A New Approach to Anomaly Detection based on Possibility Distributions

Joseph Ndong
Department of Mathematics and Computer Science,
University Cheikh Anta Diop of Dakar, Sénégal
Email: joseph.ndong@ucad.edu.sn

***Abstract*—This paper presents a new approach for anomaly detection based on possibility theory for normal behavioral modeling. Combining subspace identification algorithms and Kalman filtering techniques could be a good basis to find a suitable model to build a decision variable where, a new decision process can be applied to identify anomalous events. A robust final decision scheme can be built, by means of possibility distributions to find the abnormal space where anomalies happen. Our system uses a calibrated state space dynamical linear model where the model's parameters are found by the principal component analysis framework. The multidimensional Kalman innovation process is used to build the unidimensional decision variable. Thereafter this variable is clustered and possibility distributions are used to separate the clusters into normal and abnormal spaces when anomalies happen. We had studied the false alarm rate *vs.* detection rate trade-off by means of the Receiver Operating Characteristic curve to show the high performance obtained via this new methodology against other approaches. We validate the approach over different realistic network traffic.**

***Index Terms*—Anomaly detection, GMM, probability-possibility theory, subspace identification, PCA, Kalman filter.**

I. Introduction

Kalman filter based techniques first calibrate a Maximum-Likelihood based model for normal behavior modeling for the entropy reduction step [10][11][12]. Thereafter the decision variable is obtained as the filter innovation process. Analyzing residual for anomaly detection can be a good approach, since in favorable conditions, this process is assumed to be a zero mean gaussian white noise. However, if we believe that anomalies can cause low, high or abrupt changes in the traffic, this can attempt to appear in different statistical properties in the residual, making us to believe that, this signal is instead an ensemble of normal distributions. So, it will be interesting to take into account the residual process and, try to build a few set of (normal/abnormal) clusters. Finally, our attention can be put on the abnormal clusters to track anomalies.

Principal component analysis (PCA) approach [7][8][15] provides very good model of normal behavior with strong differentiation with abnormal behavior. However it is weaken by its high sensitivity to non-stationarity and parameter settings. Whereas Kalman filtering approach is inherently more robust to some level of non stationarity in the data because of its feedback structure. However, the main weakness in the approach proposed initially in [10] is within the Maximum Likelihood estimation that fails in capturing the essential properties of the normal behavior. The previous analysis lead us to believe that combining a PCA based normal model with Kalman filtering step, can be a good basis for building a suitable decision variable where possibilistic test could be applied for anomaly detection. In this work, we show that *subspace identification* algorithm can be used in combination of a Kalman filter to build the decision variable.

In this work, we are interested in anomaly detection based robust unsupervised clustering. If we assumed that, generally, anomalies might be rare, one can build a *few* number of clusters and try to find them in some of these classes. There are two major informations which seem to be relevant for detecting true anomalous events, and which we want to exploit here, to build a robust anomaly detector. One can determine clearly the posterior probability of a data sample being distributed in the different clusters, but we have no idea of the probability of generating the clusters themselves. Thus, using possibility distribution to estimate the degree a cluster can be seen as ”possible”, should be a great interest for anomaly detection. Thus we follow [4] to characterize the unknown probabilities of generating a set of clusters by *simultaneous confidence intervals* with a given confidence level $1-\alpha$. Thereafter these intervals will be used to calculate possibility distributions (degree of possibility) for each cluster. This operation will have the ability to separate the different classes into normal and abnormal sub spaces. It will be, at the same time, necessary to have at hand the possibility distributions for the data sample to recover a critical value of the cluster possibility degree (which we will use to determine the normal and abnormal clusters).

The organization of this paper is as follows. Section III deals with the methodology we adopt in our anomaly detection scheme. In Section IV, we validate our approach by showing efficient results. Section V concludes the work and fix some ideas for future study.

II. Related Works

In our knowledge, this work presents the first approach that deals with possibility theory to build an anomaly detector for communication networks. Generally, in the literature, the proposed approaches are based on Bayesian inference i.e., probabilistic solutions. We have developed recently some techniques for anomaly detection using statistical approaches. In [23], the proposed method to detect anomalous events is based on gaussian mixture model (GMM) for clustering. Thereafter, we proposed a hidden markov model (HMM) coupled with the Viterbi algorithm to subdivide the space into two other sub spaces: the first containing a few number of cluster data corresponding to the abnormal sub space and the second one containing the majority of the data and corresponding to the normal space. However in that study, there is a great challenge to calibrate properly the GMM since it is necessary to run the model several times to achieve model convergence. The same problem occurs when searching for the best parameters of the HMM in order to learn about spacial and temporal correlations between the GMM clusters, in order to classify them into two or more states. In [27] the monitoring system is also based on the coupling of a GMM and HMM and the same problems arise. The searching of the best number of the HMM parameters need a thorough calibration of several models and the choice of the best model is based on the transition matrix. One should have high probabilities in the main diagonal of this matrix to decide the model selection. However ”high probability” was not defined more suitably. Our present work deal with these problems by proposing a new scheme based on possibility theory to separate the GMM clusters into two sub-spaces corresponding to the normal and abnormal regions. Our model does not necessitate multiple re-calibrations and the methodology to build a threshold to separate the space into other sub spaces is more reliable. In this work, we show the advantage to use the framework of possibility theory which is more reliable to characterize the probability of generating the clusters themselves, since we do not know their real distributions. This operation have the great advantage to mark a cluster as normal or abnormal. This findings was not achieve in the previous studies.

III. Normal behavior modeling

An anomaly detector is generally built using a normal behavior model. As network traffic is a dynamical signal, one would like to build a dynamical normal behavior model. A classical approach to model dynamical signal is using Linear Time Invariant State-Space (LTISS) [17] model, representing input-output multivariate data sequences, as shown in the following difference equations:

$$\begin{cases} x_{t+1} = Ax_t + Bu_t + w_t \\ y_t \; = Cx_t + Du_t + v_t \end{cases} \qquad (1)$$

In (1), the system state x_t, the measurable output y_t and the input u_t are multi-dimensional vectors of appropriate dimensions. The noise processes are assumed to be uncorrelated zero-mean gaussian white-noise processes with covariance matrices $cov(w_t) = Q$ and $cov(v_t) = R$, respectively. The input signal and the process noise are assumed to be statistically independent.

A. How to build the Decision Variable ?

Calibrating a normal behavior model need to finding the values (A, B, C, D) and (Q, R) that fit better a learning set containing signals gathered over a period where, no anomalies have happened. To calibrate these system quantities, we follow the methodology described in [16], where subspace identification algorithms are presented to be a valuable tool to identify the state space parameters. Sub space algorithms have the ability to provide accurate state space models for multivariate linear systems and to retrieve system related matrices as sub spaces of projected data matrices. This means that the Kalman filter states can be recovered from the given input-output data. The identification problem is essentially characterized by the extraction of these matrices from input-output data, by using QR factorization and Singular Value Decomposition (SVD). In this work, we use the sub space identification algorithms based on Multivariable Output-Error State Space (MOESP) approach. The main idea for models based on MOESP method is to reconstruct the *past* input-output and *future* input-output data. The multi-dimensional output response Y and input U are first transformed into block Hankel matrices. Then the MOESP algorithm performs the compression of a compound matrix using the input and output Hankel matrices, into a lower triangular matrix by means of orthogonal transformations and QR decomposition. Thereafter, the column space of specific sub matrices of the resulting lower triangular factor approximates the column space of the extended observability matrix in a convenient way. Thereafter PCA can be computed by means of SVD technique and solution of a set of linear equations can then be performed to find the deterministic components. There is many raisons to use sub space model identification methods (SMI) for state space parameter learning: i) when correctly implemented, SMI algorithms are fast, despite the fact that they use QR and SVD decomposition. They are faster than classical identification methods, such as Prediction Error Methods, because they are not iterative ii) numerical robustness is guaranteed precisely due to the well-understood algorithms obtained from numerical linear algebra iii) the user will never be confronted with problems such as: lack of (slow) convergence, numerical instability, local minima and sensitivity of initial estimates iv) the reduced model can be obtained directly from input-output data, without having to compute first the high order model, one is always inclined to obtain models with as low as an order as possible.

In subspace model identification approach, one key step is the approximation of a structural subspace from spaces defined by Hankel matrices, constructed from the input-output data. For the LTISS system, the matrix pair $\{A, C\}$ is assumed to be observable, which implies that all modes in the system can be observed in the output y_t and can thus be identified; this also implies that the rank of the extended observability matrix is equal to N. The system $\{A, (BQ^{1/2})\}$ is assumed **to be controllable i.e., the modes** of the system $\{A, Q^{1/2}\}$ are assumed to be stable. That structured subspace is the extended observability matrix Γ_i (where i denotes the number of block rows), which is defined as:

$$\Gamma_i = \begin{bmatrix} C & CA & CA^2 & \ldots & CA^{i-1} \end{bmatrix}^T \qquad (2)$$

From the LTISS, we can re-organize the data by the following

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-349-0

algebraic relationships:

$$Y_{k,i,j} = \Gamma_i X_{k,j} + H_i U_{k,i,j} + T_i E_{k,i,j} \tag{3}$$

For simplicity, we can rewrite the above equation as:
$Y = \Gamma_i X + H_i U + T_i E$.
In (3), $Y_{k,i,j}$, $U_{k,i,j}$, and $E_{k,i,j}$ are block Hankel matrices with i block rows and j block columns of the form:

$$\mathbf{Y_{k,i,j}} = \begin{bmatrix} y_k & y_{k+1} & \cdots & y_{k+j-1} \\ y_{k+1} & y_{k+2} & \cdots & y_{k+j} \\ \vdots & \vdots & \vdots & \vdots \\ y_{k+i-1} & y_{k+i} & \cdots & y_{k+j+i-2} \end{bmatrix} \tag{4}$$

The user-defined subscript i should be large enough, i.e larger than the order N of the system. The Hankel matrices $U_{k,i,j}$, and $E_{k,i,j}$ are defined in the same way.
H_i and T_i are Toeplitz matrices defined as:

$$\mathbf{H_i} = \begin{bmatrix} D & 0 & \ldots & 0 \\ CB & D & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ CA^{i-2}B & \ldots & CB & D \end{bmatrix} \tag{5}$$

$$\mathbf{T_i} = \begin{bmatrix} I & 0 & \ldots & 0 \\ CK & I & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ CA^{i-2}K & \ldots & CK & I \end{bmatrix} \tag{6}$$

The state sequence matrix $X_{k,j}$ is defined as:

$$X_{k,j} = [x_k + x_{k+1} + x_{k+2} + \ldots + x_{k+j-1}] \tag{7}$$

In MOESP algorithm, one has to determine the *extended observability matrix* by means of *orthogonal projections* on sub spaces span by U columns. Thereafter, one can extract the measurement matrix C by using the first block from the extended observability matrix Γ_i, and the state matrix A is obtained by using the following formulas:

$$\Gamma_{2:i+1} = \begin{bmatrix} CA & \ldots & CA^i \end{bmatrix}^T = \Gamma_i A \tag{8}$$

More precisely: $A = \Gamma_i^{\dagger}\Gamma_{2:i+1}$, where $(.)^{\dagger}$ denotes the Moore-Penrose pseudo inverse matrix.

To find the observability matrix, one has to first split the input-output data into distinct past and future input-output sequences, and thereafter build a lower triangular matrix by means of orthogonal transformations using the *QR* factorization as follows:

$$\begin{bmatrix} U_{1,i,j} \\ U_{i+1,i,j} \\ Y_{1,i,j} \\ Y_{i+1,i,j} \end{bmatrix} = \begin{bmatrix} R_{11} & 0 & 0 & 0 \\ R_{21} & R_{22} & 0 & 0 \\ R_{31} & R_{32} & R_{33} & 0 \\ R_{41} & R_{42} & R_{43} & R_{44} \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_1^T \\ Q_1^T \\ Q_1^T \end{bmatrix} \tag{9}$$

This *QR* factorization is only valid for the case where the input signal is a zero-mean white noise, as in our case. For an arbitrary input signal, see [17] for an appropriate *QR* factorization. Thereafter SVD is computed as:

$$[R_{42} \quad R_{43}] = \hat{Q}_s \hat{\Sigma}_s \hat{V}_s^T + \hat{Q}_N \hat{\Sigma}_N \hat{V}_N^T = \Gamma_i X \prod_{U^T}^{\perp} \tag{10}$$

where $X \prod_{U^T}^{\perp}$ has full rank N and $\prod_{U^T}^{\perp}$ denotes the orthogonal projections on the lines of the null space of U.

Γ_i is estimated from $\hat{Q}_s$, which has the N principal left singular vectors corresponding to the most significant singular values. After finding an approximation of the Toeplitz H_γ matrix (see [17]), the matrices B and D are computed from least-square solution of the following over determined system [18]:

$$\begin{bmatrix} R_{31} & R_{42} \end{bmatrix} \cong H_\gamma \begin{bmatrix} R_{11} & R_{22} \end{bmatrix} \tag{11}$$

After finding the above matrices by calibrating a predictive model by means of PCA, the model described by equation 1 is re-used, and we perform Maximum Likelihood using a Kalman filter, in order to build the *decision variable* using the multi-dimensional innovation process obtained as output of the Kalman filter. The ***one-dimensional*** decision variable (DV) process is obtained by applying the formulas:

$$decisionvariable = e(t)^T V e(t) \tag{12}$$

where the matrix V (obtained as output of the Kalman filter) is the inverse of the **variance** of the multi-dimensional innovation process $e(t)$, T denotes the transpose.

The Maximum Likelihood framework can be built by running the *predictor-corrector* iterative algorithm using two steps: *prediction* comes in the *time update* phase, and *correction* in the *measurement update* phase. Due to lack of space, we do not put in the text the different equations related to these two steps of the Kalman filter. The reader can find the calibration in our previous works in [23][27] and in other studies [11][12][21].

B. How to build to normal subspace ?

We aim in this paper to learn residuals (i.e.,, the innovation process as output of the Kalman filter) for anomaly detection. Generally, it is assumed that the Kalman residual is a zero mean white gaussian noise. But it is often false to consider this assertion as a whole property of this process. In place, we are assuming that the real distribution of the innovation process is a mixture of normal distributions. We can simply calibrate a gaussian mixture model (GMM) [25], to organize the data in few number of clusters (i.e gaussian components). Anomalies might then appear is some of these gaussian components, and if one can carefully extract the potentially "abnormal" clusters (the remaining being labelled as "normal"), a basic test should be applied to detect the anomalous events. We will see in this work that, this aim can be achieved via the use of possibility distributions. First, we will use the sophisticated High Dimensional Data Clustering (HDDC) method presented by Bouveyron and al. [26], which is robust to find the best number of clusters and model parameters with low complexity.

1) Clustering operation: Generally, measured observations in communication networks are high dimensional. A popular approach to perform unsupervised clustering is to use gaussian mixture model (GMM), which rely on the assumption that each class can be represented by a gaussian density. This method supposes that observations $\{x_1, \ldots, x_N\}$ are independent realizations of a random vector $X \in \mathbb{R}^p$ with density :

$$f(x, \theta) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \Sigma_k) \tag{13}$$

where π_k denotes the mixture proportion of the kth component and ϕ is the gaussian density parametrized by the mean μ_k and the covariance matrix Σ_k. The classical approach (the well-know quadratic discriminant analysis-QDA) requires the estimation of a very large number of parameter (proportional to p^2, the number of variables in the dataset), and therefore faces to numerical problems in high dimensional spaces. In addition, classical gaussian mixture models show a disappointing behavior when the size of the training dataset is too small compared to the number of parameters to estimate. To avoid overfitting, it is necessary to find a balance between the number of parameters to estimate and the generality of the model. The HDDC acts in this way and the approach assumes that high-dimensional data live around specific sub spaces with a dimension lower than p. Bouveyron and al. have introduced a new parametrization of the gaussian mixture model which takes into account the specific sub space around which each cluster is located and, therefore limits the number of parameters to estimate. Many kind of models are proposed and the estimation of the model's parameters is done via the Expectation-Maximization (EM) algorithm and, some variants as Classification EM (CEM) for faster convergence and Stochastic EM (SEM) to avoid initialization problem. The intrinsic dimension of each cluster is determined automatically with the scree test of Catell [20], where we search a break in the curse corresponding to a local maxima. The best number of clusters can be derived by means of the Bayesian information criterion (BIC) [24]. When running the HDDC for a given model r, some additional parameters are added that may increase the likelihood. This operation can cause overfitting which can be avoid by the BIC criterion which introduces a penalty term for the number of parameters in the model. With the set of estimated models, the one with the lower value of BIC is preferred.

After finding K clusters for the multi-dimensional innovation process (kalman residual), we applying the result to the unidimensional decision variable built in (12), to put it into K clusters. It is simple to achieve this, because the HDDC gives the cluster labels (as a sequence of N mixing symbols $[1, 2, \ldots, K]$). At the same time, the HDDC clustering phase gives us a $n \times K$ matrix representing the posterior probabilities t_{ik} that the observation i belongs to the cluster k, which can be used to calculate the possibility distribution for the data sample as defined in (17).

2) Possibility theory as a tool to build normal space: Our aim here is to infer possibility distribution from data to build the normal space. Dubois and Prade have built a procedure [1][2][3] which produces the most specific possibility distribution among the ones dominating a given probability distribution. In this paper, this method is generalized to the case where the probabilities (of generating the clusters) are **unknown**. We assume the above clusters have been generated from an unknown probability distribution. It is proposed to characterize the probabilities of generating the different clusters by *simultaneous confidence intervals* with a given confidence level $1 - \alpha$. A procedure for constructing a possibility distribution is described, insuring that the resulting possibility distribution will dominate the true probability distribution in at least $100(1 - \alpha)$ of the cases.
We will also use a procedure of computing possibilities for data sample, in the case where we have at hand the probability distributions of generating the data sample inside a cluster. This second kind of possibility distribution helps to label a cluster as normal or abnormal.

To build a possibility measure related to a cluster, we consider the parameter vector $p = (p_1, p_2, \ldots, p_K)$ of probabilities characterizing the unknown probability distribution of a random variable X on $\Omega = \{\omega_1, \ldots, \omega_K\}$. Let n_k denotes the number of observations of cluster k in a sample of size N. Then, the random vector $n = (n_1, \ldots, n_K)$ can be considered as a *multinomial* distribution with parameter p. A confidence region for p at level $1-\alpha$ can be computed using *simultaneous confidence intervals* as described in [4]. Such a confidence region can be considered as a set of probability distributions.

A consistency principle between probability and possibility was first stated by Zadeh [5] in an unformal way: "*what is probable should be possible*". This requirement is translated via the inequality:

$$P(A) \leq \Pi(A) \qquad \forall A \subseteq \Omega \tag{14}$$

where P and Π are, respectively, a probability and a possibility measure on a domain $\Omega = \{\omega_1, \ldots, \omega_K\}$. In this case, Π is said to dominate P. Transforming a probability measure into a possibilistic one then amounts to choosing a possibility measure in the set $\Im(P)$ of possibility measures dominating P. This should be done by adding a strong order preservation constraint which ensures the preservation of the shape of the distribution:

$$p_i < p_j \Leftrightarrow \pi_i < \pi_j \qquad \forall i, j \in \{1, \ldots, K\}, \tag{15}$$

where $p_i = P(\{\omega_i\})$ and $\pi_i = \Pi(\{\omega_i\})$, $\forall i \in \{1, \ldots, K\}$. It is possible to search for the most specific possibility distribution verifying (14) and (15) (a possibility distribution π is more specific than $\pi^{'}$ if $\pi \leq \pi^{'}, \forall i$). The solution of this problem exists, is unique and can be described as follows. One can define a strict partial order P on Ω represented by a set of compatible linear extensions $\Lambda(\mathsf{P}) = \{l_u, u = 1, L\}$. To each possible linear order l_u , one can associate a permutation σ_u of the set $\{1, \ldots, K\}$ such that:

$$\sigma_u(i) < \sigma_u(j) \Leftrightarrow (\omega_{\sigma_u(i)}, \omega_{\sigma_u(j)}) \in l_u, \tag{16}$$

The most specific possibility distribution, compatible with $p = (p_1, p_2, \ldots, p_K)$, can then be obtained by taking the maximum over all possible permutations:

$$\pi_i = \max_{u=1,L} \sum_{\{j|\sigma_u^{-1}(j)\le\sigma_u^{-1}(i)\}} p_j \tag{17}$$

The permutation σ is a bijection and the reverse transformation σ^{-1} gives the rank of each p_i in the list of the probabilities sorted in the ascending order. The number of permutations L depends on the duplicated p_i in p. It is equal to 1 if there is **no duplicate** p_i**,** $\forall i$ and for this case P is a *strict linear order* on Ω.

In the case of searching possibilities for the data cluster themselves, we do not know the probabilities p, and then we aim to build *confidence intervals* for each of the cluster c_i. In interval estimation, a scalar population parameter is typically estimated as a range of possible values, namely a confidence interval, with a given confidence level $1-\alpha$.

To construct confidence intervals for multinomial proportions, it is possible to find simultaneous confidence intervals with a joint confidence level $1-\alpha$. The method attempts to find a confidence region $\mathcal{C}_n$ in the parameter space $p = (p_1, \ldots, p_K) \in [0;1]^K | \sum_{i=1}^{K} p_i = 1$ as the Cartesian product of K intervals$[p_1^-, p_1^+] \ldots [p_K^-, p_K^+]$ such that we can estimate the coverage probability with:

$$\mathbb{P}(p \in \mathcal{C}_n) \ge 1-\alpha \tag{18}$$

We can use the Goodman formulation in a series of derivations to solve the problem of constructing the simultaneous confidence intervals [6]. Let

$$A = \chi^2(1-\alpha/K, 1) + N \tag{19}$$

where $\chi^2(1-\alpha/K, 1)$ denotes the quantile of order $1-\alpha/K$ of the chi-square distribution with one degree of freedom, and $N = \sum_{i=1}^{K} n_i$ denotes the size of the sample. We have also the following quantities:

$$B_i = \chi^2(1-\alpha/K, 1) + 2n_i, \tag{20}$$

$$C_i = \frac{n_i^2}{N}, \tag{21}$$

$$\Delta_i = B_i^2 - 4AC_i. \tag{22}$$

Finally, the bounds of the confidence intervals are defined as follows:

$$[p_i^-, p_i^+] = \left[\frac{B_i - \Delta_i^{\frac{1}{2}}}{2A}, \frac{B_i + \Delta_i^{\frac{1}{2}}}{2A}\right] \tag{23}$$

It is now possible, based on these above interval-valued probabilities, to compute the most possibility distributions of the data inside a cluster, dominating any particular probability measure. Let P denotes the partial order induced by the intervals $[p_i] = [p_i^-, p_i^+]$:

$$(\omega_i, \omega_j) \in \mathsf{P} \Leftrightarrow p_i^+ < p_j^- \tag{24}$$

As explained above, this partial order may be represented by the set of its compatible linear extensions $\Lambda(\mathsf{P}) = \{l_u, u = 1, L\}$, or equivalently, by the set of the corresponding permutations$\{\sigma_u, u = 1, L\}$. Then for each possible permutation σ_u associated to each linear order in $\Lambda(\mathsf{P})$, and each cluster ω_i, we can solve the following linear program:

$$\pi_i^{\sigma_u} = \max_{p_1,\ldots,p_K} \sum_{\{j|\sigma_u^{-1}(j)\le\sigma_u^{-1}(i)\}} p_j \tag{25}$$

under the constraints:

$$\begin{cases} \sum_{i=1}^{K} p_i = 1 \\ p_k^- \le p_k \le p_k^+ \qquad \forall k \in \{1, \ldots, K\} \\ p_{\sigma_u(1)} \le p_{\sigma_u(2)} \le \cdots \le p_{\sigma_u(K)} \end{cases} \tag{26}$$

Then, we can take the distribution of the cluster c_i dominating all the distributions π^{σ_u}:

$$\pi_i = \max_{u=1,L} \pi_i^{\sigma_u} \qquad \forall i \in \{1, \ldots, K\} \tag{27}$$

Finally we propose to build a measure of possibility distribution π_{normal} as a threshold, and then a cluster will be considered as normal if its possibility distribution satisfies :

$$\pi_i \ge \pi_{normal}, \tag{28}$$

Otherwise it is ranged in sub space potentially suspicious. Our attention will be placed in this sub space for anomaly detection. To find the possibility distribution π_{normal}, we use the *a posteriori* probability of each gaussian component (cluster) k [26]:

$$\Pr(k|x_t, \theta) = \frac{\pi_k \phi(x_t|\mu_k, \sum_k)}{\sum_{n=1}^{K} \pi_n \phi(x_t|\mu_n, \sum_n)} \tag{29}$$

which gives us, for each data point x_t the probability distribution $p = (p_1, p_2, \ldots, p_K)$ (for each data point the constraints $\sum_{i=1}^{K} p_i = 1$ is always obtained from (29)).

Thereafter, we can use (17) to calculate the corresponding possibility distribution of each data point x_t of the sample x. We obtain a matrix π_K^N of dimension $K \times N$ (remember K is the number of clusters and N is the length of the data sample x). We take the **max** for each column (each column containing the possibility distribution for data point x_t). Then we obtain a second matrix π_1^N and finally we use (30) to derive the threshold π_{normal} :

$$\pi_{normal} = max(\pi_1^N) \tag{30}$$

IV. Model Evaluation

A. Experimental data: Abilene and SWITCH networks

In this work, we used a collection of data coming from the Abilene network and SWITCH one. The Abilene backbone has 11 Points of Presence(PoP) and spans the continental US. The data from this network was collected from every PoP at the granularity of IP level flows. The Abilene backbone is composed of Juniper routers whose traffic sampling feature was enabled. Of all the packets entering a router, 1% are sampled at random. Sampled packets are aggregated at the 5-tuple IP-flow level and aggregated into intervals of 10 minute bins. The raw IP flow level data is converted into a PoP-to-PoP level matrix using the procedure described in [7]. Since the Abilene backbone has 11 PoPs, this yields a traffic matrix with 121 OD flows. Each traffic matrix element corresponds to a single OD flow, however, for each OD flow we have a seven week long time series depicting the evolution (in 10 minute bin increments) of that flow over the measurement period. All the OD flows have traversed 41 links. Synthetic anomalies are injected into the OD flows by the methods described in [7], and this resulted in 97 detected anomalies in the OD flows. The anomalies injected in the Abilene data are small and high *synthetic volume anomalies*. We used exactly the same Abilene data as in [14]. So for a full understanding on how the **ground-truth** is obtained (based on EWMA and Fourier algorithms) , we refer the reader to [14].

The second collection of data we used for our experiments is a set of three weeks of Netflow data coming from one of the peering links of a medium-sized ISP (SWITCH, AS559). Anomalies in the data were identified using available manual labelling 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. We refer the reader to [8][22] for a full view of this data set.

B. Validation

To validate our approach, we first run the MOESP algorithm in order to find the LTISS parameters and thereafter we perform a Kalman filter to perform entropy reduction and to retrieve the innovation process [23]. Thereafter the *unidimensional decision variable* is built as explained in Section III-A. As a second step, we calibrate a gaussian mixture model, using the HDDC approach, for the purpose of clustering the multivariate innovation process. The use of gaussian mixture models seems to be relevant if we assume that the innovation process is a mixture of normal distributions, instead of a simple uncorrelated gaussian white noise. It is important to note that the HDDC clustering operation is done on the multidimensional innovation matrix and not on the unidimensional decision vector, but we aim to clustering this univariate process. The HDDC method gives as output a single vector consisting of the unique sequence (class label) of symbols (alphabet) making possible to know the length of each cluster and the data belonging to it. We have used this class label to do the clustering of the decision variable. To validate our clustering model, we run the HDDC clustering operation for a set of r components ($r \in \{2, \ldots, 9\}$) and we select the model with the lowest value of the BIC.

The first result is about the calibration of the GMM model. In this unsupervised clustering technique, we adopt the following method to find the best number of clusters. We consider 8 partitions with different number of clusters $K \in \{2, 3, \ldots, 9\}$. Since each GMM model is characterized by the mean, prior and variance vectors, the best partition is simply the one with the lowest variance vector. In our experiments, we have found $K = 3$ clusters, both for the Abilene and the UDP traffic, and $K = 4$ classes for the TCP traffic. The rest of the computations accounts for the calculations of the possibility distributions for the clusters and the data sample. In Table I we show the degree of possibility and the sample size of each cluster. To decide if a cluster in normal or not, we consider the results in Table II showing the posterior probability distributions of the data sample (given by equation (29)) easily obtained as output of the HDDC clustering, and the corresponding possibility distributions computed via equation (17), respectively for the Abilene, TCP and UDP traffic. Table II shows clearly that possibility distributions measures dominate probability distributions. So with the framework of possibility theory, we could reinforce methodology based on bayesian inference. At this point, we can easily derive the critical possibility distribution π_{normal}, calculated via (30), which is used to determine the normal clusters. The table is truncated because $N \in \{480; 1008\}$, and they show that there's for each data point (at time t) a cluster for which the possibility distribution is equal to 1. Then we obtain $\pi_{normal} = 1$ if we apply equation (30). Finally, a cluster i will be considered as normal if its possibility distribution π_i^S satisfies $\pi_i \geq 1$ as defined in (28). Now, if one applies equation (27), he/she obtains for the Abilene case, the vector $\{1.0000; 0.0595; 0.0465\}$ corresponding respectively to the possibility distribution of generating the clusters #1, #2 and #3 in that order. Finally, it becomes clear that, only the cluster #1 defines the normal behavior and the remaining ones are in the abnormal domain. The same reasoning performed on the SWITCH data, gives that the clusters #1 and #3 define the normal space for the TCP traffic, while the clusters #3 defines the normal behavior for the UDP traffic. It is interesting to observe that the length of clusters, belonging to the normal subspace, is always the highest, and contains most of the data. This seams to be the normal situation in anomaly detection, since anomalies might be rare and might appear in some clusters with few data. Finally to perform the detection issue, one has just to extract, from the decision variable all the points corresponding to the data belonging to the abnormal subspace (i.e., clusters labelled as suspicious), and apply thresholding (a limit strictly superior to zero) to identify and detect the anomalous events.

We have chosen in this work, as a criterion of performance, to analyze the trade-off between the false positive rate (FPR) and the detection rate (DR). The results are shown in the ROC curves in Figure 1. Typically, the natural way to analyze a ROC curve is to calculate the area under the curve. If the area is high, it means that the DR is high (approaching 100%) and the FPR low (approaching 0%). However, there are other possibles interpretations of the ROC curve. For example, one can put the x-axis in logarithmic form in order to find different points for comparison of different curves. Then, from the results depicted in Figure 1, we can see obviously that the technique based possibility distribution performs best. On can extract reference

TABLE I
POSSIBILITY DISTRIBUTIONS π_i^S AND LENGTH OF EACH CLUSTER.

Abilene

cluster i	1	2	3	α
p_i^-	0.0424	0.9167	0.0018	0.05
p_i^+	0.0777	0.9534	0.0137	
π_i^S	**1.0000**	*0.0595*	*0.0465*	
Length cluster i	**936**	*50*	*22*	

Switch UDP

cluster i	1	2	3	α
p_i^-	0.2960	0.0932	0.4746	0.05
p_i^+	0.3993	0.1656	0.5830	
π_i^S	0.5254	*0.1656*	**1.0000**	
Length cluster i	166	*60*	**254**	

Switch TCP

cluster i	1	2	3	4	α
p_i^-	0.3058	0.0196	0.3337	0.1754	0.05
p_i^+	0.4145	0.0631	0.4441	0.2693	
π_i^S	**1.0000**	*0.0631*	**1.0000**	*0.3325*	
Length cluster i	**172**	*17*	**186**	*105*	

TABLE II
PROBABILITIES DISTRIBUTIONS OF THE DATA SAMPLE (DECISION VARIABLE) AND CORRESPONDING POSSIBILITY DISTRIBUTIONS, ($\alpha = 0.05$).

Abilene

time t	1	2	3	...	1007	1008
posterior probability distributions						
*cluster*1	0.0353	0.9995	0.1582	...	0.3908	0.1805
*cluster*2	0.9647	0.0004	0.0001	...	0.0001	0.0001
*cluster*3	0.0000	0.0001	0.8417	...	0.6091	0.8194
possibility distributions						
*cluster*1	0.0353	**1.0000**	0.1583	...	0.3909	0.1806
*cluster*2	**1.0000**	0.0005	0.0001	...	0.0001	0.0001
*cluster*3	0.0000	0.0001	**1.0000**	...	**1.0000**	**1.0000**

Switch TCP

time t	1	2	3	...	479	480
posterior probability distributions						
*cluster*1	0.0000	0.0000	0.0008	...	0.0000	0.9566
*cluster*2	1.0000	1.0000	0.0000	...	0.0000	0.0001
*cluster*3	0.0000	0.0000	0.9991	...	0.0000	0.0039
*cluster*4	0.0000	0.0000	0.0000	...	1.0000	0.0394
possibility distributions						
*cluster*1	0.0001	0.0000	0.0011	...	0.0001	**1.0000**
*cluster*2	**1.0000**	**1.0000**	0.0000	...	0.0002	0.0001
*cluster*3	0.0000	0.0003	**1.0000**	...	0.0001	0.0041
*cluster*4	0.0002	0.0001	0.0000	...	**1.0000**	0.0423

Switch UDP

time t	1	2	3	...	479	480
posterior probability distributions						
*cluster*1	0.0000	0.0000	0.6989	...	0.0000	0.1707
*cluster*2	1.0000	1.0000	0.0000	...	0.0000	0.8041
*cluster*3	0.0000	0.0000	0.3011	...	1.0000	0.0251
possibility distributions						
*cluster*1	0.0001	0.0001	**1.0000**	...	0.0002	0.2918
*cluster*2	**1.0000**	**1.0000**	0.0003	...	0.0001	**1.0000**
*cluster*3	0.0002	0.0000	0.5102	...	**1.0000**	0.0312

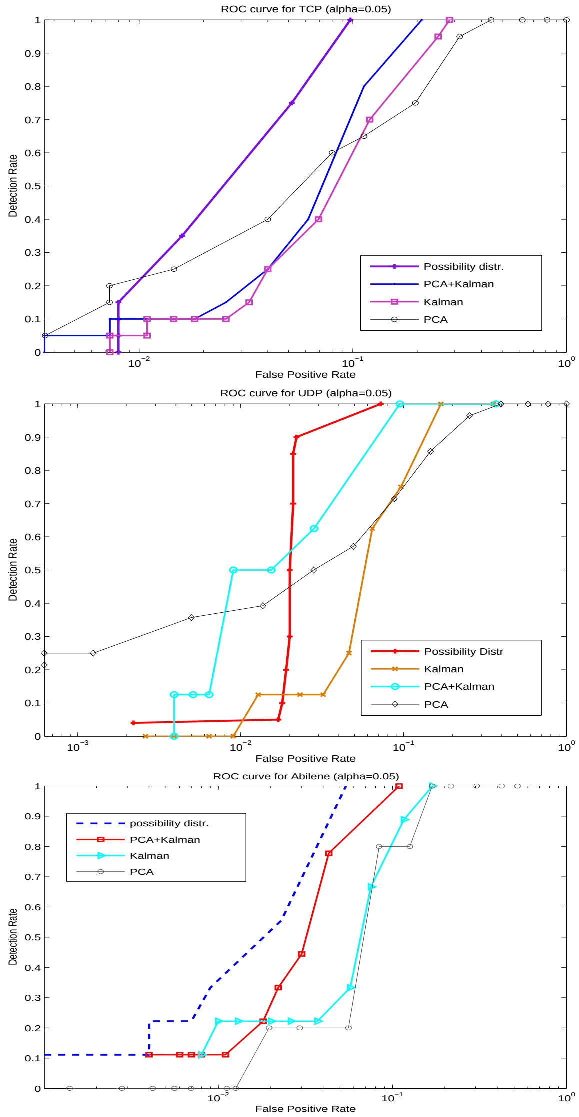

Fig. 1. ROC curve for the DR vs FPR. Top left graph for TCP, top right graph for UDP and top down for Abilene data when $\alpha = 0.05$.

points for which the FPR decrease significantly for our new scheme than for the other three techniques we had already derived in our previous works [23][27][28]. The method shows that we can achieve a DR of 100% with a FPR equal to 5%, where the best method of the three others, namely the PCA-Kalman method exhibits a FPR equal to 10%, for the Abilene data. For the SWITCH data, the new approach can achieve a probability of detection of 90% with a FPR of 2% against 7%, for the PCA-Kalman methodology, for UDP traffic. The same situation is observed for the TCP traffic.

V. CONCLUSION

In this work, we have shown the effectiveness and robustness of combining probability distributions, and possibility distributions for the purpose of anomaly detection. The robustness of the approach is achieved, in part, by the use of subspace identification algorithms (via the aid of PCA) and Kalman filtering technique, in order to build a unidimensional decision variable from multidimensional data set. Moreover, the great innovation in this paper is the use of possibility distributions to find the normal behavioral model, (by means of simple transformations from probability distributions) allowing us to extract the anomaly space. Another benefit of the solution can be found in the simplicity of the all procedure and the low complexity making easy to implement the algorithm. On the other hand, we have performed a robust and efficient high dimensional data clustering to build normal clusters with most of the data, and abnormal ones containing a few number of data where all anomalies lie. The experiments are done on different real traffic, and the ROC curve has shown high performance, compared to other techniques. It seems the main drawback of this work comes from the fact that the final decision process is based on applying manual thresholding. This problem will thus limit the applicability of the solution to dynamic and evolving systems. It will be of interest to search for more convenient technique, to automatically and dynamically adjust this threshold. We will try to address this issue soon.

REFERENCES

[1] Dubois D., Prade, H. and Sandri, S.: On possibility/probability transformations. In Proceedings of the Fourth Int. Fuzzy Systems Association World Congress (IFSA91), Brussels, Belgium, 1991, pages 50-53.

[2] Dubois, D. Foulloy, L., Mauris, G. and Prade, H. Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities. Reliable Computing, 2004, pp. 273-297.

[3] Dubois, D., Prade, H. and Sandri, S. On possibility/probability transformations. In Proceedings of the Fourth Int. Fuzzy Systems Association World Congress (IFSA91), Brussels, Belgium, 1991, pages 50-53.

[4] Masson, M., H. and Denoeux, T.: Inferring a possibility distribution from empirical data. Fuzzy Sets and Systems 157(3), 2006, pp. 319-340.

[5] Zadeh, L.,A. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 1978, pp. 3-28.

[6] Goodman, L. A.: On simultaneous confidence intervals for multinomial proportions. Technometrics, 7(2): 1965, pp. 247-254.

[7] Lakhina, A., Crovella, M. and Diot, C.: Characterization of network-wide traffic anomalies. In Proceedings of the ACM/SIGCOMM Internet Measurement Conference, 2004. pp. 201-206.

[8] Brauckhoff, D., Salamatian, K. and May, M.: Applying PCA for Traffic Anomaly Detection: Problems and Solutions. Proceedings IEEE INFOCOM, 2009 pp. 2866-2870.

[9] Ringberg, H.,Soule, A., Rexford, J., and Diot, C.:Sensitivity of PCA for Traffic Anomaly Detection. In ACM SIGMETRICS 2007.

[10] Soule, A., Salamatian, K.,Taft, N.: Traffic Matrix Tracking using Kalman Filters. ACM LSNI Workshop 2005.

[11] Soule, A., Salamatian, K. and Taft, N.: Combining Filtering and Statistical Methods for Anomaly Detection. USENIX , Association, Internet Measurement Conference, 2005, pp. 331344.

[12] Shumway, R. H. and Stoffer, D. S.: Dynamic Linear Models With Switching. Journal of the American Statistical Association, 1992, pp. 763-769.

[13] Eriksson, B., Barford, P., Bowden, R., Duffield, N., Sommers, J., Roughan, M. : BasisDetect: a model-based network event detection framework. In ACM IMC 2010.

[14] Lakhina, A., Crovella, M.,Diot, C.: Diagnosing Network-Wide Traffic Anomalies. In ACM SIGCOMM 2004.

[15] Brauckhoff, D., Salamatian, K. and May, M: Applying PCA for Traffic Anomaly Detection: Problems and Solutions. Technical Report INFOCOM 2009.

[16] Katayama, T.: subspace Methods for System Identification, Springer 2005.

[17] Verhaegen, M.: Identification of the Deterministic part of MIMO State Space Models given in Innovations Form from Input-Output Data. Journal Automatica, 1994, vol. 30 No 1.pp 61-74.

[18] Bottura, C.P, Tamariz, A.D.R, Barreto, G. and Cáceres, A.F.T: Parallel and Distributed MOESP Computational system's Modelling. Proceedings of the 10th Mediterranean Conference, 2002.

[19] Shumway, R. H. and Stoffer, D. S.: An Approach to Time Series Smoothing And Forecasting Using the EM Algorithm. Journal of Time Series Analysis, 1982, vol.3, No 4.

[20] Cattell, R.: The scree test for the number of factors. Multivariate Behavioral Research, 1966, pp. 245-276.

[21] Kailath, T., Sayed, A. H. and Hassibi B.: Linear Estimation. Prentice Hall, 2000.

[22] Brauckhoff, D., Dimitropoulos, X., Wagner, A. and Salamatian, K. : Anomaly Extraction in Backbone Networks using Association Rules. IMC09, November 46, Chicago, Illinois, USA, 2009.

[23] Ndong, J., Salamatian, K., :A Robust Anomaly Detection Technique Using Combined Statistical Methods. CNSR 2011, *IEEE Xplore* 978-1-4577-0040-8, 2011, pp: 101-108.

[24] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics, 1978, PP 461-464.

[25] Douglas A. R.: Gaussian Mixture Models. Encyclopedia of Biometrics, 2009, pp. 659-663.

[26] Bouveyron, C.,Girard, S., and Schmid, C. High-Dimensional Data Clustering, Computational Statistics and Data Analysis, vol. 52 (1), 2007, pp. 502-519.

[27] Ndong, J., Salamatian, K.,: Signal Processing-based Anomaly Detection Techniques: A Comparative Analysis. INTERNET 2011, The Third International Conference on Evolving Internet. ISBN: 978-1-61208-141-0.

[28] Ndong, J.,: Anomaly Detection: A Technique Using Kalman Filtering and Principal Component Analysis. ATAI NTC 2012 GSTF 2012.