Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 1542540, 11 pages
http://dx.doi.org/10.1155/2016/1542540

Research Article

Latent Clustering Models for Outlier Identification in Telecom Data

Ye Ouyang,[1] Alexis Huet,[2] J. P. Shim,[3] and Mantian (Mandy) Hu[4]

[1]*Columbia University, New York, NY, USA*
[2]*Nanjing Howso Technology, Nanjing, China*
[3]*Georgia State University, Atlanta, GA, USA*
[4]*Department of Marketing, The Chinese University of Hong Kong, Shatin, Hong Kong*

Correspondence should be addressed to Alexis Huet; alexis@howso.cn

Received 29 July 2016; Revised 3 November 2016; Accepted 17 November 2016

Academic Editor: Mariusz Głąbowski

Collected telecom data traffic has boomed in recent years, due to the development of 4G mobile devices and other similar high-speed machines. The ability to quickly identify unexpected traffic data in this stream is critical for mobile carriers, as it can be caused by either fraudulent intrusion or technical problems. Clustering models can help to identify issues by showing patterns in network data, which can quickly catch anomalies and highlight previously unseen outliers. In this article, we develop and compare clustering models for telecom data, focusing on those that include time-stamp information management. Two main models are introduced, solved in detail, and analyzed: Gaussian Probabilistic Latent Semantic Analysis (GPLSA) and time-dependent Gaussian Mixture Models (time-GMM). These models are then compared with other different clustering models, such as Gaussian model and GMM (which do not contain time-stamp information). We perform computation on both sample and telecom traffic data to show that the efficiency and robustness of GPLSA make it the superior method to detect outliers and provide results automatically with low tuning parameters or expertise requirement.

1. Introduction

High-speed telecom connections have developed rapidly in recent years, which has resulted in a major increase in data flow through networks. Beyond the issues of storage and management of this flow of data, a major challenge is how to select and use this mass of material to better understand a network. The detection of behaviors that differ from normal traffic patterns is a critical element, since such discrepancies can reduce network efficiency or harm network infrastructures. And because those anomalies can be caused by either a technical equipment problem or a fraudulent intrusion in the network, it is important to identify them accurately and fix them promptly. Data-driven systems have been developed to detect anomalies using machine learning algorithms and can automatically extract information from raw data to promptly alert a network manager when an anomaly occurs.

The data collected in telecom networks contains values for different features (related to network resource and usage) as well as time stamps, and those values can be modeled and processed to seek and detect anomalies using unsupervised algorithms. The algorithms use unlabeled data and assume that information about which data elements are anomalies is unknown (since anomalies in traffic data are rare and may take many forms). They do not directly detect anomalies but instead separate and distinguish data structures and patterns in order to group data from which "zones of anomalies" are deduced. The main advantage of this methodology is the ability to quickly detect previously unseen or unexpected anomalies.

Another component to be taken into consideration for understanding wireless network data behavior is time stamps. This information is commonly collected when data are generated but is not widely used in classic anomaly detection processes. However, since network load fluctuates over the course of a day, adding time-stamp attributes in an evaluation model can allow us to discover periodic behaviors. For example, a normal value during a peak period may be an

anomaly outside that period and thus remain undetected by an algorithm that does not take time stamps into account.

In this article, we use unsupervised models to detect anomalies. Specifically, we focus on algorithms combining both values and dates (time stamps) and introduce two new models to this end. The first one is the time-dependent Gaussian Mixture Model (time-GMM), which is a time-dependent extension of GMM [1] which works by considering each period of time independently. The second one is Gaussian Probabilistic Latent Semantic Analysis (GPLSA), derived from Probabilistic Latent Semantic Analysis (PLSA) [2], which combines values and dates processing together in a unique machine learning algorithm. This latter algorithm is well known in text-mining and recommender systems areas but has been rarely used in other domains such as anomaly detection. In this research, we fully implement these two algorithms with R [3] and test their ability to find anomalies and to adapt to new patterns on both sample and traffic data. We also compare the robustness, complexity, and efficiency of these algorithms.

The rest of the article is organized as follows: in Section 2, we present an overview of techniques to identify anomalies, with an emphasis on unsupervised models. In Section 3, we show different unsupervised anomaly detection models (this section defines two previously introduced unsupervised models: GPLSA and time-GMM). In Section 4, those models are compared to a sample set to highlight the differences of behavior in a simple context. In Section 5, we discuss computations performed on real traffic network data. We finally, in Section 6, draw conclusions about adaptability and robustness of GPLSA.

2. Research Background

Anomaly detection is a broad topic with a large number of previously used techniques. For a broad overview of those methods, we refer to [4].

Previous research focuses mainly on unsupervised statistical based methods such as clustering methods to perform anomaly detection [5–8]. A common assumption for statistical based methods is that the underlying distribution is Gaussian [9], although mixtures of parametric distributions, where normal-points anomalies correspond to two different distributions [10], are also possible. In clustering methods, the purpose is to separate data points and to group objects together that share similarities, and each group of objects is called a cluster. We usually define similarities between objects analytically. Many clustering algorithms that differ on how similarities between objects are measured (using distance measurement, density, or statistical distribution) exist but the most popular and simplest clustering technique is K-means clustering [11].

Advanced methods of detection combine statistical hypotheses and clustering, as seen in the Gaussian Mixture Model (GMM) [1]. This method assumes that all data points are generated from a mixture of K Gaussian distributions; parameters are usually estimated through an Expectation-Maximization (EM) algorithm, where the aim is to iteratively increase likelihood of the set [12]. Some studies have used GMM for anomaly detection problems, as described in [13–15]. Selecting the number of clusters K is not easy: Although methods to automatically select a value of K do exist (a comparison between different algorithms is presented in [16]), the selection is usually chosen manually by researchers and refined after performing different computations for different values.

In telecom traffic data, time stamps are a component to be considered when seeking for traffic anomalies. This information, referred to as contextual attributes in [4], can dramatically change the results of anomaly detection. For example, a value can be considered normal in a certain context (in a peak period) but abnormal in another context (in off-peak periods), and the differentiation can only be made clear when each value has a time stamp associated with it. An overview of outlier detection for temporal data can be found in [17], which comprises ensemble methods (e.g., [18, 19]), time-series models (e.g., with ARIMA or GARCH models in [20]), and correlation analysis [21, 22].

Clustering methods for temporal anomaly detection can automatically take into account and separate different types of behavior from raw time-series data, which allows for some interesting results. One way to incorporate time stamps is to consider the original GMM (i.e., a mixture of K Gaussian distributions), but to weigh each distribution differently, depending on time. This method was first introduced for text-mining [2, 23] with a mixture of categorical distributions and named Probabilistic Latent Semantic Analysis (PLSA). Its actual form (with Gaussian distribution), GPLSA, is used for recommendation systems [24]. No published article that applies GPLSA for anomaly detection has been found.

In the next section, we present five anomaly detection models for traffic data. The first three models are classic models: Gaussian model, time-dependent Gaussian, and GMM, which do not combine clustering and contextual detection and are expected to have several disadvantages. The two remaining models take clustering and time stamps into consideration: the fourth model is a time-dependent GMM, where a GMM is independently determined for each time slot; the fifth model is Gaussian Probabilistic Latent Semantic Analysis (GPLSA) model, which is solved by optimizing all parameters related to clusters and time in a unique algorithm.

3. Presentation of Models

In this section, five different models are defined: Gaussian, time-dependent Gaussian, GMM, time-dependent GMM, and GPLSA. We use the same following notations for all:

(i) W is a traffic data set. This set contains N values indexed with i. N is usually large, that is, from one thousand to one hundred million. Each value is a vector of $\mathbf{R}^P$, where p is the number of features. Furthermore, each feature is assumed to be continuous.

(ii) D is the time-stamp set of classes. This set also contains N values. Since we are expecting a daily cycle, each value d_i corresponds to each hour of the day, consequently standing in $\{1, \ldots, 24\}$.

(iii) $X = (W, D)$ are observed data.

Table 1: Anomaly detection methods compared.

	No date	Date
No clustering	Gaussian	Time-Gaussian
Clustering	GMM	(i) Time-GMM (ii) GPLSA

(iv) For clustering methods, we assume that each value is related to a fixed (although unknown) cluster, named Z. It is a "latent" set, since it is initially unknown. We assume that number of clusters K is known.

An example of traffic data retrieved is shown as follows:

date	Feat. 1	$\cdots$	Feat. p	W	D
04/13 0:00	1069	$\cdots$	2.4	$(1069, \ldots, 2.4)$	1
04/13 0:30	1004	$\cdots$	2.3	$(1004, \ldots, 2.3)$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
05/04 23:30	997	$\cdots$	2.7	$(997, \ldots, 2.7)$	24.

(1)

For each model, the aim is to estimate parameters with maximum likelihood. When the direct calculation is intractable, an EM algorithm is used to find a local optimum (at least) of the likelihood. A usual hypothesis of independence is added, which is needed to compute the likelihood of the product over the set:

(H) The set of triplets $(W_i, Z_i, D_i)_i$ is an independent vector over the rows i. Note that if the model does not consider D or Z, we remove this set in the hypothesis.

The different models are shown in Table 1, grouped according to their ability to consider time stamps and clustering. In the following, for each model, each hypothesis is listed on the form (X2), where X is current model paragraph followed by the hypothesis number.

3.1. Gaussian Model. In the Gaussian model, the whole data set is assumed to come from a variable that follows a Gaussian distribution. Consequently, each part of the day has a similar behavior and there are no clusters. Mathematically (note that same letter is used for set and variable) the following occurs:

(A1) Each variable W_i follows Gaussian distribution with mean and variance m, Σ. Here, m is a p-vector and Σ is a variance-covariance matrix of size p. They are both independent of i.

Parameters are easily estimated with empirical mean and variance.

3.2. Time-Dependent Gaussian Model. A time component is added to this model, as opposed to the Gaussian model, which does not include a time component. Each time of the day is considered independently, following a particular Gaussian distribution. This allows us to take dependence of time into account:

(B1) For each $s \in \{1, \ldots, 24\}$, each conditional variable W_i such that $D_i = s$ follows a Gaussian distribution with mean and variance m^s, Σ^s.

As for the Gaussian model, parameters are estimated with empirical mean and variance for each class of dates.

3.3. Gaussian Mixture Model. Compared to the Gaussian model, in the GMM, data is assumed to come from a mixture of Gaussian distributions rather than one single Gaussian distribution. The number of clusters K is fixed in advance.

(C1) Each record belongs to a cluster $Z_i = k \in \{1, \ldots, K\}$ with probability α_k.

(C2) Each variable $(W_i \mid Z_i = k)$ follows a Gaussian distribution with mean and variance m_k, Σ_k.

Therefore, each record belongs to an unknown cluster. The task is to estimate both probability for each cluster and the parameters of each Gaussian distribution. To solve this problem, the following decomposition is done:

$$P(W_i) = \sum_k P(W_i \mid Z_i = k) P(Z_i = k). \quad (2)$$

The parameters can be successively updated with an EM algorithm (see [23] for details).

3.4. Time-Dependent Gaussian Mixture Model. Combining the models described in Sections 3.2 and 3.3, we obtain the time-dependent GMM, which includes both clustering and time-dependence. As in Section 3.3, the EM algorithm is used to estimate parameters.

(D1) For each $s \in \{1, \ldots, 24\}$, each record such that $D_i = s$ belongs to a cluster $Z_i = k \in \{1, \ldots, K\}$ with probability $\alpha_{k,s}$.

(D2) For each $s \in \{1, \ldots, 24\}$, each variable $(W_i \mid Z_i = k)$ such that $D_i = s$ follows a Gaussian distribution with mean and variance m_k^s, Σ_k^s.

3.5. Gaussian Probabilistic Latent Semantic Analysis Model. The GPLSA model is based on the classic GMM but introduces a novel link between data values and time stamps. In time-GMM, the different classes of dates are considered independently, whereas GPLSA introduces dependence between latent clusters and time stamps but only within those two variables. That is, in knowing latent cluster Z, we assume there is no more dependence on time. This assumption allows making the problem computationally tractable. Explicitly, the following occurs:

(E1) For each $s \in \{1, \ldots, 24\}$, each record such that $D_i = s$ belongs to a cluster $Z_i = k \in \{1, \ldots, K\}$ with probability $\alpha_{k,s}$.

(E2) Each variable $(W_i \mid Z_i = k)$ follows a Gaussian distribution with mean and variance m_k, Σ_k.

(E3) For all i, $P(W_i \mid D_i, Z_i) = P(W_i \mid Z_i)$.

To solve this problem, the following decomposition is done (the assumption (E3) is used for the first factor of the sum):

$$P\left(W_i \mid D_i = s\right) = \sum_k P\left(W_i \mid Z_i = k\right) P\left(Z_i = k \mid D_i = s\right). \tag{3}$$

The EM algorithm can be adapted in this case to iteratively increase the likelihood and estimate parameters in order to obtain exact update formulas. The complete calculus to derive these formulas is given in the Appendix. We let $f(\cdot \mid m, \Sigma)$ equal the density of a Gaussian with parameters m and Σ. Also, we define E_s as the set of indexes i, where $d_i = s$. The following algorithm describes the steps to get final parameters:

Step 1. At time $t = 1$, let some initial parameters $m_k^{(t-1)}$, $\Sigma_k^{(t-1)}$, and $\alpha_{k,s}^{(t-1)}$ for all k, s.

Step 2. For all k, i, compute the probability $Z_i = k$ knowing $W_i = w_i$, $D_i = d_i$, and parameters

$$T_{k,i}^{(t)} := \frac{f\left(w_i \mid m_k^{(t-1)}, \Sigma_k^{(t-1)}\right) \alpha_{k,d_i}^{(t-1)}}{\sum_{l=1}^{K} f\left(w_i \mid m_l^{(t-1)}, \Sigma_l^{(t-1)}\right) \alpha_{l,d_i}^{(t-1)}}. \tag{4}$$

Step 3. For all k, s, compute (here $\#E_s$ stands for the length of E_s)

$$S_{k,s}^{(t)} = \sum_{j=1}^{\#E_s} T_{k,E_s(j)}^{(t)}. \tag{5}$$

Step 4. For all k, s, update $\alpha_{k,s}$ with

$$\alpha_{k,s}^{(t)} = \frac{S_{k,s}^{(t)}}{\sum_{l=1}^{K} S_{l,s}^{(t)}}. \tag{6}$$

Step 5. For all k, update the means with

$$m_k^{(t)} = \frac{\sum_{i=1}^{N} w_i T_{k,i}^{(t)}}{\sum_{i=1}^{N} T_{k,i}^{(t)}}. \tag{7}$$

Step 6. For all k, update the covariance matrix with (here $'$ refers to the transpose)

$$\Sigma_k^{(t)} = \frac{\sum_{i=1}^{N} \left(w_i - m_k\right)' \left(w_i - m_k\right) T_{k,i}^{(t)}}{\sum_{i=1}^{N} T_{k,i}^{(t)}}. \tag{8}$$

Step 7. Let $t = t + 1$ and repeat Steps 2 to 7 until convergence at a date T. At that date, parameters are estimated.

Step 8. For each i, the chosen cluster is k maximizing $T_{k,i}^{(T)}$.

Step 9. For each i, the likelihood of this point for the estimated parameters is

$$P\left(d_i\right) \sum_{l=1}^{K} f\left(w_i \mid m_l^{(T)}, \Sigma_l^{(T)}\right) \alpha_{l,d_i}^{(T)}. \tag{9}$$

4. Comparison of Models

All five models defined in Section 3 are implemented with R [3] into a framework that is able to perform computations and to show clustering and anomaly identification plots (using ggplot2 [25]). In this section, we apply our framework to a sample set to compare abilities to detect anomalies and check robustness of the methods. The sample set is built to highlight the difference of behaviors between models in a simple and understandable context. Consequently, only one sample feature is considered in addition to time-stamp dates.

In this set, we observe that time-GMM and GPLSA are able to detect anomalies within the set, and those methods are then potential candidates for anomaly detection in a time-dependent context. Furthermore, we show that GPLSA is more robust and allows a higher interpretation level of resulting clusters.

4.1. Sample Definition. The sample is built by superposing the three following random sets:

$$\begin{aligned} t &\longmapsto \cos\left(\frac{2\pi t}{T}\right) + \varepsilon, \\ t &\longmapsto \cos\left(\pi + \frac{2\pi t}{T}\right) + \varepsilon, \\ t &\longmapsto -2.5 + \varepsilon, \end{aligned} \tag{10}$$

where ε is independent random variables for each t sampled according to the continuous uniform distribution on $[0, 1]$ and where T has a daily period. The range of the two first functions is 24 hours, whereas the third one is only defined from 0:00 to 15:00.

Three anomalies are added on this set, defined, respectively, at 6:00, 12:00, and 18:00 with values −1.25, 0.5, and 1.65. The resulting set is shown in Figure 1.

4.2. Anomaly Identification. All five models are trained and the likelihood of each point is computed for each model. Since we expect 3 anomalies to be found in this sample set, the 3 lowest likelihood values are defined as anomalies for each model. For the clustering process, the chosen number of clusters is $K = 5$.

The results are shown in Figure 1. In (a), the whole data set is modeled as one Gaussian distribution and we found no expected anomalies. In (b), each period is determined with a Gaussian distribution, and we only discovered the anomaly at 18:00. In (c), the whole set is clustered and we only discovered the anomaly at 6:00. Finally, in (d), the time-GMM and GPLSA models are trained and the same results obtained: the 3 anomalies were successively detected.

Thus, time-GMM and GPLSA are both able to detect expected anomalies contrary to other methods.

4.3. Comparison between Time-GMM and GPLSA. The same anomalies have been detected with time-GMM and GPLSA. However, they are detected differently. We offer a summary of the comparison in Table 2.

First, GPLSA evaluates time stamps and values at once; that is, all parameters are estimated at the same time.

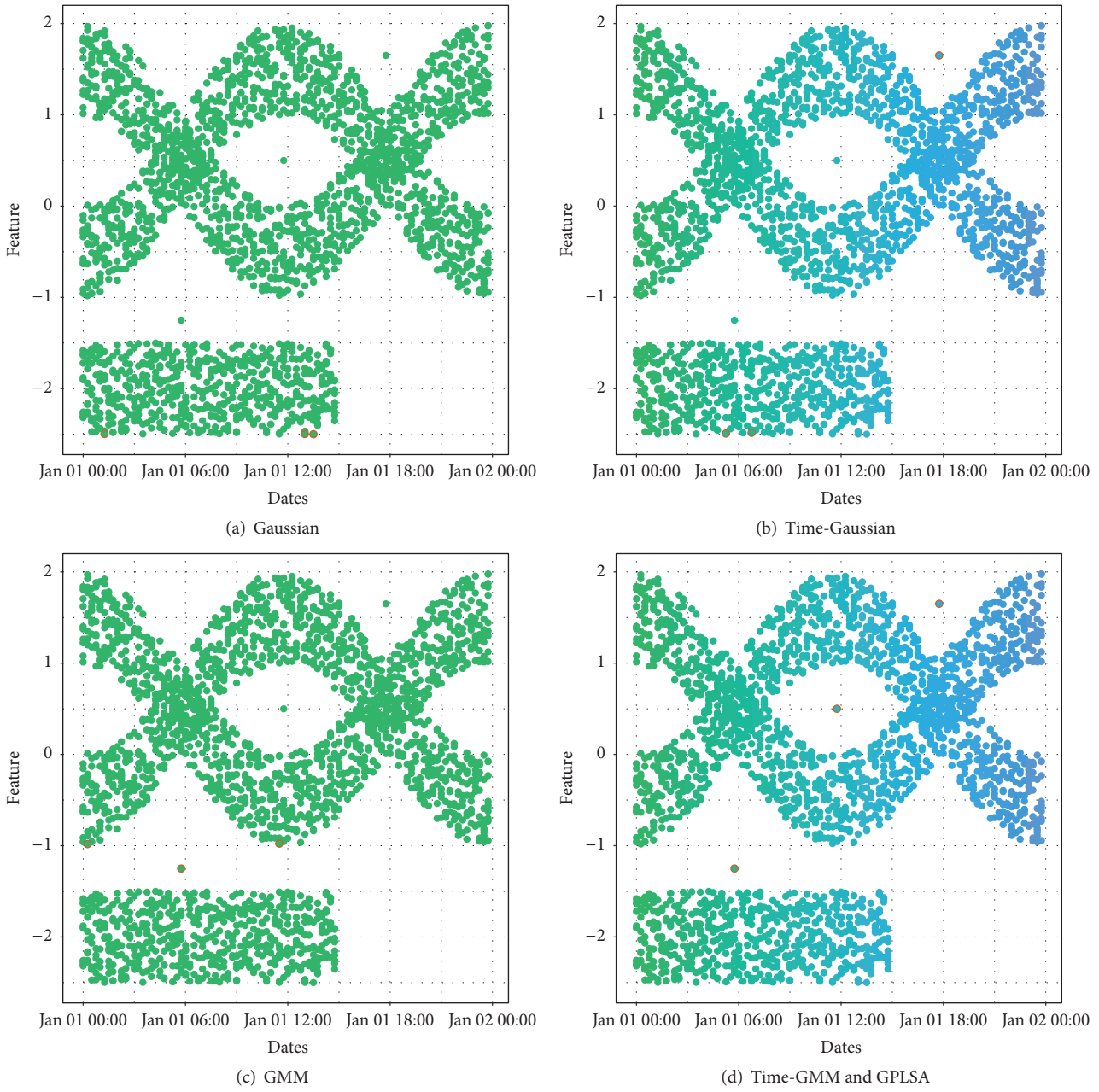


Figure 1: Anomaly detection for 5 different models in the sample set defined in Section 4. The three values with the lowest likelihood are circled in orange. Each color represents a different time-stamp class (only 1 class for (a) and (c); 24 classes for (b) and (d)).

Table 2: Comparison between time-GMM and GPLSA.

	Time-GMM	GPLSA
Cluster number	Fixed number of clusters at each date	Number of clusters can adapt to each date
Cluster relations	No relation between clusters of each date	Homogeneity of clusters across dates
Interpretability	Low	High
Data used	Only a part of data is used at each date	All data is used for each date
Nb: of param.	$(3K-1)D$	$(D+2)K$
Robustness	Medium	High

Consequently, consecutive dates can share similar clustering behaviors. With time-GMM, parameters are trained independently for each class of dates, and no relation exists between the clusters of different classes.

Second, the number of clusters in each class is soft for GPLSA (i.e., it can be different to the specified number of clusters for some class of dates). This allows the model to automatically adapt the number of clusters depending on which cluster is needed in the model. In time-GMM, each class has a specified number of clusters. This is shown in Figure 2, where the first seven hours are plotted in identified clusters for time-GMM (a) and GPLSA (b).

Third, the model is trained with the whole data for GPLSA, whereas only a fraction of data is used for each time-GMM computation. If there is a limited number of data in a class of dates, this can cause a failure to correctly estimate time-GMM parameters.

Fourth, the number of parameters needed for estimation is $(D + 2) \times K$ for GPLSA and $(3K - 1) \times D$ for time-GMM (with D number of classes and K number of clusters, and in dimension $p = 1$). Consequently, there are fewer parameters to estimate with GPLSA.

On the whole, GPLSA implies a better interpretation level (first and second points) of resulting clusters over time-GMM, combined with a higher robustness (third and fourth points).

5. Results and Discussion

In this section, anomaly detection is performed on real traffic network data. Based on the comparison of models done in Section 4, we select GPLSA to deduce anomalies and compare results with time-GMM. In Section 5.1, the collected data set is described and preprocessed; then, we apply GPLSA and show the results in Section 5.2. This Section 5.2 specifically focuses on behavior observed after applying the algorithm. Those results are compared with time-GMM results in Section 5.3. Finally, Section 5.4 highlights the ability of GPLSA to perform anomaly detection.

5.1. Data Description and Preprocessing. Data have been gathered from a Chinese mobile operator. They comprise a selection of 24 traffic features collected for 3,000 cells in the city of Wuxi, China. The features are only related to cell sites and do not give information about specific users. They represent, for example, the average number of users within a cell or the total data traffic for the last quarter of hour. The algorithm is trained over two weeks, with one value for each quarter of hour and for each cell.

We discarded the rows of data containing missing values. Only values and time stamps were taken into consideration for computations, and the identification number of cells was discarded. Some features only take nonnegative values and have a skewed behavior, and consequently, some features are preprocessed by applying the logarithm. To maintain interpretability, we do not apply feature normalization on variables. We expect that GPLSA can manage this set, even though some properties of the model are not verified, such as normality assumptions.

5.2. Computations and Results. We used the GPLSA model for the feature corresponding to the "average number of users within cell" and selected $K = 3$ clusters. Anomalies are values with the lowest resulting likelihood, computed to get (on average) 2 alerts and 8 warnings each day. Visual results are shown on Figure 3.

In (a), the three clusters are identified, whereas, in (b), a different color is used for each class of dates. In (c), the different log-likelihood values are shown. Finally, in (d), the estimation of the probability $\alpha_{k,s}$ to be in each cluster k knowing $D = s$ is plotted.

Anomalies are shown in (a), (b), and (c) and the extreme values related to each class of dates are correctly detected. In (a) and (d), identified clusters are shown in three distinct colors. The probability to be in each cluster varies across class as expected, with a lower probability in the upper cluster during off-peak hours. Also, as shown in (a), the upper cluster has a symmetric shape and the mean value is relatively similar across dates.

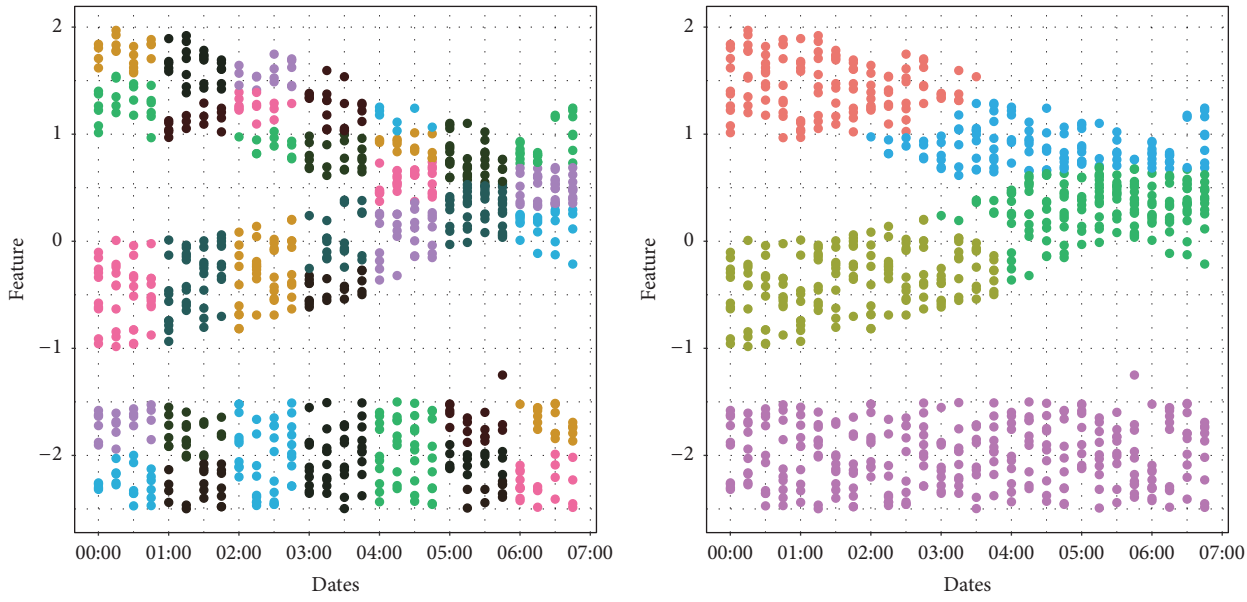
5.3. Comparison with Time-GMM. We compare results obtained in Section 5.2 with time-GMM, using the same number of clusters $K = 3$, and the same number of alerts and warnings each day. We show results on Figure 4. In (a), the three clusters are identified for each class D (between 1 and 24) and in (b), the different log-likelihood values are shown.

We observe that time-GMM correctly detects most of extreme values. Each class is related to a specific likelihood function and has its own way to represent data. We see that the cluster extents related to the highest values have a similar width for all classes on Figure 4(a) ($D = 1$ to 24). By comparing Figure 4(b) with Figure 3(c), we observe a larger "bump" (located in green during off-peak hours) for time-GMM. For these reasons, and contrary to GPLSA, anomalies are overrepresented in some classes (e.g., 3 warnings are detected for $D = 8$ for the first two days) whereas others do not contain anomalies for this time period ($D = 6$). Those results endorse the higher level of interpretation and robustness of GPLSA over time-GMM.

5.4. Discussion. According to the results, GPLSA is able to detect anomalies in a time-dependent context. We identified global outliers (e.g., on Figure 3(b) at Apr. 15 16:00 in red) as well as context-dependent anomalies (e.g., at Apr. 15 5:00 in orange). Off-peak periods are taken into consideration, and unusual values specific to those periods detected.

Gaussian hypothesis on GPLSA is not really constraining. As shown in Figure 3(a), clusters are adaptable and try to fit Gaussian distributions. They are appropriate to represent the value distribution for each class of dates and cluster.

Cluster adaptation is shown in Figure 3(d). The three clusters represent different level of values. The upper cluster represents higher values, which are more probable during peak periods. The lower cluster represents lower values, with a roughly constant probability. The third cluster in the middle is also useful to obtain a good anomaly detection behavior (results with $K = 2$ clusters are unable to correctly detect anomalies).



(a) Time-GMM (b) GPLSA

FIGURE 2: Identified clusters for 2 models in the sample set defined in Section 4 between 0:00 and 7:00. In (a), each class of one hour contains 5 clusters, and clusters are not related across hours. In (b), the whole set contains 5 clusters.

About anomaly detection itself, a threshold indicating the number of alerts to be detected can be set. This method of detection is static and relatively simple. Improving this method of detection is possible and straightforward through likelihood computations: inside a cell, an anomaly could be detected with a repetition of low likelihood scores.

6. Conclusion

In this paper, we present and compare unsupervised models to detect anomalies in wireless network traffic and demonstrated the robustness, interpretability, and ability of the GPLSA model to detect anomalies, as compared to other methods such as time-GMM. Anomaly detection was also performed and analyzed on real traffic data. We highlighted the adaptability of the GPLSA in this context to detect anomalies, even those with new patterns that are difficult to manually predict. As a result, mobile operators can have a versatile way to identify and detect anomalies, which would reduce the cost of possible aftermaths (e.g., network failure).

Improvement of this methodology could be operated. Currently, once the model is computed, anomaly detection is only based on punctual detection through likelihood values. A dynamic detection from consecutive values of likelihood could increase credibility of each alert and reduce the number of false alarms.

Furthermore, the model is only trained from a fixed data set in this research. But this could be extended by considering real-time stream data dealt with in an online context. Thus, new patterns could be updated quickly, to improve responsiveness and anomaly identification.

Appendix

A. Recall of Hypotheses for GPLSA

We recall that $X = (W, D)$ are observed data and Z are latent unobserved data.

Observed values are $(x_i)_{i\in\{1,\dots,N\}}$, where $x_i = (d_i, w_i)$. Each traffic value w_i is a vector of $\mathbf{R}^P$, where p is the number of features. Each time stamp d_i is an integer in $\{1, \dots, S\}$, with $S = 24$. In the following, levels of $\{1, \dots, S\}$ are indexed with s.

Latent values are $(z_i)_{i\in\{1,\dots,N\}}$. They take a finite number of states $k \in \{1, \dots, K\}$, where K is the defined number of clusters.

We recall the different hypotheses for GPLSA:

(H) The set of triplets $(W_i, Z_i, D_i)_i$ is an independent vector over the rows i.

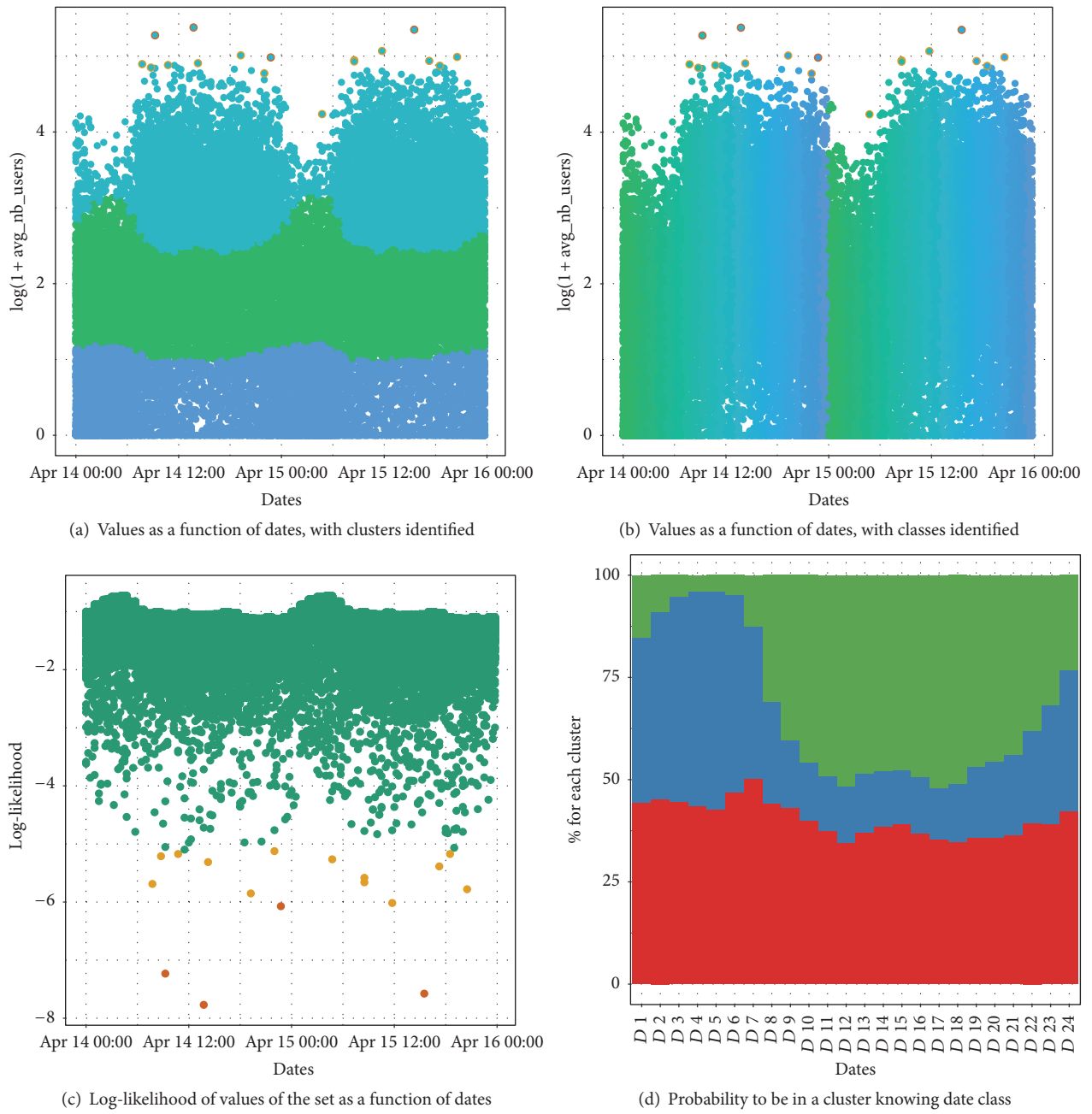
(E1) For each $s \in \{0, \dots, 23\}$, each record such that $D_i = s$ belongs to a cluster $Z_i = k \in \{1, \dots, K\}$ with probability $\alpha_{k,s}$.

(E2) Each variable $(W_i \mid Z_i = k)$ follows a Gaussian distribution with mean and variance m_k, Σ_k.

(E3) For all i, $P(W_i \mid D_i, Z_i) = P(W_i \mid Z_i)$.

Unknown parameters of the model are grouped together into $\theta := (\alpha_{k,s}, m_k, \Sigma_k;\ k \in \{1, \dots, K\},\ s \in \{1, \dots, S\})$.

Initial estimated parameters $\theta^{(0)}$ are defined as follows: all terms $\alpha_{k,s}^{(0)}$ are equal to $1/K$, and $(m_k^{(0)}, \Sigma_k^{(0)})$ are initialized using K-means clustering.



(a) Values as a function of dates, with clusters identified

(b) Values as a function of dates, with classes identified

(c) Log-likelihood of values of the set as a function of dates

(d) Probability to be in a cluster knowing date class

FIGURE 3: Anomaly detection with GPLSA from traffic data set presented in Section 5. Plots are restricted to two days in (a), (b), and (c). Red and orange points are related to the lowest likelihood values obtained, with an average of 2 red points and 8 orange points each day.

We define at each iteration t

$$\theta^{(t)} := \left(\alpha_{k,s}^{(t)}, m_k^{(t)}, \Sigma_k^{(t)};\ k \in \{1,\dots,K\},\ s \in \{1,\dots,S\}\right). \tag{A.1}$$

Estimated parameters $\theta^{(t)}$ are updated from $\theta^{(t-1)}$ iteratively using the EM algorithm. The algorithm stops when convergence of the related likelihood is reached.

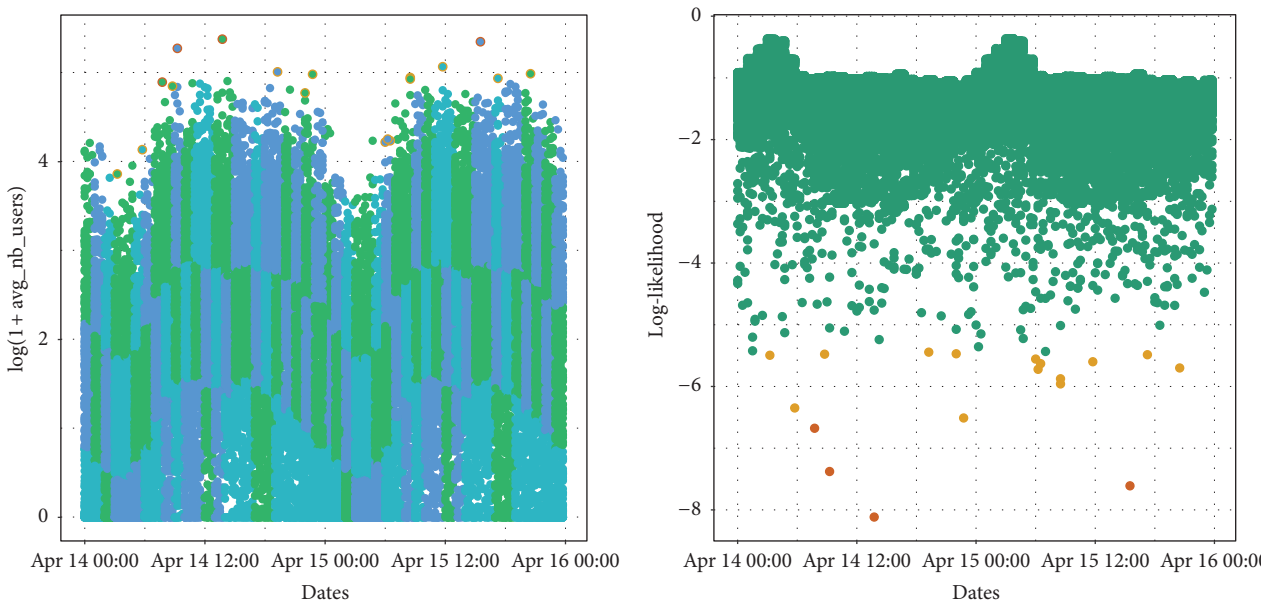
We use our hypotheses to express useful probabilities using θ. We recall that $f(\cdot \mid m, \Sigma)$ is density of a Gaussian with parameters m and Σ. Let $w_i \in \mathbf{R}^p$, $k \in \{1,\dots,K\}$ and $s \in \{1,\dots,S\}$.

From (E3), we know that $P(W_i = w_i \mid \theta,\ D_i = s,\ Z_i = k) = P(W_i = w_i \mid \theta,\ Z_i = k)$. Applying (E2), we deduce that $P(W_i = w_i \mid \theta,\ Z_i = k) = f(w_i \mid m_k, \Sigma_k)$. On the whole, we obtain

$$P\left(W_i = w_i \mid \theta,\ D_i = s,\ Z_i = k\right) = f\left(w_i \mid m_k, \Sigma_k\right). \tag{A.2}$$

Also, from (E1),

$$P\left(Z_i = k \mid D_i = s, \theta\right) = \alpha_{k,s}. \tag{A.3}$$



(a) Values as a function of dates, with clusters identified (cluster identification is independent for each time slot)

(b) Log-likelihood of values of the set as a function of dates

FIGURE 4: Anomaly detection with time-GMM from traffic data set as presented in Section 5. Plots are restricted to two days. Red and orange points are related to the lowest likelihood values obtained, with an average of 2 red points and 8 orange points each day.

This probability follows a discrete multinomial distribution that is proportional to $\alpha_{k,s}$ (where for each d, the coefficients sum to 1 over all k).

B. Recall about EM

The chosen strategy to estimate parameters θ is to find some parameters that maximize the marginal likelihood of the observed data X, as defined by

$$L(\theta; X) = P(X \mid \theta) = \sum_{Z} P(X, Z \mid \theta). \tag{B.1}$$

As the direct computations are intractable, we use EM to update parameters iteratively:

(1) Set some initial parameters $\theta^{(0)}$. For t from 0 until convergence, repeat the following steps (2) and (3).

(2) Perform the expectation step (E step):

$$Q\left(\theta \mid \theta^{(t)}\right) \coloneqq \mathrm{E}_{Z \mid X, \theta^{(t)}}\left[\log\left(L(\theta; X, Z)\right)\right] \tag{B.2}$$

which can be rewritten as

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{Z} \log P(X, Z \mid \theta)\, P\left(Z \mid X, \theta^{(t)}\right). \tag{B.3}$$

(3) Perform the maximization step (M step):

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta \mid \theta^{(t)}\right). \tag{B.4}$$

A theoretical reason to update the expected value function $Q(\cdot \mid \theta^{(t)})$ is that likelihood $L(\theta; X)$ will increase or remain constant at each step [26]. However, after convergence, the parameters can be stuck in a local maximum of the likelihood function.

C. Expectation Step of EM in the GPLSA Context

We assume we are in step t, and we want to update $m_k^{(t)}$, $\Sigma_k^{(t)}$, $\alpha_{k,s}^{(t)}$ for all k and s. From (B.3) and using hypothesis (H), we get

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} \log P\left(X_i = x_i,\ Z_i = k \mid \theta\right) \cdot P\left(Z_i = k \mid X_i = x_i, \theta^{(t)}\right). \tag{C.1}$$

For the left term, since $X_i = (D_i, W_i)$ and using equations (A.2) and (A.3),

$$P\left(D_i = d_i,\ W_i = w_i,\ Z_i = k \mid \theta\right) = f\left(w_i \mid m_k, \Sigma_k\right) \alpha_{k,d_i} P\left(D_i = d_i\right). \tag{C.2}$$

For the right term, using (A.2) and (A.3) for parameter $\theta^{(t)}$,

$$P\left(Z_i = k \mid X_i = x_i, \theta^{(t)}\right) = \frac{P\left(W_i = w_i \mid Z_i = k\right) P\left(Z_i = k \mid D_i = d_i\right)}{\sum_{l=1}^{K} P\left(W_i = w_i \mid Z_i = l\right) P\left(Z_i = l \mid D_i = d_i\right)}, \tag{C.3}$$

$$P\left(Z_i = k \mid X_i = x_i, \theta^{(t)}\right) = \frac{f\left(w_i \mid m_k^{(t)}, \Sigma_k^{(t)}\right)\alpha_{k,d_i}^{(t)}}{\sum_{l=1}^{K} f\left(w_i \mid m_l^{(t)}, \Sigma_l^{(t)}\right)\alpha_{l,d_i}^{(t)}}. \tag{C.4}$$

Then we define $T_{k,i}^{(t)}$ as $P(Z_i = k \mid X_i = x_i, \theta^{(t)})$, which is explicitly computable from (C.4).

Seen in the whole,

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} \log\left[f\left(w_i \mid m_k, \Sigma_k\right)\alpha_{k,d_i} P\left(D_i = d_i\right)\right] T_{k,i}^{(t)}, \tag{C.5}$$

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K}\left[-\frac{K}{2}\log 2\pi - \frac{1}{2}\cdot \log\left|\Sigma_k\right| \frac{1}{2}\left(x_i - m_k\right)' \Sigma_k^{-1}\left(x_i - m_k\right) + \log\alpha_{k,d_i} + \log P\left(D = d_i\right)\right] T_{k,i}^{(t)}. \tag{C.6}$$

Finally, we obtain an explicit formula for $Q(\theta \mid \theta^{(t)})$ which can be maximized.

D. Expectation Step of EM in the GPLSA Context

From the shape (C.6) of $Q(\cdot \mid \theta^{(t)})$, we can separate maximization of (m_k, Σ_k) for each k and weights $(\alpha_{k,s})_k$ for each s.

(1) For the Weights $\alpha_{k,s}$. For each fixed time stamp s, we update $(\alpha_{k,s})_k$. These are considered all together since there is a constraint: the sum over k has to be 1. From (C.6), we only have to maximize

$$G\left((\alpha_{k,s})_k\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} \log\alpha_{k,d_i} T_{k,i}^{(t)} = \sum_{k=1}^{K}\sum_{i=1}^{N} \log\alpha_{k,d_i} T_{k,i}^{(t)}. \tag{D.1}$$

For each $s \in \{1, \ldots, S\}$, we let E_s the set of indexes $i \in \{1, \ldots, N\}$ such that $d_i = s$. Therefore

$$G\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{\#E_s} \log\alpha_{k,s} T_{k,E_s(j)}^{(t)}, \qquad G\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K}\sum_{s=1}^{S} \log\alpha_{k,s}\sum_{j=1}^{\#E_s} T_{k,E_s(j)}^{(t)}. \tag{D.2}$$

We let for all k, s: $S_{k,s}^{(t)} := \sum_{j=1}^{\#E_s} T_{k,E_s(j)}^{(t)}$ to obtain

$$G\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K}\sum_{s=1}^{S} S_{k,s}^{(t)} \log\alpha_{k,s}. \tag{D.3}$$

Since s is fixed, all terms except one are constant. Consequently, we only have to maximize

$$F\left((\alpha_{k,s})_k\right) := \sum_{k=1}^{K} S_{k,s}^{(t)} \log\alpha_{k,s}. \tag{D.4}$$

Finally, we compute the derivative with respect to $\alpha_{k,s}$. Here, we remember that $\sum_{k=1}^{K}\alpha_{k,s} = 1$. To remove this constraint, we let $\alpha_{K,s} = 1 - (\alpha_{1,s} + \cdots + \alpha_{K-1,s})$. We rewrite this as

$$F\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K-1} S_{k,s}^{(t)} \log\alpha_{k,s} + S_{K,s}^{(t)} \log\left(1 - \left(\alpha_{1,s} + \cdots + \alpha_{K-1,s}\right)\right). \tag{D.5}$$

By differentiation,

$$\frac{\partial F\left((\alpha_{k,s})_k\right)}{\partial\alpha_{k,s}} = \frac{s_{k,s}^{(t)}}{\alpha_{k,s}} - \frac{s_{K,s}^{(t)}}{\alpha_{K,s}}, \tag{D.6}$$

$$\frac{\partial F\left((\alpha_{k,s})_k\right)}{\partial\alpha_{k,s}} = \frac{S_{k,s}^{(t)}}{\alpha_{k,s}} - \frac{S_{K,s}^{(t)}}{\alpha_{K,s}}. \tag{D.7}$$

If we want this value (D.6) to be zero, we get

$$\alpha_{k,s} = \alpha_{K,s}\frac{S_{k,s}^{(t)}}{S_{K,s}^{(t)}}. \tag{D.8}$$

Now, using the constraint

$$1 = \alpha_{K,s}\frac{\sum_{k=1}^{K} S_{k,s}^{(t)}}{S_{K,s}^{(t)}} + \alpha_{K,s} \tag{D.9}$$

then,

$$\alpha_{K,s} = \frac{S_{K,s}^{(t)}}{\sum_{k=1}^{K} S_{k,s}^{(t)}}. \tag{D.10}$$

This follows for all $k \in \{1, \ldots, K\}$

$$\alpha_{k,s} = \frac{S_{k,s}^{(t)}}{\sum_{l=1}^{K} S_{l,s}^{(t)}}. \tag{D.11}$$

By computing the Hessian matrix, we find that the obtained extremum is the maximum value.

(2) For the Means and Variances (m_k, Σ_k). From (C.6), we can perform computations for each fixed cluster k. Since some terms of this sum have no dependence on k, we have to maximize

$$-\frac{1}{2}\sum_{i=1}^{N}\left[\log\left|\Sigma_k\right| + \left(x_i - m_k\right)' \Sigma_k^{-1}\left(x_i - m_k\right)\right] T_{k,i}^{(t)} \tag{D.12}$$

or minimize

$$\sum_{i=1}^{n}\left[\log\left|\Sigma_k\right| + \left(x_i - m_k\right)' \Sigma_k^{-1}\left(x_i - m_k\right)\right] T_{k,i}^{(t)}. \tag{D.13}$$

We obtain the same formula as for GMM and then give the update rules with

$$m_k = \frac{\sum_{i=1}^{n} x_i T_{k,i}^{(t)}}{\sum_{i=1}^{n} T_{k,i}^{(t)}}, \qquad \Sigma_k = \frac{\sum_{i=1}^{n}\left(x_i - m_k\right)'\left(x_i - m_k\right) T_{k,i}^{(t)}}{\sum_{i=1}^{n} T_{k,i}^{(t)}}. \tag{D.14}$$

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] G. J. McLachlan and K. E. Basford, "Mixture models. Inference and applications to clustering," in *Statistics: Textbooks and Monographs*, Dekker, New York, NY, USA, 1988.

[2] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, 1999.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016, https://www.R-project.org/.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, article 15, 2009.

[5] P. Laskov, P. Düssel, C. Schäfer, and K. Rieck, "Learning intrusion detection: supervised or unsupervised?" in *Image Analysis and Processing—ICIAP 2005: 13th International Conference, Cagliari, Italy, September 6–8, 2005. Proceedings*, vol. 3617 of *Lecture Notes in Computer Science*, pp. 50–57, Springer, Berlin, Germany, 2005.

[6] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[7] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[8] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[9] V. Bamnett and T. Lewis, *Outliers in Statistical Data*, 1994.

[10] D. Agarwal, "Detecting anomalies in cross-classified streams: a Bayesian approach," *Knowledge and Information Systems*, vol. 11, no. 1, pp. 29–44, 2007.

[11] A. K. Jain, "Data clustering: 50 years beyond *K*-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[13] H. Hajji, "Statistical analysis of network traffic for adaptive faults detection," *IEEE Transactions on Neural Networks*, vol. 16, no. 5, pp. 1053–1063, 2005.

[14] Y. Ouyang, M. H. Fallah, S. Hu et al., "A novel methodology of data analytics and modeling to evaluate LTE network performance," in *Proceedings of the 13th Annual Wireless Telecommunications Symposium (WTS '14)*, IEEE, Washington, DC, USA, April 2014.

[15] M. J. Desforges, P. J. Jacob, and J. E. Cooper, "Applications of probability density estimation to the detection of abnormal conditions in engineering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 212, no. 8, pp. 687–703, 1998.

[16] M. M. T. Chiang and B. Mirkin, "Intelligent choice of the number of clusters in *k*-means clustering: an experimental study with different cluster spreads," *Journal of Classification*, vol. 27, no. 1, pp. 3–40, 2010.

[17] M. Gupta, J. Gao, C. Aggarwal, and J. Han, "Outlier detection for temporal data," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 5, pp. 1–129, 2014.

[18] S. Rayana and L. Akoglu, "Less is more: building selective anomaly ensembles," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 4, article 42, 2016.

[19] Y. Ouyang and T. Yan, "Profiling wireless resource usage for mobile apps via crowdsourcing-based network analytics," *IEEE Internet of Things Journal*, vol. 2, no. 5, pp. 391–398, 2015.

[20] J. Guo, W. Huang, and B. M. Williams, "Real time traffic flow outlier detection using short-term traffic conditional variance prediction," *Transportation Research Part C: Emerging Technologies*, vol. 50, pp. 160–172, 2015.

[21] A. Y. Lokhov, N. Lemons, T. C. McAndrew, A. Hagberg, and S. Backhaus, "Detection of cyber-physical faults and intrusions from physical correlations," https://arxiv.org/abs/1602.06604.

[22] Y. Ouyang and H. M. Fallah, "A performance analysis for UMTS packet switched network based on multivariate KPIs," in *Proceedings of the IEEE Wireless Telecommunications Symposium*, Tampa, Fla, USA, April 2010.

[23] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, Berkeley, Calif, USA, 1999.

[24] Z. Lu, W. Pan, E. W. Xiang, Q. Yang, L. Zhao, and E. H. Zhong, "Selective transfer learning for cross domain recommendation," https://arxiv.org/abs/1210.7056.

[25] H. Wickham, *Elegant Graphics for Data Analysis*, Springer Science & Business Media, 2009.

[26] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1987.