Imperial College London

Imperial College London

Department of Mathematics

Study of the Conformance Anomaly Detection Algorithm on Streamed Data

Author: Millie Deng (CID: 02003063)

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2020-2021

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Sign: Millie Deng

Date: 05/09/2021

Abstract

In this thesis, we study a newly established unsupervised anomaly detection algorithm proposed by Cochrane et al. (2020). Particularly, we theoretically prove the dependence of this approach on the underlying parameters and applied it to financial streams of data for the anomaly detection tasks. We first use the signature method to extract essential features of the time series data. Then based on the Mahalanobis distance, the concept of conformance is introduced to measure the distance among the signature of the data. Finally, we implement the Gaussian concentration inequality to identify the conformance threshold for the corpus of data with different dimensions, corpus sizes and error bounds.

The simulated Brownian motion data, which is usually used to approximate the stock behaviour, is utilized to test the effectiveness of the conformance algorithm and propose suitable input parameters. Then, we further modify this algorithm and use it to identify the anomaly trading date of the cryptocurrency order book data. It turned out that this conformance anomaly detection method is effective for identifying anomaly market behaviour.

Acknowledgements

I want to express my deep gratitude to my supervisor Dr Thomas Cass for introducing me to the promising and challenging domain of the path signature. Without his patient and valuable instruction throughout this three month, I wouldn't be able to accomplish the research.

I would also like to thank Remy Messdene, Dr Cristopher Salvi and Gordon Lee for their assistance during the summer and my family, friends, the whole MSc Mathematics and Finance team at Imperial College for their support and encouragement during this special year.

Contents

List of Figures

List of Tables

Introduction

According to Goldstein and Uchida (2016) and Grubbs (1969), the anomaly observations are outliers that appear to be significantly different from the rest of the observations concerning their features. Their appearance is also rare compare with the normal instances. The motivations of anomaly detection tasks initially focus on removing the outliers before further analysis, as some tasks like pattern recognition are sensitive to extreme behaviour. Nowadays, with more fast and meticulous methods being proposed, the researchers are interested in studying the anomaly itself — like the cause and pattern of the particular instances. Therefore, various anomaly detection applications have been proposed in different domains, including the fraud detection of financial transactions, the intrusion detection of server applications, and illness identification based on medical images or signals.

The general setup of the anomaly detection tasks is similar to the classifications problem, where an anomaly classifier is first trained and then tested to check the algorithm's performance. The method could be supervised, semi-supervised, and unsupervised, depending on whether the training and test data are labelled. One example of the supervised anomaly detection is fraud detection for credit card payments logs, where labelled data is available for each transaction. The typical supervised learning algorithms are Decision trees, Support Vector Machines (SVM). The training set of the semi-supervised anomaly detection method only contains normal instances, so the anomaly is identified if it deviated from the standard corpus. Methods like One-class SVMs and density function modelling are designed to tackle this kind of problem.

Most anomaly detection tasks require using the unsupervised learning algorithm, as no previous knowledge of the data is known. Therefore, only intrinsic features of the data are available during the identification process. In this thesis, we focus on studying the unlabelled multidimensional time-series data (streamed data), so only unsupervised learning methods will be considered.

Traditional unsupervised anomaly detection methods usually require metrics to measure the distances among the corpus of data and decide whether an event is an anomaly based on the distance. Those methods' performance relies on the arbitrary choice of the metric, which requires external information that may be hard to interpret and define. To this end, Cochrane et al. (2020) proposed a simple novel anomaly detection approach for streamed data. Based on the features of streams specified through the signature method, the conformance threshold is calculated to distinguish the anomaly instances from the normal observations. The focus of this thesis would be understanding the dependence of this approach on the underlying parameters and then applied it to the anomaly detection tasks for the simulated Brownian motion and real-market financial streams of data.

The path signature, construct by iterated integral of the path, is an informative transform that maps the multidimensional paths to the sequence of the iterated integrals (Gyurkó et al., 2013). The main reason we are interested in the signature is due to its practical interpretation of the characteristic feature of the data. Besides, the analytic and geometric properties of the iterated integral allow us to use few terms to interpret most key information of the data. That makes the signature calculation a suitable method to prepare data for the anomaly detection algorithm. Therefore, we will first transform the streamed data into paths then use the truncated signature of the path as inputs for later training and testing steps. A detailed explanation of the signature and its essential properties will be given in Chapter 1.

Then we move on to the study of the main algorithm in Chapter 2. A norm, variance norm,

is proposed to measure the distance for the vector in space V. We proved it coincides with the Mahalanobis distance for data in $\mathbb{R}^d$. Then, the conformance threshold for identifying the anomaly data is defined using the Mahalanobis distance. With the Gaussian concentration inequality, we could calculate the exact relationship of the conformance threshold q_ϵ for Gaussian data with some dimension d, corpus size n, and error bound ϵ. Then, based on this result, we proposed the conformance anomaly detection algorithm which use the empirical conformance threshold of the signature of the streamed data to identify the anomaly instances. The confusion matrix is also introduced for the evaluation of the algorithm.

In Chapter 3, the empirical test of the algorithm is implemented on Brownian motion data, a stochastic process that is usually used to simulate the logarithm stock price. One, two, and four-dimensional Brownian motion data contaminated by different drifts and variances are first generated. Then, we test different input parameters for each data group through the algorithm and evaluate the performance by calculating the confusion matrix. Based on the evaluation results for the specificity(true negative rate), sensitivity(true positive rate), and overall accuracy, we find that the conformance algorithm could effective identify the anomaly instances under proper input parameters. Therefore, we propose the most suitable parameters for the data with a particular contamination rate, which could be utilized in other empirical tests.

In Chapter 4, we modify the anomaly detection algorithm to identify the anomaly trading date for cryptocurrency order book data in the eight-month period. The level one order book data after certain transformations become a 3-dimensional data frame, and by assigning data within each date as a path, we encode it into streamed data. After finding the suitable parameter based on the conclusions in Chapter 3 and the training data's evaluation results, we plot the identified anomaly date against the logarithm mid-price and spread of the test data and find it could identify most of the anomaly dates.

Chapter 1

Path Signature

The signature of a path is essentially a vector construct by iterated integral of the path. After Chen (1958) first introduced the study of the signature for the piece-wise smooth paths, Hambly and Lyons (2010) extended the method to the continuous paths with bounded variation. The intuition of signature initially arises in the Taylor expansion of the controlled ordinary differential equations (Lyons, 2014). It has been explained by Lyons (2014) and Chevyrev and Kormilitzin (2016) that the signature determines the solution of the controlled differential equations.

Scholars have put forward various applications of the signature method in different fields. Particularly, Levin et al. (2013) introduced the signature to the study of financial time series data. Chevyrev and Kormilitzin (2016) explained the signature's ability to extract characteristic features from data and demonstrate it is suitable for various types of machine learning applications, including both supervised and unsupervised learning algorithms. Moreover, Gyurkó et al. (2013) provide a concrete example of utilizing the signature to extract underlying features of the order book data and then perform the classification based on the linear regression model.

In this thesis, we focus on using the signature to extract essential features of the streamed data. We will give a detailed explanation of the concepts and properties of the path signature in this Chapter. An introduction of the basic definitions of the path signature is given in section 1.1. Then we explain some crucial analytic and geometric properties of signature in section 1.2 and 1.3. The last section explained how we calculated the signature for the streamed data and proposed some effective transformation methods for the streamed data to reveal its specific properties. Those concepts and properties provide a base for the anomaly detection algorithm that we will explain in the next Chapter.

1.1 Definition of the Path Signature

A path X in $\mathbb{R}^d$ is a continuous mapping from some interval $[a, b]$ to $\mathbb{R}^d$, written as $X : [a, b] \to \mathbb{R}^d$. We usually use $X_t = X(t)$ to denote the dependence on the parameter $t \in [a, b]$. Assume the path is piece-wise differential and if the path has derivatives of all orders, then it is a smooth path. For example , a two-dimensional smooth path with $t \in [0, 1]$ could be

$$X_t = \{X_t^1, X_t^2\} = \{t, t^2\}$$

which could be plotted in figure 1.1. A two-dimensional non-smooth piece-wise linear path could be

$$X_t = \{X_t^1, X_t^2\} = \{t, f(t)\}$$

where function f represent a stock price as shown in figure 1.2.

To include both smooth and non-smooth paths, we suppose the path to be continuous and bounded variation in the following chapter. We find this is a suitable assumption for data we will be dealing with.

The definition of signature is based on the iterated integral of the path. For any path in d

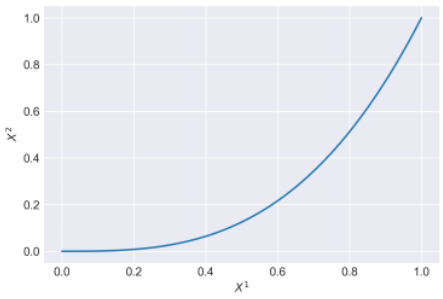


Figure 1.1: Two-dimensional smooth path

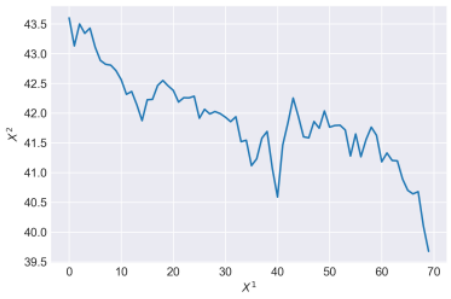


Figure 1.2: Two-dimensional non-smooth path

dimensional space $X = (X_t^1, \cdots, X_t^d) : [a, b] \to \mathbb{R}^d$, define the integral of i-th coordinate of path X at time $t \in [a, b]$ as

$$S(X)_{a,t}^i = \int_{a \le s \le t} \mathrm{d}X_s^i = X_t^i - X_a^i \tag{1.1.1}$$

which $S(X)_{a,t}^i$ is a real valued path that mapping from $[a, t]$ to $\mathbb{R}$. Then follow this single integral, we can define the double-iterated integral for coordinate $i, j \in \{1, 2, \cdots, d\}$ of path X

$$S(X)_{a,t}^{i,j} = \int_{a \le s \le t} S(X)_{a,s}^i \, \mathrm{d}X_s^j = \int_{a \le r \le s \le t} \mathrm{d}X_r^i \, \mathrm{d}X_s^j \tag{1.1.2}$$

Since $S(X)_{a,s}^i$ and X_s^j are both real-valued paths, the $S(X)_{a,t}^{i,j}$ is also a real value path that project $[a, t]$ to $\mathbb{R}$.

Follow this intuition, we could define the *k-fold iterated integral* of path X for multi-indexes $i_1, \cdots, i_k \in \{1, \cdots, d\}$ as

$$S(X)_{a,b}^{i_1,\cdots,i_k} = \int_{a \le t_1 \le \cdots \le t_k \le b} \mathrm{d}X_{t1}^{i_1} \cdots \mathrm{d}X_{t_k}^{i_k} \tag{1.1.3}$$

which again is a real-value path. Define the set that contains all the combination of the coordinate to be

$$W = \{(i_1, \cdots, i_k) \mid k \ge 1, i_1, \cdots, i_k \in \{1, \cdots, d\}\}$$

and it is called the *set of words on the alphabet* $A = \{1, \cdots, d\}$ consisting of d letters.

Denote the k-fold tensor product space of $\mathbb{R}^d$ as $(\mathbb{R}^d)^{\otimes k}$, it is the vector space generated by

$$\{e_{i_1} \otimes \cdots \otimes e_{i_k} : i_1, \cdots, i_k = \{1, \cdots, d\}\}$$

where $\{e_i : i = 1, \cdots, d\}$ is the standard basis of $\mathbb{R}^d$. We use the notation $S(X)^k$ to describe a k-tensor over $\mathbb{R}^d$:

$$\begin{aligned} S(X)_{a,b}^k &= \sum_{i_1,\cdots,i_k=1}^{d} S(X)_{a,b}^{k;i_1,\cdots,i_k} e_{i_1} \otimes \cdots \otimes e_{i_k} \\ &= \left(\int_{a \le t_1 \le \cdots \le t_k \le b} \mathrm{d}X_{t_1}^{i_1} \cdots \mathrm{d}X_{t_k}^{i_k} \right)_{i_1,\cdots,i_k \in \{1,,d\}} \end{aligned} \tag{1.1.4}$$

Then $S(X)_{a,b}^k$ belongs to k-fold tensor product space $(\mathbb{R}^d)^{\otimes k}$.

Definition 1.1.1 (Signature)**.** The signature of a path $X : [a, b] \to \mathbb{R}^d$, denoted by $\mathbb{S}(X)_{a,b}$, is the sequence of tensors (infinite series of all iterated integrals of X). It could be write as:

$$\mathbb{S}(X)_{a,b} := (1, S(X)^1_{a,b}, \cdots, S(X)^k_{a,b}, \cdots) \in \prod_{k=0}^{\infty} (\mathbb{R}^d)^{\otimes k} \tag{1.1.5}$$

We usually set the first term equal to 1 by convention and the signature can also be seen as a sequence of real numbers with multi-indexes run along the set of all words W.

Definition 1.1.2. The the signature of path $X : [a, b] \to \mathbb{R}^d$ of level $N \in \mathbb{N}$ is the truncate signature of a path X:

$$\mathbb{S}^N(X)_{a,b} := (1, S(X)^1_{a,b}, \cdots, S(X)^N_{a,b}) \tag{1.1.6}$$

$$= \left(\int_{a \le t_1 \le \cdots \le t_k \le b} \mathrm{d}X^{i_1}_{t1} \cdots \mathrm{d}X^{i_k}_{t_k} \right)_{\substack{i_1, \cdots, i_k \in \{1,,d\} \\ k=0,1,2,\cdots,N}} \tag{1.1.7}$$

Therefore, the truncated signature $\mathbb{S}^N(X)_{a,b}$ lives in the truncated tensor algebra which has dimension:

$$d_N := 1 + d + d^2 + \cdots + d^N = \frac{d^{N+1} - 1}{d - 1}.$$

1.2 Analytical Properties of the Signature

In this chapter, we will explain three essential properties of the signature. Firstly, we introduced the parameterisation invariant property that used to prove the invariant of signature for streamed data after transform in section 1.4.

We first state the definition of the reparameterisation:

Definition 1.2.1 (Reparameterisation)**.** Suppose a path $X : [a, b] \to \mathbb{R}^d$. The reparameterisation of $[c, d]$ onto $[a, b]$ is the monotonically increasing function $\sigma : [c, d] \to [a, b]$ and any path $Y = X \circ \sigma : [c, d] \to \mathbb{R}^d$ is called a reparameterisation of X.

Then the signature of reparameterisation invariant property is given as:

Lemma 1.2.2. *For the path $X : [a, b] \to \mathbb{R}^d$, suppose path $Y : [c, d] \to \mathbb{R}^d$ is the reparameterisation of X by $\sigma : [c, d] \to [a, b]$, then*

$$\mathbb{S}(X)_{a,b} = \mathbb{S}(Y)_{c,d} \tag{1.2.1}$$

Proof. Adopting the prove given by Cass (2021), we use the mathematical induction to prove this lemma. For 1-fold iterated integral, any $t \in [c, d]$

$$\begin{aligned} S(Y)^1_{c,t} &= Y(t) - Y(c) = X(\sigma(t)) - X(\sigma(c)) \\ &= X(\sigma(t)) - X(a) = S(X)^1_{a,\sigma(t)} \end{aligned}$$

Suppose for any word $w = i_1 i_2 \cdots i_m$ of length $m \le k$ with $i_1, \cdots i_m \in \{1, \cdots, d\}$, and any $t \in [c, d]$:

$$S(Y)^{m;w}_{c,t} = S(X)^{m;w}_{a,\sigma(t)} \tag{1.2.2}$$

Then for $w' = i_1 i_2 \cdots i_k i = wi$ where $i = 1, \cdots, d$:

$$\begin{aligned} S(Y)^{k+1;w'}_{c,t} &= \int_c^t S(Y)^{k;w}_{c,s} dY^i_s \\ &= \int_a^b S(X)^{k;w}_{a,\sigma(s)} dX^i_{\sigma(s)} \\ &= S(X)^{k+1;w'}_{a,\sigma(t)} \end{aligned}$$

As the signature is a sequence of iterated integral with multi-indexes run along the set of all words w, the signature is therefore parameterisation invariance. □

Then we introduce the lemma that explains the amount of information that truncated signature could preserve. The proof is also adopted from Cass (2021). Firstly, we introduce the collection of norms on the space $(\mathbb{R}^d)^{\otimes k}$. When $k = 1$ we use the Euclidean norm on $\mathbb{R}^d$ and when $k = 2, \cdots$, based on the expression of the k-tensor in equation (1.1.4)

$$\left\|S(X)^k_{a,b}\right\|_k = \left(\sum_{i_1,\cdots,i_k=1}^{d} \left(S^{k;i_1,\cdots,i_k}_{a,b}(X)\right)^2\right)^{\frac{1}{2}} \tag{1.2.3}$$

Lemma 1.2.3. *Suppose a path $X : [a, b] \to \mathbb{R}^d$. For the k-tensor $S^k(X)_{a,t}$ over $\mathbb{R}^d$ where $t \in [a, b]$ we have*

$$\|S(X)^k_{a,t}\|_k \leq \frac{L(t)^k}{k!} \tag{1.2.4}$$

where L is the length of the path:

$$L(t) := \int_a^t |\dot{X}_t| \ \mathrm{d}t \tag{1.2.5}$$

Proof. When $k = 1$, $\left\|\int_a^t dX_t\right\|_k = |\int_a^t \dot{X}_t dt| \leq \int_a^t |\dot{X}_t| dt = L$.
For some arbitrary k, suppose for any $l = 1, \cdots, k$, $t \in [a, b]$:

$$\|S(X)^l_{a,t}\|_l \leq \frac{L(t)^l}{k!} \tag{1.2.6}$$

holds. Then by equation (1.1.4) and the property of the tensor product[1]:

$$\begin{aligned}
\left\|S(X)^{k+1}_{a,t}\right\|_{k+1} &= \left\|\int_a^t \left(\int_{a\leq t_1\leq\cdots\leq t_k\leq s} \mathrm{d}X_{t_1} \otimes \cdots \otimes \ \mathrm{d}X_{t_k}\right) \otimes dX_s\right\|_{k+1} \\
&= \left\|\int_a^t S(X)^k_{a,s} \otimes dX_s\right\|_{k+1} \\
&\leq \int_a^t \left\|S(X)^k_{a,s}\right\|_k \|dX_s\|_1 \\
&\leq \int_a^t \frac{L(s)^k}{k!} dL(s) = \frac{L(t)^{k+1}}{(k+1)!}
\end{aligned}$$

Therefore, according to the mathematical induction, this lemma holds. □

As equation (1.2.4) shows, the norm of the $S(\gamma)^k_{a,b}$ is bounded by a constant value for each k and the value decreased as k becomes larger enough. Therefore, we could conclude that most of the signature information is contained by the first few terms of the signature. Typically, the signature with level 5 is sufficient to represent the features of the path.

Finally, we introduce the important *shuffle product identity* that will be used to calculate the signature's conformance in the next Chapter. The property shows that the product of two iterated integral with multi-indexes $\{i_1, \cdots, i_k\}$ and $\{j_1, \cdots, j_m\}$ could be expressed as the sum of higher order iterated integral with multi-indexes only depend on $\{i_1, \cdots, i_k, j_1, \cdots, j_m\}$.

We first give the definition of *shuffle*, it is a certain way to permuted the sets:

Definition 1.2.4 ((n, m)-shuffle)**.** Define a permutation σ of the set $\{1, 2, \cdots, n+m\}$ that satisfy

$$\sigma(1) < \sigma(2) < \cdots < \sigma(n) \text{ and } \sigma(n+1) < \sigma(n+2) < \cdots < \sigma(n+m)$$

as the set of (n, m)-shuffle denote by $Sh(n, m)$.

For multi-indexes $I = (i_1, \cdots i_l)$ and $J = (j_1, \cdots, j_m)$ where $i_1, \cdots, i_l$, $j_1, \cdots, j_m \in \{1, \cdots, d\}$, we could define the multi-index

$$(r_1, \cdots, r_k, r_{k+1}, \cdots, r_{k+m}) = (i_1, \cdots, i_k, j_1 \cdots, j_m)$$

Then the shuffle product of I and J is defined as:

[1]The third line of the prove used a property and we states it in appendix A.1

Definition 1.2.5 (Shuffle product)**.** The finite set of multi-index with length $k+m$ is the shuffle product of I and J:

$$I \sqcup\!\sqcup J = \{(r_{\sigma(1)}, \cdots, r_{\sigma(l+m)}) \mid \sigma \in Sh(l.m)\} \tag{1.2.7}$$

Theorem 1.2.6 (Shuffle product identity for k-fold iterated integral)**.** *Suppose a path* $X : [a, b] \to \mathbb{R}^d$ *and the multi-indexes* $I = (i_1, \cdots i_l)$ *and* $J = (j_1, \cdots, j_m)$ *where* $i_1, \cdots, i_l,\ j_1, \cdots, j_m \in \{1, \cdots, d\}$*:*

$$S(X)^I_{a,b} S(X)^J_{a,b} = \sum_{K \in I \sqcup\!\sqcup J} S(X)^K_{a,b} \tag{1.2.8}$$

The main step to prove this theorem is to apply *Fubini's theorem* to the product of two iterated integrals. Then it could be written as the sum of higher older iterated integrals. One simple example to illustrate this identity would be:

$$S(X)^1_{a,b} S(X)^2_{a,b} = S(X)^{1,2}_{a,b} + S(X)^{2,1}_{a,b} \tag{1.2.9}$$

Based on theorem 1.2.6, we could deduce that the product of terms of the signature can be write as the sum of the linear combination of the higher order terms.

1.3 Geometric Interpretations of the Signature

There are some straightforward interpretations of the path that the signature could demonstrate, like the increment and signed area of the path. We use two-dimensional paths to illustrate how the lower level signature capture the character of the paths, and those conclusions could be put forward to a higher signature level.

For example, as equation (1.1.1) shows, the signature of level one could only represent the increment of the path, and cannot interpret the difference of the area that the path encloses. This could be illustrated by comparing two-dimensional paths $X = (X_1, X_2)$ in figure 1.3. We could easily see that the signature of order one is the same for two paths.

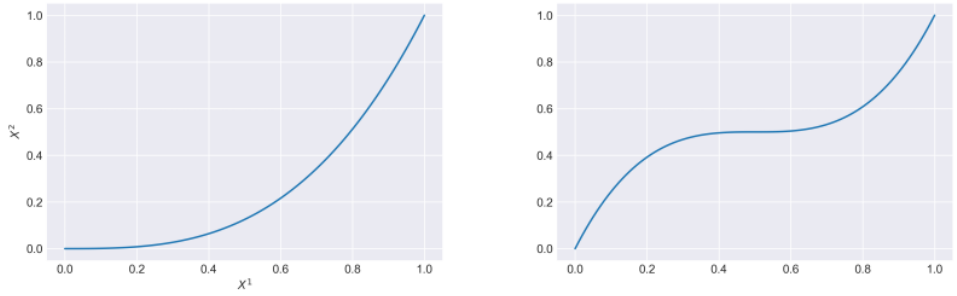


Figure 1.3: Two-dimensional smooth path comparison

Some combination of 2-fold iterated integrals have the ability to represent the signed area enclosed by the path and the chord (straight line connecting the path's beginning and endpoint). It is called *levy area*. For the two dimensional path, the levy area could be expressed as:

$$A := \frac{1}{2}\left(\int_{a<t_1<t_2<b} \mathrm{d}X^1_{t_1} X^2_{t_2} - \int_{a<t_1<t_2<b} \mathrm{d}X^2_{t_1} X^1_{t_2}\right) = \frac{1}{2}(S(X^{1,2}_{a,b}) - S(X^{2,1}_{a,b})) \tag{1.3.1}$$

The levy area for the above paths is shown in figure 1.4. We could see that paths with the counter clockwise movement enclosed the positive levy area, and those with clockwise direction enclosed the negative area.

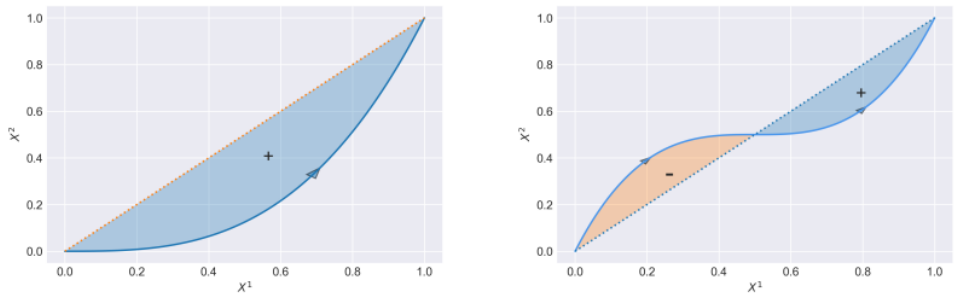


Figure 1.4: Two-dimensional smooth path Levy area

The higher level of signature could interpret other specific geometric characteristics for the paths. Specifically, the signature with order three could identify the difference between paths with the same levy areas. To sum up, those geometric interpretations demonstrate that the signature with lower level could capture certain characteristics of the paths which valid the property in section 1.2 and ensure the viability of using the first few terms of the signature to lower the dimension of the streamed data.

1.4 Streamed Data and its Transform

In financial fields, the data we obtained are usually streams of data.

Definition 1.4.1 (Streams of data)**.** The space of streams of data in a set χ is defined as

$$\mathcal{S}(\chi) := \{x = (x_1, \cdots, x_n) : x_i \in \chi, n \in \mathbb{N}\}$$

Therefore, to calculate the signature for those data, we need to map the discrete data stream into continuous time series. The most straight forward transform is using the linear interpolation to convert the discrete points into continuous paths. Cochrane et al. (2020) gives the following definition for signature of streamed data:

Definition 1.4.2 (Signature for streams of data)**.** Suppose streams of data in set $\mathcal{C} \subset \mathbb{R}^d$

$$x = (x_1, \cdots, x_n) \in \mathcal{S}(\mathcal{C})$$

Let

$$X\left(\frac{i}{n-1}\right) = x_{i+1} \text{ for } i = \{0, 1, \cdots, n-1\}$$

and linear interpolation in between. Since $x_i \in \mathbb{R}^d$, the dimension for each time stamps is equal to d. Therefore, we could denote X as $X = (X_1, \cdots, X_d) : [0, 1] \to \mathbb{R}^d$. X is a continuous and bounded variation path so we could further define the signature of order N for the streamed data x:

$$\mathbb{S}^N(x)_{a,b} = \left(\int_{a \le t_1 \le \cdots \le t_k \le b} \mathrm{d}X_{t1}^{i_1} \cdots \mathrm{d}X_{t_k}^{i_k} \right)_{\substack{i_1, \cdots, i_k \in \{1,,d\} \\ k=0,1,2,\cdots,N}} \tag{1.4.1}$$

In the Chapter3 and 4, we will calculate the signature for simulated and real market streamed data based on definition 1.4.2.

The certain transformation that maps the streamed data to another streamed data could reveal specific properties of the data. As we explain in section 1.2, the signature has the parameterisation invariance property, which allows the data after reparameterisation to still has the same signature value. Therefore, we could implement the transformation while ensuring the signature of streamed data remains unchanged. We introduce three main transform methods that will be utilised in the practical steps in Chapter 3. All transforms are tested effectively in preserving certain properties

of the paths in the work of Flint et al. (2016), and Cochrane et al. (2020).

Suppose a stream of data $x_i \in \mathbb{R}^d$

$$x = (x_1, \cdots, x_n)$$

Add Time Transform

The *add time transform* is simply adding an extra time dimension into the stream of data:

$$X^{add-time} = ((t_0, x_0), \cdots, (t_i, x_i), \cdots, (t_n, x_n)) \tag{1.4.2}$$

This transform could be effective for the financial data stream as the time stamp is usually a essential information for financial time series data, and adding this time value allow the signature to fetch more precise feature of the data.

Invisibility Transform

The invisibility transform given in Cochrane et al. (2020) is a transformation that maps the data in $\mathbb{R}^d$ into $\mathbb{R}^{d+1}$:

$$X^{invis} = (\hat{x}_0, \cdots, \hat{x}_i, \cdots, \hat{x}_{n+1}) \tag{1.4.3}$$

$\hat{x}_0 = (x_i, 0)$ and for $i = 1, \cdots, n+1$

$$\hat{x}_i = (x_{i-1}, 1)$$

The invisibility transform is used to preserve the absolute value of the streamed data after calculating the signature.

Lead-lag Transform

The evaluation of quadratic variation is an essential property for financial data. To capture the quadratic variation for the paths, we need to perform the *lead-lag transform* which is similar to the calculation of the Levy area. The lead-lag transform maps data in $\mathbb{R}^d$ dimension into a $\mathbb{R}^{2d}$ dimension:

$$X^{lead-lag} = (\hat{x}_0, \cdots, \hat{x}_i, \cdots, \hat{x}_{2n}) \tag{1.4.4}$$

for $i = 0, \cdots, n$

$$\hat{x}_{2i} = (x_i, x_i) \qquad \hat{x}_{2i+1} = (x_i, x_{i+1})$$

We use a simple example to illustrate how the lead-lag transform capture the quadratic variation of the path:

Remark 1.4.3. Suppose a one dimensional path $X = \{x_1, x_2, x_3, x_4\} = \{1, 3, 2, 4\}$, the lead-lag transform result is:

$$X^{\text{lead-lag}} = \{(1,1), (1,3), (3,3), (3,2), (2,2), (2,4), (4,4)\} \tag{1.4.5}$$

We could plot the lead-lag result in figure 1.5:

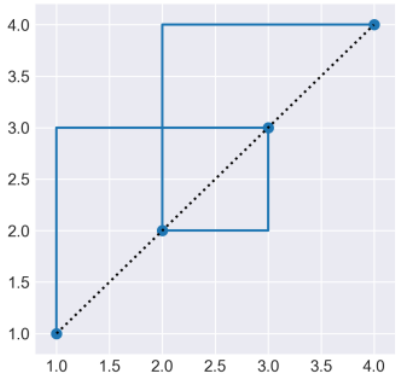


Figure 1.5: Lead-lag transform for the example

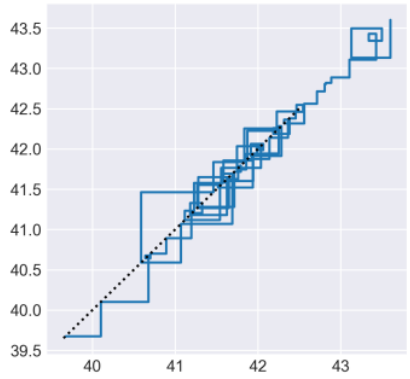


Figure 1.6: Lead-lag transform for the stock path

The absolute value for the levy area of the path after lead-lag transform is:

$$|A| = \frac{1}{2}[(3-1)^2 + (3-2)^2 + (4-2)^2] \tag{1.4.6}$$

and it coincides with the quadratic variation of the data which is:

$$|A| = \frac{1}{2}QV(X) = \frac{1}{2}\sum_{i=0}^{4}(X_{i+1} - X_i)^2 \tag{1.4.7}$$

We also draw the lead-lag transform result for the entire stock price data, the result is shown in figure 1.6.

Chapter 2

Anomaly Detection using the Conformance Distance

After calculating the signature of the streamed data, we now move on to study the anomaly detection method for the transformed data in the vector space.

Various types of norms that aim to measure the distance among vectors have been proposed to project a real or complex vector space to a non-negative real number. Traditional normal like the Euclidean norm only captures each element's value and draws less attention to the interaction among the data in higher dimension space. Inspired by the covariance, which measures the joint variability of two random variables, Cochrane et al. (2020) construct the variance norm for the vector in space V. When $V = \mathbb{R}^d$, we find this norm is the same as the Mahalanobis distance, which has been widely used in the classification of high dimensional data.

Then we move forward to searching for methods that utilized distance to identify the anomaly behaviour. Cochrane et al. (2020) proposed the use of Gaussian concentration inequality, a theory that demonstrates how much a random variable deviates from its mean. The proof of another lemma, Johnson–Lindenstrauss lemma in infinite dimension case, also utilized the concentration inequality. Inspired by its proving steps, we use the inequality to theoretically calculated a threshold q_ϵ for identifying anomalies among data with underlying parameters sample size n , dimension d and error bound ϵ. Some empirical test of this threshold is also given, and we find there is a discrepancy between the empirical and theoretical results that could cause by the approximation steps during the proof. However, the dependence of the theoretical and empirical threshold on the underlying parameters is consistent. Therefore, using the threshold to identify the anomalies is still valid, and we could establish a detailed algorithm based on this conformance threshold.

In the following sections, we will first introduce the definitions and properties of the variance norm and conformance distance. Then we apply the Gaussian concentration method to define the conformance threshold used to identify the outlying behaviours. After that, some effective evaluation methods for the anomaly detection algorithm are discussed in the third section. In the last section, we summarize the conformance algorithm and prepare for its implementation in the next Chapter.

2.1 Measuring the Distance

We first introduce the basic definition relate to the vector space that we will be working on.

Definition 2.1.1 (Dual Space)**.** Given a vector space V, the Dual space V^* is defined as the set of all linear maps $\varphi : V \to \mathbb{F}$. The φ is called a linear functional.

Remark 2.1.2. Suppose $V = \mathbb{R}^d$ and the elements of its dual space V^* to be $\varphi : \mathbb{R}^d \to \mathbb{R}$. Then φ represent linear maps that accept vectors $v = (v_1, \cdots, v_d) \in \mathbb{R}^d$ as inputs and spits out real numbers, and the dual of V has the same dimension as V which is $\mathbb{R}^d$ in this case. The linear map

$\varphi \in V^*$ can be write as

$$\varphi(x) = \langle x, v \rangle = \sum_{i=1}^{d} v_i x_i \tag{2.1.1}$$

where $v_i \in \mathbb{R}$. Suppose $\|v\|_2 = \sqrt{v_1^2 + \cdots + v_n^2}$ as the euclidean norm, an immediate corollary is that by Cauchy–Schwarz, we have

$$|\varphi(x)| \leq |\langle x, v \rangle| \leq \|v\|_2 \|x\|_2 = c\|x\|_2$$

where $c = \|v\|_2$. Therefore, we proved that $\varphi(x)$ is bounded.

Let μ be a probability measure that re-centred to have mean zero on a vector space V. Define a covariance quadratic form on the dual of V, it could be written as:

$$cov(\psi, \varphi) := \int \psi(x)\varphi(x)\mu(dx) = \mathbb{E}^{\mu}[\psi(x)\varphi(x)]$$

If we suppose $V = \mathbb{R}^d$, by the equation (2.1.1) in the Remark 2.1.2, the covariance becomes,

$$\begin{aligned} cov(\psi, \varphi) &= \mathbb{E}^{\mu}[\psi(x)\varphi(x)] \\ &= \mathbb{E}^{\mu}[(\sum_{i=1}^{d} v_i x_i)(\sum_{j=1}^{d} w_j x_j)] \\ &= v^T K w \end{aligned} \tag{2.1.2}$$

where $K = \mathbb{E}^{\mu}[x^T x]$ is the covariance matrix of x, and $v = (v_1, \cdots, v_d), w = (w_1, \cdots, w_d) \in \mathbb{R}^d$.

2.1.1 The Variance Norm and the Mahalanobis Distance

We first recall the definition of the norm and then state the variance norm defined by Cochrane et al. (2020):

Definition 2.1.3 (Norm)**.** Given a vector space V, a norm on V is a real-valued function $\|\cdot\| : V \to \mathbb{R}$ with the following properties:

- for all $x \in V$, if $\|x\| = 0$ then $x = 0$.
- $\|sx\| = |s|\|x\|$ for all $x \in V$ and all scalars s. ($|s|$ denotes the absolute value of a scalar s)
- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$

Definition 2.1.4 (Variance norm)**.** Let μ be a probability measure (re-centred to have mean zero) on the vector space V. Then covariance quadratic form $cov(\psi, \varphi)$ induces a dual norm defined for $x \in V$ by

$$\|x\|_{\mu} := \sup_{cov(\varphi,\varphi) \leq 1} \varphi(x) \tag{2.1.3}$$

According to equation (2.1.1) and (2.1.2), if $V = \mathbb{R}^d$, the variance norm can be written as:

$$\|x\|_{\mu} = \sup_{v, v^T K v \leq 1} v^T x = \sup_{v, v^T K v \leq 1} \sum_{i=1}^{d} v_i x_i \tag{2.1.4}$$

where $K = \mathbb{E}^{\mu}[x^T x]$ is the covariance matrix for centred x and $v_i \in \mathbb{R}$ for $i = 1, \cdots, d$

Equation (2.1.4) could be seen as maximizing a linear function over a ellipsoid centred at the origin (Boyd et al., 2004). To solve this optimization problem, define new variables $y = K^{\frac{1}{2}} v$ and $\widetilde{x} = K^{\frac{1}{2}} x$, then equation (2.1.4) equals to:

$$\begin{aligned} \max \quad & \widetilde{x}^T y \\ \text{s.t.} \quad & y^T y \leq 1 \end{aligned} \tag{2.1.5}$$

By solving the optimization problem, we could deduce that the solution to equation (2.1.5) is $y^* = \frac{\widetilde{x}}{\|\widetilde{x}\|_2}$. In the following chapter, we suppose the K is *positive definite* where elements in x is linear independent unless $x = 0$ (then $\|x\|_\mu = 0$) [1]. Therefore, K is invertible and the solution for origin problem is

$$v^* = \frac{K^{-1}x}{\|K^{-\frac{1}{2}}x\|_2} \tag{2.1.6}$$

Insert this result in equation (2.1.4), the variance norm for $x \in \mathbb{R}^d$ becomes:

$$\|x\|_\mu = \sqrt{x^T K^{-1} x} \tag{2.1.7}$$

This result corresponds to the definition of *Mahalanobis distance* $\|x\|_S$ which represent the distance between random vector x and its mean. Now we prove that the variance norm is actually a norm:

Proof. For the first property in definition 2.1.3, suppose there exists some $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, that

$$\|x\|_\mu = \sqrt{x^T K^{-1} x} = 0 \tag{2.1.8}$$

Since we suppose the $K = \mathbb{E}^\mu[x^T x]$ is positive definite unless $x = 0$. Therefore, the quadratic form

$$x^T K^{-1} x \geq 0$$

and the equivalent only holds when $x = 0$.

For the second statement,

$$\begin{aligned} \|sx\|_\mu &= \sqrt{s^2 x^T K^{-1} x} \\ &= |s|\sqrt{x^T K^{-1} x} \\ &= |s|\|x\|_\mu \end{aligned} \tag{2.1.9}$$

Now we prove the last property (adapted the prove given by Costa (2018)). By Cholesky decomposition for the positive definite matrix K (when $x \neq 0$), there exist a unique triangular matrix U with positive diagonal entries such that

$$K = UU^T$$

If we suppose $\widetilde{x} = U^{-1}x$, $\widetilde{y} = U^{-1}y$ and $\|\cdot\|_2$ to be the Euclidean norm, then:

$$\begin{aligned} \|x\|_\mu &= \sqrt{x^T (UU^T)^{-1} x} = \sqrt{x^T (U^T)^{-1} U^{-1} x} \\ &= \sqrt{(U^{-1}x)^T U^{-1} x} = \|\widetilde{x}\|_2 \end{aligned} \tag{2.1.10}$$

Then replace x by y and $x + y$, we have

$$\begin{aligned} \|y\|_\mu &= \|\widetilde{y}\|_2 \\ \|x + y\|_\mu &= \|\widetilde{x} + \widetilde{y}\|_2 \end{aligned} \tag{2.1.11}$$

Since by the triangular inequality for the Euclidean norm we have :

$$\|\widetilde{x} + \widetilde{y}\|_2 \leq \|\widetilde{x}\|_2 + \|\widetilde{y}\|_2 \tag{2.1.12}$$

Then by equations (2.1.10) and (2.1.11), we have:

$$\|x + y\|_\mu \leq \|x\|_\mu + \|y\|_\mu \tag{2.1.13}$$

□

[1] This assumption is plausible as we assume the data to follow i.i.d distribution in later analyze.

Another important property that we will use is the *Lipschitz continuity* of the variance norm. Zantedeschi et al. (2016) proved the Mahalanobis distance of a centred pair $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$ with some positive semi-definite matrix M

$$\|x_2 - x_1\|_\mu = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$

is k-Lipschitz continuous function. Adapting the proof, we first state the definition of multi-variate Lipschitz continuity to prepare for the calculation of the Lipschitz constant for the variance norm.

Definition 2.1.5 (Multi-variate Lipschitz continuity)**.** The function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is k_n-lipschitez with respect to a norm $\|\cdot\|_n$ for any $x_1, x_2, x_1', x_2' \in \mathbb{R}^d$:

$$|f(x_1, x_2) - f(x_1', x_2')| \leq k_n \left\| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \right\|_n \tag{2.1.14}$$

and if f is differential, then the best constant k_n could be estimated by:

$$k_n = \sup_{x_1, x_2, x_1', x_2' \in \mathbb{R}^d} \left(\frac{|f(x_1, x_2) - f(x_1', x_2')|}{\left\| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \right\|_n} \right) \tag{2.1.15}$$

$$= \sup_{x_1, x_2 \in \mathbb{R}^d} \|\nabla f(x_1, x_2)\|_n \tag{2.1.16}$$

Lemma 2.1.6. *For $x_1, x_2 \in \mathbb{R}^d$, the variance norm*

$$\|x_1 - x_2\|_\mu = \sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}$$

with $K = UU^T = \mathbb{E}^\mu[(x_1 - x_2)(x_1 - x_2)^T]$ is k–Lipschitz function with respect to the norm $\|\cdot\|_2$ where

$$k = \sqrt{2}\|V\|_2$$

and $V = (U^{-1})^T$

Proof. By definition 2.1.5, $k = \sup_{x_1, x_2 \in \mathbb{R}^d} \|\nabla f(x_1, x_2)\|_2$, and since we have

$$\begin{aligned} \frac{\partial \sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}}{\partial x_1} &= \frac{1}{2\sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}} \frac{\partial}{\partial x_1} \left((x_1 - x_2)^T K^{-1} (x_1 - x_2) \right) \\ &= \frac{2K^{-1}x_1 - 2K^{-1}x_2}{2\sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}} \\ &= \frac{K^{-1}(x_1 - x_2)}{\sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}} \end{aligned}$$

and

$$\frac{\partial \sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}}{\partial x_2} = \frac{K^{-1}(x_2 - x_1)}{\sqrt{(x_1 - x_2)^T K^{-1} (x_1 - x_2)}}$$

Suppose $K = UU^T$ and $V := (U^{-1})^T$, then

$$K^{-1} = (U^T)^{-1} U^{-1} = (U^{-1})^T U^{-1} = VV^T$$

Therefore, when f is the variance norm of a pair (x_1, x_2), the Lipschitz constant k equals to:

$$
\begin{aligned}
&\sup_{x_1,x_2\in\mathbb{R}^d} \| \bigtriangledown f(x_1,x_2)\|_2 \\
&= \sup_{x_1,x_2\in\mathbb{R}^d} \sqrt{\left\| \frac{\partial\sqrt{(x_1-x_2)^TK^{-1}(x_1-x_2)}}{\partial x_1}\right\|_2^2 + \left\| \frac{\partial\sqrt{(x_1-x_2)^TK^{-1}(x_1-x_2)}}{\partial x_2}\right\|_2^2} \\
&= \sup_{x_1,x_2\in\mathbb{R}^d} \sqrt{\left\| \frac{K^{-1}(x_1-x_2)}{\sqrt{(x_1-x_2)^TK^{-1}(x_1-x_2)}}\right\|_2^2 + \left\| \frac{K^{-1}(x_2-x_1)}{\sqrt{(x_1-x_2)^TK^{-1}(x_1-x_2)}}\right\|_2^2} \\
&= \sup_{x_1,x_2\in\mathbb{R}^d} \sqrt{2\left\| \frac{K^{-1}(x_1-x_2)}{\sqrt{(x_1-x_2)^TK^{-1}(x_1-x_2)}}\right\|_2^2} \\
&= \sup_{x_1,x_2\in\mathbb{R}^d} \sqrt{2\left\| \frac{VV^T(x_1-x_2)}{\sqrt{(x_1-x_2)^TVV^T(x_1-x_2)}}\right\|_2^2} \\
&= \sup_{x_1,x_2\in\mathbb{R}^d} \sqrt{2\left\| V\frac{V^T(x_1-x_2)}{\|V^T(x_1-x_2)\|_2}\right\|_2^2} \\
&= \sup_{x_1,x_2\in\mathbb{R}^d} \sqrt{2\|V\|_2^2\left\| \frac{V^T(x_1-x_2)}{\|V^T(x_1-x_2)\|_2}\right\|_2^2} \\
&\leq \sqrt{2}\|V\|_2
\end{aligned}
$$

Therefore, we have that the variance norm of a pair $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$ is a Lipschitz continuity function where the $k = \sqrt{2}\|V\|_2$ □

2.1.2 The Variance Norm for the Signature

Now we are ready to apply the variance norm on the signature of streamed data. Let $\mathcal{C} \subset \mathbb{R}^d$ be a finite corpus of streams of data, $X \in \mathcal{C}$ is an element with dimension $\mathbb{R}^d$. Let $\mathbb{S}^N$ be the signature of level $N \in \mathbb{N}$. Define $\|\cdot\|_\mu$ as the variance norm associated with the centred empirical measure μ of $\{\mathbb{S}^N(X) : X \in \mathcal{C}\}$. With the definition of variance norm equation (2.1.4), we have that for any $w \in \mathbb{S}^N(X)$

$$
\begin{aligned}
\|w\|_\mu^2 &= \sup_{cov(\varphi,\varphi)\leq 1} \varphi(w)^2 \\
&= \sup_{\varphi\in\mathbb{R}^{d_N}\setminus\{0\}} \frac{\varphi(w)^2}{cov(\varphi,\varphi)}
\end{aligned}
\tag{2.1.17}
$$

Based on equation (2.1.7), the variance norm could be written as the Mahalanobis distance:

$$\|w\|_\mu = \sqrt{w^TK^{-1}w}$$

or

$$\|w\|_\mu^2 = \langle w, K^{-1}w\rangle$$

where K is the empirical covariance matrix $K = \mathbb{E}^\mu[w^Tw]$. By the shuffle product identity, the product of terms of signature could be express as the sum of higher order signature terms. Therefore, for terms with multi-index I, J, $w_i = \mathbb{S}^I(X)$ and $w_j = \mathbb{S}^J(X)$ in signature w,

$$K_{I,J} = E^\mu[\mathbb{S}^I(X)\mathbb{S}^J(X)] = \mathbb{E}^\mu[\sum_{M\in I\sqcup\sqcup J} \mathbb{S}^M(X)]$$

or Cochrane et al. (2020) write it as

$$K_{i,j} := \langle e_i \sqcup\sqcup e_j, \mathbb{E}^\mu\left[\mathbb{S}^{2N}(X)\right]\rangle$$

for $i, j = 1, \ldots, d_N$.

2.1.3 The Definition of Conformance

Now we are ready to use the variance norm defined above to measure the distance between groups. As we mainly work on the $\mathbb{R}^d$ space, the Mahalanobis distance is used express the new measure *conformance*.

Definition 2.1.7 (Conformance)**.** Let μ be a probability measure on a vector space $\mathbb{R}^d$. Define the conformance of a vector x to μ as function $f(x;\mu) : \mathbb{R}^d \to \mathbb{R}$

$$f(x;\mu) := \inf_{y \in supp(\mu)} \|x-y\|_\mu = \inf_{y \in supp(\mu)} \sup_{v, v^T K v \leq 1} v^T(x-y) = \inf_{y \in supp(\mu)} \sqrt{(x-y)^T K^{-1}(x-y)} \tag{2.1.18}$$

Where $K = \mathbb{E}^\mu[(x-y)^T(x-y)]$

2.2 Determine the Threshold for Anomaly Behaviours

In this part, we will be using the Gaussian concentration inequality theorems to find the largest acceptable conformance distance for the data within the same Gaussian distribution group.

There are different versions of Gaussian concentration inequality and here we demonstrate the definition given by Boucheron et al. (2013):

Definition 2.2.1 (Gaussian concentration inequality)**.** Let $X = (X_1, \cdots, X_d)$ be a vector of d independent standard normal random variables. Let $f : \mathbb{R}^d \to \mathbb{R}$ denote an L-Lipschitz function. Then, for all $t > 0$,

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) \geq t\} \leq e^{-t^2/(2L^2)} \tag{2.2.1}$$

2.2.1 Identify the Conformance Threshold

Corpus $\mathcal{C}$ contain vectors of data $x_i \in \mathbb{R}^d$ where $i = 1, \cdots, n$, we are interested in studying how large the conformance distance r need to be so that for a vector $x \in \mathbb{R}^d$ that independent from the corpus $\mathcal{C}$, the probability that the x's conformance to corpus $\mathcal{C}$ is greater than a *error bound* ϵ.

We could express this statement in the formula (2.2.2):

$$\mathbb{P}\{\inf_{x_i \in \mathcal{C}} \|x - x_i\|_\mu < r\} \geq \epsilon \tag{2.2.2}$$

In the proof of the *Johnson–Lindenstrauss lemma*, Boucheron et al. (2013) introduce the use of Gaussian concentration inequality. The proving steps inspired us of how to find the boundary of r in our problem, so we will first look into the Johnson–Lindenstrauss lemma and its proof [2].

Lemma 2.2.2. *Let A be a infinite subset of $\mathbb{R}^D$ with cardinality n. Assume that for some $v > 1$, $X_{i,j} \in \mathcal{G}(v)$ [3] and let ϵ,$\sigma \in (0,1)$, if $d > 100v^2\epsilon^{-2}log(n/\sqrt{\sigma})$, then the linear mapping W from $\mathbb{R}^D \to \mathbb{R}^d$ with:*

$$W_i(\alpha) = \sum_{j=1}^{D} \alpha_j X_{i,j} \tag{2.2.3}$$

is a ϵ-isometry on A which is for any $a, a' \in A$:

$$(1-\epsilon)\|a-a'\|^2 \leq \|f(a) - f(a')\|^2 \leq (1+\epsilon)\|a-a'\|^2$$

According to the Boucheron et al. (2013) in section 2.9, W is a ϵ-isometry on A if and only if the random variable $\sup_{\alpha \in T} |\|W(\alpha)\|^2 - 1|$ is highly concentrated around its mean. This could be proved using Theorem 2.2.3:

[2] The complete proof of the Johnson–Lindenstrauss lemma is given in appendix A, we only explain the relevant proving part in this section

[3] $X_{i,j}$ follow the sub-Gussian disturbution which is proposed by Buldygin and Kozachenko (1980)

Theorem 2.2.3. *For $M = d\sup_{\alpha\in T}\|W(\alpha)\|^2$ and all $t > 0$:*

$$\mathbb{P}\{M - \mathbb{E}M \geq 2\sqrt{2t\mathbb{E}M} + 2t\} \leq e^{-t} \tag{2.2.4}$$

where

$$T = \{\frac{a-a'}{\|a-a'\|}, (a,a') \in A \times A \text{ and } a \neq a'\}$$

Proof. We first assume the T is a finite set. Define function $f : \mathbb{R}^{dD} \to \mathbb{R}$

$$f(x) = \sup_{\alpha\in T} \Sigma_{i=1}^{d}(\Sigma_{j=1}^{D}\alpha_j x_{i,j})^2 \tag{2.2.5}$$

Then $M = f(x)$. The equation (2.2.5) could be seen as maximizing a linear function over a ball. We could deduce that $\sqrt{f}$ is 1-Lipschitz function, adopting the definition 2.2.1, we have for any $t > 0$:

$$\begin{aligned}\mathbb{P}\{M \geq \mathbb{E}M + 2\sqrt{2t\mathbb{E}M} + 2t\} &\leq \mathbb{P}\{M \geq (\mathbb{E}\sqrt{M} + \sqrt{2t})^2\} \\ &\leq \mathbb{P}\{\sqrt{M} - \mathbb{E}\sqrt{M} \geq \sqrt{2t}\} \\ &\leq e^{-t}\end{aligned}$$

□

Similar approach could be adopted to find the boundary of r satisfy equation (2.2.2) if we replace the infimum in the conformance by the sum of all variance norm of x to x_i. Note that for

$$\{r : r \in A\} \subset \{r : r \in A'\}$$

we have

$$\mathbb{P}(r \in A) \leq \mathbb{P}(r \in A')$$

Therefore, since

$$\{r : \sum_{x_i\in\mathcal{C}} \sqrt{(x-x_i)^T K^{-1}(x-x_i)} < n \times r\} \subset \{r : \inf_{x_i\in\mathcal{C}} \sqrt{(x-x_i)^T K^{-1}(x-x_i)} < r\} \tag{2.2.6}$$

we could further write the equation (2.2.2) as

$$\begin{aligned}\mathbb{P}\{\inf_{x_i\in\mathcal{C}} \|x-x_i\|_\mu < r\} &= \mathbb{P}\{\inf_{x_i\in\mathcal{C}} \sqrt{(x-x_i)^T K^{-1}(x-x_i)} < r\} \\ &\geq \mathbb{P}\{\sum_{x_i\in\mathcal{C}} \sqrt{(x-x_i)^T K^{-1}(x-x_i)} < n \times r\}\end{aligned} \tag{2.2.7}$$

and $K = \mathbb{E}[(x-x_i)^T(x-x_i)]$. We have proved that the variance norm is a k-Lipschitz function, and as the Lipschitz continuity preserve over sum, the Lipschitz constant for the sum of variance norm:

$$\sum_{x_i\in\mathcal{C}} \sqrt{(x-x_i)^T K^{-1}(x-x_i)}$$

is

$$nk = n\sqrt{2}\|V\|_2 = \sqrt{2}n\sigma_{max}(V)$$

where $\sigma_{max}(V)$ is the largest singular value of $V = (U^{-1})^T$ and U is the lower triangular matrix where $K = UU^T$

The expectation of the sum of variance norm is:

$$n\alpha := \mathbb{E}(\sum_{x_i\in\mathcal{C}} \sqrt{(x-x_i)^T K^{-1}(x-x_i)}) = n\mathbb{E}(\sqrt{(x-x_i)^T K^{-1}(x-x_i)})$$

then by Gaussian concentration inequality:

$$\begin{aligned}
&\mathbb{P}\{\sum_{x_i\in\mathcal{C}}\sqrt{(x-x_i)^TK^{-1}(x-x_i)}-n\alpha<n\times r-n\alpha\}\\
&=1-\mathbb{P}\{\sum_{x_i\in\mathcal{C}}\sqrt{(x-x_i)^TK^{-1}(x-x_i)}-\alpha\geq n\times r-\alpha\}\\
&\geq 1-e^{-\frac{(r-\alpha)^2}{2k^2}}\geq\epsilon
\end{aligned}$$

Solving the equation we get [4], if r satisfy

$$r\geq\alpha+k\sqrt{2\ln\frac{1}{1-\epsilon}} \tag{2.2.8}$$

then it satisfy equation (2.2.2)

2.2.2 The Conformance Threshold for Gaussian Variables

If we further suppose that x_i, $i=\{1,\cdots,n\}$ follow i.i.d centred Gaussian distribution $\mathcal{N}(0,\sigma^2\mathbb{I}_d)$ and x is a (fixed) Gaussian sample that independent from x_i. Then

$$x-x_i\sim\mathcal{N}(0,\sigma^2\mathbb{I}_d) \tag{2.2.9}$$

Therefore, $K=\sigma^2\mathbb{I}_d$ and $U=\sigma\mathbb{I}_d$ so the Lipschitz constant for the variance norm is:

$$k=\sqrt{2}\|V\|_2=\frac{\sqrt{2}}{\sigma} \tag{2.2.10}$$

and we have

$$Y=(x-x_i)^TU^T\sim\mathcal{N}(0,\mathbb{I}_d) \tag{2.2.11}$$

therefore,

$$(x-x_i)^TK^{-1}(x-x_i)=YY^T\sim\mathcal{X}_d^2 \tag{2.2.12}$$

and the square root of the chi-square distribution is chi-distribution, so the expectation for the variance norm is the expectation for a chi-distribution random variable, then we have

$$\alpha=\mathbb{E}[\sqrt{(x-x_i)^TK^{-1}(x-x_i)}]=\frac{\sqrt{2}\Gamma(\frac{1}{2}(d+1))}{\Gamma(\frac{1}{2}d)}$$

Then,

$$r\geq\frac{\sqrt{2}\Gamma(\frac{1}{2}(d+1))}{\Gamma(\frac{1}{2}d)}+k\sqrt{2\ln\frac{1}{1-\epsilon}} \tag{2.2.13}$$

Therefore, we deduce the lower bound of the conformance distant for identifying the anomalies, we denote it as

$$q_\epsilon=\frac{\sqrt{2}\Gamma(\frac{1}{2}(d+1))}{\Gamma(\frac{1}{2}d)}+k\sqrt{2\ln\frac{1}{1-\epsilon}} \tag{2.2.14}$$

If the conformance distance r of an independent Gaussian vector in $\mathbb{R}^d$ to a Gaussian corpus $\mathcal{C}$ is smaller than this threshold q_ϵ, we regard it to be a normal element in that group (share same distribution), otherwise we define it as an anomaly. In the next section, we will performing some empirical test and propose how to utilize this result in anomaly detection tasks.

[4] Two solutions could be deduce through the function but we only keep the one that gives us the lower bound of r

2.2.3 Empirical Test of the Conformance Threshold

We will test whether the empirical simulation could deduce the same threshold distance as q_ϵ in this section.

To approximate the probability of the conformance distance of one Gaussian sample $x \in \mathbb{R}^d$ to a group of Gaussian samples. We first use the conformance distance of 100 Gaussian samples to a group of Gaussian samples to construct the empirical distribution. This process is repeated ten times to eliminate the randomness.

Then the average of the ϵ quantile $\hat{q}_\epsilon$ for each empirical distribution (correspond to the theoretical threshold with error bound ϵ) is calculated to compare with the theoretical result q_ϵ. The empirical and theoretical results are compared for $d = 2, 3, 4, 5$, $n = [500, 2000]$ and $\epsilon = 0.9$. Part of the results are shown in the chart 2.1:

dimension	threshold type	$n = 500$	$n = 600$	$n = 700$	$n = 800$	$n = 900$	$n = 1000$
2 dim	empirical	2.227	2.474	2.181	2.146	2.187	2.124
	modified theoretical	4.288					
3 dim	empirical	2.701	2.594	2.758	2.631	2.839	2.664
	modified theoretical	4.631					
4 dim	empirical	3.107	3.285	3.299	3.042	3.204	3.079
	modified theoretical	4.915					
5 dim	empirical	3.594	3.446	3.411	3.337	3.516	3.626
	modified theoretical	5.163					

Table 2.1: Comparison of theoretical and empirical conformance threshold for Gaussian sample

We could see a discrepancy between the empirical and theoretical threshold for data with each dimension. Several potential sources could induce this difference:

Firstly, there are several approximation steps during the proof (like equation (2.2.7)), which made the theoretical threshold higher than the empirical one. We are confident that if a Gaussian sample's conformance value is larger than the theoretical threshold, it must be an anomaly. However, the actual boundary does not need to be that high as we could find a more precise one through the empirical quantile of the data sample. Secondly, the numerical approximation for the empirical test, like the inverse calculation of the cor-variance matrix, may not be accurate enough. Therefore, we need to conduct more detailed research to find the exact source of this discrepancy in the future.

Although the discrepancy exists, we still find that the dependence of empirical conformance threshold $\hat{q}_\epsilon$ on dimension d and n is the same as the theoretical result in equation (2.2.14). By plotting the empirical results in figure 2.1, we could see for both empirical and theoretical conformance thresholds:

- As dimension d increased, the conformance threshold q_ϵ increased
- For Gaussian data with fixed dimension d, the change of sample size n has negligible influence on the value of conformance threshold q_ϵ, as it is always around a certain value.

Therefore, we could identify the conformance threshold for the sample data with certain characteristics despite the sample size, which means that it is feasible to utilize this method to distinguish the anomaly instances from the normal corpus.

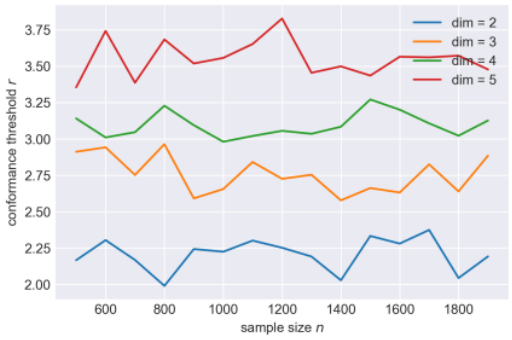


Figure 2.1: The empirical conformance threshold for Gaussian sample with different dimensions

Based on this result, we are ready to introduce the conformance algorithm for anomaly detection tasks. Cochrane et al. (2020) proposed that we could first split the corpus into two halves then identify the suitable conformance threshold for the data.

Let $\mathcal{I} \subset \mathbb{R}^d$ be a finite corpus of Gaussian data with size $2n$. If we split the corpus randomly into two equal-size halves and denote the two parts by $\mathcal{I}_1, \mathcal{I}_2$. For each random point $x \in \mathcal{I}_1$, we could calculate its conformance distance r to subset $\mathcal{I}_2$. Then based on the proportion that the data has been contaminated, we could define the suitable threshold to identify the anomaly instances by changing the error bound ϵ. The detailed algorithm will be introduced in section 2.4.

2.3 The Evaluation Method for Anomaly Detection Algorithms

After proposing the outline of how to using the conformance threshold to identify anomaly instances, we also need to discuss how to evaluate the performance of this method. As a special classification problem, the anomaly points usually have a low occurrence, so it is not sufficient to evaluate the method only by the overall accuracy.

For example, in fraud detection tasks for credit card data, the focus is on correctly identifying the anomaly transaction instead of correctly identifying the standard transaction, as missing any fraud transaction could result in significant loss for the financial institution. Therefore, more evaluation calculations need to be considered to evaluate the anomaly detection algorithm's performance.

The *Confusion matrix* as a special type of 2×2 evaluation matrix is designed to represent the prediction result against the actual label. The value in each position of the matrix is listed in figure 2.2.

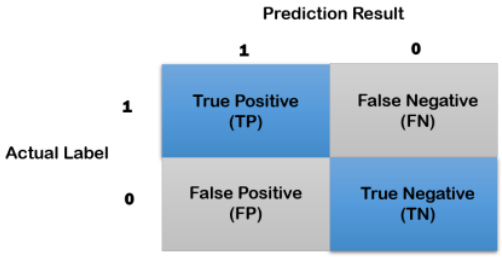


Figure 2.2: Confusion matrix

The *True Positive* (TR) represents the model correctly identify the anomaly points as anomalies. The *False positive* (FP) represents the model wrongly predict the normal points as anomaly points. The *False negative* (FN) represents when the model identifying the anomaly points as the normal points while the *True negative* (TN) is an outcome where the model identify the normal points correctly.

Several quantities that used the value in the confusion matrix are proposed to evaluate the performance of the model:

Sensitivity or *True positive rate* (TPR) is the proportion that labelled anomaly data is test anomaly:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.3.1}$$

Specificity or *True negative rate* (TNR) is the proportion that normal data are tested as normal:

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2.3.2}$$

Accuracy (ACC) is the proportion of correct predictions among all data:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{2.3.3}$$

If we changed the conformance threshold and let the true positive increase, the true negative rate would decrease. Therefore, there is a trade-off between the two values, and we hope to find the most suitable threshold that yields the best result.

This trade-off could be visualized by the *Receiver Operating Characteristic* (ROC) curve, which measures the performance of the model at various threshold settings. It is created by assigning the true positive rate on the x-axis and the false positive rate on the y-axis. The *area under the curve* (AUC) measure the ability of the algorithm to identify the anomaly as the higher AUC, the better the performance of the model.

In Cochrane et al. (2020)'s work, the ROC-AUC is plotted for each conformance anomaly detection task (using the conformance distance as the scores) to compare the effectiveness of this model to other models. However, in this thesis, we define the conformance threshold q_ϵ through the error bound ϵ, and we are interested in finding the suitable ϵ that yields the best result. Therefore, we will directly compare the sensitivity, specificity and accuracy results for different ϵ instead of plotting the ROC-AUC curve.

2.4 The Conformance Anomaly Detection Algorithm

In this section, we summarise the conformance method for unsupervised anomaly detection problems. From any given stream of data, we would first split it to train and test set where the train set is used to define the conformance threshold, and the testing set is used to validate the effectiveness of this detection algorithm. We assign the label for normal corpus to be $y = 0$ and anomaly as $y = 1$.

Algorithm 1 Identify the conformance threshold for the anomaly

Input: training data X_{train}, group index y_{train}
the transform method, signature levels k, error bounds ϵ

Output: the anomaly threshold q_ϵ, predicted label for part of the training data $\hat{y}^1_{train}$,
the evaluation parameters

begin

1. Applying certain transformation to the training data X_{train} and calculate its signature as $S(X_{train})$ with level k. Then centred the $S(X_{train})$ by applying Min-Max normalisation
2. Randomly split the $S(X_{train})$ and its correspond label y_{train} into two equal size halves $S(X^1_{train})$ and $S(X^2_{train})$, y^1_{train} and y^2_{train}
3. Find the conformance distance of each elements x in $S(X^1_{train})$ to subset $S(X^2_{train})$. Specifically, calculate the variance norm (Mahalanobis distance) of x to each element in $S(X^2_{train})$ and then find the smallest result $f(x, S(X^2_{train}))$ to be the conformance value
4. Look at the right tail of the empirical distribution of the conformance value. For a given error threshold ϵ, find the conformance value correspond to the ϵ quantile of the empirical distribution and set it as the conformance threshold q_ϵ to identify the anomaly behaviour
5. Predict label $\hat{y}^1_{train}$ for elements in X^1_{train}. If the conformance of x to $S(X^2_{train})$ larger than threshold q_ϵ, we define it as the anomaly data
6. Check the accuracy of the algorithm for training data by calculating the sensitivity, specificity, and accuracy using the predicted label $\hat{y}^1_{train}$ and original label y^1_{train}

end

In the next step, the *conformance threshold* q_ϵ is utilized to identify the anomaly streams in test data X_{train} and check the performance of the method. Note that since we proved in section 2.2.2 that the sample size have negligible influence on the conformance threshold, the corpus size N_{train} of testing data could be different from the N_{test}.

Algorithm 2 Identify anomaly data using the conformance threshold

Input: Part of the training data X^2_{train}, testing data X_{test} and the correspond group index y_{train}
the anomaly threshold q_ϵ

Output: The predict label for testing data $\hat{y}_{test}$, the evaluation parameters

begin

1. Applying same transformation as the training data on testing data X_{test}, calculating its signature as $S(X_{test})$ and applying the Min-Max normalisation to $S(X_{test})$.
2. Calculating the conformance distance of each element in $S(X_{test})$ to $S(X^2_{train})$ similar to the training process
3. For elements x' in $S(X_{test})$ that the conformance score $f(x', S(X^2_{train}))$ is larger than the anomaly threshold q_ϵ, we predict it to be an anomaly data and assign its label as $\hat{y}_{test} = 1$. The rest is labeled as $\hat{y}_{test} = 0$
4. Comparing the accuracy of the algorithm for testing data by calculating the sensitivity, specificity and accuracy using the predicted label $\hat{y}_{test}$ and original label y_{test}.

end

The algorithms 1 and 2 will be implemented in Chapter 3 on streamed data to study how inputs in terms of the transform method, signature levels and error bound ϵ influence the performance of the model. Then, the most suitable input parameters combination will be selected to detect the anomaly instances on the financial market data in Chapter 4.

Chapter 3

Evaluation on Brownian Motion Data

In the Chapter 2, we demonstrated the proof of using the conformance threshold to identify anomalies among Gaussian data and then proposed a conformance anomaly detection algorithm based on that theory. In this Chapter we will continue to test this method on anomaly detection tasks for the streamed data.[1]

In the financial field, the log-normal moneyness of option price is usually model by Brownian motion, a stochastic process with Gaussian property. Therefore, the simulated Brownian motion data would be an ideal example to test the conformance algorithm. By generating a group of standard one-dimensional Brownian Motion paths, we find that this method could identify paths with relatively extreme behaviour, namely the path with larger drift or variance compared to the main corpus.

Then, we move on to study the performance of this conformance method. As states in the section 2.4, the results could be affected by various underlying inputs parameters, in terms of the corpus size n and dimension d, the transform method applied on the original data and the error bound ϵ. We hope to understand how the conformance threshold and prediction performance depend on those variables and suggest the most effective input parameters for different data groups.

Specifically, we generate 1-dimensional, 2-dimensional, and 4-dimensional data contaminated by a small proportion of anomaly paths (either by drift or variance), partitioning it to obtain the training and testing corpus, and applied the conformance anomaly detection algorithm on those data. Following the evaluation way mentioned in section 2.3, the sensitivity, specificity, and accuracy are then calculated using the predicted label and the original label. As we focus on the anomaly detection task, the sensitivity that represents how many anomaly instances have been correctly identified should be high, so we will first consider this parameter during the evaluation.

For one-dimensional Brownian motion, we focus on studying how the error bound ϵ and the transform method influence the performance of the conformance model. The Isolation forest model, one of the most popular unsupervised learning algorithms, is also applied to the one-dimensional Brownian motion data as a comparison. We then selected the most suitable inputs based on the evaluation result and applied them to the two and four-dimensional Brownian motion anomaly detection tasks. We also tried different training and testing size to check whether sample size n influence the performance of the model. The evaluation result shows that as long as we select the suitable underlying parameters, this algorithm could effectively identify the anomalies among Brownian motion data with different sizes and dimensions.

[1]The experiments are conducted on a Surface Pro 6 equipped with Intel Core i5-8250U CPU and 8 GB RAM.

3.1 Anomaly Detection on One-Dimensional Brownian Motion Data

We first give the definition of the standard Brownian motion:

Definition 3.1.1 (Standard Brownian motion)**.** A standard Brownian motion W_t also known as the Wiener process, is a stochastic process that has following properties:

- $W_0 = 0$
- The increments of W is independent. For any $t > 0$ and $u \geq 0$, $W_{t+u} - W_t$ are independent from the W_s where $s \leq t$
- The increment of W is Gaussian: $W_{t+u} - W_t \sim \mathcal{N}(0, u)$
- W_t is continuous in t

We could see that the Brownian motion is closely related to the Gaussian distribution. Therefore, as we proved the feasibility of identifying the conformance threshold for the Gaussian sample in section 2.2.2, the Brownian motion sample would be a suitable example to test our conformance anomaly detection algorithm.

3.1.1 Identify the Anomaly Paths

We first generate the one dimensional standard Brownian motion sample to test our conformance method. For each Brownian motion path, suppose the time span $t \in [0, 1]$, we equally divide it into $t_1, \cdots, t_{100}$. Follow the definition 3.1.1, the sample W_t for each time stamp t_i could be generated through equation:

$$W_{t_i} = W_{t_{i-1}} + Z \tag{3.1.1}$$

where $Z \sim \mathcal{N}(0, t_i - t_{i-1})$

We repeat this process to generate $N_{\text{train}} = 1000$ sample paths in total (see figure 3.1).

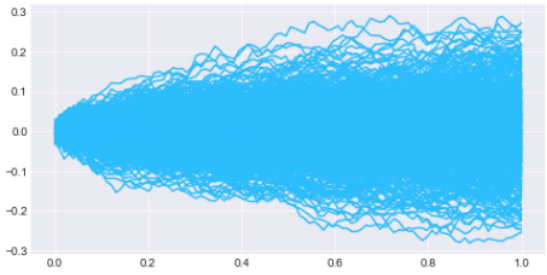
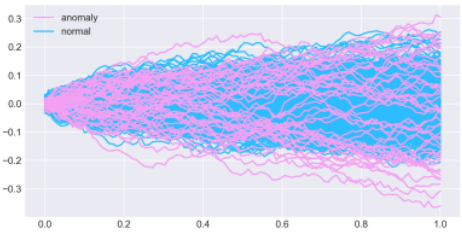


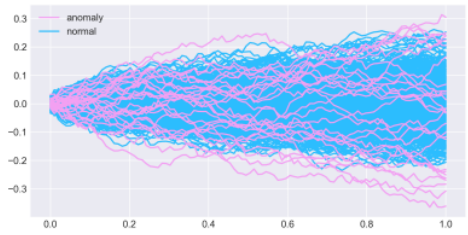
Figure 3.1: Simulation of 1000 standard Brownian motion paths

Some transformation methods mentioned in section 1.4 are first adopted on the data before the signature calculation steps. Then following the algorithm 1 in section 2.4, we calculated the threshold and labeled the anomaly paths. As the anomaly paths usually take up a few amounts of the total population, we set the error bounds to be $\epsilon = 0.9, 0.95$ to approximate the actual scenario.

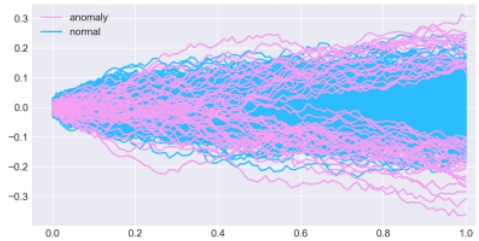
We calculate the results for signature level $k \in \{1, \cdots, 5\}$. As the results for different signature levels are similar, we only present the result of $k = 4$ in figure 3.2 for illustration. The pink paths identified anomaly paths and the normal instances are in blue.



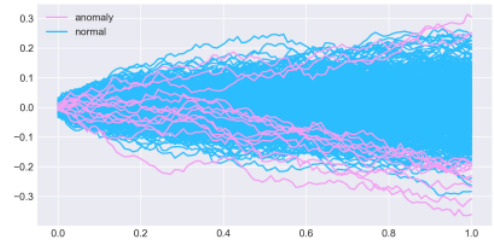
$\epsilon = 0.9$ with time transform



$\epsilon = 0.95$ with time transform



$\epsilon = 0.9$ with Invisible transform



$\epsilon = 0.95$ with Invisible transform

Figure 3.2: One-dimensional Brownian motion anomaly detection results

We could conclude that the identified anomaly paths have two general behaviours:

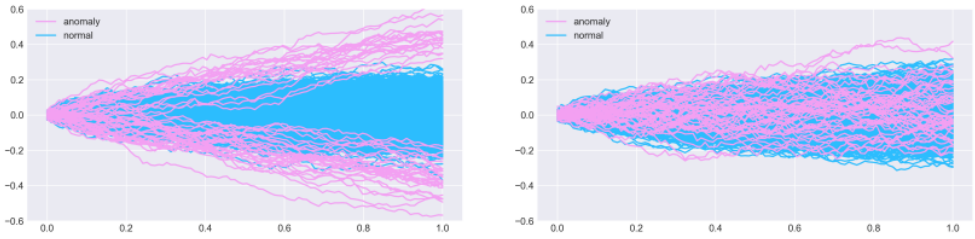
- The anomaly paths fluctuated more frequently than the normal ones so their variances are larger than the general corpus.
- The anomaly paths are generally spread outside the main streams and when $\epsilon = 0.95$, the final points are usually local outside the $(-0.2, 0.2)$.

Therefore, we could generate the anomaly path by relaxing some restrictions on the parameters of the standard Brownian motion.

Definition 3.1.2 (Brownian motion with drift)**.** A Brownian motion with drift parameter $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$ is a stochastic process that has following properties:

- $W_0 = 0$
- Independent increments: For any $t > 0$ and $u \leq 0$, $W_{t+u} - W_t$ are independent from the W_s where $s < t$
- The increment of W follow Gaussian distribution: $W_{t+u} - W_t \sim \mathcal{N}(\mu, \sigma^2 u)$
- W_t is continuous in t

Follow the conclusion and definition 3.1.2, we set the drift for the anomaly paths to be $\mu_a = 0.4$ and variance to be $\sigma_a^2 = 1.5$ while the normal corpus will still be the standard Brownian motion with $\mu_n = 0$ and $\sigma_n^2 = 1$. The *contamination rate* (the proportion of anomalies in total) will maintain low which is around 0.05. The generated Brownian paths are shown in figure 3.3.



Drift contamination Variance contamination

Figure 3.3: Contaminated one-dimensional Brownian paths

In the next section, we will use this contaminated data to evaluate the performance of our conformance anomaly detection algorithm.

3.1.2 One-Dimensional Brownian Motion with Contamination

We first define the training data size to be $N_{\text{train}} = 1000$. Since in the conformance anomaly detection algorithm, the conformance threshold is defined through the conformance distance of half of the whole dataset. Therefore, we set the test size to be half of the training set, which is $N_{\text{test}} = 500$. Then by algorithms 1 and 2 in section 2.4, the anomaly threshold is identified and the performance of the model in terms of sensitivity, specificity and overall accuracy are calculated for train and test data.

The input parameters we will be tested are:

- signature levels $k \in \{1, 2, 3, 4, 5\}$
- error bounds $\epsilon \in \{0.85, 0.9, 0.95, 0.99\}$
- transform method: add time transform and lead-lag transform

However, for $\epsilon \in \{0.85, 0.99\}$, the evaluation results in terms of the true positive rate are all below 0.6 which are too lower to be accepted, therefore, we will only demonstrate the results for $\epsilon \in \{0.9, 0.95\}$.

One famous anomaly detection algorithm, Isolation forest, is applied here as a benchmark method to compare with the performance of our algorithm.

A Brief Introduction of the Isolation Forest

The Isolation forest algorithm is an unsupervised anomaly detection method that directly targets anomalies. Based on the idea that anomaly instances usually have extreme behaviour, it is easier to isolate them from the normal corpus by constructing decision trees.

For example, we could define a decision tree to identify the anomaly paths among the data contaminated by larger drift and variance. Suppose the normal data are the path with variance smaller than 1.5 and drift smaller than 0.4. The X, Y, Z represent one-dimensional Brownian path where the drift and variances are: $\mu_X = 0.5$, $\mu_Y = 0.2$, $\mu_Z = 0.2$, $\sigma_Y = 2$, $\sigma_Z = 1$

The decision tree in figure 3.4 demonstrate how the Isolation forest identify the anomaly path. We could see that the anomaly data should be found close to the root of the tree which means it is more easily to be separate from the norm corpus.

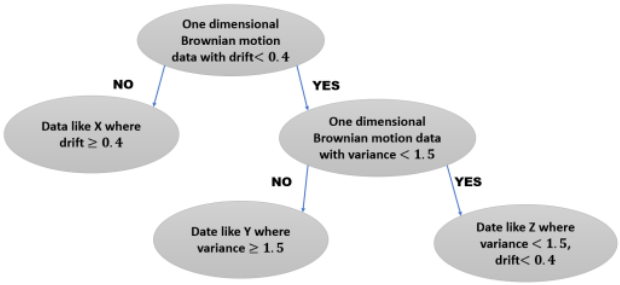


Figure 3.4: Isolation forest example

Therefore, this method is suitable for detecting the anomaly Brownian motion paths with higher variance and drift. By implementing the method directly on the train and test Brownian motion data, we could get the predicted label of the train and test set. Then we calculated the sensitivity (true positive rate), specificity (true negative rate), accuracy of the test set and compared it with the evaluation results of the conformance method.

Evaluation Result for Data Contaminated by the Drift

Suppose the drift for normal sample paths is $\mu_n = 0$, define the contamination percentage to be 5% of the total numbers of the corpus and let the anomaly's drift to be $\mu_a = 0.4$. The evaluation results for the testing group regrading the sensitivity and specificity for different underlying parameters are shown in figure 3.5 and 3.6, and the overall accuracy is shown in the table 3.1

Signature level	1	2	3	4	5
$\epsilon = 0.9$; Add-time Transform	0.92	0.92	0.92	0.93	0.95
$\epsilon = 0.95$; Add-time Transform	0.97	0.97	0.97	0.97	0.97
$\epsilon = 0.9$; Lead-lag Transform	0.94	0.94	0.95	0.95	0.97
$\epsilon = 0.95$; Lead-lag Transform	0.97	0.96	0.97	0.97	0.98
Isolation forest	0.94				

Table 3.1: Accuracy for one-dimensional paths contaminated by different drift

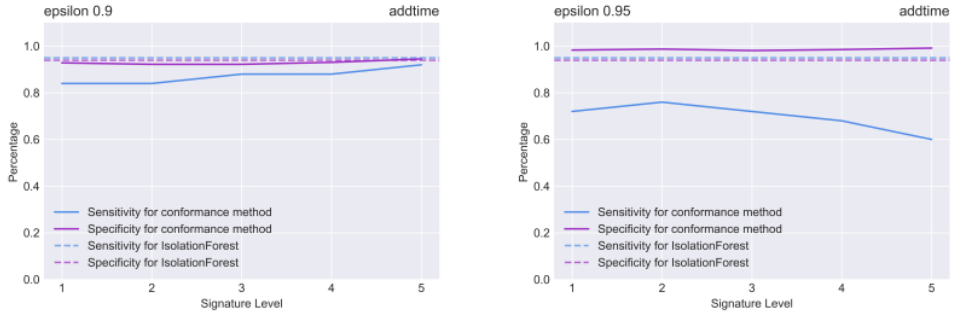


Figure 3.5: Evaluation result of drift contaminated BM paths with add time transform

As the figure 3.5, 3.6 and table 3.1 show, setting the error bound as 0.9 sacrificed a little accuracy in overall accuracy and specificity (the percentage of correctly identify the normal paths) but largely improve the sensitivity (the percentage of correctly identifying the anomaly points). As our goal is to identify the anomaly data from the normal corpus, it is extremely important to increase the sensitivity of the algorithm. Therefore, we conclude that the anomaly detection method has the

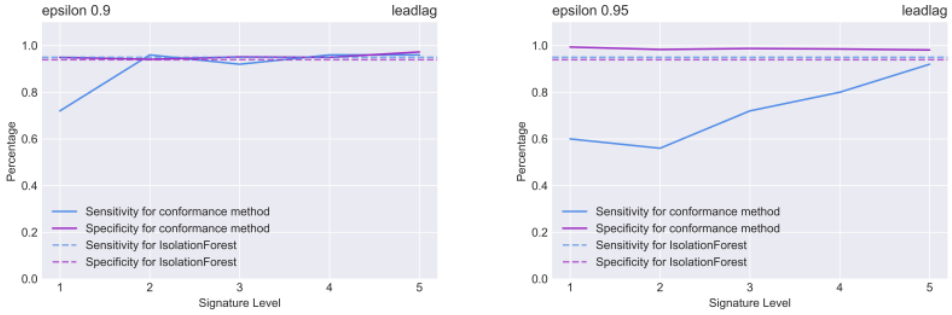


Figure 3.6: Evaluation result of drift contaminated BM paths with lead-lag transform

better performance when choosing $\epsilon = 0.9$.

Comparing the performance of the two transformation methods, we could see the lead-lag transform sightly outperforms the add time transform for signature level larger or equal than 3.

The Isolation Forest method's performance on identifying the anomaly paths is showing by the dashed line. We could see the conformance method's sensitivity and specificity are similar to (slightly better than) the Isolation Forest when data transform by lead-lag and signature level larger than 3. Also, the overall accuracy of the conformance method is higher than the Isolation Forest as demonstrated in the table 3.1. Therefore, we conclude that the conformance algorithm with underlying parameters: error bound $\epsilon = 0.9$, signature level larger than two and lead-lag transform, has better performance in detecting the one dimension Brownian motion paths contaminated by drift compared to the Isolation Forest method.

Evaluation Result for Data Contaminated by the Variance

Suppose the variance for normal sample paths to be $\sigma_n = 1$ and anomaly to be $\sigma_a = 1.5$. The contamination rate is again 5%. The performance results of the model are shown in figure 3.7, 3.8 and table 3.2

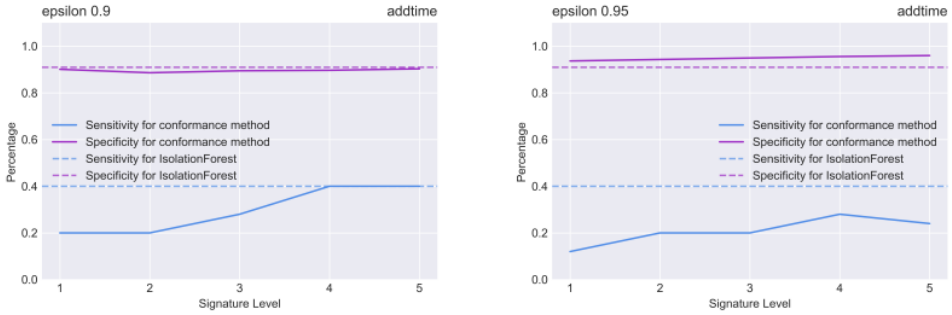


Figure 3.7: Evaluation result of variance contaminated BM paths with time transform

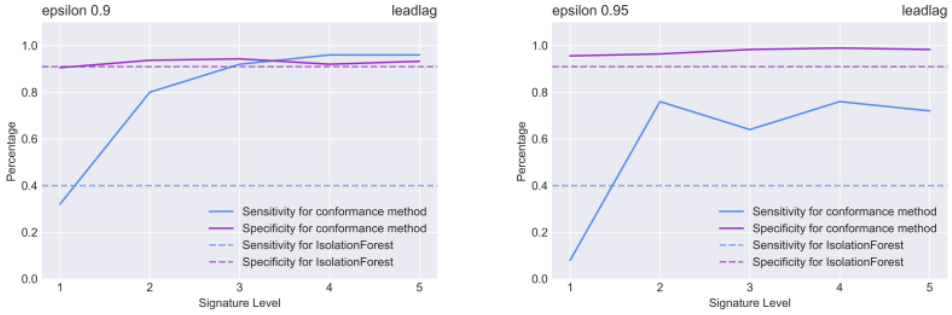


Figure 3.8: Evaluation result of variance contaminated BM paths with lead-lag transform

Signature level	1	2	3	4	5
$\epsilon = 0.9$; Add-time Transform	0.87	0.85	0.86	0.87	0.88
$\epsilon = 0.95$; Add-time Transform	0.90	0.91	0.91	0.92	0.92
$\epsilon = 0.9$; Lead-lag Transform	0.88	0.93	0.94	0.92	0.94
$\epsilon = 0.95$; Lead-lag Transform	0.91	0.95	0.96	0.97	0.97
Isolation forest	0.9				

Table 3.2: Accuracy for one-dimensional paths contaminated by different variance

For the error bound ϵ, we could see that setting it as 0.9 again result in much higher sensitivity on identifying the anomaly paths.

As we introduced in section 1.4, the lead-lag transform could interpret the quadratic variance of the path. The evaluation results demonstrate that this interpretation is preserved in our anomaly detection algorithm as the sensitivity for data with lead-lag transform is significantly better than the add time transform. Particularly, the sensitivity for data with lead-lag transform reaches higher than 90% for all the train and test groups (when signature level larger than 3).

While for the Isolation Forest method, the sensitivity is only 40% and the accuracy is 90% which are all worse than the conformance method with lead-lag transformation.

Conclusion for One Dimensional Brownian Motion data

Based on the above analysis, for data with 5% contamination, setting the error bound ϵ as 0.9 render the best evaluation results compare with other values in $\{0.85, 0.9, 0.95, 0.99\}$. Our conformance anomaly detection method out-performance the Isolation forest when applying the lead-lag transformation and signature level higher than three especially for the data contaminated by different variance. The add time transform is also effective when identifying the anomaly among drift contaminated data, but it is not suitable for identifying the anomaly for data contaminated by different variances.

3.2 Higher Dimensional Brownian Motion Data

As we proved in the section 2.2, this conformance method is effective for data with various dimensions and corpus size. Therefore, we continue to evaluate the model's performance on the 2-dimensional and 4-dimensional Brownian motion data.

As demonstrated in the section 3.1.2, the lead-lag transform and $\epsilon = 0.9$ work effectively for data with 5% contamination. Therefore, we will stick with the above parameters and study whether

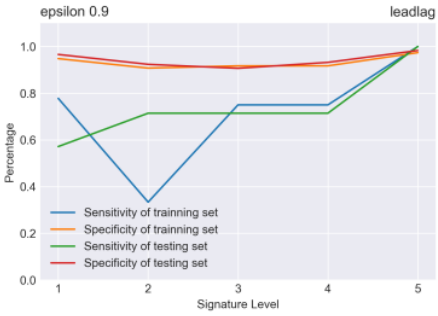
the different corpus sizes would influence the conformance threshold and prediction performance.

3.2.1 Two-Dimensional Brownian Motion with Different Sample Size

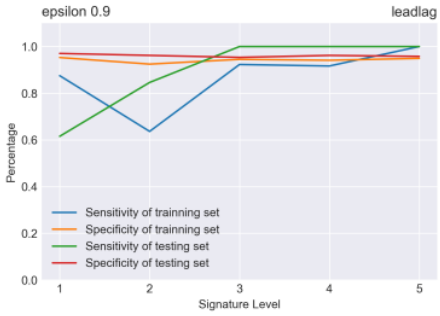
We select four different train and test sample sizes in terms of 250 and 125 ; 500 and 250; 750 and 375 ; 1000 and 250. The sensitivity, specificity of both train and test group are shown in figure 3.9 and the overall accuracy is shown in table 3.3.

	Signature level	1	2	3	4	5
group 1	$N_{\text{train}} = 250$	0.91	0.91	0.93	0.94	0.94
	$N_{\text{test}} = 125$	0.92	0.94	0.94	0.93	0.94
group 2	$N_{\text{train}} = 500$	0.94	0.94	0.94	0.94	0.95
	$N_{\text{test}} = 250$	0.91	0.93	0.93	0.94	0.94
group 3	$N_{\text{train}} = 750$	0.91	0.93	0.93	0.93	0.94
	$N_{\text{test}} = 375$	0.9	0.9	0.92	0.92	0.93
group 4	$N_{\text{train}} = 1000$	0.92	0.94	0.95	0.95	0.95
	$N_{\text{test}} = 500$	0.89	0.90	0.91	0.92	0.95

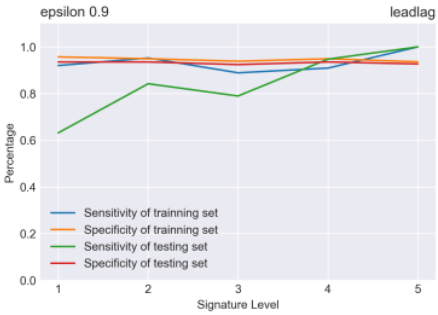
Table 3.3: Accuracy for two-dimensional paths contaminated by different drift



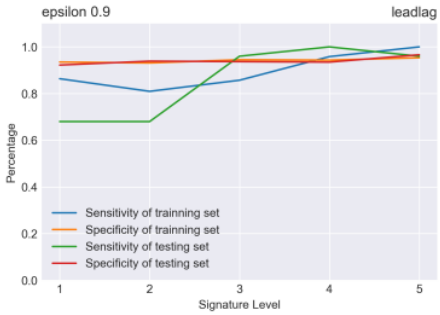
$N_{\text{train}} = 250$ and $N_{\text{test}} = 125$



$N_{\text{train}} = 500$ and $N_{\text{test}} = 250$



$N_{\text{train}} = 750$ and $N_{\text{test}} = 375$



$N_{\text{train}} = 1000$ and $N_{\text{test}} = 500$

Figure 3.9: Evaluation result of drift contaminated two-dim BM paths with different group size

Signature level	1	2	3	4	5
$N_{\text{train}} = 250$; $N_{\text{test}} = 125$	0.42	2.51	8.52	117.26	97.77
$N_{\text{train}} = 500$; $N_{\text{test}} = 250$	0.28	2.13	7.13	32.03	234.63
$N_{\text{train}} = 750$; $N_{\text{test}} = 375$	0.26	2.13	6.80	21.63	657.53
$N_{\text{train}} = 1000$; $N_{\text{test}} = 500$	0.18	2.11	5.63	20.51	157.26

Table 3.4: Conformance thresholds for two-dimensional paths contaminated by different drift

We could see that the specificity of data with two dimensions is similar to one dimension, while the sensitivity and accuracy slightly decreased but still higher than 80% and 92% when signature level equal to 3, 4, 5 and sample size larger than 250. Therefore, we believe the conformance algorithm work effectively in identifying two-dimensional Brownian motion data with extreme drift if sample size larger than 250.

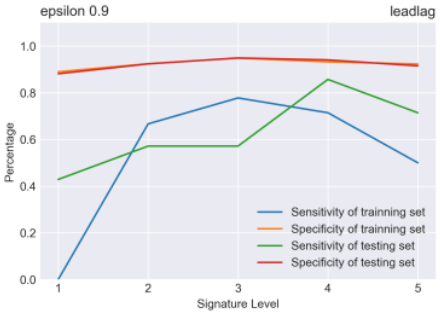
The conformance thresholds result in table 3.4 shows that as the sample size increase, the threshold tends to stabilize at a certain level for signature level smaller than 4. This correspond to the theoretical prove in the section 2.2.2 that the corpus size have little influence on the conformance threshold. The unstable conformance thresholds value for signature equal to 5 shows that we don't need too high signature level to capture the features of the data. Therefore, together with the evaluation results, signature of level 3 and 4 would be suitable for capturing the feature of high dimensional Brownian motion data.

When the sample size is equal to 250, the conformance threshold is slightly larger, which might cause by the randomness of the empirical sample. This further results in low sensitivity and specificity as the figure 3.9 shown. Therefore, the sample size should be larger than 250 to avoid the inaccurate result.

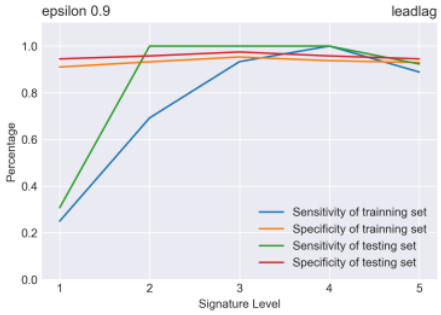
For the two-dimensional Brownian motion data contaminated by different variance, we also test 4 different corpus size and demonstrate the performance result in 3.10 and table 3.5:

	Signature level	1	2	3	4	5
group 1	$N_{\text{train}} = 250$	0.84	0.91	0.93	0.92	0.90
	$N_{\text{test}} = 125$	0.86	0.91	0.93	0.94	0.91
group 2	$N_{\text{train}} = 500$	0.87	0.92	0.95	0.94	0.93
	$N_{\text{test}} = 250$	0.91	0.96	0.98	0.96	0.94
group 3	$N_{\text{train}} = 750$	0.89	0.94	0.94	0.94	0.94
	$N_{\text{test}} = 375$	0.86	0.94	0.93	0.92	0.94
group 4	$N_{\text{train}} = 1000$	0.87	0.92	0.94	0.94	0.94
	$N_{\text{test}} = 500$	0.86	0.92	0.94	0.92	0.92

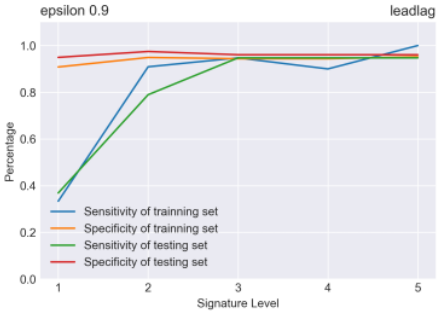
Table 3.5: Accuracy of two-dimensional paths contaminated by different variance



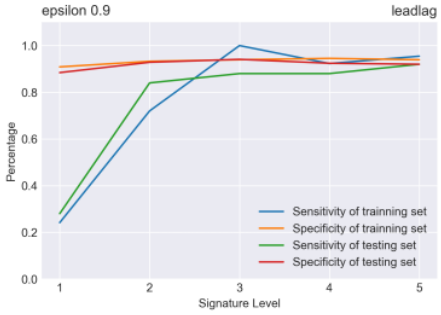
$N_{\text{train}} = 250$ and $N_{\text{test}} = 125$



$N_{\text{train}} = 500$ and $N_{\text{test}} = 250$



$N_{\text{train}} = 750$ and $N_{\text{test}} = 375$



$N_{\text{train}} = 1000$ and $N_{\text{test}} = 500$

Figure 3.10: Evaluation result of variance contaminated two-dim BM paths with different group size

Signature level	1	2	3	4	5
$N_{\text{train}} = 250$; $N_{\text{test}} = 125$	0.36	2.26	10.84	185.85	158.60
$N_{\text{train}} = 500$; $N_{\text{test}} = 250$	0.29	2.50	8.42	27.91	200.11
$N_{\text{train}} = 750$; $N_{\text{test}} = 375$	0.24	2.43	8.35	22.80	613.22
$N_{\text{train}} = 1000$; $N_{\text{test}} = 500$	0.22	1.92	5.75	22.07	158.88

Table 3.6: Conformance thresholds for two-dimensional paths contaminated by different variance

The conformance method with the lead-lag transform again demonstrate its excellent ability in identify the data with abnormal variances for group size larger than $N_{\text{train}} = 250$. The sensitivity and overall accuracy slightly went down when the group size equal to $N_{\text{train}} = 250$. Therefore, similar to the two dimension data with drift contamination, the sample size needs to be larger to ensure the stable performance of this conformance method on variance contamination data.

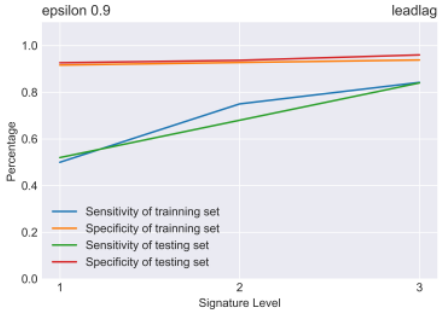
For the conformance threshold in table 3.6,we could again see the converge tendency for signature value smaller to equal to 4 as the sample size growth larger. This again correspond to the conclusion between the threshold and underlying parameters in section 2.2.2.

3.2.2 Four-Dimensional Brownian Motion Evaluation Result

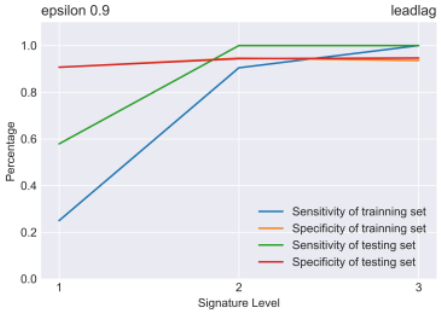
Based on the conclusion we derived for one and two dimensional Brownian motion data, we have that the group size larger than $N_{\text{train}} = 250$, signature level of 3, 4 and $\epsilon = 0.9$ generally render better results.

Therefore, we repeated the steps in section 3.2 by assigning the group size to be $N_{\text{train}} = 750$ and $N_{\text{test}} = 375$, $\epsilon = 0.9$, transforming the data using the lead-lag and calculating the signature level up to 3. The sensitivity and specificity results for drift and variance contaminated data with lead-lag transform are shown in figure 3.11.

The results shows that the sensitivity and specificity for data calculates through signature level equal to 3 are all larger than 80%. Therefore, we conclude that the conformance threshold anomaly detection method functions well for the four-dimensional Brownian motion data either with drift or variance contamination.



Contaminated by drift



Contaminated by Variance

Figure 3.11: Evaluation result of four-dimensional Brownian motion paths

3.2.3 Conclusion for Higher Dimensional Brownian Motion Data

For 2-dimensional Brownian motion data, the evaluation results for different sample sizes n show that it does not influence the performance as long as the size exceeds 250. This corresponds to the proof in section 2.2.2. Although for data with different sample size ($n \geq 500$), the sensitivity all slightly decreased compared to the one-dimensional case, the sensitivity and accuracy are higher than 90% when the signature level larger than 2 which is still a satisfactory result.

As the dimension increased to four-dimensional, the performance of this algorithm again decreased particularly for data contaminated by different drift. However, as the sensitivity and accuracy are still higher than 80% for both training and testing sets, we could conclude that this conformance model can identify various sources of anomalies among Brownian motion data with different dimensions if we select the proper input parameters. Therefore, based on the above result, we are ready to apply this method to more general data set for the anomaly detection tasks in Chapter 4.

Chapter 4

Anomaly Detection on Financial Streamed Data

With the development of computer-based trading in the financial market, the frequency of order placement and order cancellation has increased significantly. Therefore, the task of capturing the characteristics of the market becomes a challenging topic. In the work of Gyurkó et al. (2013), the signature method is adopted on streamed data (the crude oil future order book data) for classification problems. Follow this intuition, we hope to implement the conformance anomaly detection method on the order book data to identifying the atypical market behaviour across the different trading dates.

It has been shown by Baur and Dimpfl (2021) that the volatility of the cryptocurrency's price, like the Bitcoin price, is almost ten times the major exchange rate. Therefore, as the conformance algorithm is good at detecting the anomaly variance in the streamed data, we would like to test whether it is suitable to detect the anomaly behaviour for the cryptocurrency data.

We aim to use the historical market limit orders for bids and asks of four cryptocurrencies in terms of Bitcoin, Zcash, Ethereum, and Litecoin to detect the anomaly behaviour daily. Follow the order book processing steps suggest by Gyurkó et al. (2013), after having the level one order book data with the best bid and ask price and their corresponding amount, we calculated the mid-price, spread, and imbalance, which are more suitable to represent the characteristics of the data. Then the data is grouped by date and transform into streams of data. The train and test data is further defined according to the month of the order.

Then we modified the conformance algorithm proposed in section 2.4 and selected the suitable input parameters based on the conclusion in section 3.2 and prediction results on the training data. Then we test the algorithm's performance on the test data set and visualized the result by highlighting the anomaly date on the mid-price and the spread curve. We could see that the algorithm identified most of the data with extreme order, which demonstrates this method's effectiveness.

4.1 Input Data

The original data we obtain is historical market limit orders for both bid and ask (every 10 minutes) from 2020-01-01 to 2020-08-30[1]. Four popular cryptocurrencies available on Gemini are selected for evaluation. The level one order book data, a four-dimensional time-series data ($d = 4$), is generated using the original transaction data. In particular, it could be written as:

$$X := (P^a_{t_i}, P^b_{t_i}, V^a_{t_i}, V^a_{t_i})^N_{i=0}$$

where:

- $P^a_{t_i}$ best ask price : the lowest quoted offer price among sellers at some time stamps t_i

[1]The order book data of cryptocurrencies is obtain from https://www.cryptodatadownload.com/data/gemini/

- $P^b_{t_i}$ best bid price : the highest quoted offer price among buyers at some time stamps t_i
- $V^a_{t_i}$ the outstanding orders amount at the best ask price at some time stamps t_i
- $V^b_{t_i}$ the outstanding orders amount at the best bid price at some time stamps t_i

To capture the character of the order through the ask and bid price and order, we could transform the streams of order book data into

$$\hat{X} := (p_{t_i}, s_{t_i}, d_{t_i})_{i=0}^{N}$$

where:

- the logarithm of *mid-price* shows the average of the current bid and ask prices being quoted

$$p_{t_i} = \ln \frac{P^a_{t_i} + P^b_{t_i}}{2}$$

- the *spread* is the different between the ask and bid price:

$$s_{t_i} = P^a_{t_i} - P^b_{t_i}$$

- the *imbalance* represent the relatively difference of buy and sell orders amount:

$$d_{t_i} = \frac{V^a_{t_i} - V^b_{t_i}}{V^a_{t_i} + V^b_{t_i}}$$

The extreme change of the mid-price, the spread or the imbalance could all be the sign of the anomaly market behaviour, so the $\hat{X}$ is a suitable transformation of the original order transaction data.

4.2 Modify the Conformance Anomaly Detection Algorithm

Before applying the conformance anomaly detection algorithm on $\hat{X}$, we first need to clarify the selection of the input parameters. Follow the conclusion in section 3.2, the lead-lag transform and signature of level 3 demonstrate excellent performance in identifying the anomaly paths of the 4-dimensional Brownian motion contaminated data, so we will continue applying those parameters while calculating the signature of the input data.

The selection of suitable ϵ in section 3.2 is based on the evaluation results. However, we do not have the actual label for the cryptocurrencies data. Therefore, we decided to identify the suitable ϵ by plotting the anomaly detection result [2] for the training set and checking whether dates with extreme behaviour have all been identified. As in the original algorithm, we only provide the label for one half of the training set, so the algorithm need to be modified to make sure all the training data are labelled.

The first four steps are the same, and we only need to add one extra step that repeat the calculation of the conformance threshold for another group of training data:

[2]The imbalance cannot be clearly visualized by 2-dimensional plot, and we hope to include a more sophisticated method to demonstrate the effectiveness of the anomaly detection method on the order book data in the future.

Algorithm 3 Identify the conformance threshold for the anomaly of all training data

Input: training data X_{train}, group index y_{train}, the error bounds ϵ

Output: the anomaly threshold q_ϵ, predicted label for the training data $\hat{y}^1_{train}$, $\hat{y}^2_{train}$

begin

1. Applying lead-lag transformation to the training data X_{train} and then calculate the signature of level 3 as $S(X_{train})$, Then centred the $S(X_{train})$ by applying Min-Max normalization

2. Randomly split the $S(X_{train})$ and its correspond label y_{train} into two equal size halves $S(X^1_{train})$ and $S(X^2_{train})$, y^1_{train} and y^2_{train}

3. Find the conformance distance of each elements x in $S(X^1_{train})$ to subset $S(X^2_{train})$ Specifically, calculate the variance norm (Mahalanobis distance) of x to each element in $S(X^2_{train})$ and then find the smallest value $f(x, S(X^2_{train}))$ to be the conformance result

4. Look at the right tail of the empirical distribution of the conformance result. For a given error bound ϵ, find the conformance score correspond to the ϵ quantile of the empirical distribution and set it as the conformance anomaly threshold q_ϵ

5. Repeat the step 3 and 4 by flipping the position of X^1_{train} and X^2_{train} and get the conformance anomaly threshold q'_ϵ. Then, define the overall threshold as

$$\bar{q}_\epsilon = \frac{q_\epsilon + q'_\epsilon}{2}$$

6. If the conformance scores larger than q_ϵ or q'_ϵ, we define it as the anomaly data. Then we get the predict label $\hat{y}^1_{train}$ for elements in $S(X^1_{train})$ and $\hat{y}^2_{train}$ for elements in $S(X^2_{train})$.

7. Check the results by plotting the spread, logarithm mid-price and highlighting the anomaly date.

end

We first set the ϵ to be 0.95, which could be further adjusted based on the plot of the training set. After selecting the suitable ϵ, the algorithm of the test data X_{test} also need to be adopted as:

Algorithm 4 Identify anomaly data using modified conformance threshold

Input: training date X_{train}, testing data X_{test}, group index y_{train} the anomaly threshold $\bar{q}_\epsilon$

Output: The predicted label for the testing data $\hat{y}_{test}$

begin

1. Applying lead-lag transformation on testing data X_{test}, calculating the signature of level 3 as $S(X_{test})$ and apply Min-Max normalisation to $S(X)_{test}$.

2. Calculating the conformance distance of each element in $S(X_{test})$ to $S(X_{train})$ similar to the training process

3. For elements x' in $S(X_{test})$ that have larger conformance score $f(x', S(X_{train}))$ than the anomaly threshold $\bar{q}_\epsilon$, we predict it to be an anomaly data and set the label $\hat{y}_{test} = 1$. The rest is labeled as $\hat{y}_{test} = 0$. Then we demonstrate the results by plotting the spread, logarithm mid-price and highlighting the predicted anomaly date.

end

Note that to reduce the randomness of the data splitting process, we could repeat steps 3, 4, 5 in algorithm 3 for m times. Then the conformance threshold could be assigned as the average of the m values :

$$\frac{\sum_{i=0}^{m} \bar{q}^i_\epsilon}{m}$$

However, in the numerical test, we find the algorithm 3 is sufficient to identify the anomaly instances, so we leave the test of this repeat experiment in the future analysis of other data set.

4.3 Numerical Test Results

In this section, we discuss the anomaly detection results on four types of cryptocurrencies order book data. We select the January to May data as the training set and June to August as the testing set.

4.3.1 Anomaly Detection Results for Bitcoin

Bitcoin is one of the earliest cryptocurrency designed by Satoshi Nakamoto. It is a decentralized digital currency that can be sent directly from user to user without intermediaries. Therefore, its order behaviour is less relevant to the macroeconomic or financial development and is suitable for using only the historical data to identify the anomaly instances.

Through the modified conformance algorithm 3, the anomaly date is identified and highlighted by the pink line plot against the log mid-price and spread in figure 4.1 and 4.2.

For the logarithm of mid-price, the most significant drop that happened on March 12th have been identified. The extremely high differences between ask and bid in February and May have been marked as anomalies. Therefore, we believe that this conformance threshold is suitable for identifying the anomaly behaviours of the Bitcoin prices.

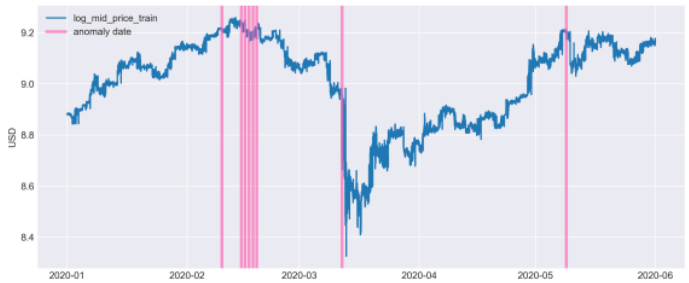


Figure 4.1: Anomaly detection result for BTC train data against the Log Mid price

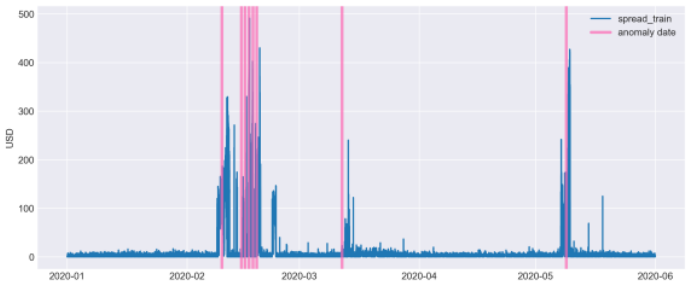


Figure 4.2: Anomaly detection result for BTC train data against the spread

Applying that conformance threshold into the algorithm 4, we plot the result for both the training (blue) and testing (yellow) data in figure 4.3 and 4.4. We find that the algorithm successfully identified the most significant increases of the Bitcoin mid-price on July 27th and the dates with high spread are also marked as anomalies. This result demonstrates the effectiveness of this method on identifying the anomaly behaviour of Bitcoin order book data.

Figure 4.3: Anomaly detection result for BTC against the Log Mid price

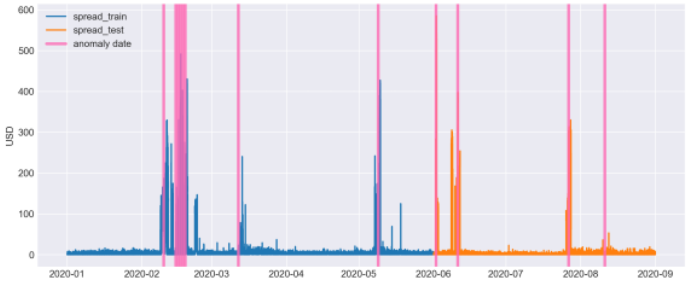


Figure 4.4: Anomaly detection result for BTC against the spread

4.3.2 Anomaly Detection Results for Ethereum

Ether, similar to Bitcoin, is the native cryptocurrency of the Ethereum platform. The Ethereum platform is also a decentralized, open-source block-chain. Therefore, it is suitable to use the conformance threshold of the historical order data to identify the anomaly behaviour.

Applying the algorithm 3 on the training data, the detecting results are shown in figure 4.5 and 4.6. The dates (March 13th to 18th) with the notable mid-price drop and large spread have been identified. As no other part has the significant changes in a short period, we believe this threshold is a good boundary to separate the anomaly instances from the normal group.

Figure 4.5: Anomaly detection result for ETH train data against the Log Mid price

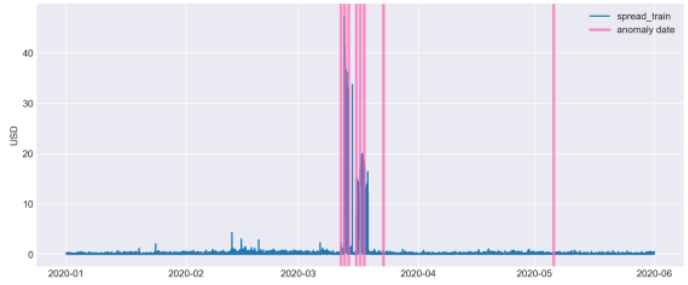


Figure 4.6: Anomaly detection result for ETH train data against the spread

For the testing data from June to August, we find the mid-price continually increased and there is no extreme change in the mid-price. Also, the spread is constantly low during this period. Therefore, we believe that all the behaviour is normal during the testing period. This correspond to the anomaly detection results showing in figure 4.7 and 4.8, as there is no anomaly date identified for the training set.

Figure 4.7: Anomaly detection result for ETH against the Log Mid price

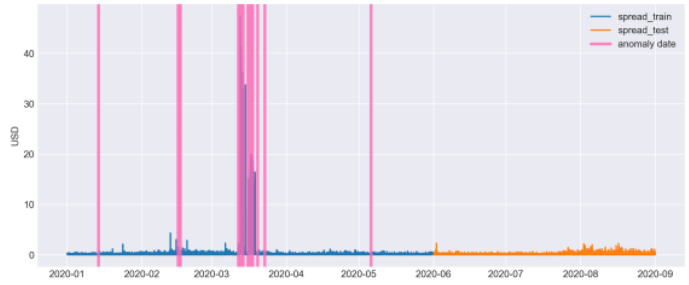


Figure 4.8: Anomaly detection result for ETH against the spread

4.3.3 Anomaly Detection Results for Litcoin

The Litcoin is an alternative cryptocurrency based on the model of Bitcoin. Therefore, its behaviour and price fluctuation is similar to the bitcoin.

The spread of the Litcoin is extremely high in the middle of January and March and those sudden up and down have been successfully identified through our conformance algorithm as the figure 4.9 and 4.10 show.

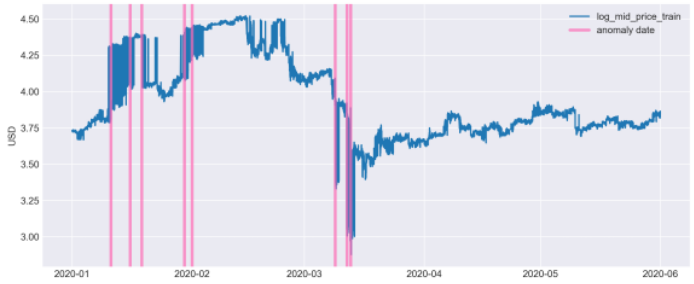


Figure 4.9: Anomaly detection result for LTC train data against the Log Mid price

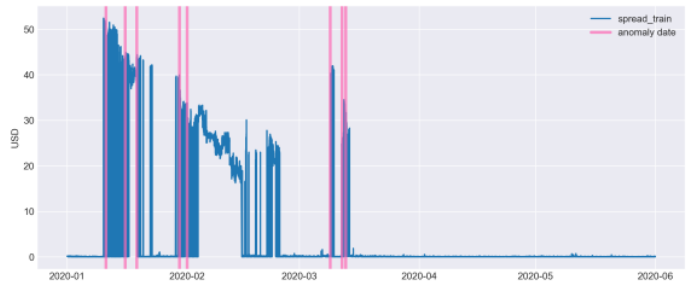


Figure 4.10: Anomaly detection result for LTC train data against the spread

Based on the threshold identified through the testing data, our conformance anomaly detection method shows that there is no anomaly date from June to August. We could see that in figure 4.11 and 4.12, the mid-price from June to August is relatively stable, and the spread is continually low. Therefore, we believe the detection result is valid as there is no anomaly date during this period that.

Figure 4.11: Anomaly detection result for LTC against the Log Mid price

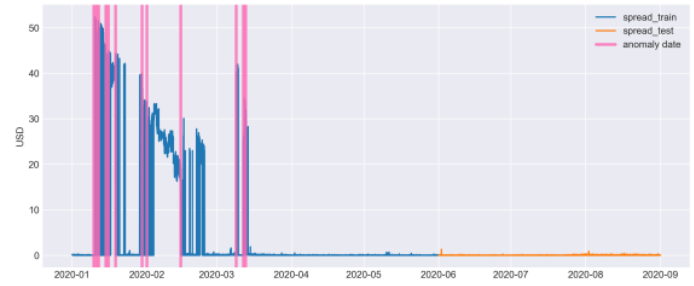


Figure 4.12: Anomaly detection result for LTC against the spread

4.3.4 Anomaly Detection Results for Zcash

Zcash is also a cryptocurrency based on Bitcoin. It uses cryptography to provide enhanced privacy for users. Therefore, the order and price changing pattern is also similar to the Bitcoin.

The logarithm mid-price of the Zcash fluctuate remarkably during the first three months as the figure 4.13 shows. Also, the spread of those dates is extremely high as demonstrated in figure 4.14. Since those dates with extreme behaviour have all been identified by the conformance algorithm as the figures show, we believe the conformance threshold is suitable for identifying the anomaly behaviour of the Zcash.

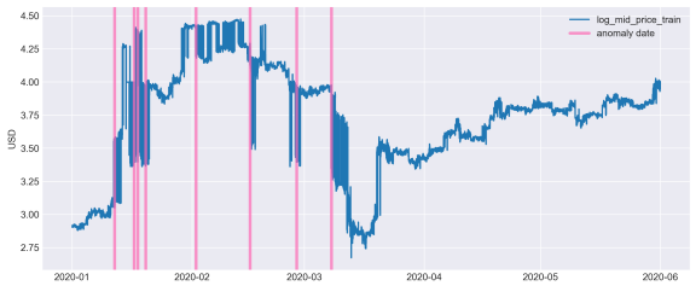


Figure 4.13: Anomaly detection result for ZEC train data against the Log Mid price

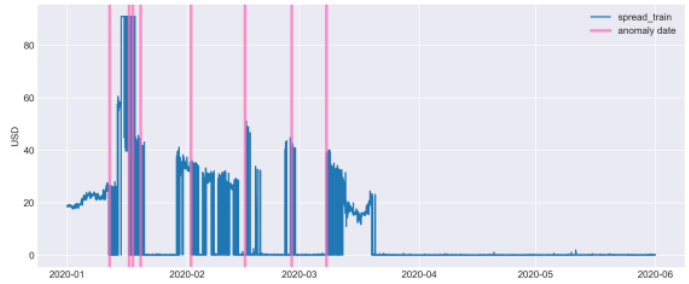


Figure 4.14: Anomaly detection result for ZEC train data against the spread

Using the conformance threshold to identify the anomaly instance in the test data (From June to August), the result shows one anomaly date — July 27th. The mid-price increase and the spread of that date are high, which could be regarded as anomaly behaviour. However, the dates around July 27th also demonstrate similar extreme behaviour and one date in mid-July experiences a sudden increase of both variables. Those dates have not been identified as the anomaly by the conformance algorithm.

By looking at the conformance threshold comparison shown in table 4.1, we could see that since the fluctuation of Zcash and Litcoin is more extreme than the Bitcoin, the conformance thresholds of those two are much higher than the Bitcoin. This high threshold made the detection not so sensitive to the anomaly as the boundary is too high. We could change the conformance threshold by lower the ϵ. However, that would result in too many anomalies identified for the training test, which is also an inaccurate result. Therefore, we consider including more data in future studies to ensure we capture the suitable threshold for the anomaly data.

coin name	Bitcoin	Ethereum	Litcoin	Zcash
conformance threshold	2740	3563	453896	146595

Table 4.1: Conformance threshold comparison for the cryptocurrencies order book data

Figure 4.15: Anomaly detection result for ZEC against the Log Mid price

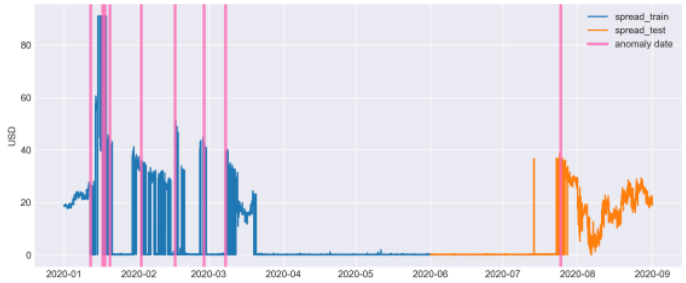


Figure 4.16: Anomaly detection result for ZEC against the spread

Conclusion

In this thesis, we study the conformance anomaly detection method proposed by Cochrane et al. (2020). Both theoretical proof and empirical experiments are implemented to understand this algorithm's dependence on the underlying parameters. We also proved that this algorithm could effectively identify anomalies among Brownian motion and real market financial datasets as long as we select suitable input parameters.

In Chapter 2, the conformance threshold for the Gaussian sample is proved and tested empirically. The exact relationship of the threshold q_ϵ to other parameters (the dimension d, error bound ϵ and variance σ^2 of the data) is theoretically deduced using the Gaussian concentration inequality. We also find that the corpus size n does not influence the threshold so that a constant threshold would be identified regardless of the sample size. One problem arises during the empirical test that the empirical conformance threshold of the simulated Gaussian data is generally smaller than the theoretical threshold. The potential cause of this discrepancy could be the approximation during the theoretical proving steps and the less accurate numerical approximation during the practical implementation. More works need to be conducted in the future to determine the exact sources of the discrepancy. Nevertheless, the conformance threshold's dependence on data dimension and corpus size is consistent for the theoretical and empirical result, making the empirical application of using the conformance distant to identify the anomaly behaviour still valid.

In Chapter 3, the evaluation result for the one-dimensional Brownian motion data shows that the conformance method with the lead-lag transformation demonstrates better performance than the Isolation Forest when identifying the data contaminated by different variances and drifts. The result also indicates that the error bound ϵ should be slightly smaller than one minus the contamination rate. Adopting the above parameter into the two-dimensional Brownian motion data analysis, we proved that the sample size n has negligible influence on the accuracy. Besides, the sample size should be larger than 250 as too small datasets may result in unsatisfied sensitivity (true positive rate). Therefore, we conclude that the conformance algorithm could effectively identify the anomaly path of different dimension Brownian motion data for the sample size larger than 250.

To apply this conformance algorithm to real-market order book data analysis, we further modify the algorithm to make sure all the training data will be labelled. Then, applying this modified algorithm to the training data, we could identify the suitable conformance threshold by changing the ϵ and used the threshold to identify the dates with anomaly behaviour in the test data. The results of the modified algorithm on the order book for four cryptocurrencies demonstrate its satisfactory ability to identify the anomaly date with extreme behaviour. However, we only demonstrate this by highlighting the anomaly date in the logarithm mid-price and spread plot. More sophisticated methods need to be proposed in future works to validate the effectiveness of this anomaly detection method. Also, some anomaly dates have not been identified for the Zcash testing data set, which might be induced by the training set's high volatility. A possible solution could be enlarging the sample size, but we leave it for further exploration due to the limitation of acquiring the order book data.

Appendix A

Additional Proofs

A.1 Property of the Tensor Product

For a vector $a, b \in (\mathbb{R}^d)^{\otimes k}$:

$$a = \sum_{i_1,\cdots,i_k=1}^{d} a^{i_1\cdots i_k} e_{i_1} \otimes \cdots \otimes e_{i_k}$$

$$b = \sum_{i_1,\cdots,i_k=1}^{d} b^{i_1\cdots i_k} e_{i_1} \otimes \cdots \otimes e_{i_k}$$

Then we have that

$$\|a \otimes b\|_{k+l} = \|a\|_k \|b\|_l$$

where

$$\|a\|_k = \left(\sum_{i_1,\cdots,i_k=1}^{d} (a^{i_1 i_2 \cdots i_k})^2 \right)^{\frac{1}{2}}$$

A.2 Proof of Johnson-Lindenstrauss Lemma for the Infinite Case

Follow the first part of proof given in section 2.2, define $M' = d \inf_{\alpha \in T} \|W(\alpha)\|^2$ and adopt similar approach as theorem 2.2.3, we have for all $t > 0$,

$$\mathbb{P}\{M - \mathbb{E}M \geq 2\sqrt{2t\mathbb{E}M} + 2t\} \leq e^{-t} \tag{A.2.1}$$

and for all $t > \frac{1}{2}$

$$\mathbb{P}\{M - \mathbb{E}M \leq -2\sqrt{2t\mathbb{E}M}\} \leq e^{-t} \tag{A.2.2}$$

Then we have

$$V := \sup_{\alpha \in T} \left(\|W(\alpha)\|^2 - 1 \right) = \frac{M}{d} - 1 \tag{A.2.3}$$

and

$$V := \sup_{\alpha \in T} \left(-\|W(\alpha)\|^2 + 1 \right) = -\frac{M'}{d} + 1 \tag{A.2.4}$$

Then for any $t > 0.5$, with a double application of theorem 2.2.3 and above result, we could obtain with probability at least $1 - 2e^{-t}$,

$$\sup_{\alpha \in T} \left| \|W(\alpha)\|^2 - 1 \right| = max(V, V') \leq \max(\mathbb{E}V, \mathbb{E}V') + 2\sqrt{\frac{2(1 + \mathbb{E}V)t}{d}} + \frac{2t}{d} \tag{A.2.5}$$

Define quantity $\Delta = d \max(\mathbb{E}V, \mathbb{E}V')^2$, the equation (A.2.5) can be written as:

$$\sup_{\alpha \in T} \left| \|W(\alpha)\|^2 - 1 \right| = \max(V, V') \leq 2\sqrt{\frac{\Delta}{d}} + 2\sqrt{\frac{2t}{d}} + \frac{4t}{d} \tag{A.2.6}$$

This holds with probability at least $1-2e^{-t}$. This then proved the random variable $\sup_{\alpha\in T}|\|W(\alpha)\|^2-1|$ is highly concentrated around its mean which therefore shows that W is a ϵ-isometry on A.

A.3 Another Approach to Identify the Conformance Threshold for Gaussian Data

In this section, we will use the Gaussian Isoperimetric inequality to identify the conformance threshold of the anomalies. We could deduce the same result as in section 2.2 Chapter 2.

We first clarify some notations. For $x \in \mathbb{R}^n$, the probability density function of the standard Gaussian distribution $\varphi_n(x)$ is:

$$\varphi_n(x) = \frac{1}{\sqrt{2\pi}} \exp(-\|x\|^2/2) \tag{A.3.1}$$

where $\|\cdot\|$ is the Euclidean norm.
The cumulative distribution function of one dimensional Gaussian distribution is:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$$

Let $A \in \mathbb{R}^n$ be a Borel set, then its n-dimensional Gaussian measure is:

$$\gamma_n(A) = \int_A \varphi_n(x)dx \tag{A.3.2}$$

Definition A.3.1 (Gaussian Isoperimetric Inequality)**.** Let $A \in \mathbb{R}^n$ be a Borel set, then for any $h > 0$

$$\Phi^{-1}(\gamma_n(A^h)) \geq \Phi^{-1}(\gamma_n(A)) + h$$

where $A^h = \{x \in \mathbb{R}^d : \|x-a\|_\mu < h \text{ for some } a \in A\}$

Following this definition, we could further deduce a lemma as follow:

Lemma A.3.2. *The Gaussian Isoperimetric inequality is equivalent to the following statement: let $A \in \mathbb{R}^n$ be a Borel set, and $H \in \mathbb{R}^n$ be a half-space, such that $\gamma_n(A) = \gamma_n(H)$, then for any $h > 0$,*

$$\gamma_n(A^h) \geq \gamma_n(H^h) \tag{A.3.3}$$

Now we apply this lemma to introduce an important corollary. Let $\mathcal{I} \in V$ be a finite corpus of vector data. Here we assume our data to be Gaussian. We could then split the corpus randomly into two equal size halves and denote them as $\mathcal{I}_1, \mathcal{I}_2$.

Suppose we want to study the elements $x \in \mathcal{I}_2$'s conformance distant to $\mathcal{I}_1$. In section 2.2, we transform the problem of finding the conformance threshold into the quantile of the sum of the Mahalanobis distance. To simplify the notation, we define sum of the Mahalanobis distance as:

$$f(x, \mathcal{I}_1) := \sum_{x_i \in \mathcal{I}_1} \sqrt{(x-x_i)^T K^{-1}(x-x_i)}$$

where $K = \mathbb{E}[(x-x_i)^T(x-x_i)]$. We have proved that the Mahalanobis distance is a k-Lipschitz function, and as the Lipschitz continuity preserve over sum, the Lipschitz constant for the sum of Mahalanobis distance is

$$L := nk = n\sqrt{2}\|U\|_2 = \sqrt{2}n\sigma_{max}(U)$$

where $\sigma_{max}(U)$ is the largest singular value of U in $K = UU^T$

Then we find the conformance distance R so that the probability of x's sum of Mahalanobis distance to $\mathcal{I}_1$ smaller than R is $\frac{1}{2}$, we could write it as:

$$\mathbb{P}(f(x, I_1) \leq R) = \frac{1}{2}, x \in I_2 \tag{A.3.4}$$

The R could be seen as a medium of the sum of Mahalanobis distance.

Let $A = \{x \in I_2 : f(x, I_1) \leq R\}$ then for any $y \in A^r$, there exists an $x \in A$, such that $|y-x| < r$. As the sum of the Mahalanobis is L-Lipschitz function, $|f(y, I_1) - f(x, I_1)| < rL$. Then $f(x, I_1) \leq R$ is equivalent to $f(y, I_1) < r + R$, so we have:

$$\begin{aligned} \mathbb{P}(f(y, I_1) - R < rL) &\geq \gamma_n(A^r) \\ &\geq \Phi(\Phi^{-1}(\gamma_n(A)) + r) \\ &= \Phi(\Phi^{-1}(\frac{1}{2}) + r) \\ &= \Phi(r) \end{aligned} \tag{A.3.5}$$

Similarly we have

$$\mathbb{P}(f(y, I_1) - R > -rL) \geq \Phi(r) \tag{A.3.6}$$

Therefore,

$$\mathbb{P}(|f(y, I_1) - R| \geq rL) \leq 2(1 - \Phi(r)) \leq 2e^{-\frac{r^2}{2}} \tag{A.3.7}$$

and we could have

$$\mathbb{P}(f(y, I_1) \geq R + rL) \leq \exp(-\frac{r^2}{2}) \leq 1 - \epsilon \tag{A.3.8}$$

If we calculated the $R+rL$, and replace the medium R by the mean α, we could find it is the same as the result in section 2.2 equation (2.2.8): [1]

$$r \geq \frac{\alpha + L\sqrt{2 \ln \frac{1}{1-\epsilon}}}{n} \tag{A.3.9}$$

Therefore, we validate the theoretical result of using the conformance threshold to identify the anomaly behavior.

[1] r is divided by n to be consistent with the result in section 2.2

Bibliography

Baur, D. G. and Dimpfl, T. (2021), 'The volatility of bitcoin and its role as a medium of exchange and a store of value', *Empirical Economics* pp. 1–21.

Boucheron, S., Lugosi, G. and Massart, P. (2013), *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press.

Boyd, S., Boyd, S. P. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge university press.

Buldygin, V. V. and Kozachenko, Y. V. (1980), 'Sub-gaussian random variables', *Ukrainian Mathematical Journal* **32**(6), 483–489.

Cass, T. (2021), Week 2: Analytical properties of the signature. lecture notes, Advanced topics in data science, Msc Mathematics and Finance, Imperial College London, delivered March 2021.

Chen, K.-T. (1958), 'Integration of paths–a faithful representation of paths by noncommutative formal power series', *Transactions of the American Mathematical Society* **89**(2), 395–407.

Chevyrev, I. and Kormilitzin, A. (2016), 'A primer on the signature method in machine learning', *arXiv preprint arXiv:1603.03788* .

Cochrane, T., Foster, P., Lyons, T. and Arribas, I. P. (2020), 'Anomaly detection on streamed data', *arXiv preprint arXiv:2006.03487* .

Costa, E. (2018), 'Proving that mahalanobis norm is a norm indeed', Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/2602778 (version: 2018-01-13).

Flint, G., Hambly, B. and Lyons, T. (2016), 'Discretely sampled signals and the rough hoff process', *Stochastic Processes and their Applications* **126**(9), 2593–2614.

Goldstein, M. and Uchida, S. (2016), 'A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data', *PloS one* **11**(4), e0152173.

Grubbs, F. E. (1969), 'Procedures for detecting outlying observations in samples', *Technometrics* **11**(1), 1–21.

Gyurkó, L. G., Lyons, T., Kontkowski, M. and Field, J. (2013), 'Extracting information from the signature of a financial data stream', *arXiv preprint arXiv:1307.7244* .

Hambly, B. and Lyons, T. (2010), 'Uniqueness for the signature of a path of bounded variation and the reduced path group', *Annals of Mathematics* pp. 109–167.

Levin, D., Lyons, T. and Ni, H. (2013), 'Learning from the past, predicting the statistics for the future, learning an evolving system', *arXiv preprint arXiv:1309.0260* .

Lyons, T. (2014), 'Rough paths, signatures and the modelling of functions on streams', *arXiv preprint arXiv:1405.4537* .

Zantedeschi, V., Emonet, R. and Sebban, M. (2016), 'Lipschitz continuity of mahalanobis distances and bilinear forms', *arXiv preprint arXiv:1604.01376* .

Deng_Millie_02003063

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

PAGE 54

PAGE 55