A General Variance-based Outlier Mining Model
Hongwei Qi  Junping Zhang  Jue Wang
Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences,
PO Box 2728 Beijing 100080, China
Email: {hongwei.qi, junping.zhang, jue.wang}@mail.ia.ac.cn

Abstract
To solve the outlier mining problems where the outliers are highly intermixed with normal data, a general Variance-based Outlier Mining Model (VOMM) is presented, in which the information of data is decomposed into normal and abnormal components according to their variances. With minimal loss of normal information in our model, outliers are viewed as the top k samples holding maximal abnormal information in a dataset. And then, the principal curve that is a smooth nonparametric one-dimensional curve passing through the “middle” of the dataset and that provides a good nonlinear summary of the data is introduced as an algorithm of our model. Being a general approach to outlier mining, our proposed VOMM is proved to be feasible by experiments carried out on abnormal returns detection in stock market. Additionally, the model is potentially applicable to many other domains such as credit cards, telecommunication, medicine and traffic etc.

Keywords
Outlier Mining, Principal Curve, Stock Market

1. Introduction
Being an integral part of data mining and having attracted much attention recently, outlier mining aims at revealing the nontrivial knowledge, i.e., abnormal information, hidden in data [1].

Unfortunately, there is no exact definition on outliers till now except for many similar “descriptions” about their appearance [1-5]. The one given by Jiawei Han – Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers [1] – is probably the most representative. Inspired by the ideas in the above descriptions, the usual strategy of outlier mining is to find a model for a given dataset. And then samples inconsistent with the model are viewed as outliers. Based on the above strategy, numerous outlier mining models are obtained successively, which can be further categorized into four approaches: the statistical approach [4], the distance-based approach [6-7], the deviation-based approach [8-9], and the density-based approach [10].

However, in many domain datasets, outliers are highly intermixed with normal data. In stock market, for example, the closing price of each trading day includes not only long-term trends but also daily random fluctuations [11]. Therefore, the differences between the normal and abnormal data are not evident enough for the previous outlier mining models to detect outliers. And so an alternative outlier mining approach is needed here for such complex datasets above.

Generally speaking, data in those sets are affected by normal and abnormal factors simultaneously. In most situations, the normal factors play a dominant role in data generation, and make most samples behave as normal data. However, the sporadic abnormal factors can cause outliers. For example, most of the daily closing prices controlled mainly by long-term trends are normal data. But some hideous events, like vicious insider trading may result in abnormal closing price on the trading day.

Therefore, in this paper, we present a general Variance-based Outlier Mining Model (VOMM) that divides the information of data into normal and abnormal components according to their variances. With minimal loss of normal information in our model, outliers are viewed as the top k samples holding maximal abnormal information in a dataset. And then, the principal curve that is a smooth nonparametric one-dimensional curve passing through the “middle” of the dataset and that provides a good nonlinear summary of the data is introduced as an algorithm of our model. Compared with the traditional outlier mining models, the VOMM has the following advantages. First, it holds the information of both normal data and outliers. Second, the outliers are detected while the model is being constructed. Last, the VOMM can sort the outliers by their variances of abnormal component in descending order. Being a general approach to outlier mining, our proposed VOMM is proved to be feasible by experiments carried out on abnormal returns detection in stock market.

The remainder of the paper is organized as follows. In section 2, we present the Variance-based Outlier Mining Model. The principal curve outlier mining algorithm is introduced in section 3. Experiments on two real datasets of stock market are given in section 4. We conclude our study in the last section.

2. Variance-based Outlier Mining Model (VOMM)
Outliers are viewed as samples holding more abnormal information in a dataset. Hence, the first step of outlier mining is to construct a model that can decompose the information of the data into normal and abnormal components.

Definition 2.1 Let \( D \subseteq \mathbb{R}^d \) be a sample space and \( X \) be a random vector drawn from \( D \). Then \( X \) is called decomposable if there exists a function on \( D \) such that

\[
X = f(w) + \epsilon
\]

where \( w \in W \), \( W \) is a set of parameters, and \( \epsilon \) is a d-dimensional residual vector.

The outlier mining task is to decompose the information of \( X \) into that of \( f(w) \) and \( \epsilon \). The normal component \( f(w) \) is stable and deterministic, whereas the abnormal component \( \epsilon \) is random and influenced by various sources. As an alternative approach, the task turns into recovering \( f(w) \) based on the noisy observation \( X \), and the objective is to minimize \( \| X - f(w) \| \) under the restriction of low computational complexity of \( f(w) \).
Definition 2.2 Consider the vector $X$ and parameter $w$ defined in Definition 2.1. Let $f(w) = E(x \mid w)$ be a vector function, where

$$f_j(w) = E(x_j \mid w), \quad j = 1, 2, \cdots, d. \quad (2)$$

According to formula (2), $f(w) = E(x \mid w)$ is an expected estimate of $x$ on the condition of $w$.

Definition 2.3 (Expected Squared Distance Function)

Let the squared Euclidean distance between $f(w)$ and $x$ be

$$\varepsilon(x, w) = \sum_{j=1}^{d} (x_j - E(x_j \mid w))^2. \quad (3)$$

Then the Expected Squared Distance Function of $f(w)$ is defined as

$$\Delta(f) = E[\|X - E(X\mid w)\|^2]. \quad (4)$$

Theorem 2.4 (VOMM) Consider the $X \in R^d$ and $w$ defined in Definition 2.1. The total variance in the $d$ coordinates can be decomposed into the estimate variance and the residual variance on the condition of $w$, that is:

$$\sum_{j=1}^{d} Var(x_j) = \sum_{j=1}^{d} Var(E(x_j \mid w)) + \sum_{j=1}^{d} E((x_j - E(x_j \mid w))^2). \quad (5)$$

The proof of Theorem 2.4 is shown in Appendix 1.

Obviously, the estimate variance represents the normal information of data, while residual variance reflects that of abnormal component.

To maximally estimate the normal information of data, we should minimize the residual variance under the condition of Theorem 2.4. Thus, the definition of outlier is naturally obtained as follows:

Definition 2.5 (Outlier Definition)

Under the condition of Theorem 2.4, with the object of minimizing the Expected Squared Distance Function (4), the outlier in dataset $\{x_1, x_2, \cdots, x_n\}, x_i \in R^d$ drawn from $D$ is defined as:

$$Outlier_i = \underset{x_i}{\operatorname{argmax}} \|x_i - E(x\mid w)\|^2, \quad i = 1, 2, \cdots, n. \quad (6)$$

where $n$ is the number of the samples.

Note that the $i$ in Definition 2.5 has a meaning different from that of the $j$ in Theorem 2.4 in that the former is the sequence number of the samples while the latter is the coordinate index of the vector.

The meaning of Definition 2.5 is that an outlier is a sample with maximal residual variance. In practical task, if there’s need to detect a set of outliers, one of the following two techniques can be adopted:

(1) Sort the samples in descending order by the value of \( \|x_i - E(x_i \mid w)\|^2 \) \( (i = 1, 2, \cdots, n) \) and choose the top $k$ ($k < n$) samples as outliers. Note that this technique can descendingly sort the outliers.

(2) Given a threshold value $\theta$, samples with $\|x_i - E(x_i \mid w)\|^2 > \theta$ $(i = 1, 2, \cdots, n)$ are outliers.

3. The principal curve outlier mining algorithm for the VOMM

Since the VOMM is a general theoretical framework for outlier mining, corresponding algorithms should be constructed. Intuitively, there are two guidelines for constructing the algorithms:

(1) Given a dataset, if we know the natural distribution of the data in advance, then it is the best choice to derive algorithm based on the distribution function.

(2) However, in most situations, we are not familiar with the nature of the data. In line with Vapnik’s principle [12] that “when solving a given problem, try to avoid solving a more general problem as an intermediate step”, we should derive the algorithm directly from the data and avoid solving a more complex problem, namely, estimating the distribution of data.

As this paper tries to solve complex outlier mining problems in which outliers are highly intermixed with normal data, in this section the principal curve outlier mining algorithm that directly based on data is introduced.

3.1. The Principal Curve

We first introduce the basic principle of the principal curve briefly in our own way. The principal curve, presented by Trevor Hastie [13], is a smooth nonparametric one-dimensional curve passing through the “middle” of a distribution.

Definition 3.1 [14] A one-dimensional curve in $R^d$ is a continuous function $f: \Lambda \mapsto R^d$, where $\Lambda = [a, b] \subset R^d$.

The curve $f$ can be considered as a vector of $d$ functions of a single variable $\lambda \in [a, b] \subset R ^d$, i.e., $f(\lambda) = (f_1(\lambda), f_2(\lambda), \cdots, f_d(\lambda))$, where $f_1(\lambda), f_2(\lambda), \cdots, f_d(\lambda)$ are called the coordinate functions.

Definition 3.2 (Projection index) [13] For any $x \in R^d$, the corresponding projection index $\lambda_j(x)$ is defined by

$$\lambda_j(x) = \sup \{\lambda : \|x - f(\lambda)\| = \inf \|x - f(\tau)\|\}, \quad (7)$$

where $f(\lambda)$ is a curve in $R^d$ parameterized by $\lambda \in R^d$.

The projection index $\lambda_j(x)$ of $x$ is the value of $\lambda$ for which $f(\lambda)$ is closest to $x$. If there are a number of such points, we pick the largest of such values of $\lambda$. Consequently, the projection point of $x$ to $f$ is $f(\lambda_j(x))$ (see Figure 1).

![Figure 1. Projecting points to a curve](image-url)

Definition 3.3 [13] The squared Euclidean distance between curve $f$ and $x$ is defined as the squared distance from $x$ to its projection point on $f$, i.e.,

$$\varepsilon(x, f) \overset{\text{def}}{=} \|x - f(\lambda_j(x))\|^2 \quad (8)$$

Definition 3.4 (Distance Function for curve) [13] Consider $X \in R^d$ defined in Definition 2.1. The Distance Function of a curve $f$ is defined as the expected squared distance between $X$ and $f$, i.e.,

$$\Delta(f) \overset{\text{def}}{=} E[\|X - f(\lambda_j(X))\|^2] \quad (9)$$
Definition 3.5 (The principal curve) [14] The smooth curve \( f(\lambda) \) is a principal curve if the following holds:

(a) \( f \) does not intersect itself; (b) \( f \) has finite length inside any bounded subset of \( \mathbb{R}^d \); (c) \( f \) is self-consistent, that is,

\[
f(\lambda) = E(X | \lambda_r(X) = \lambda) \quad \forall \lambda \in \mathbb{R}^d
\]  

(10)

The self-consistent property here means that each point on the curve is the conditional mean of the points projecting there. Thus, the principal curve is a smooth nonparametric self-consistent curve, which passes through the “middle” of a distribution and provides a good one-dimensional nonlinear summary of the data [14] (see Figure 1).

3.2. The principal curve outlier mining algorithm for distribution

The principal curve can be used as an achievement of algorithm of the VOMM.

Corollary 3.6 Consider the \( X \in \mathbb{R}^d \) and \( w \) defined in Definition 2.1. If \( f(\lambda) \) is a principal curve and let \( w = \lambda_r(x) \), it holds that:

\[
\sum_{j=1}^d \text{Var}(x_j) = E \| x - f(\lambda_r(x)) \|^2 + \sum_{j=1}^d \text{Var}(f_j(\lambda_r(x))).
\]  

(11)

The proof of Corollary 3.6 is shown in Appendix 2.

Formula (11) means that the total variance in the \( d \) coordinates is decomposed into the estimate variance explained by the true curve and the residual variance in the expected squared distance from a point to its true position on the curve.

From the aspect of the principal curve, which is the minimin point of the Distance Function (9) [13], the outlier in dataset \( \{x_1, x_2, \cdots, x_n\}, x_i \in \mathbb{R}^d \) can be represented as follows based on Definition 2.5:

\[
\text{outliers} = \arg \max_{x_i} \| x_i - f(\lambda_r(x_i)) \|^2, \quad i = 1, 2, \cdots, n.
\]  

(12)

If we need to mine a set of outliers, they are considered as the top \( k \) (\( k < n \)) samples descendingly ordered by \( \| x_i - f(\lambda_r(x_i)) \|^2 \).

Based on the Definitions and Corollary mentioned above, we can form an algorithm of outlier mining for a distribution.

Step 0: Let original curve \( f^{(0)}(\lambda) \) be the first principal component line for \( X \). Set \( j = 0 \).

Step 1: Set \( \lambda_{\lambda_{f^{(0)}}}(x) = \max \{ \lambda : \| x - f(\lambda) \| = \min \{ \| x - f(\tau) \| \} \} \) for all \( x \in \mathbb{R}^d \).

Step 2: Set \( f^{(1)}(\lambda) = E[X | \hat{\lambda}_{f^{(1)}}(X) = \lambda] \).

Step 3: Stop if \( (1 - \Delta f^{(1)}(\lambda))/\Delta f^{(1)}(\lambda) < \delta \). Otherwise, let \( j = j + 1 \), go to step 1. Where \( \delta \) is a threshold value.

Step 4: Sort the samples in descending order by the value of \( \| x_i - f(\lambda_r(x_i)) \|^2 \).

Step 5: Choose the top \( k \) (\( k < n \)) samples as outliers.

3.3. The principal curve outlier mining algorithm for datasets

So far, we have considered the principal curve outlier mining algorithm for a multivariate probability distribution. However, we usually have a finite dataset \( \{x_1, x_2, \cdots, x_n\}, x_i \in \mathbb{R}^d \) drawn from the distribution. Therefore, the algorithm needs to be extended to dataset.

(1) In Step 2, at most one point projects to a given point on the curve according to the definition of self-consistency. Using this one point in the averaging would result in a curve that visit all the data points after the first iteration. To tackle this problem, \( E[X | \hat{\lambda}_{f^{(1)}}(X) = \lambda^{(j)}] \) is estimated by averaging samples whose corresponding projection indices are the local neighborhood of \( \lambda^{(j)} \) (See Figure 2). We use the locally weighted running-lines smoother [15] for local averaging in this paper. The locally weighted operator is defined as follows:

\[
\theta_n = \begin{cases}
(1 - (\frac{\lambda_n - \lambda^{(j)}}{\beta_n - \lambda^{(j)}}))^{1/3} & \text{if } \lambda_n - \lambda^{(j)} \leq \beta_n - \lambda^{(j)} \\
0 & \text{otherwise}
\end{cases}
\]  

(13)

Where \( B \) is the maximal neighbor numbers, \( \lambda^{(j)} \) is the projection index of a point on curve that needs to be adjusted, \( \lambda_n \) is the projection index close to \( \lambda^{(j)} \).

4. Experiments

Being a general approach to outlier mining, the VOMM and its corresponding algorithm can be applied in many domains. In this paper, experiments carried out on abnormal stock returns detection are shown.

4.1. Datasets

Two datasets are chosen from stock market of China; one is the set of the returns of INDEXSH (Integrate Index in Shanghai Stock Exchange), and the other is that of a personal stock, namely Stock-1, both during the period of 01/01/1998 to 12/31/2001. Note that only nickname of the personal stock is given here for the sake of secrecy, but the real dataset is still available from the authors.

4.2. Data preprocessing

There are two steps for data preprocessing.

(1) Constructing 5, 20 and 60 aggregate days’ return sets.

In general, abnormal returns of different causes have different duration. Therefore, it is more reasonable to distinguish short-term and long-term abnormal returns. We detect the short-term abnormal ones from the daily and 5 aggregate days’ returns, and the long-term ones from the 20 and 60 aggregate days’ returns. The 5, 20 and 60 aggregate days’ returns can be constructed from daily returns set \( R = \{r_1, r_2, \cdots, r_n\} \), using the method of rolling average, that is,

\[
r_i^M = \sum_{j=1}^{i+M-1} r_j, M = 5,20,60, \quad i = 1, \ldots, n - M + 1,
\]  

(15)
where \( r_i^M \) is the M aggregate days’ return, and \( n \) is the number of the daily return samples. Hence \( R^M = \{ r_1^M, r_2^M, \ldots, r_n^M \} \), \( M = 5, 20, 60 \).

(2) Adding temporal sequence to returns

Given one-dimensional returns dataset \( R \) (or \( R^M, M = 5, 20, 60 \)). In order to preserve the temporal sequence information of the returns, we adopt the TC-PCA (Time Constraint Principal Component Analysis) approach presented by K. Reinhardt el al [16], which expands the dimensionality of the data by using

\[
r_i^4 = \eta^4 (1, \cdots, n) \tag{16}
\]

Hence \( R_\eta = \{(\eta^4 \cdot 1, r_i), (\eta^4 \cdot 2, r_i), \ldots, (\eta^4 \cdot n, r_i)\} \), and the extra dimension represents a scalars ordering as time constraint. Similarly, we can get \( R_\eta^M \) from \( R^M \) (\( M = 5, 20, 60 \)) using TC-PCA. The scale factor \( \eta \) is introduced to control the weighting imposed by the arbitrary choice of incorporating the order information. The TC-PCA transforms the temporal extended return vectors into two orthogonal coordinate system, and the \( \eta \) is tuned to make the first principal component line almost parallel to the time axis.

4.3. Causes of abnormal returns

We classify the factors that can cause abnormal returns as follows. Macro factors include relevant political, economic, and social events beyond the stock market. Background factors of listed companies are the business events of companies. Other than macro factors and background factors of listed companies, market manipulation or other uncertain factors are closely related with both the regulation of trading and the efficiency of the market [17], and therefore should be viewed as the most concerned factors.

4.4. Experimental results

(1) INDEXSH

We detect 25, 20, 15 and 10 outliers, called respectively as set \( O_{25}, O_{20}, O_{15} \) and \( O_{10} \) from the temporal extended \( R_\eta \) and \( R_\eta^M \) (\( M = 5, 20, 60 \)) of INDEXSH. It is worthy noting that the numbers of the outliers, 25, 20, 15 and 10, are chosen by domain knowledge. Furthermore, two parameters, time constraint factor \( \eta \) in formula (16) and neighbor number \( B \) in formula (13), should be analyzed in the experiments.

We observe in experiments that different \( \eta \) have very little influence on experimental results. Without loss of generality, we set \( \eta = 2.0 \).

We also use different neighbor number \( B = (B_1, B_2, \ldots, B_p) \) to observe the influence of the smoothing operation, formula (13) of the principal curve outlier mining algorithm, on the results of outlier detection. Let \( O_{n_i}^k \ (i = 1, 2, \cdots, p) \) be the outlier set detected from \( R_\eta \) over the \( B_i \). And let \( N(O_{25}^k, O_{20}^{k+1}) \) be the number of different samples between \( O_{25}^k \) and \( O_{20}^{k+1} \), that is, the number of samples in set \( O_{25}^k \) but not in \( O_{20}^{k+1} \). Hence \( N_{20} = |N(O_{25}^k, O_{20}^{k+1})| \cdot N(O_{20}^{k+1}, O_{15}^{k+2}) \cdot \ldots \cdot N(O_{5}^{k+6}, O_{2}^{k+7}) \) \). Similarly, we can get \( N_{20} \), \( N_{15} \) and \( N_{10} \) from \( R_\eta^M \) (\( M = 5, 20, 60 \)) over \( B \) respectively. The change of \( N_{25}, N_{20}, N_{15}, N_{10} \) with \( B \) is illustrated in Figure 3, which shows \( B \) has no significant influence on the results of the outlier detection.

![Figure 3. The influence of smoothing parameter on the results of outlier detection in INDEXSH.](image)

Therefore, we first set \( B = 50 \), and get outlier sets \( O_{25}, O_{20}, O_{15} \) and \( O_{10} \) from the temporal extended \( R_\eta \) and \( R_\eta^M \) (\( M = 5, 20, 60 \)) (Figure 4 is the illustration of \( O_{25} \)). Second, merge these 70 outliers, which are the sum of outliers in sets \( O_{25}, O_{20}, O_{15} \) and \( O_{10} \), if they happen at the same time. Last, check the stock market announcements [18] to find out what events cause these outliers. Table 1 is a description of the relation between outliers and events in stock market.

![Figure 4. The principal curve generated for the temporal extended database \( R_\eta \) of INDEXSH.](image)

### Table 1 The external events of INDEXSH

<table>
<thead>
<tr>
<th>No</th>
<th>Time</th>
<th>Duration</th>
<th>( B )</th>
<th>( P )</th>
<th>( R )</th>
<th>( O )</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19980113</td>
<td>V</td>
<td>1 2 0 0</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>The Currency of Southeast Asian countries was impacted.</td>
</tr>
<tr>
<td>2</td>
<td>19980811</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>The impact of Asia Financial Crisis on Southeast Asia; (2) An unprecedented big flood occurred in China on Aug.2000.</td>
</tr>
<tr>
<td>3</td>
<td>19990510</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>United NATO attacked Chinese Embassy in Yugoslavia on May 9.</td>
</tr>
<tr>
<td>4</td>
<td>19990630</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>The six policies of “invigorating the market” caused the 5.19 market.</td>
</tr>
<tr>
<td>5</td>
<td>19990701</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>“Securities Act” formally took effect from 7.1.</td>
</tr>
<tr>
<td>6</td>
<td>20000110</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>General Office of the State Council announced the notice about several suggestions of setting up the risk investment mechanism.</td>
</tr>
<tr>
<td>7</td>
<td>20000216</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>The secondary market rationed new shares and stocks were released.</td>
</tr>
<tr>
<td>8</td>
<td>20001124</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>China Securities Regulatory Commission issued notice on strengthening market supervision and beating market manipulation behavior, meanwhile nine inspection bureaus were to be established and 236 illegal cases in violation of rules and regulations were to be investigated, etc.</td>
</tr>
<tr>
<td>9</td>
<td>20010722</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>The Peoples Bank of China investigated the funds in violation of rules and regulations.</td>
</tr>
<tr>
<td>10</td>
<td>20010912</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>(1) The reduction of state-owned stocks; (2) Yinchaun Guangxia fake incident; (3) The 911 incident occurred in U.S.A.</td>
</tr>
<tr>
<td>11</td>
<td>20011023</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>The reduction of state-owned stocks was suspended.</td>
</tr>
</tbody>
</table>

Note: “Y” means there is an abnormal return in one of the daily, 5, 20 and 60 aggregate days’ return sets.
From Table 1, we find that all the 11 abnormal returns, one is related with political event, 3 with economic events. However, 7 are caused by the macro-manipulative policies on stock market. Besides, during the period of 1998 to 2001, Bank of China dropped its interest rate four times altogether and considerable increase appeared in the INDEXSH on the days when interest rate dropped. But no abnormal returns happened. It’s thus clear that although macro-political and macro-economic factors can cause outliers, macro-manipulative policies on stock market are the main causes of the abnormal returns of INDEXSH.

Results of analysis above indicate that government policies do influence the change of stock price to a great extent. To avoid too many abnormal returns, the government should reduce macro-manipulative policies on stock market.

(2) Stock-1

The aim of this experiment is to find out what kind of abnormal returns market manipulation on stocks results in. Stock-1 went through market manipulation during the period of August 1998 to February 2001, and was later punished by China Securities Regulatory Commission. Let $\eta = 2.0$. Similar to experiment on INDEXSH, Figure 5 and Table 2 ($B = 80$) are the results of Stock-1.

**Figure 5. The influence of smoothing parameter on the results of outlier detection in Stock-1.**

**Table 2 The external events of Stock-1.**

<table>
<thead>
<tr>
<th>No</th>
<th>Time</th>
<th>Duration</th>
<th>Personal stock information or Macro-factor</th>
<th>Time happens</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19980113</td>
<td>Y</td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19980610</td>
<td>Y</td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>19980811</td>
<td>Y</td>
<td>Big Board fluctuated abnormally.</td>
<td>19980811-19980828</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>19980902</td>
<td>Y</td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>19981216</td>
<td>Y</td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>19981221</td>
<td>Y</td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>19990531</td>
<td>V</td>
<td>Shenzhen Trade and Business Investment Holding Company transferred its 26.11% shares to the company of Stock-1.</td>
<td>19990524</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>20000114</td>
<td>V</td>
<td>The company’s declaration: company’s stock price has been fluctuating abnormally recently and there is no information that should be but not revealed.</td>
<td>20000120</td>
<td>C</td>
</tr>
<tr>
<td>9</td>
<td>20000221</td>
<td>V</td>
<td>The company’s declaration: company’s stock price has been fluctuating abnormally recently and there is no information that should be but not revealed.</td>
<td>20000224</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>20000627</td>
<td>V</td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>11</td>
<td>20000616</td>
<td>V</td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>12</td>
<td>20010110</td>
<td>V</td>
<td>The company’s declaration: company’s stock price has been fluctuating abnormally recently and there is no information that should be but not revealed.</td>
<td>20010115</td>
<td>C</td>
</tr>
<tr>
<td>13</td>
<td>20010412</td>
<td>V</td>
<td>Three shareholders of the company signed a contract with Loncin Group Company on April 20, 2001, transferring their 97% shares to Loncin.</td>
<td>20010427</td>
<td>B</td>
</tr>
<tr>
<td>14</td>
<td>20010903</td>
<td>V</td>
<td>The company declaration: the achievement suffers loss.</td>
<td>20010928</td>
<td>B</td>
</tr>
<tr>
<td>15</td>
<td>20010108</td>
<td>Y</td>
<td>The above contract signed with Loncin Group Company was canceled.</td>
<td>20010929</td>
<td>B</td>
</tr>
<tr>
<td>16</td>
<td>20010123</td>
<td>V</td>
<td>Big Board fluctuated abnormally.</td>
<td>2001023-20011029</td>
<td>A</td>
</tr>
</tbody>
</table>

Note: “Type” stands for the types of causes of abnormal returns, where “A” represents the macro-factors; “B” the background factors of listed companies and “C” the market manipulation or other uncertain factors.

There are 16 abnormal returns in Stock-1, of which 2 are “A” type related with macro-factors; 4 are “B” type related with background factors of listed company. However, it is interesting to note that 10 are “C” type related with market manipulation or other uncertain factors, and the possibility of market manipulation in this stock is very high. So we are quite sure that this stock was once market manipulated and it’s proved true.

5. Conclusions

In this paper, to solve the outlier mining problems where the outliers are highly intermixed with normal data, we have introduced a general Variance-based Outlier Mining Model and its corresponding algorithm. The model is successfully applied to abnormal returns detection problem of stock market.

Conclusions include the following:

(1) Gaussian model is the traditional approach of abnormal stock returns detection. The returns, however, are generally non-normal, which show themselves as high excess kurtosis with significant skewness [19].

(2) The VOMM presented in this paper is a theoretical framework for outlier mining, and many other appropriate algorithms except for the principal curve one can be constructed for it.

(3) Being a general approach to outlier mining, our proposed VOMM is also applicable in many other domains, such as fraud detection in credit cards or telecommunication services, intrusion detection in networking and abnormal situation detection in traffic data etc.

(4) It is worthy noting that the purpose of this paper is to detect and analyze outliers from historical data and the content of the prediction isn’t concerned here.

Acknowledgement

This research is supported by National Key Project for Basic Research in China (G1998030508).

Appendix 1. Proof of Theorem 2.4

During the Proof of Theorem 2.4, the following Proposition is used.

**Proposition [20]:** Let $X$ be any real-valued random variables such that $E[X] < \infty$, and let $W_1, W_2, \cdots, W_m$ be random variables. It holds that

$$E(E(x | W_1, W_2, \cdots, W_m)) = E(x).$$

For convenience, denote by $W = (W_1, W_2, \cdots, W_m)$ the random variables in the Proposition above.

**Proof (Theorem 2.4)** For the simplicity, let $d = 1$, then we only need to prove that:

$$Var(x) = E((x-E(x|W))^2) + Var(E(x|W)).$$

Proof: $Var(x)$ is $Var(x-E(x|W)+E(x|W))$.

$$= Var(x-E(x|W))+Var(E(x|W))+2Cov((x-E(x|W)), E(x|W))$$

$$= Var(E(x|W))^2+2Var(E(x|W))+2Cov((x-E(x|W)), E(x|W))$$

$$= E((x-E(x|W))^2)+Var(E(x|W))+2Cov((x-E(x|W)), E(x|W))$$

then the problem turns into proving

$$Cov((x-E(x|W)), E(x|W)) = 0.$$
\[
\begin{align*}
\text{Cost}(x - E(X|x_0), E(X|x_0)) &= E(x - E(x|x_0) - E(X|x_0)) - E(x - E(x|x_0) - E(x)) \\
&= E((x - E(x|x_0)) - E(X|x_0)) - 0 \\
&= \int \cdots \int (x - E(X(w_1, \ldots, w_n))) - E(X(w_1, \ldots, w_n)) f(x, w_1, \ldots, w_n) \\
dw_1 \cdots dw_n \\
&= \int \cdots \int (x - E(X)) f(x, w_1, \ldots, w_n) dw_1 \cdots dw_n \\
\end{align*}
\]

Therefore Theorem 2.4 holds true.

**Appendix 2. Proof of Corollary 3.6**

**Proof (Corollary 3.6)** Let \( f \) be the principal curve and \( w = \lambda_j(x) \). According to the self-consistent property of the principal curve, we get

\[
E(x|\lambda_j(x)) = f_j(\lambda_j(x)) \quad j = 1, 2, \ldots, d.
\]

Then similar to the proof of Theorem 2.4, we prove that

\[
\sum_{j=1}^{d} \text{Var}(x_j) = E\left[\|x - f(\lambda_j(x))\|^2\right] + \sum_{j=1}^{d} \text{Var}(f_j(\lambda_j(x)))
\]

holds true.

**References**

[1] Jiawei Han, Micheline Kamber, Data Mining: Concepts and techniques, Morgan Kaufmann Publishers, 2001.


