Visualization of Anomalies using Mixture Models

Tomoharu Iwata and Kazumi Saito

NTT Communication Science Laboratories, NTT Corporation
Hikaridai 2–4, Seika-cho, Soraku-gun, Kyoto, 619–0237, Japan

Abstract. Anomaly detection is important to learn from major past events and to prepare for future crises. We propose a new anomaly detection method that visualizes multivariate data in a 2- or 3-dimensional space based on the probability of belonging to a mixture component and the probability of not belonging to any components. It helps to visually understand not only the magnitude of anomalies but also the relationships among anomalous and normal samples. This may provide new knowledge in the data, since we can see it from a different viewpoint. We show the validity of the proposed method by using both an artificial and an economic time series.

1 Introduction

In recent years, a huge amount of data has been electronically accumulated. It is important to detect anomalies in data to learn from major past events and to prepare for future crises. Anomaly detection techniques are used for various such purposes as the detection of unauthorized computer use by examining the user log data [8] or the detection of anthrax outbreaks by tracking over-the-counter medication sales [7].

Many anomaly detection methods have been proposed, mainly in the field of outlier detection [2]. However, most resort to plotting anomaly scores in the data [19]. Such methods can determine to what degree a sample in data is anomalous, however, they cannot grasp relationships among anomalous and normal samples. Since anomalies happen for various reasons, it is important to know their relationships to understand their causes.

To overcome this problem, we propose a new method of the visualization of multivariable data for anomaly detection that uses a mixture model. The method visualizes data in a 2- or 3-dimensional space based on the probability of belonging to a mixture component and the probability of not belonging to any components. The results of visualization provide not only the magnitude of anomalies but also the relationship among data. This can be used to find new knowledge in the data, since we can see it from a different viewpoint.

The remainder of this paper is organized as follows. In the next section, we present our new method, and in Section 3, we explain how to apply our method to a time series. In Section 4, we show the validity of our method by using both an artificial and an economic time series. In the last section, we present concluding remarks and future works.

2 Proposed Method

In this section we present a new method to visualize data for anomaly detection. The procedure follows:

1. Given particular data, build a mixture model.
2. Estimate the probability of belonging to each mixture component.
3. Calculate the average 2-sigma value $\tilde{\alpha}$, which is used to detect anomalies.
4. Create an augmented probability vector $\mathbf{z}$ by combining the probabilities estimated in Step 2 and $\tilde{\alpha}$.
5. Visualize $\mathbf{z}$ by using a dimension reduction method.

2.1 Mixture Model

Let $\mathcal{D} = \{D(1), \ldots, D(N)\}$ be a given data set, where N is the number of data samples. A mixture model assumes that each sample $D(i)$ is generated by a finite mixture of underlying probability distributions:

$$p(D(i)|\Theta) = \sum_{k=1}^{K} P(k)p(D(i)|k, \Theta), \tag{1}$$

where $\sum_{k=1}^{K} P(k) = 1$, $P(k) \geq 0$, K is the number of components, and Θ is a vector constructed by all parameters. We estimate parameters Θ by maximum likelihood estimation:

$$\hat{\Theta} = \arg\max_{\Theta} p(\mathcal{D}|\Theta) = \arg\max_{\Theta} \prod_{i=1}^{N} p(D(i)|k, \Theta), \tag{2}$$

where $\hat{\Theta}$ is the estimated parameters by maximum likelihood method.

We can apply any kind of mixture component, such as a normal distribution model, a multinomial distribution model, or an autoregressive model. It is important to select a component that can appropriately model the given data and whose parameters can be easily estimated.

2.2 Augmented Probability Vector

Next, we explain how to obtain an augmented probability vector, which is the combination of the probability of belonging to each component and not belonging to any components. We assume that a sample that does not belong to any components is an anomaly.

We consider the joint probability density of k-th component and i-th sample, $P(k, D(i))$, as the probability of belonging to component k. We define the probability vector of i-th sample, $\mathbf{z}^*(i)$, by a vector where the element is the joint probability density as follows:

$$\begin{aligned} \mathbf{z}^*(i) &= (p(k = 1, D(i)), \ldots, p(k = K, D(i))) \\ &= (P(k = 1)p(D(i)|k = 1), \ldots, P(k = K)p(D(i)|k = K)). \end{aligned} \tag{3}$$

$\mathbf{z}^*(i)$ represents the probabilities of belonging to each component of the i-th sample. Anomalies have low probabilities of belonging to any components. Therefore, by augmenting a small value $\tilde{\alpha}$ to $\mathbf{z}^*$ and normalizing it so that their sum is 1, the vector represents the probabilities of belonging to each component (by 1st $\sim$ K-th elements) and not belonging to any components (by $(K+1)$-th element). Then, the vector $\mathbf{z}(i)$ is:

$$\mathbf{z}(i) \propto (P(k=1)p(D(i)|k=1), \ldots, P(k=K)p(D(i)|k=K), \tilde{\alpha}), \tag{4}$$

where $\sum_{k=1}^{K+1} \mathbf{z}(i)_k = 1$. We call this an augmented probability vector. We determine $\tilde{\alpha}$ based on the average 2-sigma value, that is, the average of probability densities on 2-sigma of all components:

$$\tilde{\alpha} = \sum_{k=1}^{K} P(k)p(\mu_k \pm 2\sigma_k|k), \tag{5}$$

where μ_k and σ_k stand for the mean and standard deviation of conditional probability $P(D(i)|k)$. The probability that a sample exists in the 2-sigma area is 0.9545 in the case of a normal distribution. Therefore, most samples have higher joint probability densities than the average 2-sigma value, and so we consider them normal. Since a few samples have lower probabilities of belonging to any components than an average 2-sigma value, we consider them anomalies. Here we use an average 2-sigma value, however, an average 3-sigma value or others can be defined in the same way. To reduce false alarm, we could adopt an average 3-sigma value. We can choose $\tilde{\alpha}$ in accordance with the application.

2.3 Visualization

Last, we explain how to visualize the augmented probability vectors and how to study the results of visualization. An augmented probability vector is a vector of $K+1$ dimension. We need to reduce the dimension of this vector to 2 or 3 to visualize it. There are many dimensionality reduction methods, such as multidimensional scaling [15], spring model[9], and cross-entropy directed embedding (CE) [18]. We can apply any method to reduce the dimension of augmented probability vectors, in this paper we used the CE method.

Cross-entropy embedding (CE) is a nonlinear embedding method that embeds samples to minimize energy function based on cross-entropy between similarities among original data and proximity among embedded data. Let $s_{i,j}$ be the similarity between $z(i)$ and $z(i)$, calculated by cosine similarity:

$$s_{i,j} = \frac{\mathbf{z}(i)'\mathbf{z}(j)}{|\mathbf{z}(i)||\mathbf{z}(j)|}. \tag{6}$$

Let $\mathbf{r}_i$ be the coordinate of i-th sample in the embedding space, and the proximity between $\mathbf{r}_i$ and $\mathbf{r}_i$ can be defined as follows:

$$\rho(\mathbf{r}_i, \mathbf{r}_j) \propto \exp(-\frac{1}{2} \parallel \mathbf{r}_i - \mathbf{r}_j \parallel^2). \tag{7}$$

Then, the cross-entropy between similarity $s_{i,j}$ and proximity $\rho(\mathbf{r}_i, \mathbf{r}_j)$ is:

$$E_{i,j} = s_{i,j}\text{log}(p(\mathbf{r}_i, \mathbf{r}_j)) + (1 - s_{i,j})\text{log}(1 - \rho(\mathbf{r}_i, \mathbf{r}_j)). \tag{8}$$

$E_{i,j}$ attains its minimum value when $\rho(\mathbf{r}_i, \mathbf{r}_j) = s_{i,j}$, that is, when similar samples are closely located and dissimilar samples are located far away. An objective function with a regularization term to be minimized is:

$$J = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E_{i,j} - \frac{\mu}{2} \sum_{i=1}^{N} \parallel \mathbf{r}_i \parallel^2, \tag{9}$$

where μ is a predetermined constant. By virtue of CE, samples that have similar augmented probability vectors are closely located.

As a result of visualization, samples that are classified into the same component are closely located and form clusters. Usually, there are a few anomalies. Therefore, anomalies are located far from the normal samples that form clusters. The results tell us to which component the anomalies belong, and elucidate the relationship among data. In this way, the visualization of augmented probability vectors makes it possible to understand not only the magnitude of anomalies but also their characteristics and their relationships among data.

It is natural to characterize each sample $D(i)$ by a vector $\mathbf{z}(i)$ by focusing on fitting into each component model in which probabilistic structures of variables are specified. Another commonly used visualization method is based on values of the sample itself, using a dimensionality reduction method such as multidimensional scaling [12]. However, we believe that it is difficult to find some essential characteristics by directly applying such a visualization method because it does not use the probabilistic structure of variables.

3 Application to a Time Series

Here, we explain in detail how to apply our method to a time series. We used a vector autoregressive (AR) mixture model [17]. AR models are widely used for time series analysis [10].

3.1 AR Mixture Model

Let $\mathbf{Y} = (\mathbf{y}(1), \ldots, \mathbf{y}(T))$ be a given time series, where T is the number of time points and $\mathbf{y}(t)$ is a d-dimensional vector of variables at time t. Also, let $\mathbf{x}(t) = (\mathbf{y}'(t-1), \ldots, \mathbf{y}'(t-\tau))'$ be an input vector at time t to the AR model, where τ is an order of the AR model and $'$ means the transpose of the vector.

An AR mixture model assumes that $\mathbf{y}(t)$ is generated from the K mixture of the AR models:

$$p(\mathbf{y}(t)|\mathbf{x}(t)) = \sum_{k=1}^{K} P(k)p(\mathbf{y}(t)|\mathbf{x}(t), k), \tag{10}$$

$$p(\mathbf{y}(t)|\mathbf{x}(t),k) = \frac{1}{(2\pi\sigma_k^2)^{d/2}}\exp(-\frac{(\mathbf{y}(t)-\mathbf{A}_k\mathbf{x}(t))'(\mathbf{y}(t)-\mathbf{A}_k\mathbf{x}(t))}{2\sigma_k^2}), \qquad (11)$$

where $\mathbf{A}_k$ is the AR model parameters of component k, and σ_k^2 is the variance of component k. Here, we assume that the variance of all variables are the same and independent in each component.

3.2 Estimation of Parameters

The estimation procedure for the parameters is as follows:

Step1 Let $K^* = 1$, and estimate A_1 by minimizing the square error:

$$E_1 = \sum_{t=\tau+1}^{T} (\mathbf{y}(t) - \mathbf{A}_1\mathbf{x}(t))'(\mathbf{y}(t) - \mathbf{A}_1\mathbf{x}(t)).$$

Step2 Select s $(s \leq K^*)$, and then split the parameters:

$$\mathbf{A}_s = \mathbf{A}_s + \mathit{\Delta}\mathbf{A}, \quad \mathbf{A}_{k+1} = \mathbf{A}_s - \mathit{\Delta}\mathbf{A}.$$

Step3 Let $K^* = K^* + 1$.
Step4 Estimate the parameters by maximizing the log likelihood:

$$L_{K^*} = \sum_{t=\tau+1}^{T} \log \sum_{k=1}^{K^*} P(k)p(\mathbf{y}(t)|\mathbf{x}(t),k).$$

Step5 If $K^* < K$, then return to Step2.

In Step2, we have three choices: the component s that has the highest $P(s)$, the component s with the highest σ_s, or randomly. In our experiments, we randomly selected s for simplicity. In Step4, it is impossible to estimate parameters by maximizing L_{K^*} analytically, so we approximately estimate parameters using an EM algorithm [6]. The E- and M-Steps in the EM algorithm are:

– E-Step:

$$P(k|\mathbf{x}(t),\mathbf{y}(t)) = \frac{P(k)p(\mathbf{y}(t)|\mathbf{x}(t),k)}{\sum_{k^*=1}^{K^*} P(k^*)p(\mathbf{y}(t)|\mathbf{x}(t),k^*)}, \qquad (12)$$

– M-Step:

$$P(k) = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} P(k|\mathbf{x}(t),\mathbf{y}(t)), \qquad (13)$$

$$\sigma_k^2 = \frac{\sum_{t=\tau+1}^{T} P(k|\mathbf{x}(t),\mathbf{y}(t))(\mathbf{y}(t)-\mathbf{A}_k\mathbf{x}(t))'(\mathbf{y}(t)-\mathbf{A}_k\mathbf{x}(t))}{d\sum_{t=\tau+1}^{T} P(k|\mathbf{x}(t),\mathbf{y}(t))}, \qquad (14)$$

$$\mathbf{A}_k = \left(\sum_{t=\tau+1}^{T} P(k|\mathbf{x}(t),\mathbf{y}(t))\mathbf{y}(t)\mathbf{x}(t)'\right)\left(\sum_{t=\tau+1}^{T} P(k|\mathbf{x}(t),\mathbf{y}(t))\mathbf{x}(t)\mathbf{x}(t)'\right)^{-1}. \qquad (15)$$

By iterating the above E- and M-steps, we obtain a local optimal solution.

3.3 Average 2-sigma value and Augmented Probability Vector

The average 2-sigma value in an AR mixture model $\tilde{\alpha}_{MAR}$ is the average of probability densities on 2 sigma of K AR components as follows:

$$\begin{aligned}\tilde{\alpha}_{MAR} &= \sum_{k=1}^{K} P(k)\frac{1}{(2\pi\sigma_k^2)^{d/2}}\exp(-\frac{\sum_{j=1}^{d}(2\sigma_k)^2}{2\sigma_k^2}) \\ &= \sum_{k=1}^{K} P(k)\frac{1}{(2\pi\sigma_k^2)^{d/2}}\exp(-2d). \qquad (16)\end{aligned}$$

Then the augmented probability vector $z(t)$ in the AR mixture model is:

$$\mathbf{z}(t) = (P(k=1)p(\mathbf{y}(t)|\mathbf{x}(t), k=1), \ldots, P(k=K)p(\mathbf{y}(t)|\mathbf{x}(t), k=K), \tilde{\alpha}_{MAR}). \qquad (17)$$

Clearly, $\tilde{\alpha}_{MAR}$ is independent of time t.

4 Evaluation of Our Proposed Method

4.1 Artificial Time Series

We evaluated our method by applying it to an artificial time series with 400 sampling points. Until time 200, the time series was generated using A_1, which was a two-variable one-order AR model parameter, and from time 201, generated using A_2, which was another AR model parameter. That is,

$$\mathbf{x}(t) = A_i\mathbf{y}(t) + U, \qquad (18)$$

$$i = \begin{cases} 1 & \text{if } 1 \leq \mathrm{t} \leq 200, \\ 2 & \text{if } 201 \leq \mathrm{t} \leq 400, \end{cases} \qquad (19)$$

where U is 2-dimensional white noise with covariance matrix Σ. Then to artificially introduce an anomaly, we added 5 to the 1st variable at time 100. We used following parameters:

$$A_1 = \begin{pmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -0.5 & 1.3 \\ 0.3 & 0.3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}, \qquad (20)$$

where Σ is time invariant. Figure 1 is an example of this time series.

Figure 2 shows the 3-dimensional visualization results of this time series using our method with $\tau = 1$ and $K = 1, 2$. In the case of $K = 1$ (Figure 2(a)), most samples are clustered in the lower left. The anomalous sample at time 100 is not located furthest from the cluster, that is, it is not visualized as the most anomalous. In the case of $K = 2$ (Figure 2(b)), there are 2 clusters in the lower left and the lower right, in which most samples are located. Unlike $K = 1$, the anomaly is located furthest from the others, so it is visualized as the anomaly. Our method can detect anomalies with an appropriate K.

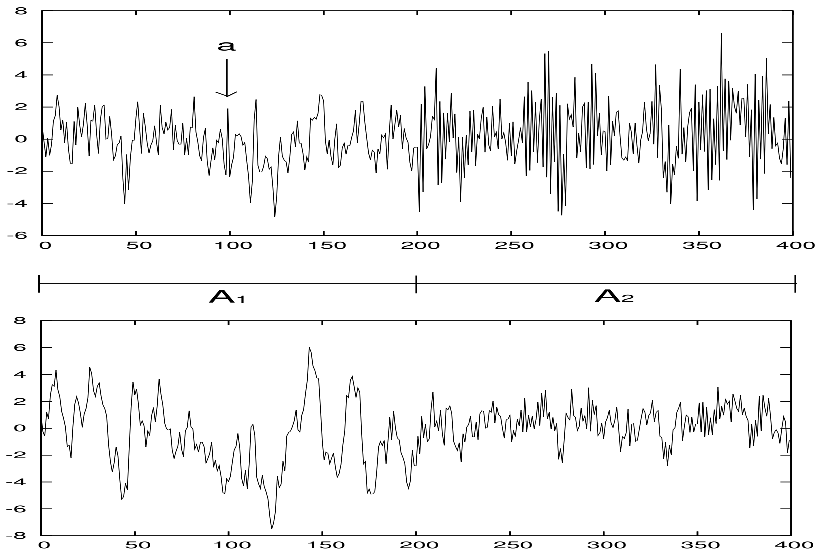


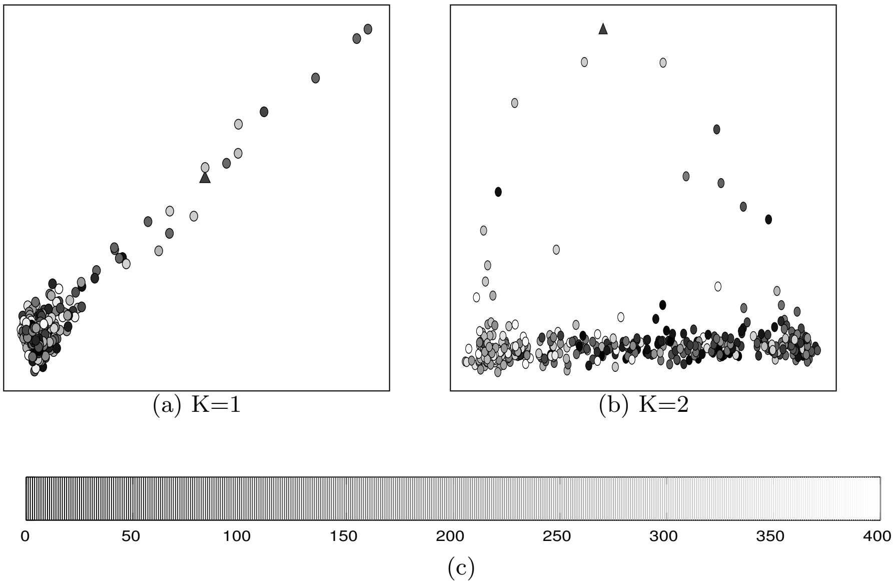
Fig. 1. Two-variable artificial time series: (upper) the 1st variable, (lower) the 2nd variable. 'a' in the upper graph represents the time point of anomaly that we added.

Figure 2(b) also shows the efficiency of the mixture model and the characteristics of the given samples. In the lower right, mainly samples until 200 (blackish points) are located, and in the lower left, mainly samples after 200 (whitish points) are located. These result indicate that the estimation was successful on some level. However, there are many points in the middle of the 2 clusters, implying that it is difficult for the model to classify some samples.

4.2 Japanese Economic Time Series

Next, we evaluated our method using real data from a monthly economic time series in Japan covering the period 1983-2003. There is a total of 240 points, and the six measured variables are: monetary base, national bond interest rate, wholesale price index, industrial produce index, machinery orders, and yen to dollar exchange rate (Figure 3). Monetary base represents the average outstanding balance deposited in the Bank of Japan. The wholesale price index represents an average of the selling prices of domestic products, excluding services. The industrial produce index measures the goods produced by the economy. The raw data are non-stationary and transformed to a stationary series through several operations, including adjustments for seasonality and trends [13].

We visualized this time series with $\tau = 1$ and $K = 1 \sim 4$ (Figure 4). One of the biggest economic events during this period was the Russian economic crisis of August 1998, which caused a large hedge fund company to fail, and the Bank of Japan rapidly raised interest rates in September 1998. At $K = 1 \sim 4$, the month



(a) K=1

(b) K=2

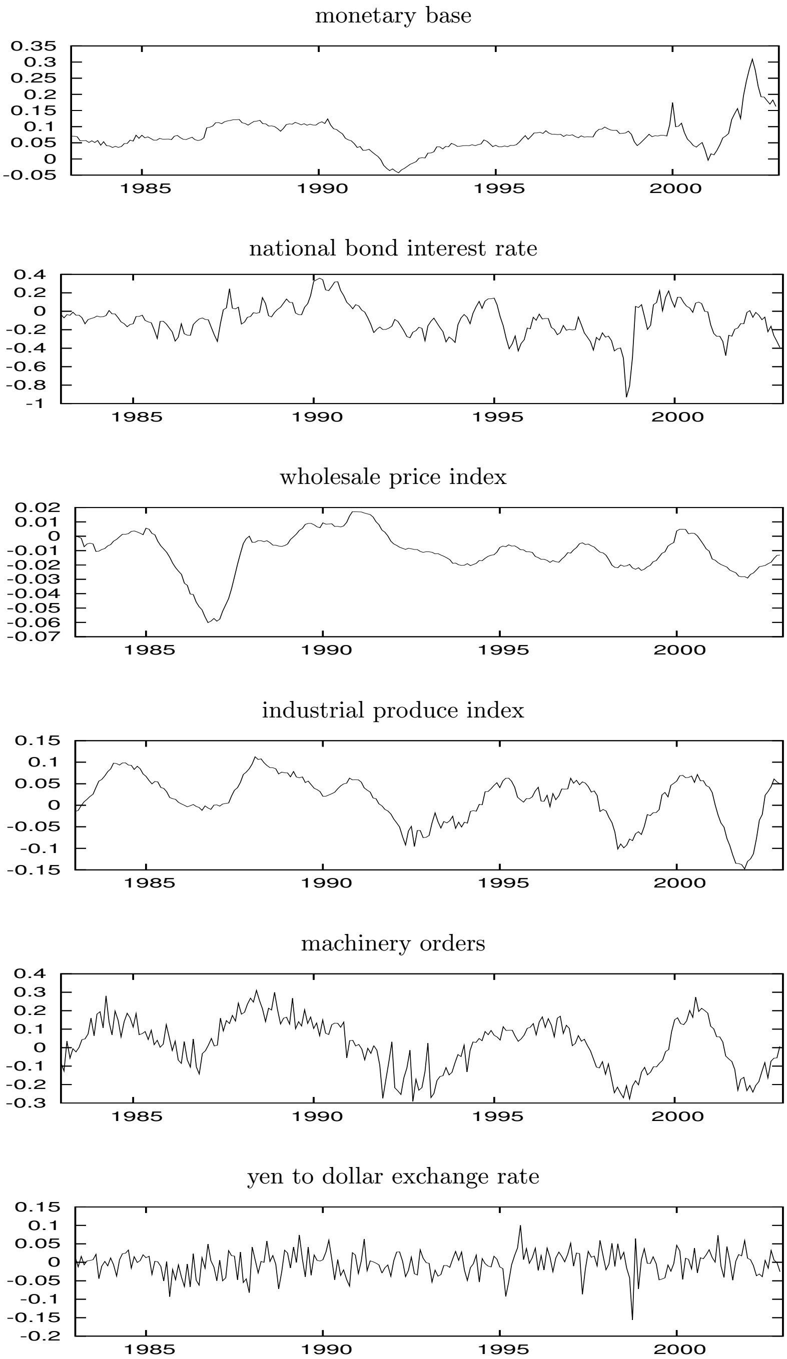
(c)

Fig. 2. Results of 3D visualization of anomaly in artificial time series. △ is artificial anomaly. (a) K=1, (b) K=2, and (c) the color-map for time of points. The point changes color from black to white going thorough time.

September 1998 (⊕) is visualized as an anomaly. However, in the case of $K = 1$, some samples of other months are also located near September 1998. Also, in the cases of $K = 2$ and $K = 3$, other samples are visualized as anomalies. On the other hand, in the case of $K = 4$, there is only one anomaly, i.e. the sample of September 1998. If $K = 1$, the model is a linear model, and thus it is impossible to appropriately model the data. As K increases, the model becomes flexible, and then it becomes less likely to show false alarms.

5 Related Work

Anomaly detection methods can be broadly classified into 2 methods: profiling and discriminating [7]. A profiling method builds a model from normal samples, and samples which deviate significantly from the model are deemed anomalies. A discriminating method builds a model from anomalous samples, and samples which match the model are deemed anomalies. If a lot of anomalous samples are given or the model of anomaly is known, a discriminating method is useful. In general, however, it is difficult to obtain many anomalous samples, and the model of the anomaly cannot be specified. So, we adopted a profiling method that builds a mixture model from data and samples which deviate significantly from the model based on the 2-sigma value are assumed to be anomalies.


monetary base

national bond interest rate

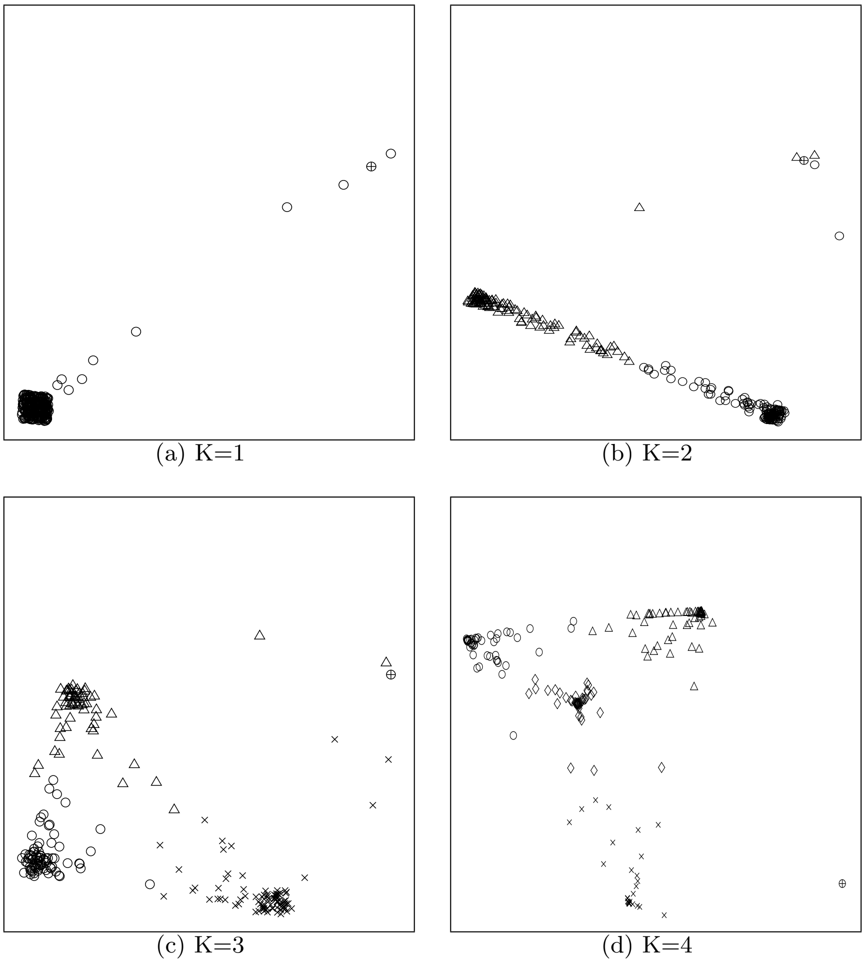
wholesale price index

industrial produce index

machinery orders

yen to dollar exchange rate

Fig. 3. Japanese economic time series



(a) K=1

(b) K=2

(c) K=3

(d) K=4

Fig. 4. Results of 3D visualization of anomalies in an economic time series. ⊕ is the sample on 9/1998, ◯s are samples classified to component 1, △s to component 2, ×s to component 3, and ◇s to component 4.

The main characteristics of our method give the relationships among data as well as the magnitude of anomalies using visualization. Davidson's method [5] is similar to our method because it visualizes anomalies with the relationships among data. However, he focuses on anomalies in clustering problems, and considers samples located near cluster boundaries as anomalies. It is useful to detect vague samples that are difficult to classify when we execute clustering. Our method can also detect such anomalies, although we do not explicitly consider them anomalies. We consider samples that do not match any mixture components as anomalies, therefore, our method can apply usual anomaly detection problems. In this point, our method is very different from Davidson.

We used a mixture of vector AR models for anomaly detection in a time series. The components of the mixture model can include more complex models such as ARMA [3]. Next, $P(k)$ can be estimated from inputs, using a gated network [16] or a threshold [14]. If a mixture AR model is inadequate to model a given time series, our method can be modified easily for other mixture models.

There are also discriminating methods using AR models. Chen et al. considered 4 types of anomalies: (1) additive outliers, (2) innovational outliers, (3) temporary changes, and (4) level shifts. They detected them using 4 types anomaly models [4]. Experiments using an artificial time series showed that our method can detect simple additive outliers. We believe that our method can cope with other types of outliers because of the flexible expressive power of mixture models. However, this claim must be confirmed by further extensive experiments.

6 Conclusions

We proposed a new method to visualize data for anomaly detection using a mixture model. This method visualizes samples based on the probability of belonging to a component and not belonging to any components. Since our visualization method is quite different from conventional approaches, it provides a new view of data and helps to find new knowledge. We also showed the validity of our method by applying it to an artificial time series and an actual economic time series.

Although our results have been encouraging to date, a number of directions remain in which we must extend our approach before it can become a useful tool for anomaly detection. First, we need to determine the number of components K. In the application to an economic time series, we incremented K from $K = 1$ until only a few anomalies are visualized clearly. It is necessary to determine K automatically by using a quantitative score such as AIC [1] or MDL [11]. Second, we must avoid getting trapped in a local minimum. Visualization results of the same data can be different. In the application we estimated parameters several times and selected parameters whose likelihood were the highest. However, if there are many local minima in the data, such a procedure is ineffective. We will study these problems in the future.

References

1. Akaike, H. : A new look at the statistical model identification. IEEE Transactions on Automatic Control, **19** (1974) 716–723.
2. Barnett, V., Lewis, T. : Outliers in statistical data. 2nd ed. John Wiley & Sons, New York (1984).
3. BrockWell, P. J., Davis, R. A. : Introduction to Time Series and Forecasting. Springer-Verlag (2002).
4. Chen, C., Liu, L. : Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association **88** (1993) 284–297.
5. Davidson, I., Ward, M. : A particle visualization framework for clustering and anomaly detection. Proceedings of KDD Workshop on Visual Data Mining (2001).
6. Dempster, A. P., Laird, N. M., Rubin, D. B. : Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B **39** (1977) 1–38.
7. Fawcett, T., Provost, F. : Activity monitoring: Noticing interesting changes in behavior. Proceedings of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (1999) 53–62.
8. Gorldenberg, A. : Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceedings of the National Academy of Sciences of the United States of America, **99** (2002) 5237–5240.
9. Kamada, T., Kawai, S. : An algorithm for drawing general undirected graphs. Information Processing Letters **31** (1989) 7–15.
10. Lutkepohl, H. : Introduction to multiple time series analysis. Spinger-Verlag (1993).
11. Rissanen, J. : Modeling by shortest data description. Automatica **14** (1978) 465–471.
12. Rorvig, M. : Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC document sets. Journal of the American Society for Information Science **50** (1999) 639–651.
13. Sato, S. : Stepwise prediction for economic time series by using vector autoregressive model. Science of Modeling, AIC2003, ISM Report on Research and Education **17** (2003) 225–233.
14. Tong, H., Lim, K. S. : Threshold autoregression, limit cycles and cyclical data. Journal of the Royal Statistical Society, Series B **42** (1980) 245–292.
15. Torgerson, S. : Theory and methods of scaling. Wiley, New York (1958).
16. Weigend, A., Mangeas, M., Srivastava, A. : Nonlinear gated experts for time series: discovering regimes and avoiding overfitting. International Journal of Neural Systems **6** (1995) 373–399.
17. Wong, C. S., Li, W. K. : On a mixture auto regressive model. Journal of the Royal Statistical Society, Series B **62** (2000) 95–115.
18. Yamada, T., Saito, K., Ueda, N. : Cross-entropy directed embedding of network data. ICML2003 (2003) 832–839.
19. Ye, N., Chen, Q. : An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. Quality and Reliability Engineering International **17** (2001) 105-112.