Random Error Reduction in Similarity Search on Time Series: A Statistical Approach

Wush Chi-Hsuan Wu [#1], Mi-Yen Yeh [#2], Jian Pei [*3]

[#] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*
[*] *School of Computing Science, Simon Fraser University, Burnaby, BC, Canada*
[1] wush978@gmail.com, [2] miyen@iis.sinica.edu.tw, [3] jpei@cs.sfu.ca

***Abstract*—**
Errors in measurement can be categorized into two types: systematic errors that are predictable, and random errors that are inherently unpredictable and have null expected value. Random error is always present in a measurement. More often than not, readings in time series may contain inherent random errors due to causes like dynamic error, drift, noise, hysteresis, digitalization error and limited sampling frequency. Random errors may affect the quality of time series analysis substantially. Unfortunately, most of the existing time series analysis methods do not address random errors, possibly because random error in a time series, which can be modeled as a random variable of unknown distribution, is hard to handle. In this paper, we tackle this challenging problem. Taking similarity search as an example, which is an essential task in time series analysis, we develop MISQ, a statistical approach for random error reduction in time series analysis. The major intuition in our method is to use only the readings at different time instants in a time series to reduce random errors. We achieve a highly desirable property in MISQ: it can ensure that the recall is above a user-specified threshold. An extensive empirical study on 20 benchmark real data sets clearly shows that our method can lead to better performance than the baseline method without random error reduction in real applications such as classification. Moreover, MISQ achieves good quality in similarity search.

I. Introduction

Time series analysis is widely used in many applications. In general, a time series is a series of readings recorded at a sequence of time instants. The quality of time series data depends on the quality of the readings.

Due to limitation of data collection equipment and methods, readings in time series data are often prone to various errors. In general, errors can be categorized into two types: *systematic errors* that are predictable, and *random errors* that are inherently unpredictable and have null expected value [1]. Often, systematic errors can be removed by calibration of the measurement equipment. However, random error is always present in a reading.

Since random errors have the property that the mean of many separate measurements approaches 0, random errors can be reduced by obtaining multiple independent measurements and using the mean of them. In practice, it is however often infeasible to obtain multiple independent measurements on one target attribute at an instant. Consequently, we still need to develop effective methods to reduce random errors in the data processing and analysis phase.

Most of the existing time series analysis methods do not address random errors, possibly because reducing random errors in the data analysis phase is far from trivial. Conceptually, random errors in a time series can be modeled as a random variable of unknown distribution. It is very challenging to systematically remove random errors.

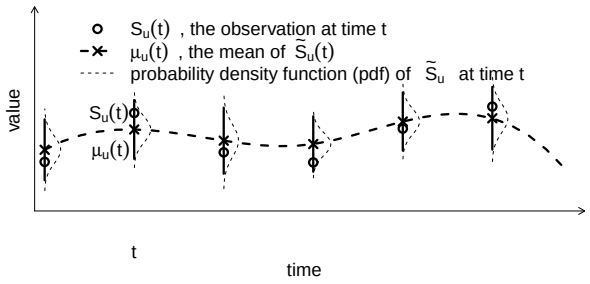


Fig. 1: Modeling time series with random errors.

In this paper, we tackle the problem of reducing random errors in time series analysis. We take the similarity search on time series as a concrete example, since it is essential for time series analysis. To take random errors into account, we model a time series $\tilde{S}_u$, as illustrated in Fig. 1, as an ordered sequence of continuous random variables. At timestamp t, the observation $S_u(t)$ recorded is a sample from an unknown random variable $\tilde{S}_u(t)$ with an unknown probability density function (pdf). The distance between two time series is consequently a random variable with unknown distribution as well.

Technically, a time series can be further regarded as a series of expected values (mean values), each being blurred with a random variable of mean 0 and unknown variance. The problem of measuring the similarity between two time series becomes approaching the distance between the two time series of mean values, as illustrated in Fig. 2. To make our discussion concrete, we use Euclidean distance in this paper. In general, any distance measure may be used, though some technical details may need to be adjusted accordingly.

In this paper, we develop **MISQ** (for mean distance queries), a similarity search method for time series. The major intuition behind the feasibility of random error reduction is that, by assuming that the readings in a time series are collected through the same equipment under the proper working condition, the random errors incurred in those readings follow the same unknown distribution. By proper deliberation of those readings, we can remove the effect of random errors

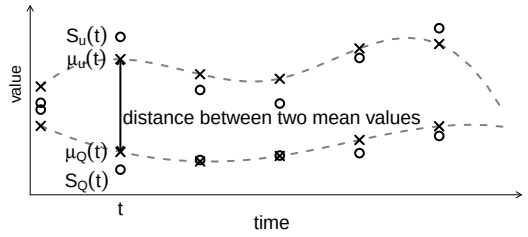


Fig. 2: Measuring the similarity between two time series using the mean values.

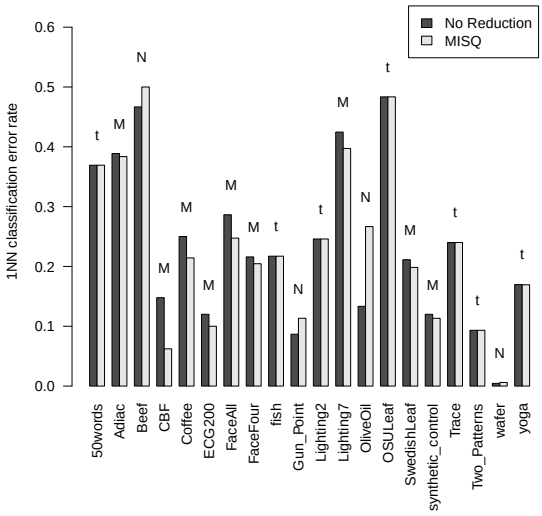


Fig. 3: The 1NN classification error rates for MISQ and using Euclidean distance without considering random errors.

in similarity search systematically.

One may doubt whether reducing random errors or not would make any difference in real world applications. To answer the concern, we use 1NN classification as a test, where a time series is classified by its nearest neighbor in a training set. On 20 real data sets in the UCR time series data repository [2], we compare the 1NN classification using MISQ to reduce random errors with that not considering random errors. Fig. 3 reports the error rates of the two methods. For each data set, we mark M and N to indicate if reducing random errors using MISQ leads to lower error rates or not, respectively, and t for a tie.Using MISQ to reduce random errors leads tolower error rates on 9 data sets (45%), and has higher error rates on 4 data sets (20%).[1] On the other 7 data sets (35%), reducing random errors or not does not lead to observable difference. The experimental results clearly show that reducing random errors is beneficial in many cases.

To the best of our knowledge, we are the first to tackle random errors in similarity search on time series without assuming any knowledge about the distribution of the errors. We make several contributions. First, we empirically show that reducing random errors can gain substantial performance on real data sets, which justifies the need of random error reduction in time series analysis. Second, we formulate the problem of similarity search on time series by considering random errors, and develop MISQ, a statistical method. Last, by an extensive empirical study, we evaluate MISQ systematically and demonstrate its power in time series analysis. MISQ can efficiently compute the high quality mean distance between two time series without significantly more cost than any faster moving average methods.

The rest of the paper is organized as follows. We formulate the problem and discuss the related work in Section II, and develop the techniques of mean distance estimation and query processing with error controlled, respectively, in Sections III and IV. We report the experiment results in Section V, and conclude the paper in Section VI.

II. Problem Definition and Related Work

A. Problem Definition

Definition 1 (Time series): Denote by $\tilde{S}_u$ a **time series** with index u, which is a sequence of random variables, as illustrated in Fig. 1. At time t, where $t = 1, 2, \ldots, l$, the random variable $\tilde{S}_u(t)$ is

$$\tilde{S}_u(t) = \mu_u(t) + V_u\tilde{\varepsilon}_u(t), \tag{1}$$

where $\mu_u(t)$ is the mean (the value without noise) of the time series and $V_u\tilde{\varepsilon}_u(t)$ describes the uncertain noise. ■

We make the following assumptions in this paper.

- $\mu_u(t)$ is an unobserved, smooth, and uniformly bounded function (those cross signs linked by a grey dashed line in Fig. 1.)
- V_u is an unknown positive constant invariant to time t. In general, V_u can also be a smooth function of t, but the case is more complicated and beyond the scope of this paper.
- $\tilde{\varepsilon}_u(t)$ is an independent identically distributed (i.i.d.) random variable with mean 0 and variance 1. As a result, $E\left(\tilde{S}_u(t)\right) = \mu_u(t)$ and $Var\left(\tilde{S}_u(t)\right) = V_u^2$.
- At each time instant t, there is only one known observation $S_u(t)$ (marked with a small circle in Fig. 1).
- The pdf of $\tilde{S}_u(t)$ has a width determined by the unknown constant V_u and the same shape as the pdf of $\tilde{\varepsilon}_u(t)$.
- Since $\tilde{\varepsilon}_u(t)$ is assumed to be i.i.d., the correlation between $\tilde{S}_u(t)$ and $\tilde{S}_u(t-1)$ only depends on the mean function $\mu_u(t)$.

Definition 2 (Distance): The **distance** between two time series $\tilde{S}_Q$ and $\tilde{S}_u$ of length l is

$$D(\tilde{S}_Q, \tilde{S}_u) = \sum_{t=1}^{l} (\tilde{S}_Q(t) - \tilde{S}_u(t))^2, \tag{2}$$

where $D()$ is the square of the Euclidean distance. ■

Apparently, $D(\tilde{S}_Q, \tilde{S}_u)$ is a random variable. Instead of directly modeling the distribution of $D(\tilde{S}_Q, \tilde{S}_u)$, we consider the *mean distance*, which excludes the effect of random errors.

[1] Since the UCR data sets do not come with the ground truth about errors, we cannot analyze the details why MISQ does not work well on those data sets. We believe that using MISQ to reduce random error becomes less effective when system errors are dramatically larger than random errors.

Definition 3 (Mean distance): The **mean distance** between two uncertain time series $\tilde{S}_Q$ and $\tilde{S}_u$ of length l is.

$$MD(\tilde{S}_Q, \tilde{S}_u) = D(\mu_Q, \mu_u) = \sum_{t=1}^{l} (\mu_Q(t) - \mu_u(t))^2. \quad (3)$$

■

In this paper, we consider two types of query on time series.

Definition 4 (Queries): Given a reference time series $\tilde{S}_Q$ and a set of time series $\mathcal{T}$, an **exact match query** (also called **exact query**) retrieves those time series $\tilde{S}_u \in \mathcal{T}$ such that $D(\mu_Q, \mu_u) = 0$; a **threshold similarity query** (also called **threshold query**) retrieves those time series $\tilde{S}_u \in \mathcal{T}$ such that $D(\mu_Q, \mu_u) \leq r$, where $r > 0$ is a user specified distance threshold. For the sake of simplicity, we omit $\mathcal{T}$ hereafter if it is clear from context. ■

Apparently, when the distance threshold r is set to 0, a threshold similarity query becomes an exact match query. As aforementioned, μ_Q and μ_u are unobserved values in practice. Thus, we cannot compute $D(\mu_Q, \mu_u)$ directly. Instead, we can use only the observation values to *estimate* it, of which the method will be introduced in later sections, and apply statistical hypothesis testings to determine if a candidate time series qualifies a query.

We design the hypothesis testing procedure for exact queries and threshold queries as follows.

Definition 5 (Null hypotheses): An exact query retrieves those $\tilde{S}_u \in \mathcal{T}$ that do not reject the **null hypothesis** $H_0 : D(\mu_Q, \mu_u) \leq 0$. Here, we use $\leq$ instead of $=$ is for the convenience of the testing. Since distance is always non-negative, there is no difference between using $=$ and $\leq$.

A threshold query retrieves those $\tilde{S}_u$ that do not reject the **null hypothesis** $H_0 : D(\mu_Q, \mu_u) \leq r$. ■

We consider two types of error.

Definition 6 (Two types of error): A **type I error** happens if a statistical test rejects a true null hypothesis (H_0). In our case, a type I error happens if an exact match query fails to retrieve a time series exactly matching the reference time series, or a threshold similarity query fails to retrieve a time series whose distance to the reference time series is less than or equal to the threshold. In other words, a low type I error rate implies a high *recall*.

A **type II error** happens when a test fails to reject a false null hypothesis (H_0). In our case, a type II error happens if an exact match query retrieves a time series that is in fact not exactly matching the reference time series, or a threshold similarity query retrieves a time series whose distance to the reference time series is in fact larger than the threshold. A low type II error rate implies a high *precision*. ■

One major challenge is how we can estimate the mean distance with only one observation at each timestamp for each time series. One may think about using some common de-noising methods, such as *moving average*, to guess the mean values at different instants first. However, those approaches are usually heavily parameter-dependent. That is, one needs to decide how many observations should be considered simultaneously to get the mean value. Such parameters are hard to decide in practice. Different from those methods, in this paper we develop a parameter-free, difference-based estimator for mean distance using only observations in time series.

Another challenge is how we can ensure that the estimated mean distance is controlled at a given confidence level. In this regard, we devise another estimator to compute the variance of the mean distance estimator, which is to decide if the estimated mean distance is *significantly* above a given threshold r or not. With the variance estimator and the user given confidence threshold $1 - \alpha$, we can compute the lower bound of the confidence level, LCI, of each estimated mean distance. Benefitting from the LCI, we can control the type I error rates. Specifically, we only report the time series whose LCI of the mean distance to the query time series is not greater than r.

For both exact queries and threshold queries, we can interpret the hypothesis testings using a *confidence interval* determined by a user given *confidence level* of the estimated $D(\mu_u, \mu_Q)$. Since the testing is one-tailed, we can compare the lower bound of the confidence interval, denoted by $LCI(D(\mu_Q, \mu_u))$, with the given distance threshold (0 for an exact query and r for a threshold query). We will show that, if the confidence level is set to $1 - \alpha$, where $\alpha \in [0, 1]$, the type I error rate most of the time is not greater than α.

Based on the above discussion, let us formally restate the two types of query using hypothesis testing with type I error rate controlled according to α as follows.

Definition 7 (Queries, using hypothesis testings): Given a reference time series $\tilde{S}_Q$, a set of time series $\mathcal{T}$, a user specified confidence level $1 - \alpha \in [0, 1]$, an **exact match query** retrieves all time series $\tilde{S}_u \in \mathcal{T}$ such that

$$LCI(D(\mu_u, \mu_Q)) \leq 0. \quad (4)$$

Moreover, a **threshold similarity query** retrieves all time series $\tilde{S}_u \in \mathcal{T}$ such that

$$LCI(D(\mu_u, \mu_Q)) \leq r, \quad (5)$$

where $r > 0$ is a given distance threshold. ■

B. Related Work

Similarity search in time series databases, as an important function in many applications, has drawn wide attention in the recent decades. Many studies investigate how to search efficiently and accurately under the widely-used Euclidean distance [3–5] and many other similarity measurements, such as [5–9]. Those methods do not consider errors incurred in the reading of each time instant yet.

Some recent studies tackle random errors in time series by modeling time series as uncertain data. Specifically, in [10–12], a type of probabilistic query is investigated that finds the uncertain time series whose distances to a reference one are not greater than a given distance threshold with a high enough probability. Different from our study, those methods all assume that both the mean and the variance of each uncertain variable are either known in advance [11, 12], or can be estimated with a set of observations at each timestamp [10]. Consequently, the lower bound or the distribution of the Euclidean distance

between two series can be computed. In real applications, however, the assumption is often hard to meet.

Aßfalg *et al.* [10] assume that multiple observations are collected at each timestamp. They treat those observations as a realization of the corresponding unknown random variable of uncertainty, and approximate the density of the distance with the combinations of the observations. For example, given two observations of $\tilde{S}_u(t)$ as $1, 2$ and two observations of $\tilde{S}_Q(t)$ as $3, 4$ at a specific timestamp t, the random variable $dist(\tilde{S}_u(t), \tilde{S}_Q(t))$ is defined as drawing a sample from the set $\{dist(1,3), dist(1,4), dist(2,3), dist(2,4)\}$ with uniform probability. In some applications, practically we can only obtain a single record at each time instant. To respect such practical constraints, in this paper, we only assume a single observation at each time instant for a time series.

Lian *et al.* [11] and Yeh *et al.* [12] model time series with random errors as we do in Fig. 1. They assume that the probability distribution of the random variable at each timestamp is unknown, and the mean and variance values are known in advance. They approximate the density distribution of $\tilde{D}(\tilde{S}_u(t), \tilde{S}_Q(t))$ by a normal distribution according to the *Central Limit Theorem* [13], and apply different methods to speed up searching for qualified candidates. Yeh *et al.* [12] further extend the similarity queries to streaming uncertain time series summarized with wavelet synopses. In both studies, however, the authors do not consider estimating the mean and the variance values by practical methods. In this paper, we not only consider how to make the estimation, but also the accuracy of the estimation.

Smruti and Karin [14] design a distance measurement that depends only on the difference of two uncertain time series. The measurement converges to the Euclidean or dynamic time warping distance when the magnitude of errors (uncertainty) is small. The computation involves the distribution of the errors that should be known in advance. They do not describe the variance of the distance measurement. In contrast, our approach focuses on the mean distance and controls the type I errors of the distance measurement. More importantly, our method does not rely on any knowledge of the distribution of the uncertainty.

We use the mean distance in this paper, which is important in practice and also used in some other studies, such as [15].

III. Mean Distance Estimation and Its Reliability

In this section, we first discuss how to measure the mean distance between two time series using a difference-based estimator. Then, we demonstrate how to find the variance and the asymptotic distribution of the mean distance estimator. The statistics derived will be used to process queries with hypothesis testing and control the type I error rate.

A. The Mean Distance Estimator

Assuming that $\mu_Q(t)$ and $\mu_u(t)$ are uniformly bounded, and $\tilde{\varepsilon}_Q$ and $\tilde{\varepsilon}_u$ are i.i.d. distributed, we establish the following approximation when the length of time series l is large:

$$\frac{D(S_Q, S_u)}{l} \approx \frac{D(\mu_Q, \mu_u)}{l} + V_u^2 + V_Q^2, \tag{6}$$

where $D(S_Q, S_u)$ is the distance between observations of $\tilde{S}_Q$ and $\tilde{S}_u$.

Proof: To show Eq. (6), we decompose

$$(\tilde{S}_Q(t) - \tilde{S}_u(t)) = (\mu_Q(t) + V_Q\tilde{\varepsilon}_Q(t) - \mu_u(t) - V_u\tilde{\varepsilon}_u(t))$$

into two parts: the certain part $\mu(t) = \mu_Q(t) - \mu_u(t)$, and the random part $\tilde{\varepsilon}(t) = V_Q\tilde{\varepsilon}_Q(t) - V_u\tilde{\varepsilon}_u(t)$. We assume that the certain part is bounded, that is, $\forall t$, $\mu(t) \leq M$, where M is a constant, and $\tilde{\varepsilon}(t)$ is i.i.d. Then we evaluate the variance of the distance between two uncertain time series as follows.

$$\begin{aligned} & Var\left(\frac{D(\tilde{S}_Q, \tilde{S}_u)}{l}\right) \\ = & \frac{1}{l^2}\sum_{t=1}^{l} Var\left[(\mu(t) + \tilde{\varepsilon}(t))^2\right] \quad \because \tilde{\varepsilon}(t) \text{ is i.i.d.} \\ = & O(\frac{1}{l}) \quad \because \mu(t) \text{ is bounded.} \end{aligned}$$

According to *Chebyshev's inequality*,

$$\begin{aligned} & Pr\left(\left\|\frac{D(\tilde{S}_Q, \tilde{S}_u)}{l} - E(\frac{D(\tilde{S}_Q, \tilde{S}_u)}{l})\right\| \geq \delta\right) \\ \leq & \frac{Var\left(\frac{D(\tilde{S}_Q, \tilde{S}_u)}{l}\right)}{\delta^2} \\ = & O(\frac{1}{l\delta^2}). \end{aligned}$$

As l approaches infinity, the expression in the big O function approaches 0. By the definition of *convergence in probability* and the computational formula for the variance, we can obtain

$$\frac{D(\tilde{S}_Q, \tilde{S}_u)}{l} \xrightarrow{p} E(\frac{D(\tilde{S}_Q, \tilde{S}_u)}{l}) = \frac{D(\mu_Q, \mu_u)}{l} + V_Q^2 + V_u^2.$$

This can be used to obtain Eq. (6):

$$\frac{D(S_Q, S_u)}{l} \approx \frac{D(\mu_Q, \mu_u)}{l} + V_u^2 + V_Q^2.$$

■

Since V_u^2 and V_Q^2 are unknown in practice, we estimate them using a non-parametric difference-based estimator proposed by von Neumann [16] and Rice [17] as follows.

We assume that μ_u varies smoothly, that is,

$$\mu_u(t) \approx \mu_u(t-1) \text{ and } \mu_Q(t) \approx \mu_Q(t-1). \tag{7}$$

This assumption holds on many applications, such as hourly temperature readings from sensors. We then have the following equation based on Definition 1.

$$V_u \times (\tilde{\varepsilon}_u(t) - \tilde{\varepsilon}_u(t-1)) = \tilde{S}_u(t) - \tilde{S}_u(t-1). \tag{8}$$

By squaring Eq. (8) and following the *Law of Large Number*, where the average of samples drawn from a large number of trials should be close to the expected value, V_u^2 and V_Q^2 can then be estimated by

$$\begin{aligned} \hat{V}_u^2 &= \frac{1}{2l}\sum_{t=1}^{l}(S_u(t) - S_u(t-1))^2, \\ \hat{V}_Q^2 &= \frac{1}{2l}\sum_{t=1}^{l}(S_Q(t) - S_Q(t-1))^2. \end{aligned} \tag{9}$$

Based on Eq. (6) and Eq. (9), we can estimate the mean distance $D(\mu_u, \mu_Q)$ by the following estimator.

$$\begin{aligned}&\hat{D}\left(\mu_u, \mu_Q\right)\\ =\ & \sum_{t=1}^{l}\left(S_u(t)-S_Q(t)\right)^2-\frac{1}{2}\sum_{t=1}^{l}\left(S_u(t)-S_u(t-1)\right)^2\\ &-\frac{1}{2}\sum_{t=1}^{l}\left(S_Q(t)-S_Q(t-1)\right)^2. \end{aligned} \quad (10)$$

From Eq. (10), we can see the main advantage of the mean distance estimator $\hat{D}\left(\mu_u, \mu_Q\right)$: it is parameter-free and simple, and needs only the observation values. The reliability of this estimator depends on Eq. (7) and the i.i.d. assumption of $\tilde{\varepsilon}_u(t)$ and $\tilde{\varepsilon}_Q(t)$. Next we show how to measure the reliability by computing the variance of the mean distance estimator.

B. Variance and Asymptotic Distribution of the Mean Distance Estimator

In order to understand the reliability of the mean distance estimator $\hat{D}\left(\mu_u, \mu_Q\right)$, we compute its variance.

According to Eq. (10), the estimator is defined as a function of $S_u(t)$ and $S_Q(t)$, the observation values of $\tilde{S}_u(t)$ and $\tilde{S}_Q(t)$, respectively. To evaluate its variance, we replace the observations S_u and S_Q by their corresponding random variables $\tilde{S}_u$ and $\tilde{S}_Q$, respectively. As a result, we have a variable $\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)$ that models all possible values of $\hat{D}\left(\mu_u, \mu_Q\right)$.

Theorem 1: The variance of $\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)$ can be evaluated using the following estimator.

$$\begin{aligned}\hat{Var}\left(\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)\right) = \quad & 4\left(\hat{V}_u^2+\hat{V}_Q^2\right)\hat{D}\left(\mu_u, \mu_Q\right)\\ &+l\cdot\left(\hat{V}_u^4+4\hat{V}_u^2\hat{V}_Q^2+\hat{V}_Q^4\right), \end{aligned} \quad (11)$$

where $\hat{V}_u^2$ and $\hat{V}_Q^2$ are defined in Eq. (9), and $\hat{D}(\mu_u, \mu_Q)$ is defined in Eq. (10).

Proof: According to Eq. (10), we rewrite the estimator variable in a vector form as follows.

$$\begin{aligned}&\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)\\ =\ & \begin{pmatrix} 1 & -1 & -1 \end{pmatrix}\begin{pmatrix} \sum_{t=t_s}^{t_e}\left(\tilde{S}_u(t)-\tilde{S}_Q(t)\right)^2\\ \frac{1}{2}\sum_{t=t_s}^{t_e}\left(\tilde{S}_u(t)-\tilde{S}_u(t-1)\right)^2\\ \frac{1}{2}\sum_{t=t_s}^{t_e}\left(\tilde{S}_Q(t)-\tilde{S}_Q(t-1)\right)^2 \end{pmatrix}\\ =\ & \begin{pmatrix} 1 & -1 & -1 \end{pmatrix}\mathbb{X}. \end{aligned} \quad (12)$$

Using Eq. (12) and following the basic property of covariance matrix, the variance of $\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)$ can be written as

$$Var\left(\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)\right) = \begin{pmatrix} 1 & -1 & -1 \end{pmatrix} Cov(\mathbb{X}) \begin{pmatrix} 1\\ -1\\ -1 \end{pmatrix}. \quad (13)$$

Based on Eq. (7) and the assumption that $\tilde{\varepsilon}_u(t)$ and $\tilde{\varepsilon}_Q(t))$ are i.i.d., we can derive the covariance matrix $Cov(\mathbb{X}) = E\left((\mathbb{X}-E\mathbb{X})(\mathbb{X}-E\mathbb{X})^T\right)$. As a result, Eq. (13) is

$$\begin{aligned}Var\left(\tilde{\hat{D}}\left(\mu_u, \mu_Q\right)\right) \ =\ & 4(V_u^2+V_Q^2)D(\mu_u, \mu_Q)\\ &+l\cdot\left(V_u^4+4V_u^2V_Q^2+V_Q^4\right). \end{aligned} \quad (14)$$

By substituting V_u, V_Q and $D(\mu_u, \mu_Q)$ with their estimators in Eqs. (9) and (10), respectively, we prove the theorem. ■

Apparently, the variance is easy to compute according to Theorem 1.

To control the type I error rate, we need to investigate the asymptotic distribution of $\tilde{\hat{D}}(\mu_Q, \mu_u)$ as well.

Theorem 2: Suppose 1) $\mu_Q(t)$ and $\mu_u(t)$ are uniformly bounded, 2) $\tilde{\varepsilon}_Q(t)$ and $\tilde{\varepsilon}_u(t)$ are i.i.d. and uniformly bounded random variables, and 3) $\mu_Q(t) \approx \mu_Q(t-1)$ and $\mu_u(t) \approx \mu_u(t-1)$. Let l be the length of the time series. Then,

$$\lim_{l\to\infty}\frac{\tilde{\hat{D}}(\mu_Q, \mu_u)-D(\mu_Q, \mu_u)}{\sqrt{l}} = N(0, \frac{Var\left(\tilde{\hat{D}}(\mu_Q, \mu_u)\right)}{l}), \quad (15)$$

where $N(\cdot)$ is the normal distribution.

Proof: According to the assumption defined in Section II, Eq. (7), and Eq. (10), $\hat{D}(\mu_Q, \mu_u)$ can be written as

$$\begin{aligned}&\hat{D}\left(\mu_u, \mu_Q\right)\\ =\ & \sum_{t=1}^{l}\left(\tilde{S}_u(t)-\tilde{S}_Q(t)\right)^2-\frac{1}{2}\sum_{t=1}^{l}\left(\tilde{S}_u(t)-\tilde{S}_u(t-1)\right)^2\\ &-\frac{1}{2}\sum_{t=1}^{l}\left(\tilde{S}_Q(t)-\tilde{S}_Q(t-1)\right)^2\\ \approx\ & \sum_{t=1}^{l}\left(\mu_Q(t)-\mu_u(t)\right)^2\\ &+2\sum_{t=1}^{l}\left(\mu_Q(t)-\mu_u(t)\right)\left(\tilde{\varepsilon}_Q(t)-\tilde{\varepsilon}_u(t)\right)\\ &+\sum_{t=1}^{l}\left(\tilde{\varepsilon}_Q(t)-\tilde{\varepsilon}_u(t)\right)^2-\frac{1}{2}\sum_{t=1}^{l}\left(\tilde{\varepsilon}_Q(t)-\tilde{\varepsilon}_Q(t-1)\right)^2\\ &-\frac{1}{2}\sum_{t=1}^{l}\left(\tilde{\varepsilon}_u(t)-\tilde{\varepsilon}_u(t-1)\right)^2\\ \approx\ & D(\mu_Q, \mu_u)+\sum_{t=1}^{l}[2\left(\mu_Q(t)-\mu_u(t)\right)\left(\tilde{\varepsilon}_Q(t)-\tilde{\varepsilon}_u(t)\right)\\ &-2\tilde{\varepsilon}_Q(t)\tilde{\varepsilon}_u(t)]\\ &+\sum_{t=1}^{l}[\tilde{\varepsilon}_Q(t)\tilde{\varepsilon}_Q(t-1)+\tilde{\varepsilon}_u(t)\tilde{\varepsilon}_u(t-1)]. \end{aligned} \quad (16)$$

Let

$$\begin{aligned}Z(t) = 2\left(\mu_Q(t)-\mu_u(t)\right)\left(\tilde{\varepsilon}_Q(t)-\tilde{\varepsilon}_u(t)\right)-2\tilde{\varepsilon}_Q(t)\tilde{\varepsilon}_u(t)\\ +\tilde{\varepsilon}_Q(t)\tilde{\varepsilon}_Q(t-1)+\tilde{\varepsilon}_u(t)\tilde{\varepsilon}_u(t-1). \end{aligned} \quad (17)$$

Then, Theorem 2 holds if $\frac{1}{\sqrt{l}}\sum_{t=1}^{l}Z(t)$ converges in distribution to $N(0, \frac{Var\left(\hat{D}(\mu_Q, \mu_u)\right)}{l})$. Although $Z(t)$ is neither identical

nor independent distributed, the convergence can be proved as follows.

Let

$$k = \left\lfloor l^{1/4} \right\rfloor,$$
$$h = \left\lfloor \frac{l}{k+1} \right\rfloor,$$
$$W(i) = \sum_{t=(i-1)k+i}^{ik+i-1} Z(t), i = 1, 2, ..., h,$$
$$W'(i) = Z(hk+i), i = 1, 2, ..., h, \text{ and}$$
$$R = \sum_{t=hk+h+1}^{l} Z(t). \quad (18)$$

As $(t_e - t_s)$ approaches infinity and according to *Chebshev's inequality*, $\frac{1}{\sqrt{l}}R$ and $\frac{1}{\sqrt{l}}\sum_{i=1}^{h} W'(i)$ converge in probability to 0. Note that $W(i)$ is an independent sequence that satisfies *Lindeberg's Condition*, since $Z(t)$ is a uniformly bounded random variable. The key point is that $Pr(|W(i)| \geq O(l^{1/4})) = 0$ and $Var\left(\sum_{i=1}^{h} W(i)\right) \geq O(l)$. After applying the *Central Limit Theorem* and adjusting the coefficient, we have that $\frac{1}{\sqrt{l}}\sum_{i=1}^{h} W(i)$ converges in distribution to $N(0, \frac{Var(\hat{D}(\mu_Q,\mu_u))}{l})$. Finally,

$$\frac{1}{\sqrt{l}}\sum_{t=1}^{l} Z(t) = \frac{1}{\sqrt{l}}\sum_{i=1}^{h} W(i) + \frac{1}{\sqrt{l}}\sum_{i=1}^{h} W'(i) + \frac{R}{\sqrt{l}}.$$

According to *Slutsky's Theorem*, $\frac{1}{\sqrt{l}}\sum_{t=1}^{l} Z(t)$ converges in distribution to $N(0, \frac{Var(\hat{D}(\mu_Q,\mu_u))}{l})$. The theorem is proved. ■

Please note that, on the one hand, the assumptions in Theorem 2, such as $\tilde{\varepsilon}_Q(t)$ and $\tilde{\varepsilon}_u(t)$ being uniformly bounded, are sufficient, but not necessary. We make such stronger assumptions for the convenience in the proof. On the other hand, the theorem is strong enough for our purpose, since its conditions are satisfied in many applications. For example, the readings of sensors should be uniformly bounded by the storage format.

As $D(\mu_Q, \mu_u)$ is an unknown but deterministic value, we can conclude from Eq. (15) that

$$\hat{D}(\mu_Q, \mu_u) \rightarrow N(0, Var\left(\tilde{\hat{D}}(\mu_Q, \mu_u)\right)). \quad (19)$$

One may wonder, in practice, how close the distribution of

$$\frac{\hat{D}(\mu_Q, \mu_u)}{\sqrt{\hat{Var}\left(\tilde{\hat{D}}(\mu_Q, \mu_u)\right)}} \quad (20)$$

is to a normal distribution under different distributions of errors. To address this concern, we report the following simulation. We generate 1000 samples of $\tilde{\varepsilon}_Q$ and $\tilde{\varepsilon}_u$ of the following four types of distribution: standardized student t-distribution with a degree of freedom 3, standard normal distribution, standardized continuous uniform distribution, and standardized Bernoulli distribution that is an example of discrete random variable, respectively. We set $\mu_u(t) = 0, \mu_Q(t) = 4$ and $V_u = V_Q = 1$ for all types of distributions. In addition, we vary the time series length (l) from 2 to 10000.

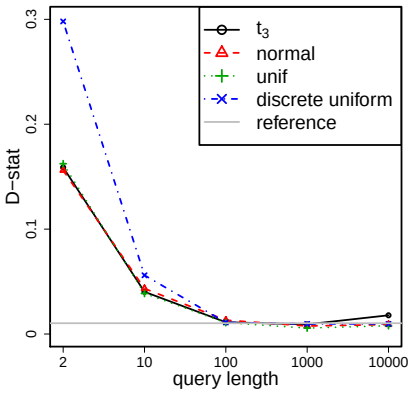


Fig. 4: The D-stat between the empirical distribution of Eq. (20) and $\Phi(\cdot)$ when the length of the series varies.

By applying the *Kolmogorov-Smirnov test*, we show the distance

$$\textit{D-stat} = \sup_x |F(x) - \Phi(x)|$$

of the empirical cumulative distribution function of the 10000 samples of Eq. (20), denoted as $F(x)$, and the cumulative distribution function of standard normal distribution, denoted as $\Phi(\cdot)$. The results are shown in Fig. 4. For reference, in the same figure we also show the distance between the empirical cumulative distribution function of the 10000 samples drawn from a standard normal distributed random variable and the theoretical $\Phi(\cdot)$ values.

As shown in Fig. 4, the distance decreases as the length of time series increases, no matter which distribution of $\tilde{\varepsilon}_Q$ and $\tilde{\varepsilon}_u$ is. In fact, there is no significant difference when the length of time series is over 100. Note that although the student t-distribution and normal distribution are not uniformly bounded, which is an assumption in Theorem 2, Fig. 4 shows that the asymptotic distribution of $\tilde{\hat{D}}(\mu_Q, \mu_u)$ can still be well approximated by the standard normal distribution.

Based on Eq. (11) and the above theoretical and empirical analysis, we can obtain the asymptotic distribution, which can be modeled as a normal distribution, of the mean distance estimator under different types of uncertainty errors. The information is the key concept used to control the type I error rate of the query, which we will show in the next section.

IV. Query Processing

In many existing methods for distance-based queries on certain time series, finding a lower bound on the distance between two time series plays an important role, since the lower bound can help to guarantee a zero type I error rate , that is, not missing any qualified time series in the answer set. For queries on uncertain time series, missing some qualified

time series is inevitable due to the unknown uncertainty. To solve the problem, we propose to control the type I error rate to a given upper bound. Theorem 2 shows that the difference between the estimated distance and the true distance tends to be a normal distribution, which can be leveraged to find the confidence interval of the estimated mean distance and the lower bound of the confidence interval.

A. Exact Match Queries

As stated in Definition 7, we need to find the lower bound, $LCI(D(\mu_u, \mu_Q))$, of the confidence interval of $D(\mu_u, \mu_Q)$ given a confidence level of $1-\alpha$. According to Theorem 2, we can compute the confidence interval with a $1-\alpha$ confidence level approximately as follows.

Suppose the time series length l is large enough. We have

$$Pr\left(\frac{\hat{D}(\mu_Q,\mu_u) - D(\mu_Q,\mu_u)}{\sqrt{Var\left(\hat{D}(\mu_Q,\mu_u)\right)}} \geq \Phi^{-1}(1-\alpha)\right) \approx \alpha, \quad (21)$$

where Φ is the cumulative distribution function of the standard normal distribution. After transposing the equation, we have

$$\begin{aligned} &Pr(X) \approx \alpha, \quad (22)\\ &\text{where } X = \quad D(\mu_Q,\mu_u) \leq \hat{D}(\mu_Q,\mu_u)\\ &\qquad\qquad - \quad \Phi^{-1}(1-\alpha)\sqrt{Var\left(\hat{D}(\mu_Q,\mu_u)\right)} \end{aligned}$$

In practice, we do not know the variance $Var\left(\hat{D}(\mu_u,\mu_Q)\right)$. Thus, we replace it by its estimators in Eq. (11). That is, $LCI(D(\mu_Q,\mu_u))$ corresponding to the $1-\alpha$ confidence level is defined as

$$\begin{aligned} &LCI(D(\mu_Q,\mu_u))\\ =\ &\hat{D}(\mu_Q,\mu_u) - \Phi^{-1}(1-\alpha)\sqrt{\hat{Var}\left(\hat{D}(\mu_Q,\mu_u)\right)} \quad (23) \end{aligned}$$

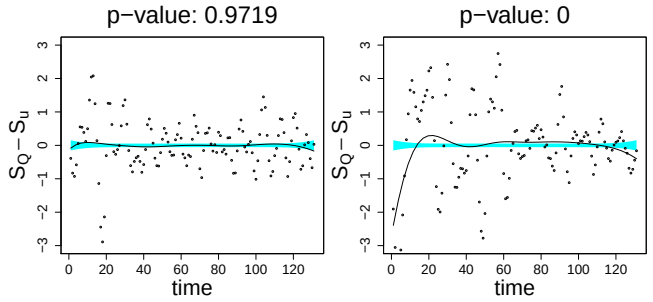
With Eq. (23), we process the exact match query by retrieving all uncertain time series $\tilde{S}_u$ that satisfy the inequality Eq. (4).

In the following theorem, we prove that the type I error rate of the exact match is controlled to α.

Theorem 3: When the length of the time series is large enough, the type I error rate of the exact match query is up to α if we retrieve all uncertain time series $\tilde{S}_u$ that satisfy $LCI(D(\mu_Q,\mu_u)) \leq 0$.

Proof: According to Definition 4, an exact match query wants $\tilde{S}_u$ having $D(\mu_Q,\mu_u) = 0$. As a result, the type I error rate is

$$\begin{aligned} &Pr\left(LCI(D(\mu_Q,\mu_u)) > 0\right)\\ \approx\ &Pr\left(\hat{D}(\mu_Q,\mu_u) - \Phi^{-1}(1-\alpha)\sqrt{Var\left(\hat{D}(\mu_Q,\mu_u)\right)} > 0\right)\\ =\ &Pr\left(\frac{\hat{D}(\mu_Q,\mu_u)}{\sqrt{Var\left(\hat{D}(\mu_Q,\mu_u)\right)}} > \Phi^{-1}(1-\alpha)\right)\\ \approx\ &\alpha. \quad (24) \end{aligned}$$



(a) (b)

Fig. 5: Two examples of the "no-effect" test.

Therefore, we can approximately control the type I error rate of the exact match query. ■

B. Threshold Similarity Query

Similarly, we use the same lower bound defined in Eq. (23) to process the threshold queries defined in Definition 7. We prove its type I error rate control as well.

Theorem 4: When the length of the reference time series is large enough, the type I error rate of the threshold query is no more than α if we retrieve all time series $\tilde{S}_u$ that satisfy $LCI(D(\mu_Q,\mu_u)) \leq r$, where r is a user specified distance threshold.

Proof: Under condition $D(\mu_Q,\mu_u) \leq r$, the type I error rate is

$$\begin{aligned} &Pr\left(LCI(D(\mu_Q,\mu_u)) > r\right)\\ \approx\ &Pr\left(\frac{\hat{D}(\mu_Q,\mu_u) - r}{\sqrt{Var\left(\hat{D}(\mu_Q,\mu_u)\right)}} > \Phi^{-1}(1-\alpha)\right)\\ \leq\ &Pr\left(\frac{\hat{D}(\mu_Q,\mu_u) - D(\mu_Q,\mu_u)}{\sqrt{Var\left(\hat{D}(\mu_Q,\mu_u)\right)}} > \Phi^{-1}(1-\alpha)\right)\\ \approx\ &\alpha. \end{aligned} \quad (25)$$

Therefore, the type I error rate of the threshold query is also under controlled. ■

V. Experiment Results

We conducted extensive experiments on 20 real data sets in the UCR time series data repository [2] to evaluate MISQ. All experiments were run on a PC with an Intel(R) Core(TM) i7 3.07 GHz CPU and 12GB RAM using R 2.12.1 [18].

A. Settings

We compared MISQ with the confidence band method on exact queries, and with the *moving average* method on both exact and threshold queries.

The idea of the confidence band method works as follows. Given two time series with their observations S_Q and S_u,

we test if $\mu_Q - \mu_u$ is equal to 0 or not. That is, the null hypothesis is set to $\mu_Q - \mu_u = 0$. We applied the testing for no effect in nonparametric regression via kernel smoothing techniques. *No effect* here refers to $\mu_Q - \mu_u$ always being 0. The testing procedure first smooths the observations of $S_Q(t) - S_u(t)$, where $t = 1, 2, \ldots, l$ to estimate $\mu_Q - \mu_u$. Then, it estimates the variance of the noise. Finally, it evaluates the corresponding p-value.

Fig. 5 shows an example. The estimated $\mu_Q - \mu_u$ is plotted with a black solid line, while the light shade area is the confidence band of the null hypothesis. In Fig. 5(a), the confidence band contains the whole solid line, which indicates that the test is significant and with a high p-value. If the confidence band does not contain the solid line, as shown in Fig. 5(b), the p-value tends to be small and the null hypothesis can be rejected with a high confidence. For more details, please refer to chapters 3-5 in [19]. We use the default settings and implementation [20] (the built-in function *sm.regression*) in language R to do the testing in our experiments. A time series $\tilde{S}_u$ is retrieved if the p-value is greater than α.

We also compared MISQ with the moving average method that is a widely used estimator for computing the mean and removing noises on time series. It basically computes the moving average of a set of consecutive observations within a given bandwidth. Then, the mean distance between two time series can be computed using the two series of mean values. Since an appropriate bandwidth is critical for the performance of the moving average method, we implemented two methods selecting the bandwidth, one using cross validation and the other using a fixed bandwidth. Given a set of time series and a query series, the method of cross validation is to select the bandwidth that minimizes the leave-one-out residual sum of square (RSS) as suggested by [21]. In the other method, we just picked 5 as the bandwidth, which is the default value used in MATLAB [22]. We denote them as movavg_cv and movavg_5, respectively. It is noted that the original moving average method cannot control the type I error rate, since it computes the mean only. To be fair, we used the residual sum of square between the mean value and the original observations normalized by the time series length to estimate V_u for each time series. Then, we used the same way in MISQ to control the type I error rates for the moving average methods.

We did not compare MISQ with the methods in [10–12, 14]. As discussed in Section II-B, those methods assume more information than MISQ, and simply cannot work in our problem setting.

Section I shows that reducing random errors makes a difference in real applications (Fig. 3). Since we focus on the quality and accuracy of similarity search in this paper, in the following evaluation, we first compare the type I error rates and the type II error rates of MISQ and the other two methods. If the error rate is no more than the given α, we call it *under-control*. In addition, the α value is set to 0.05 by default. Then, in the last subsection, we compare the distance computation time between a pair of time series of MISQ and the moving average methods.

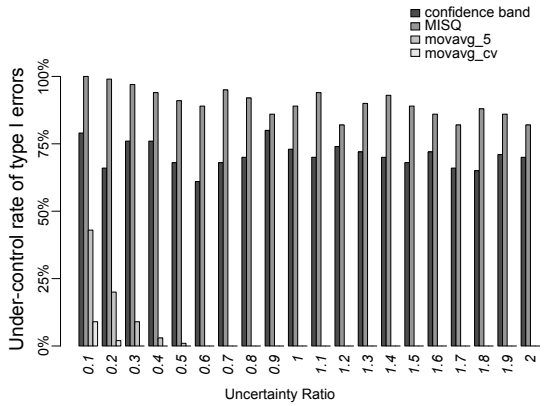


Fig. 6: Impact of uncertainty ratio on the percentage of under control considering the type I errors.

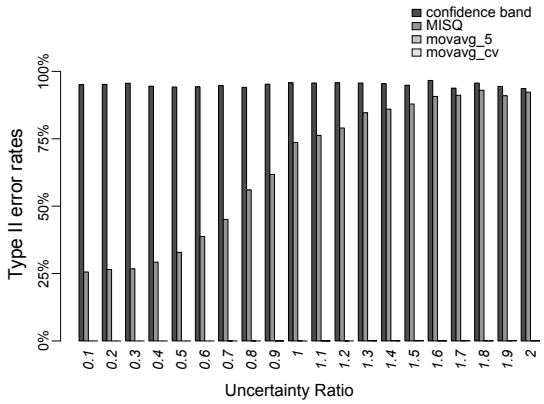
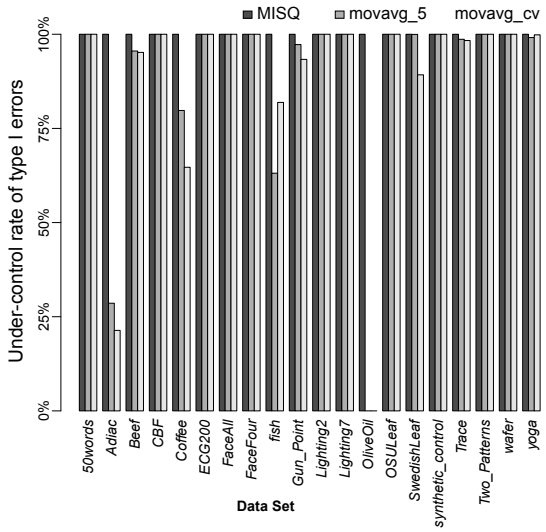


Fig. 7: Impact of uncertainty ratio on the type II errors.

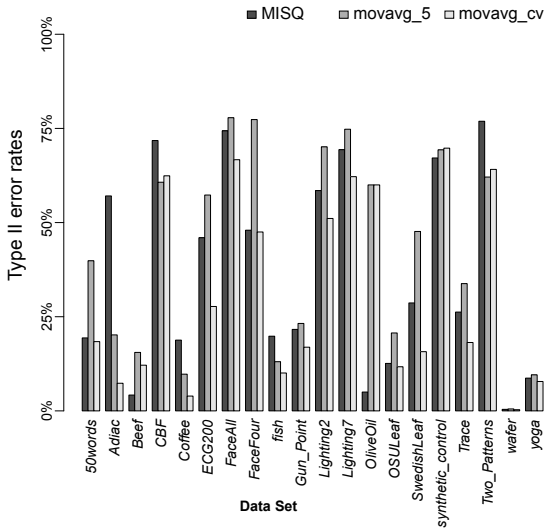
B. Exact Match Queries

We evaluated the performance on exact queries as follows. Given a real time series, we treated its original value as the mean value without noise, i.e., a sequence of μ_u. Then, we generated 101 independent series based on this one with additionally noise added each timestamp to blur the time series as our observations. The noise is normally distributed with a variance proportional to the variance of μ_u, ranging from 0.1 to 2. We call this *uncertainty ratio* hereafter. In this way, we have 101 observations whose mean values are identical to each other. From these 101 observations, we pick one as the query series time series $\tilde{S}_Q$ and the rest are the candidates. Under this settings, any tested methods should retrieve all the 100 candidates.

For each data set, we pick 5 time series and repeat the procedure above. As there are 20 data sets, we have 100 queries in total. If a query is processed with the type I error rate no more than the given α, we call it is *under-control*. We counted the under control rate of type I errors, and the results are shown in Fig. 6. MISQ controlled the type I error rate better than the confidence band method. Movavg_5 and movagv_cv hardly controlled the type I error rates.

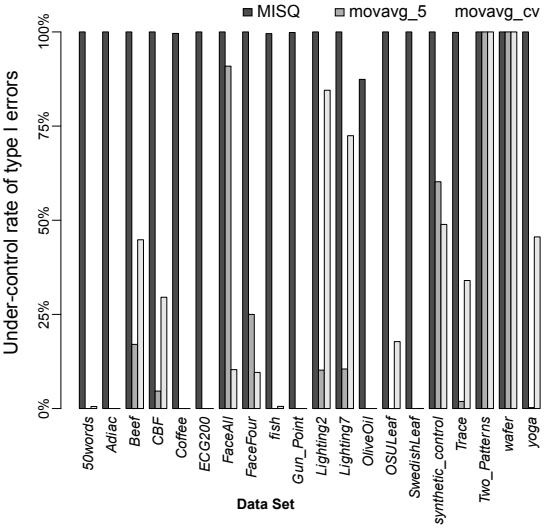


(a) The percentage of type I errors under control on different datasets

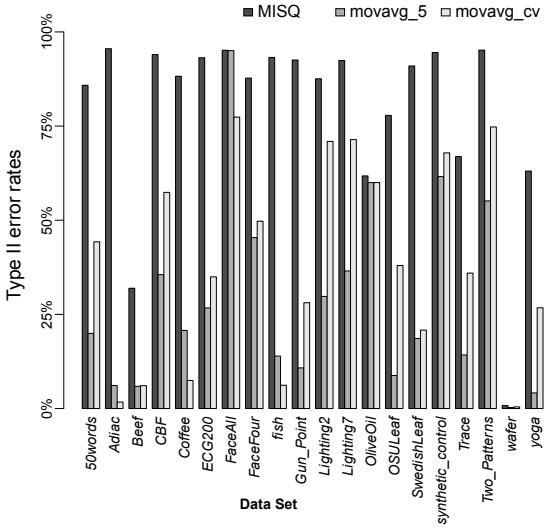


(b) The type II error rate on different datasets

Fig. 8: Normally distributed noise with uncertainty ratio=0.2.



(a) The percentage of type I error under control on different datasets

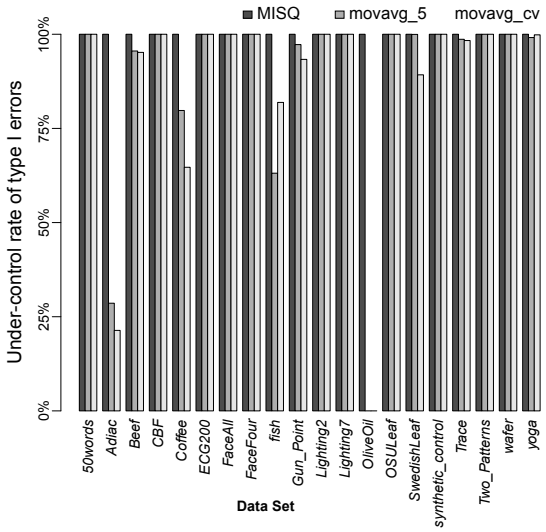


(b) The type II error rate on different datasets

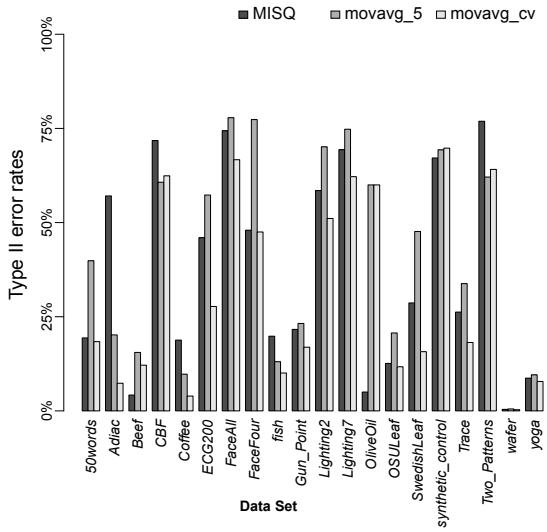
Fig. 9: Normally distributed noise with uncertainty ratio=2.

For the type II errors, the settings were the same except for one difference: we let $\mu_u(t) = \mu_Q(t) + 0.5 \times \sqrt{Var(\mu_Q)}$. Under this setting, the mean of the reference time series was different from the mean of all the candidates series, so the tested methods shall retrieve no candidate. In this way, every returned results from a method was regarded as a type II error. From Fig. 7, we see that MISQ outperformed the confidence band method. As the uncertainty ratio increases, the type II error rate of MISQ increases as well. The type II error rates of movavg_5 and movagv_cv were very small.

To understand the results, we note that MISQ directly tested if $D(\mu_Q, \mu_u) = 0$ while the confidence band method estimated the curve of $\mu_u - \mu_Q$ instead. Since the procedure of the confidence band method was indirect, its performance was affected by parameters such as the bandwidth selector. Consequently, it has poor type I and type II error rates. Interestingly, we see the trade-offs between type I error rate and type II error rate of MISQ and the moving average method. The very low type II error rate of the moving average methods was attributed to the poor performance in type I error rate control, which failed to meet our goal. MISQ sacrificed the type II error rate to control the type I error rate, since a low type I error rate is more important in many applications.

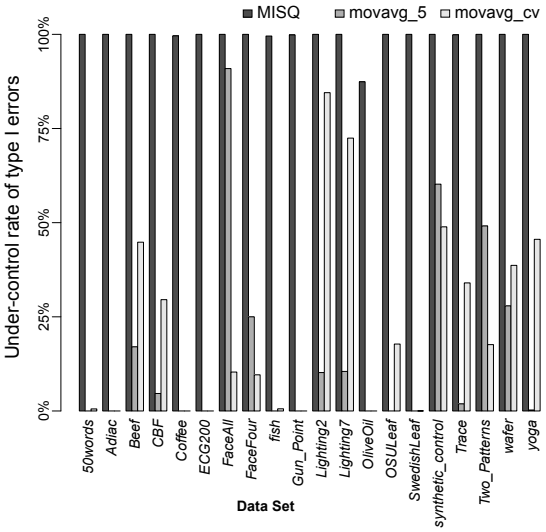


(a) The percentage of type I error under control on different datasets

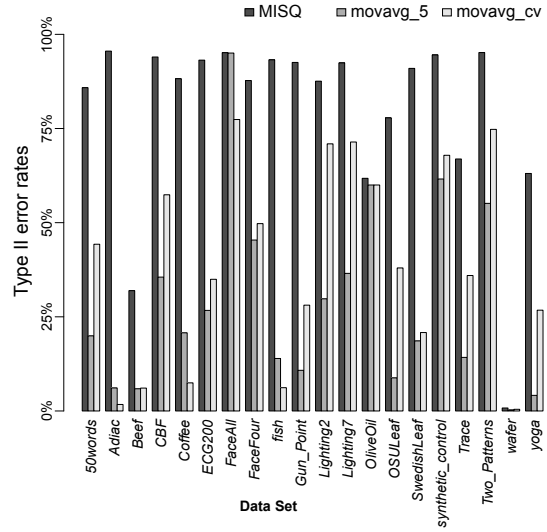


(b) The type II error rate on different datasets

Fig. 10: Uniformly distributed noise with uncertainty ratio=0.2.



(a) The percentage of type I error under control on different datasets



(b) The type II error rate on different datasets

Fig. 11: Uniformly distributed noise with uncertainty ratio=2.

C. Threshold Similarity Queries

For threshold queries, we only compared MISQ with movavg_5 and movagv_cv since the confidence band method is not able to be applied in the case when $r \neq 0$.

We introduced the uncertainty to the time series data as follows. For each time series at each timestamp, it was blurred by an i.i.d. random noise with a variance proportional to the variance of the original time series, which is the same as in the exact query experiments. To show that the proposed method is not limited by a specific distribution of the noise, we test the normal distributed and uniformly distributed noise.

The parameters used were set as follows. For each data set, we divided it into the training and the test sets as originally defined in the UCR time series data. The reference time series, i.e., the query series S_Q, was chosen from the testing set and the candidates are all members in the training dataset. The distance bound r is chosen from the $0.1, 0.2, ..., 0.9$-quantiles of the Euclidean distances between the reference time series and the candidates. For under control rate, the confidence level, i.e., $1 - \alpha$,is set to 0.95.

First, we compare the *under-control rate of type I errors* of the three methods on 20 data sets. For a query time series from the testing set, if the type I error rate is no more than the given α, we counted is as an under-control. The under-

control rate is thus the percentage of queries from the testing set that was under-control. Second, the type II error rates were averaged from the query results of all the query time series in the testing data set. Finally, since there is a trade off between the type I error rate and the type II error rate, we additionally interpolated for each method the Equal Error Rate (EER), the error rate where type I error rate equals to the type II error rate, of each method for comparison.

The under-control rate of type I errors and the corresponding type II error rates for both uniform and normal errors under two uncertain ratio values, 0.2 and 2, of each data set are shown in Fig. 8, Fig. 9, Fig. 10 and Fig. 11, respectively. From these figures, we can see that MISQ outperformed the other methods in controlling the type I error rate significantly. This supported the theoretical results derived in Section III and Section IV. For the moving average method, sometimes movavg_cv worked better while the other times movavg_5 did. This showed that the quality of moving average depended on the bandwidth selection heavily. Moreover, as the uncertainty ratio was large, e.g., when uncertainty ratio=2.0, the under-control rate of type I errors of both two moving average methods decreased significantly while MISQ still guaranteed a 100% control. On the other hand, MISQ sacrificed more type II error rates only when the uncertain ratio was large, such as 2. Note that the uncertainty ratio = 2 was quite large since it means that the variation of noise is twice that of the mean values of a time series. At a small uncertain ratio, the type II error rate of MISQ was even smaller than that of the moving average method in some data sets.

In addition, MISQ achieved good under control rates (100%) all the time in our experiments, no matter which noise distribution was used. However, the moving average method had high under-control rates of type I error for normally distributed noise but worked poorly for the uniform noise (See data sets Two_Pattern and wafer in Fig. 9(a) and Fig. 11(b).)

As there was a trade-off between the type I error rate and type II error rate, we further calculated the EER values of all methods. At different uncertainty ratios, we counted the best (smallest) EER among the 20 data sets of each method and plotted the results in Fig. 12. We can see that MISQ performed best even when the uncertainty ratio was getting larger. The movavg_5 worked better than movavg_cv when the uncertainty ratio became larger, but still it did not surpass MISQ. Therefore, we can conclude that the MISQ is the best method out of the three on the threshold query with type I error rate controlled.

D. Computation Time

Here we compared the computation cost of each method to show their efficiency. We show the computation cost on each data set of all methods for threshold queries. For each dataset, we compute the averaged processing time for each method to get the answers of a query. Using the longest processing time as a denominator, we computed the percentage of the processing time of the other two methods. The results were shown in Fig. 13. Movavg_5 was the fastest method as it just computed the average of constant number of points while

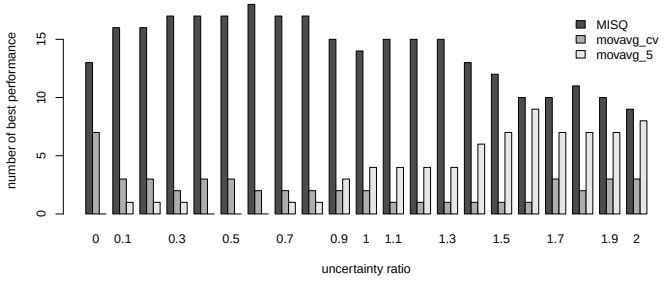


Fig. 12: The distribution of the best EER among all data sets at different uncertainty ratios.

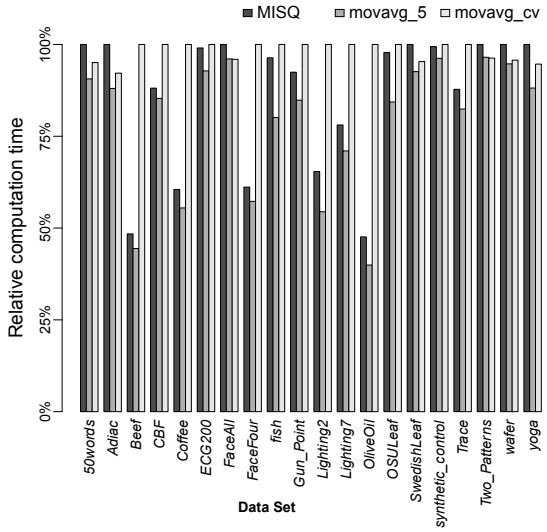


Fig. 13: The average query processing time.

movavg_cv took more time to decide the bandwidth. MISQ spent more time in computing the mean distance estimators and its variance. However, MISQ is never substantially slower than any faster methods. It shows that MISQ is efficient to control type I error rate well.

VI. Conclusions

In this paper, we presented MISQ, a statistical approach to processing similarity search on time series aiming at reducing random errors. Without any prior knowledge about the error distribution hidden in the time series, we estimated the mean distance between time series with only one observation at a timestamp. Through the statistical hypothesis testing, we provided a solution to computing exact match queries and threshold similarity queries with type I error rate effectively controlled. The results of extensive experiment on real data sets showed that, when comparing with the confidence band method and the moving average method, MISQ outperformed the two in controlling type I error rates at a cost of reasonable type II error rates. The runtime cost is low.

For future work, first we can extend MISQ to process k nearest neighbor (kNN) queries. Based on the same mean distance estimator and its variance estimator, we can compare which time series is closer/farther to the query time series in a pairwise manner. The k nearest neighbors are then those having less-than-k time series that are significantly closer to the query time series after some *multiple testing* correction. The main challenges are to reduce the significantly high type II error rates and the retrieving efficiency. Another direction is to apply the same ideas in MISQ to compute the dynamic time warping distance on time series, which is another important and widely used distance measurement in many applications, such as speech recognition.

ACKNOWLEDGMENT

The work was supported in part by the National Science Council of Taiwan, R.O.C., under Contracts NSC99-2221-E-001-010, and an NSERC Discovery Grant project. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] "Random error - from wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Random_error

[2] E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana, "The ucr time series classification/clustering," 2006. [Online]. Available: http://www.cs.ucr.edu/~eamonn/time_series_data/

[3] R. Agrawal, C. Faloutsos, and A. N. Swami, "Efficient similarity search in sequence databases," in *Proc. of FODO*, 1993, pp. 69–84.

[4] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in *Proc. of ACM SIGMOD*, 1994, pp. 419–429.

[5] F.-P. Chan, A.-C. Fu, and C. Yu, "Haar wavelets for efficient similarity search of time-series: with and without time warping," *IEEE TKDE*, vol. 15, no. 3, pp. 686 – 705, 2003.

[6] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim, "Fast similarity search in the presence of noise, scaling, and translation in time-series databases," in *Proc. of VLDB*, 1995, pp. 490–501.

[7] R. Weber, H.-J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *Proc. of VLDB*, 1998, pp. 194–205.

[8] B.-K. Yi, H. V. Jagadish, and C. Faloutsos, "Efficient retrieval of similar time sequences under time warping," in *Proc. of IEEE ICDE*, 1998, pp. 201–208.

[9] B.-K. Yi and C. Faloutsos, "Fast time sequence indexing for arbitrary lp norms," in *Proc. of VLDB*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 385–394.

[10] J. Aßfalg, H.-P. Kriegel, P. Kröger, and M. Renz, "Probabilistic similarity search for uncertain time series," in *Proc. of SSDBM*, 2009, pp. 435–443.

[11] X. Lian, L. Chen, and J. X. Yu, "Pattern matching over cloaked time series," in *Proc. of IEEE ICDE*, 2008, pp. 1462–1464.

[12] M.-Y. Yeh, K.-L. Wu, P. S. Yu, and M.-S. Chen, "Proud: A probabilistic approach to processing similarity queries over uncertain data streams," in *Proc. of EDBT*, 2009, pp. 684–695.

[13] E. W. Weisstein, "Central limit theorem," from MathWorld - A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/CentralLimitTheorem.html

[14] S. R. Sarangi and K. Murthy, "Dust: a generalized notion of similarity between uncertain time series," in *KDD*, 2010, pp. 383–392.

[15] C. C. Aggarwal and P. S. Yu, "A framework for clustering uncertain data streams," in *Proc. of IEEE ICDE*, 2008, pp. 150–159.

[16] J. von Neumann, "Distribution of the ratio of the mean square succeessive difference to the variance," *Ann. Math. Statist.*, 1941.

[17] J. Rice, "Bandwidth choice for nonparametric kernel regression," *The Annals of Statistics*, 1984.

[18] R Development Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[19] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series)*. Oxford University Press, USA, Nov. 1997. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0198523963

[20] ——, "R package `sm`: nonparametric smoothing methods (version 2.2-4)," University of Glasgow, UK and Università di Padova, Italia, 2010. [Online]. Available: http://www.stats.gla.ac.uk/~adrian/sm,http://azzalini.stat.unipd.it/Book_sm

[21] J. Racine, "An efficient cross-validation algorithm for window width selection for nonparametric kernel regression," in *COMMUNICATIONS IN STATISTICS: SIMULATION AND COMPUTATION*, 1993.

[22] "Smooth response data - matlab in r2011a mathworks documentation - curve fitting toolbox." [Online]. Available: http://www.mathworks.com/help/toolbox/curvefit/smooth.html