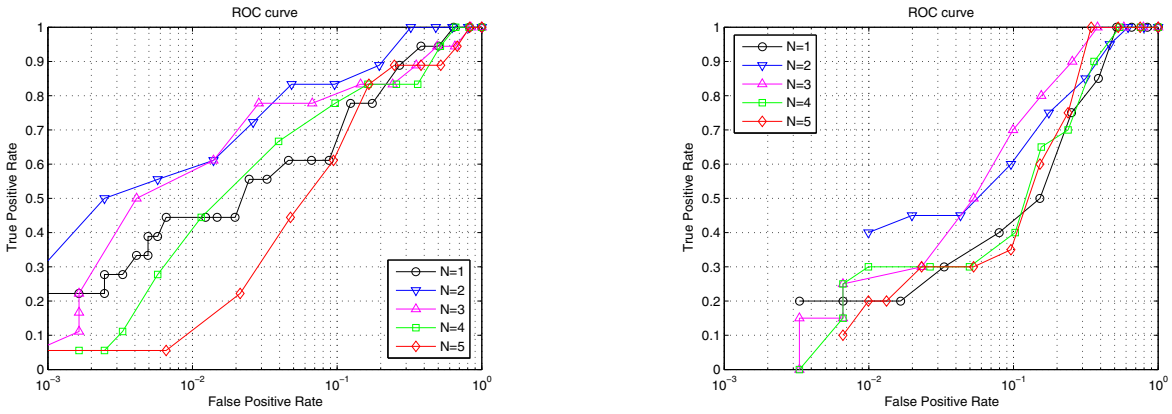
Fig. 1. PDF and normal plot of decision variable $\mathbf{D}[k]$ for classic PCA with mean removal.

Furthermore, we verified the predicted independence of error terms for standard PCA. A closer look at the proof of convergence in [8] shows that even if the convergence is robust toward non-gaussianity of the underlying variables, it is heavily dependent on the independence condition between the terms in $\mathbf{e}[k]$. We estimated the correlation between the error terms and found that there is a high correlation between some of these terms. So the hypothesis of independence should be fully rejected.

Then we repeated the tests for KL expansion with different values of the temporal correlation range N. We found that for $N \geq 2$ the bias decreases close to zero (even if the marginal gain saturates). Moreover, the correlation between the terms in $\mathbf{e}[k]$ decreases significantly with augmenting N. That validates our hypothesis that relates the bad anomaly detection performance of PCA to temporal correlation.

C. ROC curves

The Neyman-Pearson theorem about statistical tests [9] defines a fundamental trade-off between false alarm probability and true negative probability; smaller thresholds lead to lower false negative rates but larger false alarm rates and larger thresholds result in higher false negative rates and smaller false alarm rates. The Receiver Operating Characteristics (ROC) curve combining the two parameters in one value [10] captures this essential trade-off.



(a) UDP (b) TCP

Fig. 2. ROC curves resulting from application of KL expansion to the dataset (with mean removal) for different values of the temporal correlation range. Note that N = 1 corresponds to classical PCA. All curves are plotted with a logarithmic x axis for improved readability.

Fig. 2 shows two plots with ROC curves (for UDP and TCP traffic) obtained for different values of the temporal correlation range N. The figures are plotted in semilog to achieve a better comparison between different values of N. All ROC curves are obtained with enough terms to capture 95% of the total variance in the model. In all cases, no more than two expansion terms were needed for this setting.

The comparison of ROC curves shows a considerable improvement of the anomaly detection performance with use of KL expansion with $N = 2, 3$ consistently for UDP and TCP traffic and thereafter a decrease for $N \geq 4$. Particularly, for $N = 2$ the detection rate for anomalies in UDP traffic at a false alarm rate of 10% increases to 83% compared to 67% for $N = 1$. The TCP plot shows an increase in the detection rate from 43% ($N = 1$) to 70% ($N = 3$) at the same false alarm rate when the temporal correlation is taken into account.

D. Effect of non-stationarity

It remains another issue to solve: for $N \geq 4$ the performance decreases even if the bias in the decision variable decreases. We investigated this observation and found that most of the mis-detections for $N \geq 4$ happen at the end of the second week and during the third week of our dataset while they were spread in the three weeks for $N < 3$. A possible explanation is the stationarity issue: when N increases, the model contains more parameters and becomes more sensitive to the stationarity of the traffic metrics. This means that we can expect that anomaly detection performance decreases with time and this decrease is more pronounced for larger N. Non-stationarity and model recalibration are indeed important issues that need careful evaluation and analysis. It is out of the scope of this paper and will be the subject of a forthcoming paper.

V. Conclusion

This paper began with a very practical problem: how to apply the popular PCA method in real world anomaly detection. We found that direct application of the PCA method results in poor performance in terms of ROC curves; we investigated the problem and found that the main source of the problem is the bias coming from correlation in prediction error terms. After a detailed theoretical analysis, it appears that the correct framework is not the classical PCA but rather the Karhunen-Loeve expansion. We have presented the KL expansion and have provided a Galerkin method for developing a predictive model. This method has thereafter been applied to data traces from the Switch network and we have shown that an important improvement is attained when temporal correlation is considered.

Acknowledgments

The authors would like to thank Mark Crovella for his helpful discussions that have led to some aspects of this paper. We also thank Anukool Lakhina for giving access to his code and the Abilene data he used in his papers.

References

[1] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *SIGCOMM 20004*, 2004.

[2] ——, "Mining Anomalies Using Traffic Feature Distributions," in *ACM SIGCOMM 2005*, 2005.

[3] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," in *SIGMETRICS 2007*, 2007.

[4] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," *Journal of the American Statistical Association*, vol. 99, pp. 567–567, January 2004.

[5] R. M. Gray and L. D. Davisson, *An Introduction to Statistical Signal Processing*. Cambridge University Press, 2005.

[6] K. Kunisch and S. Volkwein, "Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics," *SIAM J. Numer. Anal.*, vol. 40, no. 2, pp. 492–515, 2002.

[7] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[8] D. R. Jensen and H. Solomon, "A gaussian approximation to the distribution of a definite quadratic form." *J. of the American Statistical Association*, pp. pages 898–902, 1972.

[9] R. Larsen and M. Mark, *An Introduction to Mathematical Statistics and its Applications*, 4th edition, Ed. Prentice-Hall, 2001.

[10] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *IMC '05*, 2005.