Similarity Search and Analysis Techniques for Uncertain Time Series Data

Mahsa Orang

A Thesis

In the Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy (Computer Science)

Concordia University

Montreal, Quebec, Canada

2015

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Mahsa Orang

Entitled: Similarity Search and Analysis Techniques for Uncertain Time Series Data

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Computer Science)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Fariborz Haghighat

_____ External Examiner
Dr. Davood Rafiei

_____ External to Program
Dr. Jamal Bentahar

_____ Examiner
Dr. Tien Dai Bui

_____ Examiner
Dr. Todd Eavis

_____ Thesis Supervisor
Dr. Nematollaah Shiri

Approved by _____
Dr. Volker Haarslev, Graduate Program Director

2015 _____
Dr. Amir Asif, Dean
Faculty of Engineering & Computer Science

Abstract

Similarity Search and Analysis Techniques for Uncertain Time Series Data

Mahsa Orang, Ph.D.

Concordia University, 2015

Emerging applications, such as wireless sensor networks and location-based services, require the ability to analyze large quantities of uncertain time series, where the exact value at each timestamp is unavailable or unknown. Traditional similarity search techniques used for standard time series are not always effective for uncertain time series data analysis. This motivates our work in this dissertation. We investigate new, efficient solution techniques for similarity search and analysis of both uncertain time series models, i.e., PDF-based uncertain time series (having probability density function) and multiset-based uncertain time series (having multiset of observed values) in general, as well as correlation queries in particular. In our research, we first formalize the notion of normalization. This notion is used to introduce the idea of correlation for uncertain time series data. We model uncertain correlation as a random variable that is a basis to develop techniques for similarity search and analysis of uncertain time series. We consider a class of probabilistic, threshold-based correlation queries over such data. Moreover, we propose a few query optimization and query quality improvement techniques. Finally, we demonstrate experimentally how the proposed techniques can improve similarity search in uncertain time series. We believe that our results provide a theoretical baseline for uncertain time series management and analysis tools that will be required to support many existing and emerging applications.

Acknowledgements

First and foremost, I owe my deepest gratitude to my supervisor, Dr. Nematollaah Shiri for his guidance, inspiration and encouragement throughout my study. His invaluable suggestions and discussions with me helped greatly to shape the ideas in this work. I learned a lot from him, and without his help I could not have finished my dissertation successfully. I would also like to thank my committee members: Drs. Tien Dai Bui, Todd Eavis, Jamal Bentahar and Davood Rafiei for their helpful comments at different stages of my research.

I also want to thank all members of the Computer Science and Software Engineering Department. Special thanks to Dr. Juergen Rilling and Halina Monkiewicz for their kindness and support in every aspect of student life. I would like to also extend my warmest thanks to all of my friends at Concordia University: Zahra Asadi, Bahareh Goudarzi, Wei Hann, Rahmah Brnawy, Soyoung Kim, Si Zhan, Laleh Roosta Pour, Aminata Kane, and Mahsa Mofidpoor.

My deepest appreciation goes to my immediate family and family-in-law, in particular, my parents, Mahshid and Sirous, parents-in-law, Jina and Reza, and my grandparents who gave me the opportunity and encouragement to pursue my dreams. And to my loving husband, Iman, you gave me your unconditional love and support during all these years. This journey would have been meaningless without you.

This research was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada and by Concordia University.

To Iman

Table of Contents

List of Figures

Chapter 1: Introduction

Some existing and emerging applications such as location-based services and wireless sensor networks generate and process uncertain time series, where the exact value at each timestamp is unavailable or unknown. The sources of uncertainty in time series are various and include the following:

1. *Limitations of data collection equipment and techniques, and measurement errors*: e.g., errors in sensor network readings or delays in data transmission to the server due to limited network bandwidth or limited battery power of devices [CER11, CHE03, TRA10]. For example, the positions of moving objects can be tracked using the Global Position System, and updated periodically. However, the exact positions of these objects at different times are uncertain, and are represented as a spatial range at each timestamp.

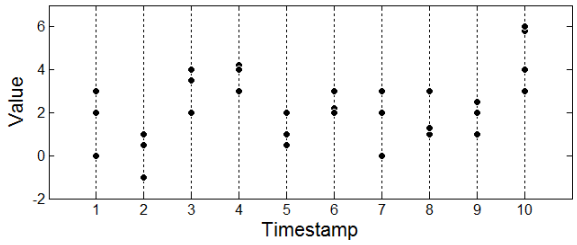
2. *Privacy concerns*: Privacy-preserving methods [AGG08] perturb data in applications such as location-based services [CHE04] and medical data analysis [LIA08]. For example, in the latter, to protect the privacy of patients, medical information is usually represented as an interval to anonymize it.

3. *Forecasting techniques*: For example, in mobile applications, future incoming data is unknown but predicted with some error [AGG09].

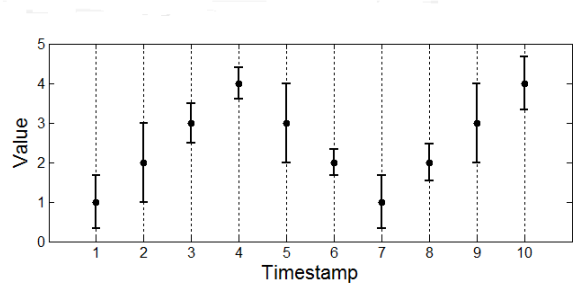
4. *Multiple readings for a measured attribute*: For instance, different sensors may read different temperatures for a specific area [ASF09].

Research in uncertain time series data analysis is relatively new and focused mostly on modeling and similarity search problems [ASF09, DAL12, DALX14, LIA08, RAJ15, SAR10, YEH09]. For the modeling problem, there are two approaches. In the first approach, each timestamp of an uncertain time series is represented as a random variable with a probability density function (PDF) [LIA08, SAR10, YEH09]. In the second approach, each timestamp is represented as a multiset of values with no assumption made or known about the underlying PDF of the data [ASF09, DALX14]. We refer to these two modeling approaches as *PDF-based* and *multiset-based*, respectively.

Figure 1 shows examples of the two modeling approaches. Figure 1 (a) shows a multiset-based uncertain time series, in which we have a multiset of observed values at each timestamp, and have no knowledge of the underlying PDF, dictating the distribution of the observed values. Moreover, the exact value measured may or may not be among these observed values.



a) Multiset-based uncertain time series



b) PDF-based uncertain time series

Figure 1. Modeling approaches for uncertain time series data.

Figure 1 (b) illustrates a PDF-based uncertain time series where, at each timestamp, there is a random variable with known mean (shown by a circle), standard deviation (shown by an interval), and PDF. For example, at timestamp 2, the random variable's mean is 2 and its standard deviation is 1. Now suppose that this random variable follows a normal distribution. In normal distribution, about 99.7% of the values are within three standard deviations [ROS09], which means the (unknown) exact value in this timestamp can vary between -1 and 5. One of the main problems in dealing with uncertain time series is similarity search. Given a dataset D of uncertain time series and an uncertain time series Q as a query, the aim is to search for uncertain time series X in D that is similar to Q, based on the notion of similarity defined. In our research, we study similarity search techniques for both PDF-based and multiset-based models.

1.1. Motivation

Existing applications, such as wireless sensor networks and location-based services, require the ability to analyze large quantities of uncertain time series. Traditional techniques are now inadequate as they were developed for standard time series, which are sequences of real numbers; new concepts and techniques are to be devised and developed for management of uncertain time series to deal with the uncertainty inherent in such data. The research on uncertain time series is new and there is no solution for some of the similarity search topics including correlation. In this work, we will address correlation analysis for uncertain time series.

Correlation analysis techniques have been developed for standard time series data in different application areas such as finance, social sciences, and engineering [NGU08, SHA04, ZHA07]. These techniques look for relationships between standard time series. Similar correlation analysis techniques are required for uncertain time series data in applications such as the following:

1. *Feature selection*: To apply machine learning techniques to uncertain time series efficiently, feature selection is a crucial preprocessing step, which can result in a

significantly reduced learning time. Correlation indicates the degree of dependency of a feature on other features. Using this information, redundant features can be identified and removed from the feature set [Hal00, WAN05].

2. *Data analysis*: Correlation can be used to analyze uncertain time series by looking for relationships among different uncertain time series. For example, suppose that in a network application, sensors are used to record the temperature of different locations. One may want to know which locations have similar temperature in a specific time frame (e.g. during a month). For this, the daily temperature of different locations can be modeled as an uncertain time series and searched for locations for which the temperature time series are highly correlated. This information can be used to analyze the relationships between temperature and climate [ASF09]. As another example in biological sequence data, correlations among microarray time series, which are known to be uncertain data [CHE06], are used to identify potential regulatory relationships among genes [RAZ13].

3. *Prediction*: As correlation can be used to identify the relationships among different uncertain time series, this information can reveal the effect that changes in the values of an uncertain time series may have on values of another uncertain time series.

4. *Pattern matching*: Correlation analysis could help identify similar patterns, moreover, the user may be interested in finding uncertain time series which are not highly correlated, e.g., in the medical domain looking for abnormalities in a test result.

Traditional correlation analysis techniques, such as the Pearson correlation [SHA04], are not adequate for uncertain time series data analysis, because they were developed for standard time series. This is illustrated in the following example. Let $x =< 0,0,0,0,0.01 >$ and $y =< 1,1,1,1,1.01 >$ be two standard time series, which we perturb to obtain the uncertain time series *X* and *Y* shown in Figure 2 (we use upper case letters for random variables and lower case for values

that are real numbers). In Figure 2, the interval at each timestamp indicates the standard deviation of the uncertain value at that timestamp (PDF-based model). Suppose that we are looking for uncertain time series in a given dataset D which are "highly" correlated to X with correlation coefficients more than 0.9. Since the Pearson correlation between x and y is 1, we expect Y to be the answer. However, the Pearson correlation between X and Y, treated as standard time series, is 0.5, and hence Y would not be considered as a so "highly" correlated uncertain time series to X. The observations above motivate this research for the development of new concepts and techniques to capture correlations for uncertain time series.

1.2. Challenges

The following are the challenges researchers faced in the development of the similarity search techniques for uncertain time series.

1.2.1. Modeling Approaches

The first challenge is the representation model of uncertain time series data, which could be PDF-based or multiset-based. Each of which requires its own particular solutions. For example, consider the traditional range query, $Eucl(x, y) \leq d$, where we are looking for standard time

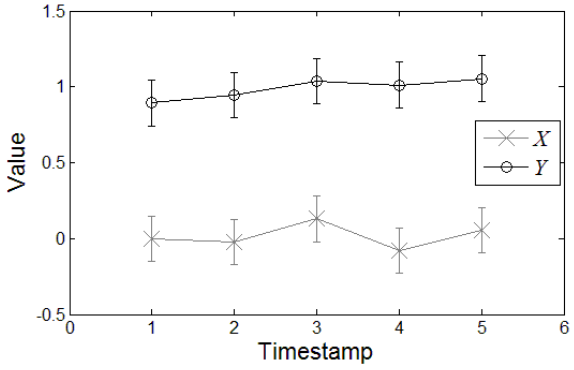


Figure 2. Uncertain time series examples.

series x the Euclidean distance of which to a given time series y is at most d. However, in the context of uncertain time series, this simple query has been addressed in different ways based on different modeling approaches [ASF09, YEH09, WU12].

1.2.2. **Probabilistic Similarity Search**

In search problems over uncertain data [ASF09, CHE03, CHE04, YEH12, DAL14], desired analysis solutions would be probabilistic, i.e., a probability is calculated and assigned to each query result. Similarity search in uncertain time series data can be very challenging due to its probabilistic nature.

1.2.3. **Multiset-based Similarity Search**

Similarity search over multiset-based uncertain time series has been challenging for excess computational cost due to the high dimensionality of uncertain time series data and multiplicity of values at each timestamp. One way to address this problem is to truncate the input uncertain time series to a much shorter length, for instance, 6 timestamps as considered in [DAL12], which limits its applications.

1.3. **Thesis Contributions**

This thesis presents suitable concepts and techniques for uncertain time series correlation analysis, and addresses the above challenges. The contributions are as follows:

- *Normalization and correlation for uncertain time series*: We first formalize the notion of normalization for uncertain time series and then introduce the notion of correlation for uncertain time series. Moreover, to address the second challenge (Section 1.2.2), we introduce a family of probabilistic, threshold-based correlation queries over such data and propose a probabilistic approach as our similarity search. Probabilistic threshold-based correlation queries consist of a dataset D, an uncertain

time series query Q, a correlation threshold c, and a probability threshold p. Given an uncertain time series Q, the goal is to search for every uncertain time series X in D such that its correlation with Q is not less than c with probability at least p.

- *Correlation analysis techniques for both models*: To address the first challenge (Section 1.2.1), we propose suitable concepts and techniques for uncertain time series correlation analysis for both PDF-based and multiset-based models. For each case, we formulate correlation for uncertain time series as a random variable and develop techniques to determine the underlying probability density function. We conducted numerous experiments to evaluate the performance of the proposed techniques under different configurations using real-life datasets. Our numerous experiments indicate that, unlike in the case of standard time series, there is a trade-off between false alarms and hit ratios, which can be controlled by the probability threshold provided by users. Our results also offer users a guideline for choosing proper threshold values.

- *Query optimization for multiset-based similarity search*: To make multiset-based similarity search feasible, i.e., the third challenge (Section 1.2.3), we propose two query optimization methods to speed up the search process. The first one is a probabilistic pruning to cut down the number of candidate results. This includes a Boolean representation technique for uncertain time series. In this representation, each observed value is replaced with a single bit. In addition to saving memory, this enjoys fast bit operations. Using this method, we introduce uncertain Boolean correlation together with an effective probabilistic pruning strategy. Second, we propose a sampling-based heuristic method that approximates the distribution of

uncertain correlation effectively and reduces the computation time significantly. Our experimental results indicate effectiveness of the proposed techniques.

- *Quality improvement of similarity search*: We also study the impact of preprocessing techniques on performance and effectiveness of the similarity measures for uncertain time series. Some existing works on uncertain time series use the same similarity measures developed for standard time series, which we refer to as traditional similarity measures. More recently, a number of new similarity measures have been proposed for uncertain time series, which we refer to as uncertain similarity measures. However, these new measures have been shown to be less effective than the traditional ones. In this work, we show that the performance of uncertain similarity measures can be improved through preprocessing techniques. We establish this through extensive experiments using the UCR benchmark data. Our results indicate that the uncertain similarity measures together with preprocessing outperform the traditional similarity measures.

We believe the proposed ideas and solutions provide a guideline for uncertain time series data analysis and lend themselves to powerful tools for uncertain time series analysis and search tasks.

1.4. Thesis Organization

The rest of this thesis is organized as follows: Chapter 2 reviews work related to uncertain time series similarity search. In Chapter 3, the notion of uncertain normalization and correlation is presented and probabilistic correlation queries are introduced. Chapter 4 provides an overview of the setup for our experiments. Chapter 5 discusses the similarity search for PDF-based and multiset-based uncertain time series. Chapter 6 introduces query optimization techniques including probabilistic pruning and sampling-based heuristic methods. Chapter 7 discusses how

preprocessing methods on uncertain time series improves the performance of existing similarity measures. Finally, Chapter 8 concludes the dissertation and presents the possible future work.

Chapter 2: Background and Related Work

For similarity search on standard time series data, which is a sequence of real numbers, efficient algorithms have been developed [AGG15, FRA15, GON15, SAK15, SHA04, KAD08, TAR14], in which, given a time series q and a set D of such time series, the goal is to identify time series in D that are "similar" to q. For this, a similarity measure is defined that captures the notion of similarity. For standard time series, there are well-known and well-defined similarity measures [WAN13], e.g., Euclidian distance [AGG15], dynamic time warping (DTW) [TAR14], Pearson correlation coefficient [NGU08]. The traditional similarity measures however are inadequate for uncertain time series [SAR10], and new suitable similarity measures are required to be developed. This has been the subject of recent research, which resulted in solution proposals that take into account different models and assumptions about the available information. In what follows, first we provide a background and review of the literature on modeling, and then we introduce our classification of similarity measures on uncertain time series.

2.1. Modeling

An uncertain time series $X =< X_1, \dots, X_n >$ is a sequence of random variables with some "statistical information" representing the uncertainty level of some real number, the exact value of which is unknown or unavailable. In the current literature, we identify two models for representing uncertain time series, PDF-based and multiset-based models, which are explained in the following sections.

2.1.1. PDF-based Model

In this model, each element at each timestamp is represented as a random variable [LIA08, SAR10, YEH09, WU12]. To be more precise, given an uncertain time series, $X =< X_1, \dots, X_n >$, each element X_i $(1 \leq i \leq n)$ can be considered as $X_i = x_i + E_{x_i}$, where x_i is the "exact" value of

the data that is unknown, and E_{x_i} is a random variable representing the error. Although in existing proposals for PDF-based uncertain time series, the random variables of different timestamps are assumed to be independent [SAR10, YEH09, WU12], there are different assumptions about the available information on random variables at different timestamps. Examples of such information include the probability distribution function [SAR10], the unknown but identical probability distribution function with known expected value and variance [YEH09], or an observed value [WU12].

2.1.2. **Multiset-based Model**

In this model, at each timestamp, there is a multiset of observed values. For uncertain time series $X =< X_1, \dots, X_n >$, each element X_i is represented as the multiset $[\![x_{i,1}, x_{i,2}, \dots, x_{i,N_{X_i}}]\!]$ where each $x_{i,j}$ $(1 \leq j \leq N_{X_i})$ is an observed value at timestamp *i,* and N_{X_i} denotes the number of observed values at this timestamp. The exact value may or may not be present in the multiset, and thus the multiset can be thought of as a realization of the unknown random variable X_i. Note that we use double square brackets ⟦...⟧ to represent multisets.

In the current literature, there are also other definitions for the cases when there are multiple values at each timestamp. In [ASF09], it is assumed that there is a set (not multiset) of observed values at each timestamp. In [DALY14], the authors define uncertain time series at each timestamp as a set of observed values, each associated with an existential probability. Moreover, if the sum of the probabilities at each timestamp is 1, the exact value exists. In [DALX14], an uncertain time series is defined as a set of standard time series. This definition captures the dependencies between observed values. In the next section, we study the relationship between PDF-based and multiset-based uncertain time series analytically.

2.1.3. Relationship between PDF-based and Multiset-based Models

As discussed in Section 2.1, in the PDF-based model, there are different assumptions about the available information on random variables at different timestamps. However, in the multiset-based model, the only available information is a multiset of observed values at each timestamp. Now the question is: can we convert one model to another model? The answer depends on the amount of information available at each timestamp, explained as follows.

When the number of observed values at each timestamp is "large" enough, a multiset-based uncertain time series can be converted to a PDF-based uncertain time series. In other words, using a large number of observed values, we can estimate the PDF of the underlying random variable. The higher the number of observed values, the lower the amount of information lost in this conversion process.

A PDF-based uncertain time series cannot always be converted to a multiset-based uncertain time series by generating observed values using the given PDF at each timestamp. First, as discussed in Section 2.1, the PDFs of random variables at different timestamps are not always available. Moreover, even if we know the PDF type (e.g., normal, exponential), we may not have exact values for the parameters of the PDF (e.g., the expected value). We study different cases for a given PDF-based uncertain time series $X =< X_1, \dots, X_n >$ in the following. Note that each X_i $(1 \leq i \leq n)$ can be written as $X_i = x_i + E_{x_i}$, where x_i is the exact value that is unknown and E_{x_i} is a random variable denoting the error.

Case 1- $E(X_i)$ *is known:* If we know the exact value for both $E(X_i)$ and $E(E_{x_i})$, we can simply calculate x_i, which is the exact value of the data, and thus we would have the underlying *certain* standard time series. Thus, if we know the exact value for $E(X_i)$, to still have an uncertain model, $E(E_{x_i})$ should be unknown. In this case, if we know the PDFs of random variables at different timestamps, we can generate a multiset of observed values

and obtain a multiset-based uncertain time series. The higher the number of observed values, the lower the amount of information lost in this conversion process.

Case 2- $E(X_i)$ *is unknown:* In this case, an observed value is used as an estimation for $E(X_i)$, and PDF-based uncertain time series cannot be converted to multiset-based uncertain time series. The reason is that if we use the PDF of X_i to generate observed values, the average of these observed values would converge to $E(X_i)$, which itself is estimated by an observed value and therefore is not exact. In this way, the produced multiset-based uncertain time series would not represent the underlying PDF-based uncertain time series.

In this section, we studied the relationship between PDF-based and multiset-based models. In general, these two models cannot be converted to each other, thus each model requires individual similarity search techniques. We next introduce a classification of existing similarity measures on uncertain time series. We will use the *uncertain similarity measure* for similarity measures on uncertain time series.

2.2. Classification of Uncertain Similarity Measures

We classify existing similarity measures for uncertain time series into two, on the basis of the output they produce, as follows:

1. ***Deterministic Similarity Measures:*** This class of similarity measures is very similar to the traditional similarity measures, and returns a real number as the similarity between two uncertain time series.
2. ***Probabilistic Similarity Measures:*** This class returns a random variable associated with a PDF as the similarity between two uncertain time series. That is, a probabilistic similarity measure assigns a probability to each possible distance between the input pair of uncertain time series.

The following sections describe these two classes in more detail.

2.3. Deterministic Similarity Measures

DUST [SAR10] is the only deterministic similarity measure devised for uncertain time series that returns a single real number as the distance between two PDF-based uncertain time series. Although the similarity measure in DUST is deterministic, it uses probability theory at the core to generalize the Euclidean and DTW distance measures for uncertain time series. The distance between uncertain time series $X =< X_1, \dots, X_n >$ and $Y =< Y_1, \dots, Y_n >$ in DUST is defined as follows:

$$DUST(X, Y) = \sqrt{\sum_{i=1}^{n} dust(X_i, Y_i)^2} \tag{1}$$

where for each i ($1 \leq i \leq n$):

$$dust(X_i, Y_i) = \sqrt{-log\big(\varphi(|X_i - Y_i|)\big) + log\big(\varphi(0)\big)} \tag{2}$$

Here $\varphi(|X_i - Y_i|)$ is the probability $p(dist(0, |X_i - Y_i|) = 0)$, that is, $\varphi(|X_i - Y_i|)$ is the probability that the exact values at timestamp i are equal, given the observed values at that timestamp. To calculate distance in DUST, we need to have the information about the probability distribution of error, and of the underlying certain time series, as well as an observed value at each timestamp. The Appendix presents our calculations of DUST distances for normal, exponential, and uniform error distributions. The calculations provide better insight about the DUST function and its use for different error distributions.

2.4. Probabilistic Similarity Measures

The probabilistic similarity measures generalize the traditional measures for standard time series, and model them as random variables. Thus, unlike the traditional measures, using which we can find the exact similarity between two standard time series, the similarity between two

uncertain time series would be a random variable with a probability distribution function. This means assigning a probability to each possible distance between the two given uncertain time series. Consequently, queries that search for similarities in uncertain time series are also probabilistic (i.e., a probability is assigned to each query result). We will revisit those probabilistic queries in Section 2.5.

To be able to use these similarity measures in search tasks, we need to know their PDFs. Finding PDFs in probabilistic similarity measures poses challenges for similarity search queries over uncertain time series, in particular when we are concerned with scalability, response time, and precision. The existing probabilistic similarity measures for uncertain time series generalize the L_p-norm, and DTW distance for standard time series. Correspondingly, we classify the probabilistic similarity measures into: (1) uncertain L_p-norm distance, and (2) uncertain DTW distance. Each class includes different approaches to model the corresponding uncertain similarity measure, based on different modeling and assumptions about the uncertain time series data dealt with. These measures are discussed in more detail as follows.

2.4.1. Uncertain L_p-norm Distance

This uncertain similarity measure has been studied for both multiset-based and PDF-based models. This measure extends L_p-norm distance for standard time series, in which p is a positive integer. For multiset-based model, Aßfalg et al. [ASF09] assume that the only available information is a set of independent observed values at each timestamp; moreover, sets at different timestamps are independent as well. Using observed values, they find a multiset of all possible L_p-norm distances between the given uncertain time series X and Y, denoted by $\widetilde{dist}_{L_p}(X,Y)$. This multiset represents a realization of uncertain L_p-norm distance, $L_p(X,Y)$. To estimate the underlying cumulative distribution function (CDF) of the elements in the multiset $\widetilde{dist}_{L_p}(X,Y)$, Aßfalg et al. use the empirical distribution function of $L_p(X,Y)$ defined as follows:

$$P\big(L_p(X,Y) \leq d\big) = \frac{\sum_{e \in \widetilde{dist}_{L_p}(X,Y)} 1\{e \leq d\}}{|\widetilde{dist}_{L_p}(X,Y)|} \tag{3}$$

where $1\{e \leq d\}$ is the indicator function, which is equal to 1 if $e \leq d$, and 0 otherwise.

For the PDF-based model, there exist two approaches for calculating the uncertain L_p-norm distance for $p = 2$ with different assumptions about the available information. Both assume that the random variables in the given uncertain time series are independent and identically distributed (i.i.d.). Yeh et al. [YEH09] propose an approach to processing similarity queries, which we simply refer to as PROUD. They assume that the expected values and variances of random variables at different timestamps of an uncertain time series are known, but their PDFs are unknown. They model uncertain Euclidean distance as a normal random variable, using the central limit theorem [ROS09]. The cumulative distribution of this random variable is defined as:

$$P(Eucl(X,Y) \leq d) = \frac{1}{2}\left(1 + erf\left(\frac{d - E(Eucl(X,Y))}{\sqrt{2Var(Eucl(X,Y))}}\right)\right)$$

where $Eucl$ is the squared Euclidean distance, and $E(Eucl(X,Y))$ and $Var(Eucl(X,Y))$ are defined as follows:

$$E(Eucl(X,Y)) = \sum_{i=1}^{n}\Big(\big(E(X_i) - E(Y_i)\big)^2 + Var(X_i) + Var(Y_i)\Big) \tag{4}$$

$$Var(Eucl(X,Y)) = 4\sum_{i=1}^{n}\big(E(X_i) - E(Y_i)\big)^2\big(Var(X_i) + Var(Y_i)\big)$$

Wu et al. [WU12] also consider uncertain squared Euclidian distance (MISQ), which they assume to have an observed value for each random variable in the given uncertain time series. Using this information, they determine a lower bound for the exact squared Euclidean distance.

2.4.2. Uncertain DTW Distance

This measure has been formalized for multiset-based uncertain time series [ASF09]. The approach used to formalize this uncertain distance is similar to that used for multiset-based uncertain L_p-norm distance [ASF09] as follows. Given uncertain time series X and Y, this approach first finds the multiset of all possible DTW distances (i.e., $\widetilde{dist_{DTW}}(X,Y)$). Using this, the CDF of the uncertain DTW distance is then approximated using the following empirical distribution function of $DTW(X,Y)$:

$$P(DTW(X,Y) \leq d) = \frac{\sum_{e \in \widetilde{dist_{DTW}}(X,Y)} 1\{e \leq d\}}{\left|\widetilde{dist_{DTW}}(X,Y)\right|}$$

Next, we study probabilistic similarity queries, which use these probabilistic similarity measures.

2.5. Probabilistic Similarity Queries

Since the output of the probabilistic similarity measures is a random variable, these similarity measures cannot be used directly for similarity search tasks. This is one of the challenges in uncertain time series similarity search. Similarity queries over uncertain time series data are probabilistic, that is, a probability is assigned to the query result. In the related literature, so far, the processing of the following probabilistic queries has been addressed.

1) **Probabilistic Range Query:** Given a set D of uncertain time series, an uncertain time series Q as a query reference, a distance threshold $d \in \mathbb{R}^+$, and a probability threshold $p \in (0,1]$, we are looking for uncertain time series X that satisfies the following:

$$P(dist(X,Q) \leq d) \geq p \quad (5)$$

Aßfalg et al. [ASF09] propose this type of query for multiset-based uncertain time series where $dist$ can be the L_p-norm or DTW distance. For PDF-based uncertain time series, Lian et al. [LIA08] and Yeh et al. [YEH09] propose this query where $dist$ is the squared Euclidean distance. In this case, we refer to a probabilistic range query as a *probabilistic threshold-based Euclidean (PTE) query*.

2) **Probabilistic Ranked Range Query:** Aßfalg et al. [ASF09] propose a ranking query for multiset-based uncertain time series. Given a set D of uncertain time series, an uncertain time series Q, and a distance threshold $d \in \mathbb{R}^+$, we are looking for an ordered list of uncertain time series $(X_1, \dots, X_m)$ that satisfies:

$$P(dist(X_i, Q) \leq d) \leq P(dist(X_{i+1}, Q) \leq d), \qquad \text{for } 1 \leq i \leq m$$

where $dist$ can be the L_p-norm or DTW distance.

3) **Threshold Similarity Query:** Wu et al. [WU12] propose another type of query that retrieves uncertain time series X from set D of uncertain time series such that $dist(\mu_Q, \mu_X) \leq d$ with the confidence level of $1 - \alpha$, where Q, ε, and α are given by the user. Here μ_Q and μ_X are the expected values of Q and X, respectively.

4) **Exact Match Query:** This query is a special case of the threshold query [WU12] where $d = 0$. Given a set D of uncertain time series, uncertain time series Q, and confidence level $1 - \alpha$, the query returns uncertain time series X such that $dist(\mu_Q, \mu_X) = 0$ with confidence level $1 - \alpha$.

The probabilistic queries above extend the range query for standard time series, and their processing poses challenges in some cases (explained in Section 1.2). Other important topics on

uncertain time series also have been studied, such as top-k nearest neighbor queries in multiset-based uncertain time series [DALX14], uncertain sliding windows over multiset-based uncertain time series data streams [DALY14], and reduction techniques. We identify two types of reduction techniques for uncertain time series: approximation [ASF09] and dimensionality reduction techniques [YEH09, QIA09, ZHA10, XU09], where the former is normally applied at each timestamp and the latter is applied to the entire uncertain time series data. Dimensionality reduction techniques extend PLA (Piecewise Linear Approximation) [QIA09] and Haar wavelet [ZHA10] for uncertain time series data.

2.6. Uncertain Vectors and Uncertain Trajectories

Uncertain time series data can also be viewed as n-dimensional uncertain vectors. Following this view, several similarity search techniques have been proposed for uncertain vectors [BER09, BOH06, KRI07, LJO07, TAO05] in applications such as biometric databases to identify individuals/objects according to features for which the exact values are unknown. As discussed by Aßfalg et al. [ASF09], the effective solutions proposed for uncertain vectors are not suitable for uncertain time series data. Other related research includes uncertain trajectories that may be considered as multidimensional uncertain time series data. Uncertain trajectories record, at each timestamp, the position of a given object [CHE04, EMR12, ZHA11]. Positions recorded at different timestamps are naturally correlated; this is different from the independence assumption often made for uncertain time series data. However, none of the related research on uncertain time series studied correlation analysis on such data.

2.7. Summary

In this chapter, we reviewed the works on similarity search on uncertain time series. We also introduced a classification of existing uncertain similarity measures. This classification shows that in the current literature, different assumptions are imposed on uncertain time series resulting

in different similarity search approaches. For example, consider the traditional range query, $Eucl(x, q) \leq d$, where we are looking for standard time series x, the Euclidean distance of which to a given time series q is at most d. However, in the context of uncertain time series, this simple query has been addressed in different ways based on different assumptions and modeling techniques, including PROUD [YEH09] and Aßfalg's method [ASF09]. Different applications consider different assumptions and have different similarity search queries. It is thus essential to investigate other types of similarity queries, including correlation queries over uncertain time series under different assumptions. Moreover, since the two models cannot be converted to each other, each model requires individual similarity search techniques.

Chapter 3: Normalization and Correlation Formulations

In this chapter, we formulate normalization for uncertain time series, referred to as *uncertain normalization*, and study its properties. Using uncertain normalization, we will introduce the notion of correlation between uncertain time series, referred to as *uncertain correlation*.

3.1. Normalization

For standard time series, it is well known that normalization makes similarity measures invariant to scaling and shifting and hence it helps better capture the similarity [SHA04]. This is desirable for uncertain time series, and indeed this is the case for the normalization we will define in this section. To establish this, we performed a probabilistic first nearest neighbor (1NN) classification using the Euclidean distance. For each uncertain time series X in the training dataset, we find the probability that X is the nearest neighbor of the test uncertain time series (to

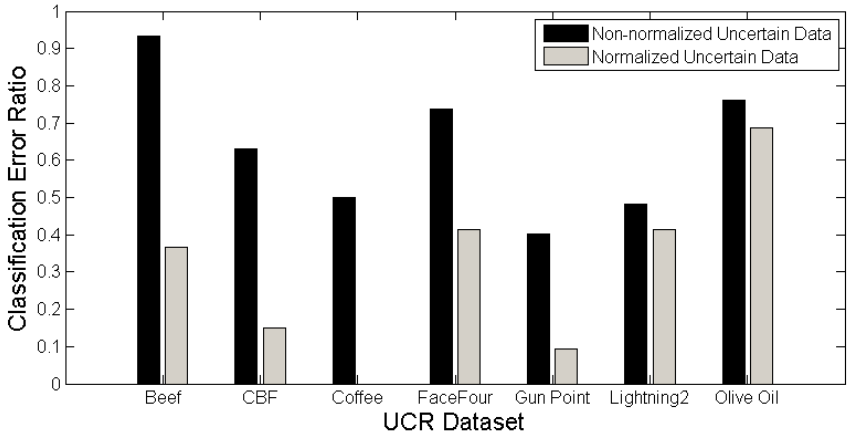


Figure 3. 1NN classification error for normalized and non-normalized data.

be classified). Classes are determined based on the highest 1NN probability. We used 7 UCR datasets [KEO] and perturbed them to obtain uncertain data for this experiment. Figure 3 shows the results: the black bars indicate the classification error of non-normalized uncertain time series and the gray bars indicate that of normalized uncertain time series. As expected, we found that the error ratio of normalized data was lower than that of raw data.

Our normalization technique extends the technique for standard time series. The normal form of a standard time series $x =< x_1, \dots, x_n >$ is defined as $\hat{x} =< \hat{x}_1, \dots, \hat{x}_n >$, in which for each timestamp *i* $(1 \leq i \leq n)$, we have:

$$\hat{x}_i = \frac{x_i - \bar{x}}{s_x} \tag{6}$$

Here $\bar{x}$ and s_x are sample mean and standard deviation of *x,* respectively, defined as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},\ s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Using this, we define the notion of uncertain normalization for both PDF-based and multiset-based models in the following sections.

3.1.1. PDF-based Normalization

In this section, we define the normal form of PDF-based uncertain time series. First, we define a general case and then a simplified case. We define the normal form of uncertain time series as follows.

Definition 3.1. Normal form of uncertain time series- *Given an uncertain time series* $X =< X_1, \dots, X_n >$*, we define* $\hat{X} =< \hat{X}_1, \dots, \hat{X}_n >$ *as its normal form, where for each i (*$1 \leq i \leq n$*), we have:*

$$\hat{X}_i = \frac{X_i - \bar{X}}{S_X}$$

in which $\bar{X}$ and S_X denote the sample mean and standard deviation of random variables at different timestamps in X, respectively. That is,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

The following lemma highlights the properties of this definition for uncertain normalization.

Lemma 1. *Given any uncertain time series $X =< X_1, \dots, X_n >$, its normal form has the following properties.*

a) $\bar{\hat{X}} = 0$

b) $S_{\hat{X}} = 1$

Proof. The proof is very similar to the standard case [SHA04]. However, note that in this case, we are working with random variables that are functions from a set of possible outcomes to a set of real numbers, $\mathbb{R}$. Here the assumption is S_X is not zero. The proof for part (*a*) is as follows:

$$\bar{\hat{X}} = \frac{\sum_{i=1}^{n} \hat{X}_i}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_X} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \bar{X}}{S_X} = 0$$

To prove part (*b*), we have

$$\sum_{i=1}^{n} \hat{X}_i^{\,2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}}\right)^2$$

$$= (n-1)\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right)^2 = n - 1$$

From this equation, we have:

$$S_{\hat{X}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{X}_i - avg(\hat{X}))^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}\hat{X}_i^{\,2}}{n-1}} = 1 \blacksquare$$

The immediate consequence of this lemma is that the normal form for uncertain time series is *idempotence*. That is, if we normalize the normal form of a given uncertain time series *X*, we would have the same result as the first application. Because using Lemma 1, we have:

$$\hat{\hat{X}}_i = \frac{\hat{X}_i - \bar{\hat{X}}}{S_{\hat{X}}} = \hat{X}_i$$

In the following, we will propose another definition for the normal form of uncertain time series and will show how this definition can simplify the computation and preserve the properties of uncertain time series.

Definition 3.2. PDF-based normal form- *Given an uncertain time series* $X = < X_1, \dots, X_n >$*, we define* $\hat{X} = < \hat{X}_1, \dots, \hat{X}_n >$ *as its normal form, where for each timestamp i* $(1 \le i \le n)$*, we have:*

$$\hat{X}_i = \frac{X_i - \bar{X}}{S_X} \tag{7}$$

in which $\bar{X}$ and S_X denote the sample mean and standard deviation of the expected values of the random variables at different timestamps in X, respectively. That is,

$$\bar{X} = \frac{\sum_{i=1}^{n} E(X_i)}{n} \quad and \quad S_X = \sqrt{\frac{\sum_{i=1}^{n} (E(X_i) - \bar{X})^2}{n-1}}$$

In other words, $\hat{X}$ is obtained from *X* by shifting *X* by the average of the expected values and then scaling by the standard deviation of the expected values. The following lemma highlights the properties of this normal form.

Lemma 2. *Let $X =< X_1, \dots, X_n >$ be an uncertain time series, and $\hat{X} =< \hat{X}_1, \dots, \hat{X}_n >$ be its normal form. Then the following statements hold:*

a) $\bar{\hat{X}} = 0$ *and* $S_{\hat{X}} = 1$.

b) *If X_i's are independent random variables, so are $\hat{X}_i$'s.*

c) *If X_i's are identically distributed, so are $\hat{X}_i$'s.*

Proof. For part (*a*), $\bar{\hat{X}}$ and $S_{\hat{X}}$ are the average and standard deviation of standard time series $E(\hat{X}) =< E(\hat{X}_1), \dots, E(\hat{X}_n) >$, respectively, which is the normal form of standard time series $E(X) =< E(X_1), \dots, E(X_n) >$. Moreover, for any standard time series, it is known that the average and standard deviation of its normal form are equal to 0 and 1, respectively [SHA04]. Parts (*b*) and (*c*) are immediate upon noting that at each timestamp *i*, $\hat{X}_i$ is a linear transformation of X_i. ■

The immediate consequences of this lemma are as follows:

1. *Preserving temporal independence*: Existing works assume that random variables at

different timestamps of uncertain time series are independent [SAR10, YEH09]. All those works can benefit from our normalization as a preprocessing step to better capture the similarity.

2. *Preserving identical distribution*: Existing works, such as PROUD [YEH09], which assumes random variables are identically distributed, can also benefit from the proposed normalization.

3. *Easy calculation of PDFs of normal forms*: If the $X_i's$ in an uncertain time series X are continuous random variables, the PDF of $\hat{X}_i$ can be simply obtained as $f_{\hat{X}_i}(x) = S_X f_{X_i}(S_X x + \bar{X})$ [ROS09]. Moreover, it holds that:

$$E(\hat{X}_i) = \frac{E(X_i) - \bar{X}}{S_X}, \text{and} \tag{8}$$

$$Var(\hat{X}_i) = \frac{Var(X_i)}{{S_X}^2} \tag{9}$$

4. *Idempotence:* The result of applying normalization on an uncertain time series X multiple times is the same as applying it once. This is because at timestamp i, we have:

$$\hat{\hat{X}}_i = \frac{\hat{X}_i - \bar{\hat{X}}}{S_{\hat{X}}}$$

and by Lemma 2, we obtain $\hat{\hat{X}}_i = \hat{X}_i$.

Due to all these properties, in this thesis, we will use Definition 3.2 for the normal form of uncertain time series.

3.1.2. **Multiset-based Normalization**

Consider a multiset-based uncertain time series $X =< X_1, \dots, X_n >$, where the only available information about each X_i is a multiset of observed values, $R_{X_i} = [\![x_{i,1}, x_{i,2}, \dots, x_{i,N_{X_i}}]\!]$, considered as X_i's realization. We use $[\![\dots]\!]$ for multisets, and N_{X_i} to denote the number of observed values for X_i. Similar to the PDF-based model (Definition 3.2), we normalize a multiset-based uncertain time series *X* by shifting *X* by the average of means and then scaling the result by the standard deviation of means. Here, at each timestamp, the mean is approximated by the average of the observed values. We define the notion of normalization for multiset-based uncertain time series formally as follows.

Definition 3.3. Multiset-based normal form- *The normal form of a multiset-based uncertain time series* $X =< X_1, \dots, X_n >$ *with* $R_{X_i} = [\![x_{i,1}, x_{i,2}, \dots, x_{i,N_{X_i}}]\!]$ $(1 \le i \le n)$ *is defined as uncertain time series* $\hat{X} =< \hat{X}_1, \dots, \hat{X}_n >$ *with* $R_{\hat{X}_i} = [\![\hat{x}_{i,1}, \hat{x}_{i,2}, \dots, \hat{x}_{i,N_{X_i}}]\!]$, *where*

$$\hat{x}_{i,j} = \frac{x_{i,j} - \bar{X}}{S_X} \tag{10}$$

Here $\bar{X}$ *and* S_X *are defined as follows:*

$$\bar{X} = \frac{\sum_{i=1}^{n} m_{X_i}}{n} \quad and \quad S_X = \sqrt{\frac{\sum_{i=1}^{n} (m_{X_i} - \bar{X})^2}{n-1}}$$

in which m_{X_i} *is the average of observed values at timestamp i, i.e.,* $m_{X_i} = \sum_{j=1}^{N_{X_i}} x_{i,j} / N_{X_i}$.

The elements of $R_{\hat{X}_i}$ are actually a realization of random variable $\hat{X}_i$ defined in (7). Since each $\hat{x}_{i,j}$ is a linear transformation of $x_{i,j}$, the defined normalization preserves the independency of the observed values, between and within timestamps. Moreover, multiset-based normalization is

idempotent, because average and standard deviation of the normal form of a multiset-based uncertain time series are 0 and 1, respectively, as shown in the following lemma.

Lemma 3. *Let* $X =< X_1, \dots, X_n >$ *be a multiset-based uncertain time series and* $\hat{X} =< \hat{X}_1, \dots, \hat{X}_n >$ *be its normal form. Then* $\bar{\hat{X}}$ *and* $S_{\hat{X}}$ *are equal to 0 and 1, respectively.*

Proof. For the average of all the observed values in $\hat{X}$, we have:

$$\bar{\hat{X}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{N_{X_i}}\hat{x}_{i,j}}{N_{X_i}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_{X_i}}\sum_{j=1}^{N_{X_i}}\frac{x_{i,j}-\bar{X}}{S_X} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{N_{X_i}}x_{i,j}}{N_{X_i}} - \bar{X}}{S_X} = 0$$

To prove that $S_{\hat{X}} = 1$, we first need to show that $\sum_{i=1}^{n} {m_{\hat{X}_i}}^2 = n - 1$, as follows:

$$\begin{aligned}\sum_{i=1}^{n} {m_{\hat{X}_i}}^2 &= \sum_{i=1}^{n}\left(\frac{\frac{1}{N_{X_i}}\sum_{j=1}^{N_{X_i}}x_{i,j} - \bar{X}}{S_X}\right)^2 = \sum_{i=1}^{n}\frac{(m_{X_i}-\bar{X})^2}{S_X^2} \\ &= \sum_{i=1}^{n}\frac{(m_{X_i}-\bar{X})^2}{\frac{\sum_{i=1}^{n}(m_{X_i}-\bar{X})^2}{n-1}} = n-1\end{aligned} \tag{11}$$

Using (11), we have the following:

$$S_{\hat{X}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(m_{\hat{X}_i} - \bar{\hat{X}}\right)^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}{m_{\hat{X}_i}}^2} = 1 \blacksquare$$

Now that we have a suitable definition for uncertain time series normalization, we can define the correlation between uncertain time series in the following section.

3.2. **Uncertain Correlation and Probabilistic Queries**

In this section, we propose a formulation of the notion of *correlation* for uncertain time series. This is done by extending the well-known Pearson correlation coefficient [SHA04] used

for standard time series data, which is the dot product of the normal forms of the given two standard time series $x =< x_1, \dots, x_n >$ and $y =< y_1, \dots, y_n >$, defined as follows:

$$Corr(x, y) = \frac{\sum_{i=1}^{n} \hat{x}_i \hat{y}_i}{n-1} \tag{12}$$

where $\hat{x}_i$ and $\hat{y}_i$ are the normal forms of x_i and y_i, defined in (6).

Unlike the correlation between standard time series that is denoted by a single value, the correlation between uncertain time series defined as follows is a random variable assessed by a PDF.

Definition 3.4. Uncertain time series correlation- *Given a pair of uncertain time series* $X =< X_1, \dots, X_n >$ *and* $Y =< Y_1, \dots, Y_n >$*, their correlation is defined as:*

$$Corr(X, Y) = \frac{\sum_{i=1}^{n} \hat{X}_i \hat{Y}_i}{n-1} \tag{13}$$

where $\hat{X}_i$ *and* $\hat{Y}_i$ *are normal forms of* X_i *and* Y_i*, respectively, defined in (7).*

For convenience, we use *uncertain correlation* to refer to the correlation between uncertain time series. Having a random variable as uncertain correlation, we define probabilistic threshold-based correlation queries, which extend the correlation range queries in the standard case. Instead of using the exact correlation between *X* and *Y*, we use the cumulative distribution function (CDF) of their correlation.

Definition 3.5. Probabilistic threshold-based correlation (PTC) queries- *Given a set D of uncertain time series, an uncertain time series Q as a query reference, a correlation threshold* $c \in (0,1]$*, and a probability threshold* $p \in (0,1]$*, PTC queries look for those*

uncertain time series X in D such that X and Q are positively correlated with probability at least p, and correlation coefficient at least c. Formally,

$$P(Corr(X,Q) \geq c) \geq p \tag{14}$$

The question that arises at this point is how to find the CDF of $Corr(X,Q)$. The answer relies on the amount of available information about random variables at different timestamps. In this thesis, we investigate this in two cases:

1. *PDF-based uncertain time series:* having PDF of each random variable at each timestamp.
2. *Multiset-based uncertain time series:* having a multiset of independent observed values for each random variable.

We study the processing of the PTC queries for these two cases. For both cases, we assume random variables at different timestamps are independent. Section 5.1 considers the PDF-based model and Section 5.2 considers the multiset-based model. The following section shows the relationship between existing work on uncertain time series and uncertain correlation.

3.3. Relationship between Correlation and Euclidian Distance Measures

If we define the uncertain correlation as dot product of the normal form of uncertain time series as in Definition 3.2, and the uncertain Euclidean distance between uncertain time series $X =< X_1, \dots, X_n >$ and $Y =< Y_1, \dots, Y_n >$ as $Eucl(X,Y) = \sum_{i=1}^{n}(X_i - Y_i)^2$, the uncertain correlation and Euclidean distance would have the following relationship.

Lemma 4. *Given two uncertain time series* $X =< X_1, \dots, X_n >$ *and* $Y =< Y_1, \dots, Y_n >$*, we have:*

$$Eucl(\hat{X}, \hat{Y}) = 2(n-1)\big(1 - Corr(X,Y)\big)$$

where X and Y are normalized as in Definition 3.2.

Proof. To prove this lemma, first we need to prove that for a given uncertain time series *X*, we have $\sum_{i=1}^{n} \hat{X}_i{}^2 = n - 1$, as follows:

$$\sum_{i=1}^{n} \hat{X}_i{}^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X}\right)^2 = (n-1)\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right)^2 = n - 1$$

Using this, we would have:

$$Eucl(\hat{X}, \hat{Y}) = \sum_{i=1}^{n} (\hat{X}_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{X}_i^2 + \hat{Y}_i^2 - 2\hat{X}_i\hat{Y}_i = 2(n-1) - 2\sum_{i=1}^{n} \hat{X}_i\hat{Y}_i =$$

$$= 2(n-1)\big(1 - Corr(X,Y)\big) \blacksquare$$

It is easy to see that this lemma holds even if one of the time series is standard. Using this lemma, we can prove that PTE queries (5) and PTC queries (Definition 3.5) can be converted to each other.

Lemma 5. *Given two uncertain time series* $X =< X_1, \dots, X_n >$ *and* $Y =< Y_1, \dots, Y_n >$*, we have:*

$$P(Corr(X,Y) \geq c) = P(Eucl(\hat{X}, \hat{Y}) \leq 2(n-1)(1-c))$$

Proof. Using the previous lemma, we have:

$$P\left(Eucl(\hat{X}, \hat{Y}) \leq 2(n-1)(1-c)\right) = P\left(2(n-1)\big(1 - Corr(X,Y)\big) \leq 2(n-1)(1-c)\right)$$

Using which, we obtain:

$$P\left(Eucl(\hat{X}, \hat{Y}) \leq 2(n-1)(1-c)\right) = P(Corr(X,Y) \geq c) \blacksquare$$

This lemma indicates that if we first normalize uncertain time series data, the result of the PTC queries would be the same as that of the PTE queries. This shows that the uncertain correlation is a principal extension of the uncertain Euclidean distance. Another important result of this lemma is that since the correlation threshold, c, is between -1 and 1, the Euclidean distance threshold, $2(n-1)(1-c)$, would be between 0 and $4(n-1)$. So the user can choose a distance threshold within this interval. On the other hand, when data is not normalized, the distance threshold could be any positive number and defining a proper distance threshold would be difficult, as it requires having a lot of information about the dataset in which we are searching. Now the question is: is there any relationship between the uncertain Euclidean distance and uncertain correlation defined in Definition 3.4? We will answer this question in Chapter 7

3.4. Summary

In summary, our definition for the normal form of uncertain time series generalizes that of the standard time series in a way that the uncertain normalization preserves the properties of the standard normalization. Moreover, multiset-based normalization generalizes PDF-based normalization. Both multiset-based and PDF-based normalization preserve properties of underlying uncertain time series (i.e., temporal independence and identical distribution). Using uncertain normalization, we introduced the notion of uncertain correlation. Having a random variable as correlation between two uncertain time series motivated our definition of probabilistic threshold-based correlation queries. Probabilistic threshold-based correlation queries consist of an uncertain time series query Q, a correlation threshold c, and a probability threshold p. Given an uncertain time series Q, the goal is to search for uncertain time series with a high enough probability that their correlation with the given uncertain time series is within a given threshold. We also studied the relationship between uncertain correlation and uncertain Euclidean distance. This relationship also held for standard time series. This shows that probabilistic similarity measures build on traditional similarity measures in a disciplined way.

Chapter 4: Experiments Setup

This chapter explains the setup of our experiments to present experimental results for different problems studied in this research more clearly. The objectives of the experiments are to study the overall performance of the proposed techniques and to compare their performance with the existing techniques. The experiments data and setup parameters follow the ones used by Dallachiesa et al. [DAL12], Sarangi et al. [SAR10], Yeh et al. [YEH09], and Aßfalg et al. [ASF09]. For all the results, we report the average over 10 different random runs. The experiments were conducted on a typical desktop PC with a 2.66 GHz CPU and 4GB of RAM. All algorithms are implemented and run in MATLAB (2013a).

4.1. Experiments Objectives

The objectives of the experiments are to answer the following research questions:

- **RQ1**: What is the overall performance of the proposed solutions with regard to data and query parameters (defined Section 4.3)?
- **RQ2**: Do the proposed solutions outperform the existing techniques (Section 4.4)?
- **RQ3:** Do the techniques for the multiset-based model yield a good approximation for uncertain correlation PDF?

To answer these research questions, we consider the following two measures as the ground truths:

- ***Ground truth I***: This is based on the result of the deterministic query defined in Section 4.4 on the dataset without uncertainty, with the same correlation threshold as the given probabilistic query.
- ***Ground truth II:*** This assumes the underlying PDF of the observed values is known at

each timestamp, using which, we can calculate the PDF of uncertain correlation.

We use ground truth *I* for evaluating *RQ1* and *RQ2*, and ground truth *II* for *RQ3*.

4.2. Datasets

Similar to prior works [ASF09, DAL12, SAR10, YEH09], to generate uncertain time series data used in our experiments, we take standard time series and perturb them using different error functions representing errors in the measurements. This data generation method allows us to use the original certain time series (with exact values) as the ground truth when evaluating the performance of the proposed solutions for uncertain data. The standard time series data includes the 20 datasets in the UCR benchmark data from real-life applications [KEO], namely, 50words, Adiac, Beef, CBF, Coffee, ECG200, Fish, FaceAll, FaceFour, Gun-Point, Lighting2, Lighting7, OSULeaf, OliveOil, SwedishLeaf, Synthetic-control, Trace, Two-Patterns, Wafer, and Yoga.

4.3. Performance Measures and Parameters

We measure the performance using hit and false alarm ratios. *Hit ratio* (or recall) is defined as the number of correct results returned to the total number of correct results. The ground truth (i.e., correct results) is based on the result of the correlation queries $Corr(x, q) \geq c$, on the dataset without uncertainty [YEH09]. *False alarm ratio* (or false discovery rate) is defined as the number of incorrect results to the total number of results returned. Another measure is F_1 score defined as follows:

$$F_1 = 2 \times (precision \times recall)/(precision + recall)$$

Recall is the hit ratio and precision is 1 minus the false alarm ratio. We repeat each experiment 10 times and report their average as the experiment result.

In our experiments, we study the effect of different parameters on the performance of the proposed solutions. These parameters belong to two categories: *data parameters* and *query parameters*. Data parameters indicate characteristics of uncertain data while query parameters are threshold values in probabilistic or deterministic queries.

4.3.1. Data Parameters

The uncertainty level in uncertain time series depends on two parameters: *standard deviation* and *probability distribution* of the random variable at each timestamp. Standard deviation in turn is based on *standard deviation ratio* and *error rate*, defined as follows.

Standard Deviation Ratio (SDR): Standard deviation of random variable at each timestamp is $\sigma \times r$ [YEH09], where σ denotes the standard deviation of the original certain time series (used as the ground truth), and r denotes the SDR, which reflects the uncertainty level in a given uncertain time series. The higher the SDR, the farther the uncertain value at each timestamp from the ground truth. As considered in earlier works [SAR10, YEH09], we use the following values for SDR: $\{0.01, 0.05, 0.1, 0.5, 1, 1.5, 2, 3, 4\}$.

Error Rate: Error rate allows different timestamps in an uncertain time series to have different uncertainty levels. This helps emulate the effect of tapering measurement accuracy within an uncertain time series. For *m*% of the timestamps, we use r as an SDR, and use $0.1 \times r$ for the rest. Following Sarangi et al. [SAR10], the value of r considered in our work varied from 0.01 to 4, and the error rate *m*% was chosen as 10, 30, 50, and 100 percent. As is done in [DAL12] for the setup of multiset-based model, the error rate we consider in the experiments for multiset-based model is 100%.

Probability Distribution: In our work, we are also interested in studying the effect of probability distribution type on the proposed solutions, for which we consider normal, uniform, and exponential distributions.

4.3.2. Query Parameters

The PTC queries (defined in Section 3.2) have two parameters: 1) probability threshold and 2) correlation threshold. In our experiments, we study the effect of these two query parameters on the performance of the proposed techniques. For the probability threshold, we considered the values 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Since the PTC queries look for highly correlated uncertain time series, we used correlation thresholds from 0.4 to 0.9.

4.4. Previous Techniques

As there is no previous work on correlation analysis for uncertain time series data, we can only compare our work with the deterministic solutions in which uncertain time series are treated as if they are (certain) standard time series [SAR10, YEH09]. In our experiments, we will compare the PTC queries with the following deterministic queries.

Definition 4.1. Deterministic threshold-based correlation (DTC) queries- *Given a set D of standard time series, a standard time series q as a query reference, and a correlation threshold c, DTC queries look for those standard time series x in D such that the Pearson correlation of x and q is at least c. Formally,* $Corr(x, q) \geq c$ *for* $c \in (0,1]$.

The correlation threshold in the DTC queries is the same as the one in the PTC queries. To compare the DTC with the PTC queries for the *PDF-based model*, the standard time series x and q would be the sequence of expected values of random variables in the corresponding uncertain time series. To compare the DTC with the PTC queries for the *multiset-based model*, the value at each timestamp of standard time series x and q would be the average of the observed values at that timestamp. In our experiments, we refer to the DTC queries as the deterministic query and to the PTC queries as the probabilistic query.

Chapter 5: Finding Correlation for PDF-based and Multiset-based Models

In this chapter, we study the processing of the PTC queries for both PDF-based and multiset-based uncertain time series in the following sections.

5.1. PDF-based Model

Given a PDF-based uncertain time series $X =< X_1, \dots, X_n >$, each X_i $(1 \leq i \leq n)$ can be written as $X_i = x_i + E_{x_i}$, where x_i is the "exact" value which is unknown, and E_{x_i} is a random variable denoting the error. Thus, the expected value of X_i would be $E(X_i) = x_i + E(E_{x_i})$. Since the exact value x_i is unknown, $E(X_i)$ would be unknown as well even if the expected value of the error is known. We thus use an observed value as an estimate for $E(X_i)$. In our work, we assume for each random variable X_i in *X,* we have its probability distribution type, its variance, and an observed value.

To answer the PTC queries, defined in Section 3.2, for PDF-based uncertain time series, we need to find PDF of uncertain correlation $Corr(X, Q)$. Let $X =< X_1, \dots, X_n >$ and $Q =< Q_1, \dots, Q_n >$ be two PDF-based uncertain time series, where $X_i's$ and $Q_i's$ are independent continuous random variables. The correlation between *X* and *Q* is defined as follows:

$$Corr(X, Q) = \frac{\sum_{i=1}^{n} \hat{X}_i \hat{Q}_i}{n - 1}$$

$(n - 1)Corr(X, Q)$ can be simply modeled as the sum of product of the random variables $\hat{X}_i$ and $\hat{Q}_i$ assessed with $f_{\hat{X}_i}(y)$ and $f_{\hat{Q}_i}(y)$, respectively. Since the normal form of each X_i (7) is a linear transformation of X_i, the PDF of $\hat{X}_i$ (and similarly of $\hat{Q}_i$) can be calculated as $f_{\hat{X}_i}(y) =$

$S_X f_{X_i}(S_X y + \bar{X})$ [ROS09]. We then find the PDF of the product of $\hat{X}_i$ and $\hat{Q}_i$, i.e., $Z_i = \hat{X}_i \hat{Q}_i$ for $1 \leq i \leq n$, noting that according to Lemma 2, when X_i's and Q_i's are assumed to be independent, so are $\hat{X}_i$'s and $\hat{Q}_i$'s, and hence the PDF of their product [ROS09] is defined as:

$$f_{Z_i}(y) = \int_{-\infty}^{+\infty} \frac{1}{|t|} f_{\hat{X}_i}(t) f_{\hat{Q}_i}\left(\frac{y}{t}\right) dt$$

At this point, the problem is reduced to sum of n independent random variables, that is:

$$(n-1)Corr(X,Q) = Z_1 + \cdots + Z_n \tag{15}$$

which can be obtained iteratively as follows. We define $Y_2 = Z_1 + Z_2$ and compute its PDF as:

$$f_{Y_2}(y) = \int_{-\infty}^{\infty} f_{Z_1}(y-t) f_{Z_2}(t)\, dt$$

We then compute PDF of $Y_3 = Y_2 + Z_3$, and so on. Finally, we can calculate the PDF of $(n-1)Corr(X,Q)$ (i.e., $Y_n = Y_{n-1} + Z_n$) and use it to find those uncertain time series that satisfy the PTC queries.

Let us consider again the example in Figure 2 showing two PDF-based uncertain time series X and Y, but instead of using the deterministic approach discussed earlier, we now use our probabilistic approach. Consider the PTC query $P(Corr(X,Y) \geq 0.9) \geq 0.5$. By calculating the PDF of correlation between X and Y, we find that $P(Corr(X,Y) \geq 0.9) = 0.7$. This shows that X and Y are highly correlated, and Y would be among the results, as expected.

We should point out that even when one of the time series is standard, PDF of $Corr(X,Q)$ in the PTC queries can still be found similarly using our approach. This provides more flexibility to the user to choose a query reference (e.g., Q in the PTC queries), which could be uncertain or standard time series. We next consider correlation of uncertain time series for a special case in which random variables are i.i.d, and show how this simplifies the processing of the PTC queries.

5.1.1. **Correlation with i.i.d. Distribution**

As a special case, suppose all random variables in the given uncertain time series X and Q in the PTC queries are i.i.d. As shown in part (*b*) and (*c*) of Lemma 2, independence and identical distribution would be preserved under our normalization technique. Thus, we consider correlation between two uncertain time series X and Q as sum of a sequence of i.i.d. random variables, i.e., $Corr(X,Q) = \sum_{i=1}^{n} \hat{X}_i \hat{Q}_i/(n-1)$. According to the central limit theorem [ROS09], as n increases, $Corr(X,Q)$ approaches normal distribution. The expected value and variance of $Corr(X,Q)$ are defined as follows:

$$E\big(Corr(X,Q)\big) = \frac{\sum_{i=1}^{n} E(\hat{X}_i)E(\hat{Q}_i)}{n-1}, \text{and}$$

$$Var\big(Corr(X,Q)\big) = \frac{\sum_{i=1}^{n}\Big(E(\hat{X}_i)^2 Var(\hat{Q}_i) + E(\hat{Q}_i)^2 Var(\hat{X}_i) + Var(\hat{X}_i)Var(\hat{Q}_i)\Big)}{(n-1)^2}$$

Moreover, given an uncertain time series X, $E(\hat{X}_i)$ and $Var(\hat{X}_i)$ at each timestamp i are defined as in (8) and (9). Knowing that $Corr(X,Q)$ has approximately normal distribution, we can easily answer the PTC queries. We next study the case that one of the input time series is standard.

5.1.2. **Correlation between Uncertain and Standard Time Series**

In this section, we study the notion of correlation between a standard and an uncertain time series, and show how to find solutions for the proposed PTC queries. We consider a standard time series $x =< x_1, \dots, x_n >$ as an uncertain time series $X =< X_1, \dots, X_n >$ with $E(X_i) = x_i$ and $Var(X_i) = 0$, for $1 \le i \le n$ and $f_{X_i}(y) = \begin{cases} 1, & y = x_i \\ 0, & Otherwise \end{cases}$.

Suppose that one of the uncertain time series, e.g., Q, in PTC queries is a standard time series,

q. To find the PDF of the uncertain correlation in PTC queries, in (15), $Z_i(1 \leq i \leq n)$ would be defined as $\hat{X}_i \hat{q}_i$, and hence the PDF of their product [ROS09] is defined as:

$$f_{Z_i}(y) = \frac{1}{|q_i|} f_{\hat{X}_i}\left(\frac{y}{q_i}\right)$$

The rest would be similar to the general case. For the i.i.d. case, we just need to replace $E(Q_i)$ and $Var(Q_i)$ by q_i and 0, respectively. In addition, if both X and Q are standard time series, expected value and variance of their correlation would be their exact Pearson sample correlation coefficient and zero, respectively. This shows that their correlation would be a certain value. For a normal distribution with zero variance, the cumulative distribution function is the Heaviside step function[1], that is:

$$P(Corr(X,Q) \geq c\ ) = \begin{cases} 1, & Corr(x,q) \geq c \\ 0, & Otherwise \end{cases}$$

Based on this, we can find the results of the PTC queries. In fact, our approach can support both standard and uncertain time series.

5.2. **Multiset-based Model**

Consider a multiset-based uncertain time series $X =< X_1, \dots, X_n >$, where the only available information about each X_i is a multiset of observed values $R_{X_i} = [\![x_{i,1}, x_{i,2}, \dots, x_{i,N_{X_i}}]\!]$, considered as X_i's realization. We use $[\![\dots]\!]$ for multisets, and N_{X_i} to denote the number of observed values for X_i. Unlike correlation between standard time series which is a real numbers, correlation between multiset-based uncertain time series is a multiset of real numbers, defined by all the possible correlation coefficients between the two given uncertain time series. This multiset of

[1] http://en.wikipedia.org/wiki/Normal_distribution#Zero-variance_limit

values can then be used to determine an approximation of the probability distribution of uncertain correlation.

Definition 5.1. Correlation multiset (CM)- *For uncertain time series* $X =< X_1, \dots, X_n >$ *with* $R_{X_i} = [\![x_{i,1}, x_{i,2}, \dots, x_{i,N_{X_i}}]\!]$, *let* T_X *be the multiset of all possible time series obtained by taking one value from each timestamp, that is:*

$$T_X = [\![< x_{1,1}, x_{2,1}, \dots, x_{n,1} >, \dots, < x_{1,N_{X_1}}, \dots, x_{n,N_{X_n}} >]\!] \quad (16)$$

The correlation multiset between two multiset-based uncertain time series X and Y is defined as follows:

$$CM(X,Y) = [\![corr(x,y): x \in T_X, y \in T_Y]\!] \quad (17)$$

where $corr(x,y) = \frac{\sum_{i=1}^{n} \hat{x}_i \hat{y}_i}{n-1}$, *and* $\hat{x}_i$ *and* $\hat{y}_i$ *are normalized as in Definition 3.3.*

Correlation multiset (CM) provides an approximation of uncertain correlation probability distribution. In this way, we can approximate $P(Corr(X,Q) \geq c)$ in the PTC queries. To estimate the true underlying CDF of the elements c_i in $CM(X,Q) = [\![c_1, \dots, c_M]\!]$, where $M = |CM(X,Q)|$, we use the empirical distribution function [SHO09] of $Corr(X,Q)$:

$$P(Corr(X,Q) \leq c) = \frac{\sum_{j=1}^{M} 1\{c_i \leq c\}}{M}$$

where $1\{c_i \leq c\}$ is the indicator function, which is equal to 1 if $c_i \leq c$, and equal to 0 otherwise. So the probability in the PTC queries is calculated as the fraction of correlation coefficients in $CM(X,Q)$, greater than or equal to *c*. More precisely:

$$P(Corr(X,Q) \geq c) = \frac{\left|[\![corr(x,q): x \in T_X, q \in T_Q, corr(x,q) \geq c]\!]\right|}{|CM(X,Q)|} \quad (18)$$

By the strong law of large numbers [ROS09], this estimation converges to the true CDF as $|CM(X,Q)|$ approaches infinity. Note that $|CM(X,Q)| = \prod_{i=1}^{n} N_{X_i} \times \prod_{i=1}^{n} N_{Q_i}$ and is huge when dealing with high dimensional data and high number of observed values. This requires computing all the correlation coefficients in $CM(X,Q)$, which is infeasible unless improved. In this thesis, we will explain our solution to overcome this complexity by providing an approximation technique for uncertain time series together with a pruning, following by a sampling-based heuristic technique. In the next section, we will present the results of our performance evaluation.

5.3. Performance Evaluation Results

In this section, we present our experimental results for both PDF-based and multiset-based models.

5.3.1. PDF-based Model

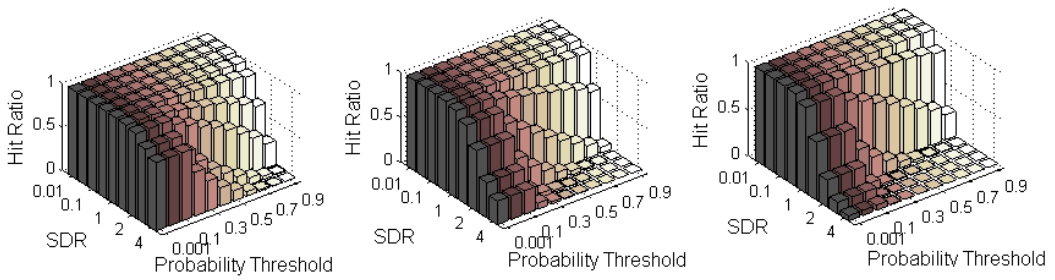
In this section, we study the performance of the probabilistic query for PDF-based model using the setup described in Chapter 4. For *RQ1*, defined in Section 4.1, the hit ratio and false alarm of the probabilistic query shown in Figure 4 and Figure 5, respectively, are for different error rates, SDRs and probability thresholds. We observe that for the SDRs higher than 0.1, the higher the error rate and the SDR, the lower the hit ratio. As discussed earlier, the expected value at each timestamp is estimated using an observed value. Moreover, recall that SDR and error rate specify uncertainty level in data, in that, the higher the uncertainty level, the farther the observed values from the exact values, and thus the lower the performance of the probabilistic query.

It should also be noted that as the probability threshold decreases, the hit ratio (Figure 4) increases and the false alarm ratio (Figure 5) decreases, in particular for high SDR values. This is

due to the fact that as we decrease the probability threshold, the probabilistic query returns more candidate results, which makes it more probable to contain the correct results. While this also increases the number of false alarms. The ratio of false alarms we found for all the error rates is low (less than 0.06).

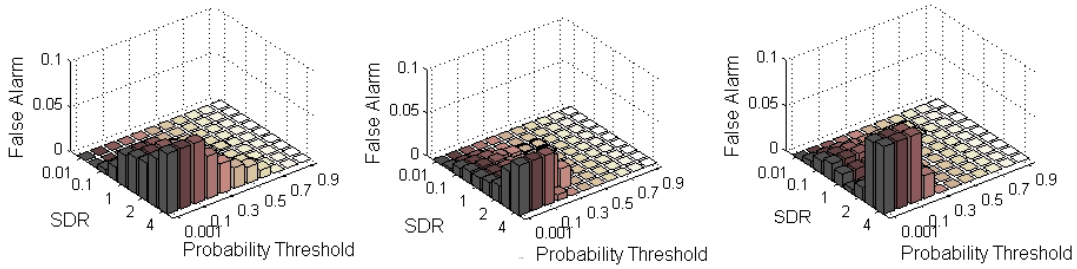
For *RQ2,* defined in Section 4.1, similar to earlier studies [YEH09, SAR10], we compared the performance of the probabilistic query with the deterministic query (Section 4.4). Figure 6 illustrates the hit ratio of the deterministic query for different error rates and SDRs. Since the false alarm ratio of the deterministic query was close to 0, we did not include the corresponding figure. As expected, the higher the SDR and the error rate, the lower the hit ratio. This is due to the fact that as uncertainty level increases, the uncertain observed value at each timestamp would be farther from the exact value. By comparing the deterministic and probabilistic queries, we observe that the probabilistic query has higher hit ratio than the deterministic query, in particular for high SDRs.

Figure 7 shows the performance of the deterministic and probabilistic queries for the i.i.d. case. Since the error rate is 100% in this case (i.e., the highest uncertainty level), both queries have lower hit ratio than any other cases. Moreover, for high SDRs, neither one returns any result. However, compared to the deterministic query, the probabilistic query has higher hit ratios for the SDR values 0.5, 1, and 1.5.



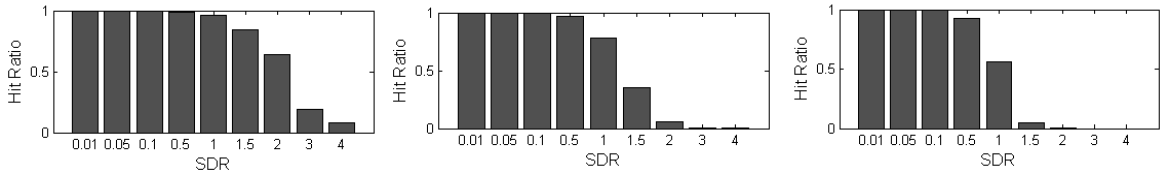
a) error rate=10% b) error rate=30% c) error rate=50%

Figure 4. Hit ratio of the probabilistic query using different error rates for Gun-Point dataset.



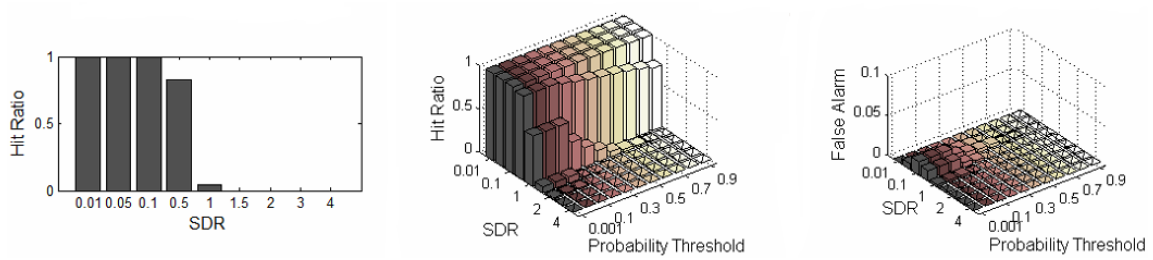
a) error rate=10% b) error rate=30% c) error rate=50%

Figure 5. False alarm ratio of the probabilistic query using different error rates for Gun-Point dataset.



a) error rate=10% b) error rate=30% c) error rate=50%

Figure 6. Hit ratio of the deterministic query using different error rates for Gun-Point dataset.



a) Hit ratio of the deterministic approach b) Hit ratio of the probabilistic approach c) False alarm of the probabilistic approach

Figure 7. Comparing the performance of the deterministic and probabilistic query for the i.i.d. case and Gun-Point dataset.

In all these experiments, we observed that in presence of high uncertainty level, the probabilistic query can find more correlated uncertain time series than the deterministic query. Another advantage of the probabilistic query (over the deterministic one) is that it allows exploring and analyzing the data by controlling the trade-off between false alarm and hit ratio. In some applications, this trade-off is important; for some applications false alarms are unacceptable or costly, while for others, it is required to have high hit ratio [YEH09].

We reported our findings for the Gun-Point dataset, when the error distribution was normal and correlation threshold was 0.5. We also studied the other five correlation thresholds from 0.4 to 0.9, and considered exponential and uniform error distributions. Moreover, we studied the effect of query and data parameters on the other 19 datasets. In all these cases, we made a similar observation as reported in this section. The complete set of experiments and results are made available to reviewers [CORX].

5.3.1.1. Experimental Results Analysis

For the i.i.d. case, we noted that deterministic and probabilistic queries have the same hit and false alarm ratios for probability threshold equal to 0.5. In this section, we will discuss its reason. Consider PDF-based uncertain time series, $X =< X_1, \dots, X_n >$ and $Q =< Q_1, \dots, Q_n >$. As discussed earlier, we will use an observed value as an estimate for the expected value at each timestamp. Thus, we would have:

$$E(Corr(X, Q)) = Corr(E(X), E(Q))$$

Where $E(X)$ is the expected value time series $E(X) =< E(X_1), \dots, E(X_n)$, and $Corr$ is defined as in (12). When probability threshold is set to 0.5, we would have: [2]

[2] For normal random variable *X*, we have: $P(X \leq x) = \frac{1}{2}\left(1 + erf\left(\frac{x - E(X)}{\sqrt{2Var(X)}}\right)\right)$

$$P(Corr(X,Q) \geq c) \geq 0.5 \Leftrightarrow erf\left(\frac{c - Corr(E(X),E(Q))}{\sqrt{2Var\big(Corr(X,Q)\big)}}\right) \leq 0 \Leftrightarrow Corr(E(X),E(Q)) \geq c$$

So X would be the candidate result of PTC queries if and only if X is the candidate result of DTC queries.

5.3.2. **Multiset-based Model**

The first objective of the experiments is to study the performance of the proposed method with regard to various data parameters and query parameters (Section 4.3). As explained earlier, the exhaustive technique, which calculates all the correlation coefficients in (18), is infeasible since its time complexity is $O(N^n)$, where N is the number of observed values at each timestamp and n is the dimension (length) of the uncertain time series. Thus, to make the similarity search feasible in different settings, similar to [DAL12], we reduced and used the input data, obtained by truncating the dataset to 50 time series of dimension 6 with 3 observed values at each timestamp. For example, given a correlation threshold c, probability threshold p, and SDR r, we need to do over 26.5 million calculations (with 50 time series) in the exhaustive technique, and in total over 15.7 billion calculations (with 9 SDR, 6 correlation thresholds, and 11 probability thresholds (Section 4.3)). This shows that even for small uncertain time series dataset, the exhaustive technique requires an excessive amount of processing time.

Figure 8 illustrates the answer for *RQ1* using the ground truth *I*. This figure shows the hit ratio and false alarm ratio of the exhaustive technique for different SDRs and probability thresholds for the Gun-Point dataset, normal distribution and correlation threshold 0.5. Generally, the hit ratio decreases and the false alarm ratio increases, as the SDR increases. Besides, the lower the probability threshold, the higher the hit ratio and false alarm ratio. This illustrates the role of the probability threshold in the trade-off between those measures.

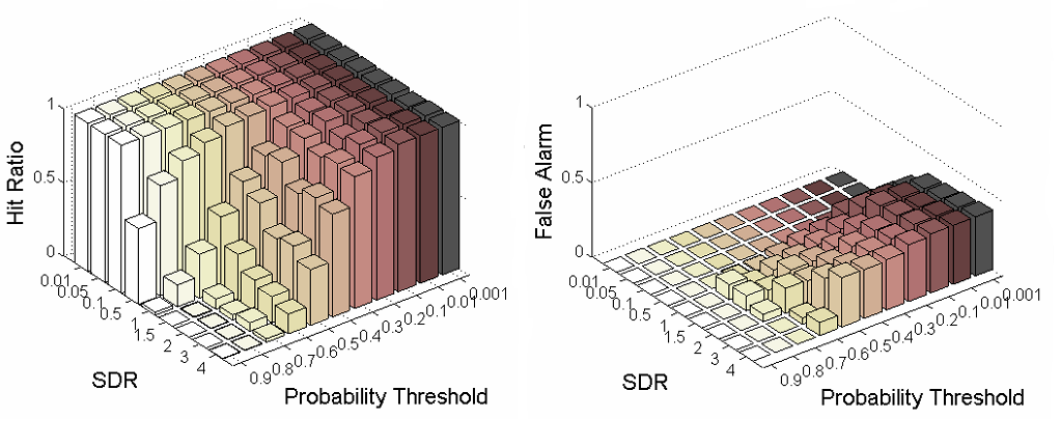


Figure 8. Hit ratio and false alarm ratio of exhaustive technique using ground truth *I* for Gun-Point dataset.

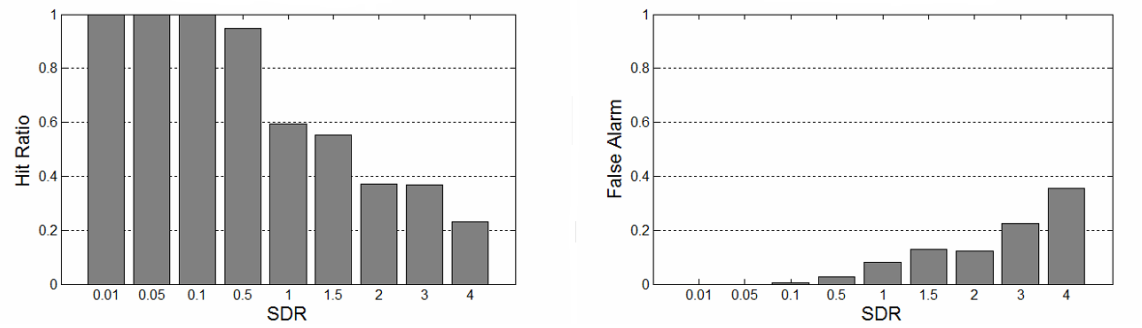


Figure 9. Hit ratio and false alarm ratio of deterministic query using ground truth *I* for Gun-Point dataset.

Moreover, for *RQ2*, we compare the probabilistic query with the deterministic query (Section 4.4). Figure 9 shows the false alarm and hit ratio of the deterministic query. The higher the SDR, the lower the hit ratio, and the higher the false alarm ratio. Unlike the deterministic query, the hit ratio of the probabilistic query can be 1 or close to 1 even for high SDRs, by choosing a proper probability threshold. For example, consider the SDR 4. For this SDR, we found the false alarm ratios for both queries are to be very close to each other. However, the hit ratio of the deterministic query is about 0.4, which is low, and in the probabilistic query, if we

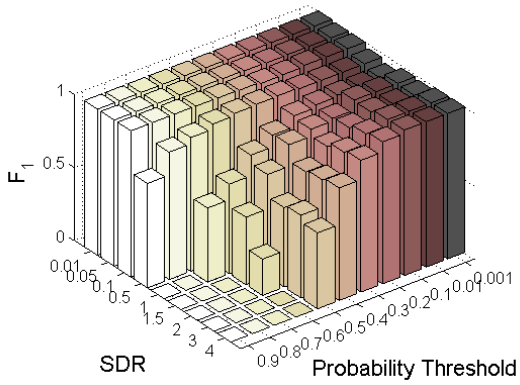


Figure 10. F_1 score of exhaustive approach using ground truth *II* for Gun-Point dataset.

choose a low probability threshold, e.g., 0.3, the hit ratio would be close to 1. That is, with the same uncertainty level, the probabilistic query outperforms the deterministic query.

Figure 10 illustrates the answer to *RQ3* for Gun-Point dataset, normal distribution and correlation threshold 0.5. It shows that for the SDRs less than 1, our approximation for the probability distribution of uncertain correlation approaches the exact probability distribution. For higher SDRs, the lower the probability threshold, the higher the F_1 score. For the cases where the F_1 score is 0, the ground truth did not return any result.

This section reports our findings for Gun-Point dataset, when error distribution is normal and correlation threshold is 0.5. We also considered five correlation thresholds from 0.4 to 0.9, for exponential and uniform error distributions. Moreover, we studied the effect of all data and query parameters on the other 19 datasets in the benchmark data. In all the cases studied, we made a similar observation as reported in this section. The complete set of results is made available to reviewers in [CORX].

5.4. **Discussion**

This dissertation has proposed suitable concepts and techniques for uncertain time series correlation analysis. Different applications consider different assumptions and have different similarity search queries. It is thus essential to investigate other types of similarity queries over uncertain time series under different assumptions. In PTC queries, the user is interested in the uncertain time series, in a database, which are positively correlated with Q, and their correlation is no less than the given threshold c. We also refer to this query as PTC-1 query. Depending on the application, other possible types of correlation queries includes (for $p \in (0,1]$):

PTC-2. $P(Corr(X,Q) \leq c) \geq p,\ c \in [-1,0)$

PTC-3. $P(|Corr(X,Q)| \geq c) \geq p, c \in (0,1]$

In PTC-2, the user is looking for uncertain time series which are negatively correlated with Q, and their correlation is no more than the threshold c. In the last one, the user is interested in uncertain time series that are highly correlated with Q (positively or negatively), and the absolute correlation is no less than c. In addition, the user may be interested in finding uncertain time series which are not highly correlated to Q, e.g., in medical domain looking for some abnormalities in a test result. We could thus have the following forms of threshold-based correlation queries:

PTC-4. $P(0 \leq Corr(X,Q) \leq c) \geq p,\ c \in [0,1)$

PTC-5. $P(c \leq Corr(X,Q) \leq 0) \geq p,\ c \in (-1,0]$

PTC-6. $P(|Corr(X,Q)| \leq c) \geq p, c \in [0,1)$

Our solution techniques for the PDF-based model, and multiset-based technique can answer all the above queries, because they can either find uncertain correlation PDF or approximate it. For the i.i.d. case of PDF-based model, since the correlation random variable is distributed normally,

its absolute value has folded normal distribution[3], and its cumulative distribution function is given by:

$$F_{|Corr(X,Q)|}(y;\mu,\sigma) = \frac{1}{2}[\text{erf}\left(\frac{y+\mu}{\sqrt{2}\sigma}\right) + \text{erf}\left(\frac{y-\mu}{\sqrt{2}\sigma}\right)]$$

For $y \geq 0$, where erf() is the error function, and μ and σ are the expected value and standard deviations of $Corr(X,Q)$. Hence in this case, we can also answer the PTC-3 and PTC-6 queries.

5.5. **Summary**

In this chapter, we presented our approach to process probabilistic threshold-based queries for both PDF-based and multiset-based models. For the former model we considered two cases; having different PDFs and identical PDFs at different timestamps. The results of our extensive experiments indicated that the probabilistic query, unlike the deterministic one, provides a trade-off between false alarm and hit ratio of the results which can be controlled by the probability threshold given by users. However, the similarity search technique proposed for multiset-based model is infeasible for real size uncertain time series unless improved. In the next chapter, we will explain our solution to overcome this complexity by providing a pruning, and a sampling-based heuristic technique.

[3] http://en.wikipedia.org/wiki/Folded_normal_distribution

Chapter 6: Query Optimization Techniques for Multiset-based Model

This chapter explains our solution to overcome the complexity of multiset-based similarity search by providing an approximation technique for uncertain time series which considers a Boolean representation together with a pruning, following by a sampling-based heuristic technique.

6.1. Probabilistic Pruning

In this section, we propose a probabilistic pruning technique which aims to cut down the number of candidates in a dataset of multiset-based uncertain time series. For this, we generalize the Boolean correlation proposed in [ZHA07] for standard time series. Given a standard time series $x =< x_1, \dots, x_n >$, its Boolean representation is a Boolean series $x^B =< x_1^B, \dots, x_n^B >$, in which

$$x_i^B = \begin{cases} 1, & x_i > \bar{x} \\ 0, & x_i \leq \bar{x} \end{cases} \quad (1 \leq i \leq n), \text{where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Let x and y be standard time series and x^B and y^B be their Boolean representations. Then their Boolean correlation [ZHA07] is defined as:

$$corr^B(x^B, y^B) = \frac{\sum_{i=1}^{n} \neg (x_i^B \oplus y_i^B)}{n} \tag{19}$$

where $\oplus$ and $\neg$ are the XOR and negation operations, respectively.

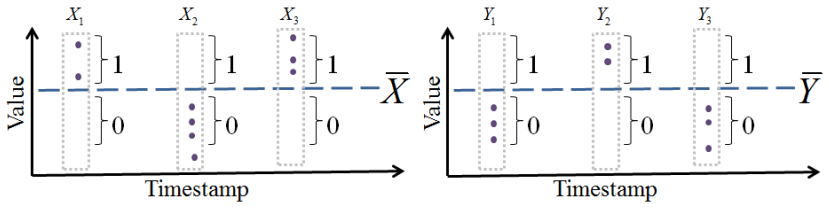


Figure 11. Uncertain time series *X* and *Y* with 3 timestamps.

We extend the standard Boolean correlation to uncertain time series, which basically replaces each observed value in an uncertain time series with a single bit. This yields a compression ratio of 32:1 (considering 32 bits for each observed value), and also allows taking advantage of fast bit operations by CPU.

Definition 6.1. Uncertain Boolean representation- *Given an uncertain time series* $X =< X_1, \dots, X_n >$ *with* $R_{X_i} = [\![x_{i,1}, x_{i,2}, \dots, x_{i,N_{X_i}}]\!]$, *we define its Boolean representation as* $X^B =< X_1^B, \dots, X_n^B >$, *with* $R_{X_i^B} = [\![x_{i,1}^B, x_{i,2}^B, \dots, x_{i,N_{X_i}}^B]\!]$, *in which*

$$x_{i,j}^B = \begin{cases} 1, & x_{i,j} > \bar{X} \\ 0, & x_{i,j} \leq \bar{X} \end{cases}, where \ \bar{X} = \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{N_{X_i}} x_{i,j} / N_{X_i} \right)}{n}$$

In this representation, each observed value would be represented as 1 if it is above the average of all the observed values, and as 0 otherwise. For example, Figure 11 shows uncertain time series X and Y with $n = 3$. For instance, consider X, its uncertain Boolean representation would be $X^B =< X_1^B, X_2^B, X_3^B >$, with $R_{X_1^B} = [\![1,1]\!]$, $R_{X_2^B} = [\![0,0,0,0]\!]$ and $R_{X_3^B} = [\![1,1,1]\!]$.

Using uncertain Boolean representation, we next present optimization techniques to speed up similarity search. The following example illustrates the idea behind this optimization. Again consider uncertain time series X and Y in Figure 11. To find the multiset $CM(X,Y)$

(Definition 5.1), we should find all the correlation coefficients between time series in T_X and T_Y defined in (16). Since the Pearson correlation detects linear dependencies between two time series, all the values in $CM(X,Y)$ would be a negative real number, thus, $P(Corr(X,Y) \geq c) = 0$, $c \in (0,1]$, and hence X would not be returned as a similar uncertain time series to Y. Now the question is: *can we prune X without having to calculate all the elements in $CM(X,Y)$?*

We show that the answer is positive and develop such a pruning technique which uses uncertain Boolean representation. For this, we next define the notion of uncertain Boolean correlation in which the XOR and negation operations are extended for random variables.

Definition 6.2. Uncertain Boolean correlation- *Let X^B and Y^B be uncertain Boolean representations of uncertain time series X and Y. Then their uncertain Boolean correlation is defined as:*

$$Corr^B(X^B, Y^B) = \frac{\sum_{i=1}^{n} \neg (X_i^B \oplus Y_i^B)}{n}$$

In the standard case, the Boolean correlation $corr^B(x^B, y^B)$ (19) is a real number indicating the Boolean correlation between the Boolean series x^B and y^B. For uncertain time series, however, Boolean correlation is a random variable. The above definition is used to define probabilistic threshold-based Boolean correlation queries.

Definition 6.3. Probabilistic threshold-based Boolean correlation (PTB) queries- *Given an uncertain time series Q as a query reference, a Boolean threshold β, and a probability threshold p^B, PTB queries return every uncertain time series X in D, such that X and Q are Boolean correlated with probability at least $p^B \in (0,1]$, and Boolean coefficient no less than $\beta \in (0,1]$. Formally,*

$$P(Corr^B(X^B, Q^B) \geq \beta) \geq p^B$$

To find the answers to the PTB queries, we define Boolean correlation multiset.

Definition 6.4. Boolean correlation multiset (BCM)- *For an uncertain time series X, let* X^B *be its uncertain Boolean representation, and* T_{X^B} *be the multiset of all possible Boolean series obtained by taking a value from each timestamp of* X^B*, that is,*

$$T_{X^B} = \left[\!\left[< x^B_{1,1}, x^B_{2,1}, \dots, x^B_{n,1} >, \dots, < x^B_{1,N_{X_1}}, \dots, x^B_{n,N_{X_n}} > \right]\!\right]$$

We regard each Boolean series in T_{X^B} *as a trend of the corresponding time series in* T_X (16)*. We define the Boolean correlation multiset between* X *and* Y *as:*

$$BCM(X, Y) = \left[\!\left[corr^B(x^B, y^B) : x^B \in T_{X^B}, y^B \in T_{Y^B} \right]\!\right]$$

Here, $corr^B(x^B, y^B)$, defined in (19), is the percentage of the number of the timestamps with the same Boolean values and indicates the degree of the similarity between the two Boolean series. Using BCM, we can find the complementary CDF of uncertain Boolean correlation between X and Q in the PTB queries, as follows:

$$P(Corr^B(X^B, Q^B) \geq \beta) = \frac{|[\![corr_B(x^B, q^B) \in BCM(X, Q) : corr^B(x^B, q^B) \geq \beta]\!]|}{|BCM(X, Q)|} \quad (20)$$

where $|BCM(X, Q)| = |CM(X, Q)|$, defined in Definition 5.1.

For instance, for uncertain time series X and Y in Figure 11, T_{X^B} includes 24 $(2 \times 4 \times 3)$ Boolean series $< 1,0,1 >$ and T_{Y^B} includes 18 $(3 \times 2 \times 3)$ Boolean series $< 0,1,0 >$. The Boolean correlation between Boolean series in T_{X^B} and the ones in T_{Y^B} is 0. Thus, the probability that X and Y are Boolean correlated is 0, i.e., $P(Corr^B(X^B, Y^B) \geq \beta) = 0$. Therefore,

the corresponding actual uncertain time series are not positively correlated and hence X is pruned and not considered (considering Y as the query reference).

To process the PTB queries $P(Corr^B(X^B, Q^B) \geq \beta) \geq p^B$, we calculate all possible Boolean correlation coefficients (20). We refer to this approach as *multiset-based Boolean approach*, since it uses the multiset of all possible Boolean correlation coefficients to find the final probability. As an alternative, to which we refer as *PDF-based Boolean approach*, to calculate this probability, we find PDF of random variable $Corr^B(X^B, Q^B)$ using the following lemma.

Lemma 6. *Given two uncertain time series X and Y with n timestamps,* $nCorr^B(X^B, Y^B)$ *has Poisson Binomial distribution.*

Proof. First, all X_i^B's and Y_i^B's $(1 \leq i \leq n)$ in uncertain Boolean representation of X and Y are Bernoulli random variables with success probabilities equal to p_i^X and p_i^Y, respectively, where p_i^X is defined as:

$$p_i^X = P(X_i^B = 1) = \frac{|[\![x_{i,j}: x_{i,j} > \bar{X}, 1 \leq j \leq N_{X_i}]\!]|}{N_{X_i}}$$

in which $\bar{X}$ is defined in Definition 6.1. Random variable $C_i^{X,Y} = \neg (X_i^B \oplus Y_i^B)$ is also a Bernoulli random variable, since the result of $C_i^{X,Y}$ is either 1 or 0, and its success probability would be:

$$p_i^{X,Y} = P\left(C_i^{X,Y} = 1\right) = P\left(X_i^B = 1\right)P\left(Y_i^B = 1\right) + P\left(X_i^B = 0\right)P\left(Y_i^B = 0\right) \qquad (21)$$

Since all $C_i^{X,Y}$'s are Bernoulli random variables, and independent of each other, with different success probabilities, we conclude that $nCorr^B(X^B, Y^B) = \sum_{i=1}^{n} C_i^{X,Y}$ would have Poisson Binomial distribution [HON13]. ■

The PDF of the Poisson binomial distribution (for probability of w successes in n trials, where trial i has success probability p_i) is as follows:

$$P(W = w) = \sum_{A \in S_w} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j)$$

where S_w is the set of all subsets of $\{1,2,\dots,n\}$ of size w, and A^c is the complement of the set A. Computing CDF of a Poisson binomial random variable could be prohibitively expensive, since S_w contains $n!/(n-w)!\,w!$ elements. Hong [HON13] proposed a simple method to derive an exact formula for CDF of Poisson binomial distribution, using the Fourier transform of the distribution characteristic function. The time to compute Hong's function is generally negligible for small number of timestamps, i.e., $n < 500$ [HON13]. Using this PDF-based Boolean approach, we can process the PTB queries $P(Corr^B(X^B, Q^B) \geq \beta) \geq p^B$ more efficiently. The following lemma shows that the probabilities obtained by both multiset-based (20) and PDF-based (Lemma 6) Boolean approaches are equal.

Lemma 7. *Given uncertain time series X and Y with n timestamps, we have:*

$$\frac{\left|\left[\!\left[corr^B(x^B, y^B) : x^B \in T_{X^B}, y^B \in T_{Y^B}, corr^B(x^B, y^B) = \beta \right]\!\right]\right|}{|BCM(X,Y)|}$$

$$= \sum_{A \in S_w} \prod_{i \in A} p_i^{X,Y} \prod_{j \in A^c} (1 - p_j^{X,Y})$$

where $w = n\beta$, S_w is the set of all subsets of $\{1,2,\dots,n\}$ of size w, and $p_i^{X,Y} = P(X_i^B = 1)P(Y_i^B = 1) + P(X_i^B = 0)P(Y_i^B = 0)$.

Proof. Let $w = n\beta$ and $S_w = \{A_1, \dots, A_k\}$ be the set of all subsets of $\{1,2,\dots,n\}$ of size w. By Lemma 6, random variable $nCorr^B(X^B, Y^B) = \sum_{i=1}^{n} C_i^{X,Y}$ (here $C_i^{X,Y} = \neg (X_i^B \oplus Y_i^B)$) has

Poisson binomial distribution with the following PDF (for the probability of w successes in n trials):

$$P\left(\sum_{i=1}^{n} C_i^{X,Y} = w\right) = \sum_{A \in S_w} \prod_{i \in A} p_i^{X,Y} \prod_{j \in A^c} (1 - p_j^{X,Y})$$

where A^c is the complement of set A and $p_i^{X,Y}$ is defined in (21). For the proof, we need to introduce some notations. Given a Boolean series $b =< b_1, \dots, b_n >$, its complement with respect to a given set $A \in S_w$ is denoted by b^A. Here b and b^A have the same values at w timestamps chosen from A's elements (and different values at other timestamps). For example, the Boolean series $< 1, 1, 0 >$ and $< 1, 0, 0 >$ are complement of each other with respect to the set $\{1,3\}$. Note that the Boolean correlation between b and b^A is equal to w/n, i.e., β. Given a Boolean series b and an uncertain time series X, we also define an uncertain time series $X(b)$ in which at timestamp i, it includes those observed values of X whose Boolean representation is b_i, that is $R_{X_i(b_i)} = [\![x_{i,j} : x_{i,j}^B = b_i, 1 \le j \le N_{X_i}]\!]$. Moreover, for a given b and $A \in S_w$, every element in $BCM(X(b), Y(b^A))$ is equal to β. We use B_n to denote the set of all Boolean series of length n. For example, $B_2 = \{< 0,0 >, < 0,1 >, < 1,0 >, < 1,1 >\}$. Using these notations, we rewrite the numerator of the left hand side of the equation in Lemma 7 as follows:

$$\left|[\![corr^B(x^B, y^B) : x^B \in T_{X^B}, y^B \in T_{Y^B}, corr^B(x^B, y^B) = \beta]\!]\right|$$

$$= \sum_{A \in S_w} \sum_{b \in B_n} |BCM(X(b), Y(b^A))|$$

Since the size of the multiset $BCM(X(b), Y(b^A))$ is $\prod_{i=1}^{n} |X_i(b_i)| \times |Y_i(b_i^A)|$, we have that:

$$\frac{\left|[\![corr^B(x^B, y^B) : x^B \in T_{X^B}, y^B \in T_{Y^B}, corr^B(x^B, y^B) = \beta]\!]\right|}{|BCM(X,Y)|} = \frac{\sum_{A \in S_w} \sum_{b \in B_n} \prod_{i=1}^{n} |X_i(b_i)| \times |Y_i(b_i^A)|}{\prod_{i=1}^{n} N_{X_i} \prod_{i=1}^{n} N_{Y_i}} \tag{22}$$

On the other hand, we know that the set B_n contains all possible Boolean series with length n. Besides, Boolean series b and b^A agree only on w timestamps chosen from the elements in A. Thus, for a given set $A \in S_w$, we have:

$$\begin{aligned} &\sum_{b \in B_n} \prod_{i=1}^{n} |X_i(b_i)| \times |Y_i(b_i^A)| \\ &= \prod_{i \in A} (|X_i(0)| \times |Y_i(0)| + |X_i(1)| \times |Y_i(1)|) \\ &\times \prod_{j \in A^c} (|X_j(0)| \times |Y_j(1)| + |X_j(1)| \times |Y_j(0)|) \end{aligned} \tag{23}$$

Moreover, according to (21), it holds that:

$$p_i^{X,Y} = \frac{|X_i(0)| \times |Y_i(0)| + |X_i(1)| \times |Y_i(1)|}{N_{X_i} N_{Y_i}}$$

$$1 - p_j^{X,Y} = \frac{|X_j(0)| \times |Y_j(1)| + |X_j(1)| \times |Y_j(0)|}{N_{X_j} N_{Y_j}}$$

Finally, from equations (22) and (23), we obtain:

$$\frac{\left|[\![corr^B(x^B, y^B) : x^B \in T_{X^B}, y^B \in T_{Y^B}, corr^B(x^B, y^B) = \beta]\!]\right|}{|BCM(X,Y)|} = \sum_{A \in S_w} \prod_{i \in A} p_i^{X,Y} \prod_{j \in A^c} (1 - p_j^{X,Y}) \quad \blacksquare$$

This lemma provides a basis for our pruning technique. Given two uncertain time series X and Q, with Q as the query reference, we can process the PTB queries $P(Corr^B(X^B, Q^B) \geq \beta) \geq p^B$ quickly using the Hong's method [HON13]. If this probability is less than p^B, we can prune away X, and avoid as many useless computations as $|CM(X,Q)|$. Note that $P(Corr^B(X^B, Q^B) \geq \beta)$ is the probability that at least as many as $n\beta$ $(\beta \in (0,1])$ of n timestamps of X and Q have the same Boolean values. Uncertain time series that do not have Boolean correlation are less likely to be much positively correlated. Thus, by choosing a proper Boolean threshold, we can prune uncorrelated or negatively correlated uncertain time series.

For the standard case, it is shown when the given standard time series have standard normal distribution, knowing the correlation threshold c, we can find the corresponding β [ZHA07]. However, in general, there is no direct relationship between the two thresholds. For the uncertain case, the results of our experiments suggest a way to find a range for a proper Boolean threshold, using which the PTB queries can have high recall and precision which can be controlled by the probability thresholds picked.

Another advantage of this solution is that for each uncertain time series X in the dataset, it replaces multiset of observed values at timestamp i with a single value $P(X_i^B = 1)$. This yields a compression ratio of $N_{X_i}: 1$ for timestamp i. However, if the number of observed values at each timestamp is less than 32 (considering 32 bits for each observed value), we can improve space utilization even more by storing the uncertain Boolean representation (Definition 6.1) of each uncertain time series X in the dataset and its average $\bar{X}$ (to be able to calculate $P(X_i^B = 1)$). This yields a compression ratio of $32: 1$.

We remark that when the input time series is a standard one, we can still benefit from the proposed pruning technique. When both input time series are standard, the proposed uncertain Boolean correlation reduces to standard Boolean correlation [ZHA07].

6.2. **Sampling-based Heuristic**

Given an uncertain time series Q, a set D of uncertain time series, a probability threshold p, and a correlation threshold c, we want to speed up the processing of the PTC queries $P(Corr(X,Q) \geq c) \geq p$ for all X in D. For example, suppose that there are d uncertain time series in D, each uncertain time series X in D is n-dimensional, and there are N observed values at each timestamp. Then to process $P(Corr(X,Q) \geq c) \geq p$, the number of pairwise correlation calculations would be $d \times N^{2n}$. Now consider just one uncertain time series in D, and suppose $N = 2$ (a very small number of observed values) and $n = 50$ (a short uncertain time series). The number of required calculations to find the corresponding probability would be 2^{100}, which is infeasible to do. Since time complexity of multiset-based similarity search is exponential in the dimension (length) of uncertain time series, and usually uncertain time series are high dimensional, reducing the number of observed values at each timestamp will not help much to reduce the processing time.

We propose to use "Sampling" which selects a subset of samples from a statistical population to estimate the characteristics of the whole population. In our case, the population would be all the elements in $CM(X,Q)$ (Definition 5.1). However, we do not want to find all the elements in this multiset. As a matter of fact, we need a method to create a subset of independent samples from $CM(X,Q)$. Having a multiset as the correlation between two multiset-based uncertain time series means that the correlation random variable $Corr(X,Q)$ is a discrete random variable equal to: $Corr(X,Q) = \frac{\sum_{i=1}^{n} \hat{X}_i \hat{Q}_i}{n-1} = \sum_{i=1}^{n} Z_i$. To find the PDF of this discrete random variable, we first

obtain the PDF of $Z_i (1 \leq i \leq n)$, which can be done using the multiset of all possible values for Z_i expressed as:

$$R_{Z_i} = [\![x.q/(n-1): x \in R_{\hat{X}_i}, q \in R_{\hat{Q}_i}]\!] = [\![z_{i,1}, \ldots, z_{i,N_{Z_i}}]\!]$$

where $N_{Z_i} = N_{X_i} N_{Q_i}$. We then find the multiset of all possible values for $Corr(X, Q) = \sum_{i=1}^{n} Z_i$, done recursively as:

$$Y_i = Y_{i-1} + Z_i, R_{Y_i} = [\![y + z: y \in R_{Y_{i-1}}, z \in R_{Z_i}]\!], 2 < i \leq n$$

where $Y_2 = Z_1 + Z_2$ and $\sum_{i=1}^{n} Z_i = Y_n$.

As mentioned earlier, we do not want to construct all the correlation coefficients. To select a subset of size S from $CM(X, Q)$, for each i $(1 \leq i \leq n)$, we take a sample from the multiset R_{Z_i} randomly and sum the n samples together. This process is repeated S times. Note that this is a sampling with replacement. Formally,

$$R_S = [\![s_{1,j} + \cdots + s_{n,j}: s_{i,j} \in R_{Z_i}, 1 \leq i \leq n, 1 \leq j \leq S]\!]$$

We remark that in our sampling technique, each sample in R_S is the average of n samples taken randomly at each timestamp from the multiset obtained by multiplication of observed values at that timestamp. To determine the sampling size S, we use the Dvoretzky-Kiefer-Wolfowitz inequality [DVO56, MAS90], which helps predict how close an empirical distribution function will be to the distribution function from which the samples are chosen. The empirical distribution function F_S for S number of observed values $x_1, \ldots, x_S$ is defined as $F_S(x) = \sum_{i=1}^{S} 1\{x_i \leq x\}$, where $x \in \mathbb{R}$ and $1\{x_i \leq x\}$ is the *indicator* function. The Dvoretzky-Kiefer-Wolfowitz inequality is defined as:

$$P(sup_{x\in\mathbb{R}}|F_S(x) - F(x)| > \varepsilon) \leq 2e^{-2S\varepsilon^2}, \forall\varepsilon > 0$$

where $sup(C)$ is the supremum of the set C of distances. To determine the minimum sampling size S, we can rewrite this inequality as:

$$S \geq ln(2/\alpha)\ /(2\varepsilon^2) \tag{24}$$

where ε is half-width of the confidence interval and determines how close the empirical distribution function would be to the corresponding distribution function, and $1-\alpha$ is the confidence level. For example, to estimate $F(x)$ within $\varepsilon = 0.01$ with 95% confidence, the inequality yields a minimum sample size of $S = 18{,}445$.

If there are N observed values at each timestamp, the complexity of the sampling technique would be $O(nN^2)$, which shows significant improvement compared with $O(N^n)$ in the exhaustive technique. The practical advantage of the sampling technique is also shown by the experiments, while providing a good approximation for the distribution function of uncertain correlation. This technique can also be used for finding correlation between a standard time series and an uncertain time series.

Similar to PTC queries, processing PTE queries (defined in Section 2.5) for multiset-based uncertain time series has been challenging for excess computational cost. One way to address this, reported in [DAL12], is to truncate the input uncertain time series to much smaller length, e.g., 6 timestamps, which seems short and limits its applications. The proposed heuristic can also be adapted for processing PTE queries as follows.

$$Eucl(X,Q) = \sum_{i=1}^{n}(X_i - Q_i)^2 = \sum_{i=1}^{n} Z_i$$

To find the PDF of this discrete random variable, we first obtain the PDF of $Z_i (1 \leq i \leq n)$, which can be done using the multiset of all possible values for Z_i expressed as:

$$R_{Z_i} = [\![(x-q)^2 : x \in R_{X_i}, q \in R_{Q_i}]\!] = [\![z_{i,1}, \ldots, z_{i,N_{Z_i}}]\!]$$

where $N_{Z_i} = N_{X_i} N_{Q_i}$. The rest is similar to the sampling based heuristic defined for uncertain correlation. We next present similarity search techniques for multiset-based uncertain time series.

6.3. **Similarity Search Techniques for Multiset-based Model**

In this section, we present the multistep algorithm for processing the PTC queries for the multiset-based model. Note that the proposed optimization techniques are independent of each other and can be used separately. Given a set *D* of uncertain time series and an uncertain time series Q, we are looking for every time series *X* in *D* which is positively correlated with Q with a probability at least p and their correlation is no less than c. That is, $P(Corr(X,Q) \geq c) \geq p$. For this, we use the probabilistic pruning introduced earlier in Section 6.1 to cut down the number of candidate uncertain time series in *D*. Using Lemma 6, we determine the probability $P(Corr^B(X^B, Q^B) \geq \beta)$. If this probability is less than the given probability threshold p^B, we prune *X*; otherwise, we use the sampling-based heuristic (Section 6.2) to estimate the probability $P(Corr(X,Q) \geq c)$. If this probability is at least p, we return *X* as a result. These steps are formally presented as an algorithm in Figure 12.

Algorithm 1: Finding uncertain time series satisfying the PTC queries

Input: Set D of uncertain time series, uncertain time series query Q, probability threshold p, and correlation threshold c

Output: uncertain time series X in D that are "highly" correlated to Q.

Procedure:

Find uncertain Boolean representation of Q

for all $X \in D$:

Find uncertain Boolean representation of X

if $P(Corr^B\ (X^B, Q^B\ ) \geq \beta) < p^B$, then prune X

else

Use sampling-based heuristic to calculate $P(Corr(X, Q) \geq c)$

if $P(Corr(X, Q) \geq c) \geq p$, then return X

end if

end else

end for

Figure 12. Similarity search for multiset-based model.

6.4. **Experimental Results**

In this section, we study the results of our performance evaluation using the setup described in Chapter 4.

6.4.1. **Probabilistic Pruning**

To measure the effectiveness of our probabilistic pruning, we calculate the hit ratio (recall) and precision (1-false alarm ratio) of the PTB queries (Definition 6.3) using ground truth I. Since the number of observed values did not have much effect on the results, we report the results for uncertain time series with 6 observed values at each timestamp. Figure 13 shows the hit ratio and

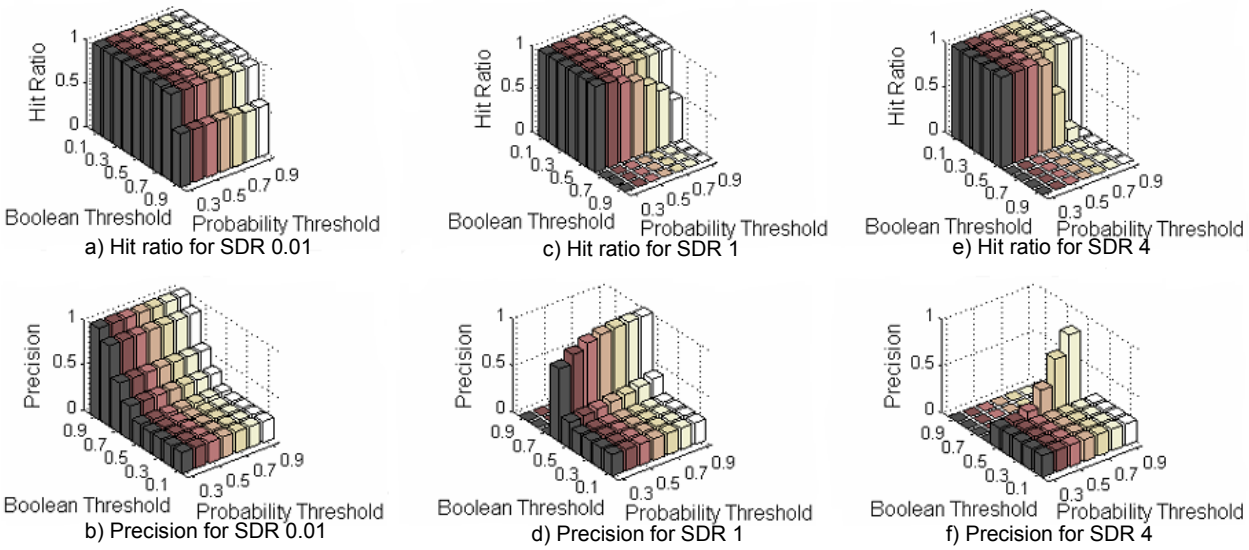


Figure 13. Hit ratio and precision of the probabilistic pruning using ground truth *I* for Trace dataset, number of observed values 6 and correlation threshold 0.5.

the precision of the PTB queries for correlation threshold 0.5 for different SDRs. In all the cases, when the Boolean threshold is less than 0.5, the PTB queries return every uncertain time series in the dataset, of which about 20% are correct (since the precision is around 0.2 and hit ratio is 1).

For the SDR 0.01 (Figure 13 (a) and (b)), PTB queries start pruning for Boolean threshold higher than 0.5. In this case, the optimum Boolean threshold would be 0.8, since the hit ratio is close to 1 and the precision is around 0.9. This means for the Boolean threshold 0.8, the PTB queries prune most of the uncertain time series that are not correct results. In this case, probability threshold does not have any effect on the performance of probabilistic pruning, since the uncertainty level is very low and the probability that the given uncertain time series are Boolean correlated is very close to either 1 or 0.

For the SDR equal to 1 (Figure 13 (c) and (d)), the PTB queries start pruning for Boolean threshold higher than 0.4. For the Boolean threshold 0.5, the hit ratio is 1, and the precision

increases as the probability threshold increases. But for the Boolean threshold 0.6, there is a trade of between hit ratio and precision. For higher Boolean thresholds, the PTB queries do not return any result. For the SDR 4 (Figure 13 (e) and (f)), the PTB queries prune uncertain time series properly only when the Boolean threshold is 0.5 and probability threshold is less than 0.6.

We also studied the effect of the correlation threshold on the probabilistic pruning performance. Figure 14 shows the hit ratio and precision of the probabilistic pruning for correlation thresholds 0.3 and 0.7. Similar to the case for correlation threshold 0.5 (Figure 13 (c) and (d)), for Boolean threshold less than 0.5, the PTB queries do not prune any data. For Boolean threshold 0.5, the probabilistic pruning has hit ratio 1. Moreover, its precision increases as the probability threshold increases. For Boolean threshold 0.6, the higher the correlation threshold, the higher the hit ratio. Thus, for the correlation threshold 0.7, we can use Boolean threshold 0.6 to prune more uncertain time series.

Figure 13 and Figure 14 show that choosing proper Boolean and probability threshold depends on the uncertainty level and correlation threshold. Moreover, the results from all UCR

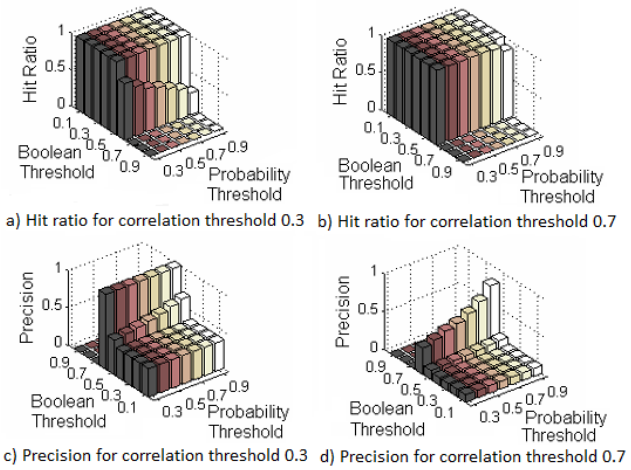
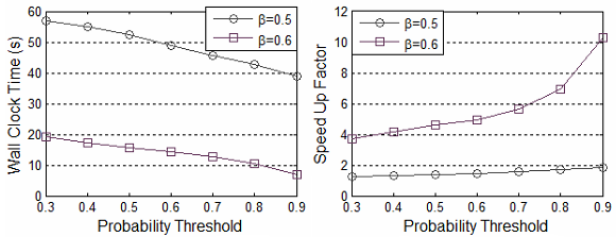


Figure 14. Hit ratio and precision using ground truth *I*, for Trace dataset, number of observed values equal to 6, and SDR 1.

datasets show that these thresholds are also dataset-dependent. However, our observation on all the UCR datasets show we can find the range for proper Boolean thresholds as follows. We can choose a small portion of the dataset, e.g. 10%, and find the results of the PTB queries. The range would be all the Boolean thresholds for which the result includes some time series in the dataset but not all of them. In other words, the range would be Boolean thresholds which help probabilistic pruning to prune some but not all uncertain time series in the dataset. For example, for the Trace dataset, when correlation threshold is 0.7 (Figure 14 (b) and (d)), we found the proper Boolean threshold to be between 0.5 (for which probabilistic pruning starts pruning) and 0.6 (which is the largest threshold for which probabilistic pruning does not prune all the uncertain time series in the dataset).

Figure 15 (a) shows wall clock time of processing the PTC queries using both probabilistic pruning and sampling-based heuristic for 6 observed values, and SDR 1 with different Boolean and probability thresholds. The higher the probability and Boolean threshold, the lower the execution time. We also processed the PTC queries using only sampling-based heuristic in 72 seconds wall clock time. Using this information, we measured the speed up factor for probabilistic pruning. The speed up factor is defined as the ratio between the execution time of the sampling-based heuristic without and with probabilistic pruning. Figure 15 (b) shows that the



a) Execution time b) Speed up factor

Figure 15. The execution time and speed up factor of the PTC queries using both probabilistic pruning and sampling-based heuristic for Trace dataset, number of observed values 6, and SDR 1.

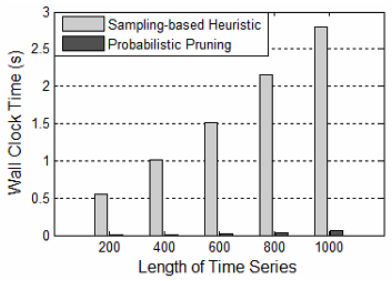


Figure 16. The execution time for number of observed values 6, and SDR 1 with different lengths.

speed up factor varies between 1 and 10. For example, according to Figure 13 (c) and (d), we now know the best thresholds for probabilistic pruning is 0.6 for the Boolean and 0.4 for the probability threshold. Using these thresholds, the speed up factor would be around 4.

Figure 16 shows the wall clock time to process the PTC queries between two uncertain time series using sampling-based heuristic and to process the PTB queries between the same two uncertain time series. Here, we varied uncertain time series length from 200 to 1000. Figure 16 shows that as the uncertain time series length grows, the execution time increases for both sampling-based heuristic and probabilistic pruning. However, the difference between the execution time of these two techniques is huge, and increases as the uncertain time series length increases. This difference also shows how much probabilistic pruning can save in the execution time. An additional remark is that varying the number of observed values had no effect on either techniques.

In this section, we reported our findings for Trace dataset, when error distribution is normal. We also considered exponential and uniform error distributions. For the other 19 datasets, we also studied the effect of all the query and data parameters on probabilistic query. In all of the studied cases, we made a similar observation as reported in this section. The complete set of results is available online in [CORX] for review.

6.4.2. **Sampling-based Heuristic**

Following inequality (24), in order to determine the minimum sampling size, we consider $\varepsilon = 0.01$ with 99% confidence level in different settings. For the sampling-based heuristic, we also study the effect of the number of observed values, *N*, at different timestamps, varied from 2 to 10, as done in [ASF09]. Figure 17 and Figure 18 show the answer to *RQ1*, and illustrate the effect of SDRs, the number of observed values (*N*), and the probability threshold on the hit ratio and false alarm ratio of the sampling-based heuristic. Only when *N*=2, we found the hit ratio to be less than other values of *N* in particular for high SDRs. The results indicate that in general, *N* does not have much effect on the performance of the sampling-based heuristic for SDRs less than 2.

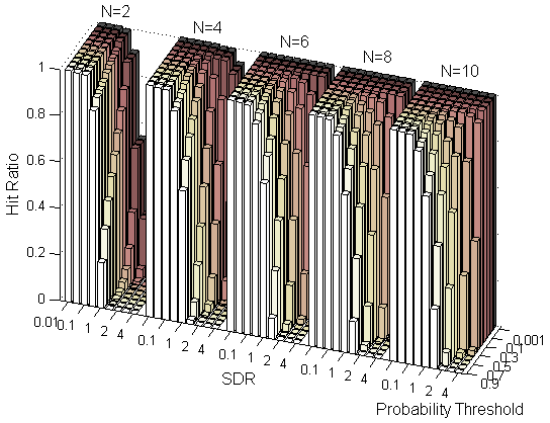


Figure 17. Hit ratio of sampling-based heuristic using ground truth *I* for Gun-Point dataset.

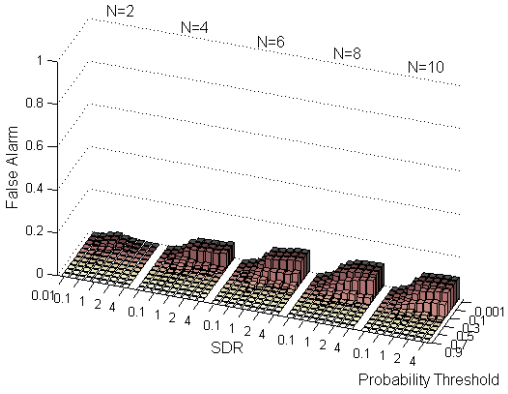


Figure 18. False alarm ratio of sampling-based heuristic using ground truth *I* for Gun-Point dataset.

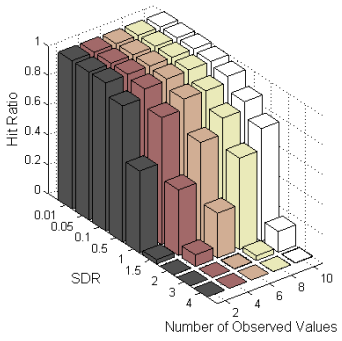


Figure 19. Hit ratio of the deterministic query using ground truth *I* for Gun-Point dataset.

For higher SDRs, a higher *N* resulted in a higher hit ratio.

We also compare the probabilistic and deterministic queries described as *RQ2* in Section 4.1. Figure 19 shows the effect of the SDR and number of observed values, *N*, on the hit ratio of the deterministic query. Since the false alarm of this approach was close to 0, we did not include the corresponding figure. As can be seen, the performance of the deterministic query depends on *N*, especially for SDRs larger than 0.1. The higher the number *N*, the higher the hit ratio. This confirms the law of large numbers [ROS09], which asserts the higher the number of observed values, the closer would become their average to the expected (i.e., exact) value.

Comparing the results of Figure 17 with Figure 19, we note that the probabilistic query is more resilient to the uncertainty level, in particular when the SDR is greater than 1. The hit ratio of the deterministic query decreases significantly for the SDRs larger than 2. However, in the probabilistic query, even for high SDRs, we can have high hit ratios by choosing a proper probability threshold.

We also study *RQ3* for this set of evaluations, for varying number of observed values at each timestamp. Figure 20 illustrates the F_1 score in the sampling-based heuristic for the ground truth *II*. For SDRs less than 1, the F_1 score is 1 or close to 1. This means that our approximation for the probability distribution of uncertain correlation approaches the exact probability distribution.

However, for higher SDRs, the lower the probability threshold and the higher the number of observed values, the higher the F_1 score.

In this section, we reported our results for the Gun-Point dataset, for when the error distribution is normal and correlation threshold is 0.5. We also studied the other five correlation thresholds from 0.4 to 0.9, for exponential and uniform error distributions. In addition, we studied the effect of all these query and data parameters on the other 19 datasets. As the results and observations were similar to those reported here for the Gun-Point dataset, we do not report them in this paper, however, they are available online for review [CORX].

Another important issue is the effect of random selection on the sampling-based heuristic. Given an uncertain dataset *D* and an uncertain time series *Q*, to find the result of the PTC queries, we use the sampling-based heuristic technique 30 times. Thus, we would have 30 sets of results. We compare and measure the similarity between these sets using the *Jaccard similarity coefficient* [LEV71], defined as:

$$J(A_1, \ldots, A_n) = \left| \bigcap_{1 \le i \le n} A_i \right| / \left| \bigcup_{1 \le i \le n} A_i \right| \tag{25}$$

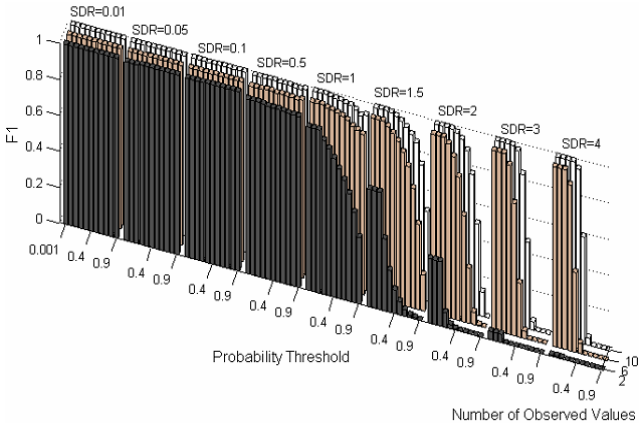


Figure 20. F_1 Score of sampling-based heuristic using ground truth *II* for Gun-Point dataset.

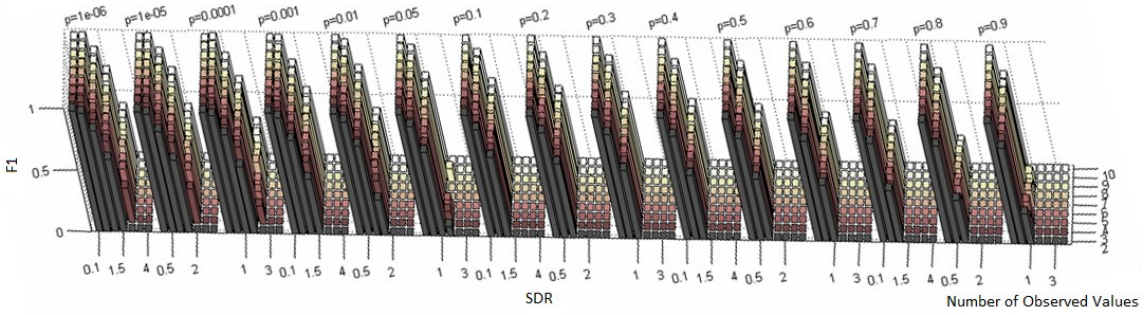


Figure 21. F_1 score of the multiset-based approach for the Gun-point dataset using ground truth *I*.

where $0 \leq J(A_1, \ldots, A_n) \leq 1$. Absolute value signs are used to indicate number of elements. The higher the Jaccard coefficient, the more similar are the given sets. In our experiments, with 9 SDRs, 6 probability and correlation thresholds, we have 324 different setups. For each setup, we measure the similarity between the 30 result sets using Jaccard similarity coefficient. In 90% of 324 setups, the Jaccard coefficient is 1, i.e., all 30 result sets are identical. The rest has an average value of 0.97, and standard deviation of 0.03. This shows the random sampling selection in the sampling-based heuristic has no or negligible effect on the result of the PTC queries. This shows that the random sampling selection does not interfere with the outcome of the sampling-based heuristic, while it makes similarity search on multiset-based uncertain time series feasible.

6.4.2.1. Multiset-based Euclidean Distance

In this section, we present our performance evaluation results for multiset-based Euclidean distance for PTE queries defined in Section 2.5. Evaluations of uncertain Euclidean distance for PDF-based model, i.e., PROUD [YEH09] and uncertain correlation revealed existence of a trade-off between hit ratio and false alarm ratio which could be controlled by a user defined probability threshold. However, for multiset-based Euclidean distance, there is no experimental evaluation on large uncertain time series, due to huge amount of required calculations. We will use the proposed sampling-based heuristic method to overcome this problem.

Figure 21 illustrates the impact of probability threshold, SDR, and number of observations on the F_1 score of the multiset-based approach. Generally, the F_1 score increases as the probability threshold decreases. For SDR higher than 1, the higher the number of observed values the higher the F_1 score. Therefore, when the uncertainty level is high, we require more information on unknown values at each timestamp (i.e., more observed values) to have higher accuracy. On the other hand, for smaller SDR values, the number of observed values does not have much effect on the F_1 score. Besides, the F_1 score increases when SDR decreases, since the lower the SDR, the closer the observed values would be to the unknown true values.

We also measured the hit ratio and false alarm for the multiset-based approach. Figure 22 shows the effect of the probability threshold, the number of observed values, and SDR on the hit ratio and false alarm for the multiset-based approach. The lower the probability threshold, the higher the hit ratio and false alarm ratio. Thus, similar to probabilistic similarity measures, in the multiset-based approach, there is always a trade-off between hit ratio and false alarm ratio which a user may wish to control.

Moreover, we observed that number of observed values does not have a significant effect on the hit ratio of the multiset-based approach for high probability thresholds. For small probability thresholds, the higher the number of observed values, the lower the hit ratio. Besides, as expected, the higher the SDR, the lower the hit ratio.

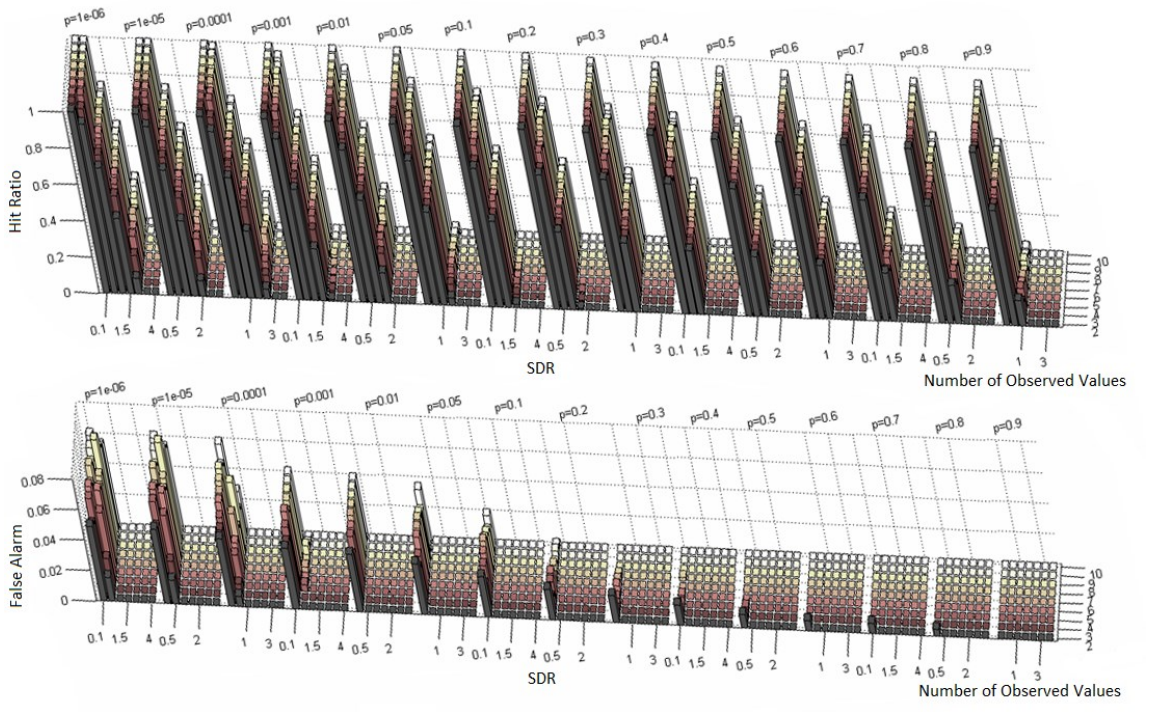


Figure 22. Hit ratio and false alarm ratio of the multiset-based approach for the Gun-point dataset using ground truth *I*.

6.5. **Summary**

In this chapter, to speed-up processing queries over multiset-based model, we proposed a probabilistic pruning which cuts down the number of candidates in the dataset. This includes a Boolean representation technique for uncertain time series, in which, each observed value is replaced with a single bit. In addition to saving memory, this enjoys fast bit operations. Moreover, we introduced another representation which replaces a multiset of observed values at each timestamp with a single value that is the probability the Boolean representation at that timestamp is equal to one. Using this, we introduced uncertain Boolean correlation together with an effective probabilistic pruning strategy. The proposed solutions are also applicable in finding correlations between uncertain and standard time series. We conducted numerous experiments

using the UCR benchmark dataset [KEO]. The results show the effectiveness of the proposed pruning.

We also introduced a sampling-based heuristic that approximates the distribution of uncertain correlation effectively and reduces the computation time significantly. This technique can also be adapted for processing PTE queries (defined in Section 2.5) over multiset-based uncertain time series. Note that the only solution suggested for this case in [DAL12] was to truncate the lengths of the uncertain time series to a small value, e.g., 6 timestamps. Thus our technique solves the problem of prohibitive similarity search for real-life size multiset-based data. Our experimental results also illustrate that while our sampling-based heuristic reduces the time complexity, it finds a good approximation for distribution of uncertain correlation.

Chapter 7: Performance Improvement of Similarity Search

As discussed in Chapter 2, a number of similarity measures have been proposed and used for uncertain time series to quantify similarities required in different analysis tasks [ASF09, DAL11, ORA12, ORA14, SAR10, YEH09, WU12]. Two approaches have been employed to develop similarity measures for uncertain time series. One approach considered "using" the similarity measures originally proposed for standard time series [DAL12, ORA14], which we refer to as *traditional similarity measures*. As for a second approach, a number of similarity measures have been proposed specifically for uncertain time series [ASF09, ORA12, SAR10, YEH09, WU12], and are obtained by "adapting" the traditional similarity measure. We refer to these as *uncertain similarity measures*. The difference between the traditional and uncertain similarity measures is in the information used to quantify the similarity. Traditional similarity measures use only a single uncertain value at each timestamp to represent the unknown exact value at that timestamp whereas uncertain similarity measures exploit more information including some statistical information that represents the uncertainty level at each timestamp.

It has been shown that traditional similarity measures outperform uncertain similarity measures in general [DAL12, ORA14]. The reason we noted for the superiority of traditional similarity measures is that they use a preprocessing step. In fact, we found missing the corresponding preprocessing step for uncertain similarity measures rather surprising, which explains our motivation here to study the impact of preprocessing on performance of uncertain similarity measures.

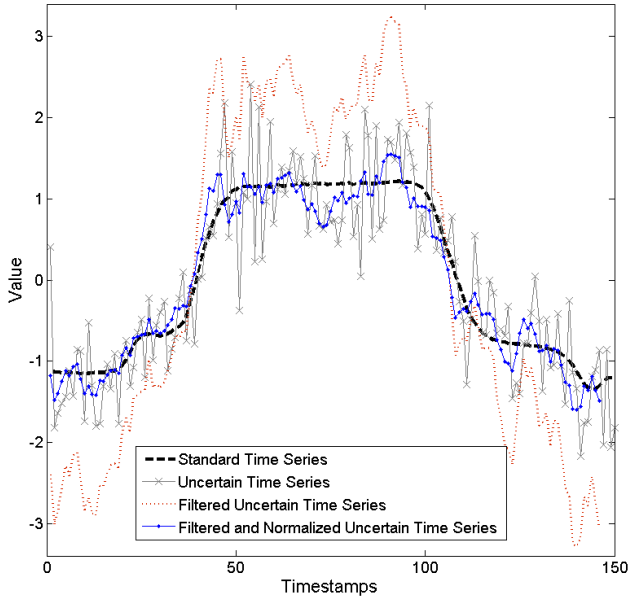


Figure 23. The impact of different preprocessing techniques on uncertain time series.

Preprocessing uncertain values leads to improved estimates of the exact unknown values and similarity search [DAL12]. Hence, the result of traditional measures on preprocessed uncertain time series data is more accurate than non-preprocessed data. Figure 23 illustrates this point. The black dashed line demonstrates a normalized standard time series for the Gun-Point dataset [KEO] with the exact values. Suppose that this data is observed as an uncertain time series (shown by the gray line marked with stars). By applying the uncertain moving average filter [DAL12], the uncertain time series would become smoother (shown in red dotted line). However, this filter changes the scale of the data, resulting in values that are farther from the exact values when compared to the uncertain values in the gray data. To solve this problem, we apply normalization, yielding the uncertain time series shown as the blue dotted line, which provides a better approximation of the exact values.

Since uncertain similarity measures also use uncertain values, e.g. the gray line in Figure 23, we expect the preprocessing of data to improve the quality of uncertain measures. In this paper

we study the impact of preprocessing on uncertain similarity measures for two classes of problems (defined in Section 2.2): probabilistic measures including PROUD [YEH09] and uncertain correlation (defined in Section 5.1.1), and deterministic measures including the DUST proposal [SAR10]. We also compare uncertain similarity measures to traditional similarity measures with and without data preprocessing.

In our study, we consider filtering and normalization as preprocessing techniques. For filtering, we consider the three methods discussed in [DAL12], which include simple moving average, uncertain moving average, and uncertain exponential moving average filters. These filters help smooth out fluctuations in uncertain time series data. For normalization, we consider two normalization methods for standard time series [SHA04] and for uncertain time series defined in Section 3.1.1. The aim of normalization methods is to transform all data in a dataset to the same baseline and scale.

In our experimental evaluation of the proposed methods, we use the UCR benchmark [KEO]. Our findings are as follows:

- We observe that the uncertain similarity measures can outperform the traditional similarity measures with and without data preprocessing. This indicates the effectiveness of uncertain similarity measures in practice.

- Our results show that preprocessing is necessary for similarity search in uncertain time series. Moreover, our results indicate that simple and uncertain moving average filters improve the performance of the probabilistic measures more than uncertain exponential moving average filter.

- We propose an enhancement for the PROUD similarity measure [YEH09], which improves its performance with and without data preprocessing.

These findings provide a better understanding of the proposed similarity measures, which in turn yield more effective techniques for analysis and mining uncertain time series [AGG13, BER13, JIA13, LIA09, MA11]. The rest of this chapter is organized as follows: Sections 7.1 and 7.2 reviews work on similarity measures and preprocessing techniques for uncertain time series. Section 7.3 studies the effect of preprocessing techniques on probabilistic and deterministic similarity measures. The results of our experiments are presented in Section 7.4.

7.1. Uncertain Similarity Measures

In this section, we review similarity measures for uncertain time series. As discussed in Section 2.2, existing similarity search methods for uncertain time series generalize the corresponding methods proposed for standard time series. They include the Euclidean distance [YEH09, SAR10] and the Pearson correlation coefficient that was introduced in this thesis. We classified similarity measures in Section 2.2 as *probabilistic similarity measures* and *deterministic similarity measures*. For the deterministic similarity measure, we study DUST [SAR10], and for the probabilistic similarity measures, we consider uncertain correlation (Section 5.1.1), which extends the Pearson correlation, and PROUD [YEH09] which extends the Euclidean distance for uncertain time series.

7.2. Preprocessing Techniques

Data preprocessing is an important step in similarity search. As shown in [DAL12] and [ORA14], the preprocessing techniques can improve the performance of traditional similarity measures. In this paper, we stress the use of these techniques for uncertain similarity measures as well. Let us begin with an overview of the preprocessing techniques including moving average filters and normalization.

7.2.1. **Moving Average Filters**

In time series data, usually the values of adjacent timestamps are correlated. Using this as a basis, moving average filters smooth time series data by averaging adjacent values. There are different variations of moving average filters, including simple, uncertain, and uncertain exponential moving average filters [DAL12].

For a given uncertain time series $x = < x_1, \dots, x_n >$ in the form of a sequence of observed values, different variations of moving average are defined as follows. In these definitions, w determines the window length (which is equal to $2w + 1$). For the simple moving average filter [DAL12], defined below, each value is substituted by average of $2w$ adjacent values.

Definition 7.1. Simple moving average (MA)- *Simple moving average returns times series* $x^{MA} = < {x_1}^{MA}, \dots, {x_m}^{MA} >$ *in which*

$$ {x_i}^{MA} = \frac{1}{2w+1} \sum_{k=i-w}^{i+w} x_k \, , 1 \leq i \leq m $$

The uncertain moving average filter [DAL12], on the other hand, is defined as a weighted average of adjacent values. The weight at each timestamp is defined using standard deviation σ_i at that timestamp. The lower the standard deviation, the higher the weight of the value of that timestamp.

Definition 7.2. Uncertain moving average (UMA)- *Uncertain moving average returns times series* $x^{UMA} = < {x_1}^{UMA}, \dots, {x_m}^{UMA} >$ *for which*

$$ {x_i}^{UMA} = \frac{1}{2w+1} \sum_{k=i-w}^{i+w} \frac{x_k}{\sigma_k}, 1 \leq i \leq m $$

Finally, as defined below, for the uncertain exponential moving average [DAL12], weights decrease exponentially.

Definition 7.3. Uncertain exponential moving average (UEMA)- *Uncertain exponential moving average returns times series* $x^{UEMA} =< {x_1}^{UEMA}, \dots, {x_m}^{UEMA} >$ *where*

$$ {x_i}^{UEMA} = \frac{\sum_{k=i-w}^{i+w} x_k (e^{-\lambda|k-i|}/\sigma_k)}{\sum_{k=i-w}^{i+w} e^{-\lambda|k-i|}} $$

where λ *controls the exponential decreasing weight factor.*

We next review normalization as another important preprocessing technique.

7.2.2. **Normalization**

As discussed in Chapter 3, normalization transforms all time series in a dataset to the same baseline and scale. This helps similarity measures better capture the similarity. The normal form of a given standard time series $x =< x_1, \dots, x_n >$ is defined as $\hat{x} =< \hat{x}_1, \dots, \hat{x}_n >$ [SHA04], in which for each timestamp i $(1 \leq i \leq n)$, $\hat{x}_i$ is defined as in (6). We introduced normalization for uncertain time series data, referred to as uncertain normalization, in Chapter 3. Given an uncertain time series $X =< X_1, \dots, X_n >$, its normal form is defined as $\hat{X} =< \hat{X}_1, \dots, \hat{X}_n >$, where for each timestamp i $(1 \leq i \leq n)$, $\hat{X}_i$ is defined as in Definition 3.2.

Normalization makes similarity measures invariant to scaling and shifting and hence helps better capture the similarity [ORA12, SHA04]. One of the situations in which we need normalization for uncertain time series is when applying weighted filtering techniques such as the uncertain moving average (Definition 7.2) or the uncertain exponential moving average filter (Definition 7.3) (as shown in Figure 23). For example, suppose we want to find the Euclidean distance between uncertain time series $x =< x_1, \dots, x_n >$ and $y =< y_1, \dots, y_n >$, in which

standard deviations in all timestamps are equal to σ. By applying the uncertain moving average to x and y, with w equal to 0 for simplicity, we obtain uncertain time series $x^{UMA} = \frac{1}{\sigma}x$ and $y^{UMA} = \frac{1}{\sigma}y$, respectively. Calculating the Euclidean distance between these two filtered uncertain time series, we get:

$$Eucl(x^{UMA}, y^{UMA}) = \frac{1}{\sigma^2} Eucl(x, y)$$

This means that the lower the standard deviation, the higher the Euclidean distance. This affects the performance of search process. To address this problem, we apply normalization. For example, by applying normalization to x^{UMA}, for each timestamp, we obtain:

$$\widehat{x^{UMA}}_i = \frac{x_i{}^{UMA} - \frac{1}{n}\sum_{j=1}^{n} x_j{}^{UMA}}{\sqrt{\frac{1}{(n-1)}\sum_{k=1}^{n}(x_k{}^{UMA} - \frac{1}{n}\sum_{j=1}^{n} x_j{}^{UMA})^2}} = \frac{\frac{x_i}{\sigma} - \frac{1}{n}\sum_{j=1}^{n}\frac{x_j}{\sigma}}{\sqrt{\frac{1}{(n-1)}\sum_{k=1}^{n}(\frac{x_k}{\sigma} - \frac{1}{n}\sum_{j=1}^{n}\frac{x_j}{\sigma})^2}} = \hat{x}_i$$

This shows that when we normalize x^{UMA}, the result does not depend on the standard deviation anymore and thus the Euclidean distance between x and y would be invariant to scaling and shifting. In this example, we used $w = 0$, however, in practice w would be larger than zero. In this example, if we consider $w > 0$, $\widehat{x^{UMA}}_i$ would become equal to $\widehat{x^{MA}}_i$ which is still independent of standard deviation. Thus, after applying some filtering techniques, data has to be normalized. However, normalization is independent of filtering and can be used to improve similarity search quality, even when the data is not filtered. In the next section, we study the effect of preprocessing techniques on uncertain similarity measures.

7.3. Applying Preprocessing Techniques to Uncertain Similarity Measures

In this section, we describe how to apply preprocessing techniques to uncertain similarity measures and why preprocessing techniques can be effective. These topics are discussed separately for the probabilistic and deterministic similarity measures in the following sections.

7.3.1. Probabilistic Similarity Measures

In this section, we discuss how the preprocessing techniques can help probabilistic similarity measures to better capture similarity. Specifically, we study uncertain correlation (defined in Section 5.1.1) and PROUD [YEH09]. In both measures, an observed value at each timestamp is used as an estimate for the expected value of the random variable at that timestamp. Given an uncertain time series $X =< X_1, \dots, X_n >$, each X_i can be written as $X_i = x_i + E_{x_i}$, where x_i is the "exact" value which is unknown, and E_{x_i} is a random variable denoting the error. Since the exact value x_i is unknown, the expected value of X_i, i.e., $E(X_i) = x_i + E(E_{x_i})$, would be unknown as well even if the expected value of the error were known. Thus, an observed value is used as an estimate for $E(X_i)$. The aim of applying preprocessing techniques to a given uncertain time series $X =< X_1, \dots, X_n >$ is to smooth the expected value time series $E(X) =< E(X_1), \dots, E(X_n) >$ and make each $E(X_i)(1 \leq i \leq n)$ closer to the unknown exact value.

7.3.1.1. Uncertain Correlation

We illustrate how preprocessing techniques can improve the performance of uncertain correlation using the following example. Suppose that for two standard time series *x* and *y* from the Gun-Point dataset [KEO], we have $Corr(x, y) \geq 0.6$. Now, suppose *x* and *y* are observed as uncertain time series *X* and *Y*. Since the Pearson correlation between their underlying standard time series is more than 0.6, we expect that *X* and *Y* satisfy the PTC query, $P(Corr(X, Y) \geq 0.6) \geq 0.8$. The gray line in Figure 24 shows the complementary cumulative distribution function

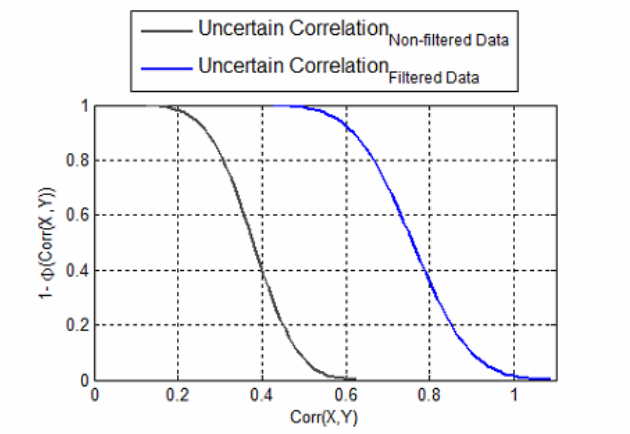


Figure 24. The impact of filtering on uncertain correlation.

of $Corr(X,Y)$, which indicates that $P(Corr(X,Y) \geq 0.6)$ is around zero. If we compute this distribution function after applying a moving average filter to *X* and *Y*, we obtain the blue line in Figure 24. As can be seen, the probability $P(Corr(X,Y) \geq 0.6)$ is around 0.9, which indicates with high probability that the correlation between *X* and *Y* is more than 0.6, as expected. In this example, without filtering techniques, we could not determine the existing correlation between uncertain time series *X* and *Y*. However, using filtering, we were able to successfully quantify their actual correlation.

7.3.1.2. PROUD

In this section, we discuss the potential effects and challenges of the preprocessing techniques on the performance of PROUD [YEH09] using the following example. Suppose that for two standard time series *x* and *y* from the Gun-Point dataset [KEO], we have $Eucl(x,y) \leq 100$. Now, we perturb *x* and *y* to get uncertain time series *X* and *Y*. Since the squared Euclidean distance between their underlying standard time series is less than 100, we expect that *X* and *Y* satisfy the PTE query $P(Eucl(X,Y) \leq 100) \geq 0.8$. The gray line in Figure 25 illustrates the cumulative distribution function of $Eucl(X,Y)$ which shows $P(Eucl(X,Y) \leq 100) = 0$. Now we apply a moving average filter and normalization technique to *X* and *Y* and again compute the cumulative

distribution function of $Eucl(X,Y)$, which is shown by the yellow line. Surprisingly, we again observe $P(Eucl(X,Y) \leq 100) = 0$, that is, the preprocessing techniques could not help PROUD to quantify the actual Euclidean distance between *X* and *Y*.

We found that the underlying reason is the way PROUD calculates the expected value of the Euclidean distance between uncertain time series, i.e., $E\big(Eucl(X,Y)\big)$ (4). To highlight the existing challenge in PROUD, we rewrite the equation (4) as follows:

$$E\big(Eucl(X,Y)\big) = Eucl\big(E(X),E(Y)\big) + \sum_{i=1}^{n}(Var(X_i) + Var(Y_i))$$

In which $E(X)$ is the expected value time series $E(X) =< E(X_1), \dots, E(X_n) >$. In other words, the expected value of the squared Euclidean distance between two uncertain time series would be the sum of the squared Euclidean distance of expected value time series and all the variances of all the timestamps. Thus, the higher the variance, the farther the expected value of PROUD from the Euclidean distance between the expected value time series. For instance, in the example in Figure 25, the variance at each timestamp is equal to 1 for both *X* and *Y* and the length of *X* and *Y* is 150. Thus, we have:

$$E\big(Eucl(X,Y)\big) = Eucl\big(E(X),E(Y)\big) + 300$$

To solve this problem, we propose an enhanced version of PROUD in the following section.

7.3.1.3. PROUDS

In this section, we introduce PROUD**S,** an enhanced version of PROUD. PROUD**S** defines the squared Euclidean distance between uncertain time series $X =< X_1, \dots, X_n >$ and $Y =< Y_1, \dots, Y_n >$, as follows:

$$Eucl(X,Y) = \sum_{i=1}^{n} (E(X_i)^2 + E(Y_i)^2 + 2X_iY_i) \quad (26)$$

Using this, if the random variables are i.i.d., according to the central limit theorem [ROS09], as n increases, $Eucl(X,Y)$ approaches normal distribution. Moreover, $E\big(Eucl(X,Y)\big)$ and $Var\big(Eucl(X,Y)\big)$ are calculated as follows:

$$E\big(Eucl(X,Y)\big) = \sum_{i=1}^{n} (E(X_i) - E(Y_i))^2, \text{and}$$

$$Var\big(Eucl(X,Y)\big) = 4\sum_{i=1}^{n} \big(E(X_i)^2 Var(Y_i) + E(Y_i)^2 Var(X_i) + Var(X_i)Var(Y_i)\big)$$

As can be seen, the expected value of the Euclidean distance is the Euclidean distance between the expected value time series, and the variances at different timestamps do not have any effect on it. Let us consider again the example in Figure 25. We use PROUD**S** on filtered and normalized uncertain time series X and Y to find the cumulative distribution function of $Eucl(X,Y)$, which is shown by the blue line in the figure. We can see that X and Y satisfy the

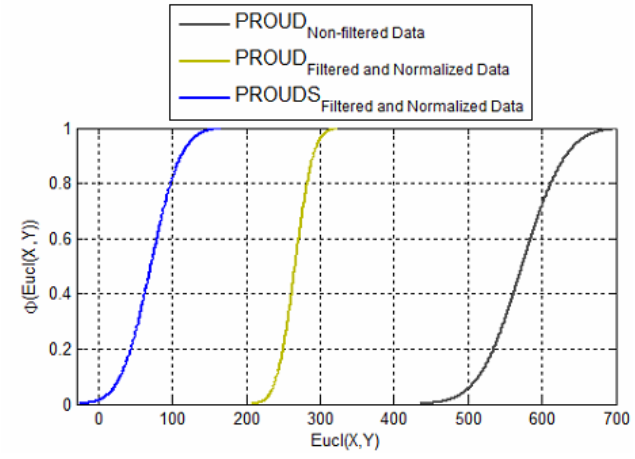


Figure 25. The impact of filtering on PROUD and PROUDS .

PTE query $P(Eucl(X,Y) \leq 100) \geq 0.8$, as expected. Another advantage of PROUD**S** is captured in the following lemma.

Lemma 8. *Given uncertain time series X and Y with normal forms $\hat{X}$ and $\hat{Y}$, the following holds.*

$$Eucl(\hat{X},\hat{Y}) = 2(n-1)(1-Corr(X,Y))$$

where $Eucl(\hat{X},\hat{Y})$ is defined as in (26) *and $Corr(X,Y)$ is defined as in Definition 3.4.*

Proof. To show this, first, we need to prove that $\sum_{i=1}^{n} E(\hat{X}_i)^2 = n-1$. From (8), we have:

$$\sum_{i=1}^{n} E(\hat{X}_i)^2 = \sum_{i=1}^{n} \left(\frac{E(X_i)-\bar{X}}{S_X}\right)^2 = \frac{\sum_{i=1}^{n}(E(X_i)-\bar{X})^2}{\sum_{i=1}^{n}(E(X_i)-\bar{X})^2/(n-1)} = n-1$$

Using (26) and Definition 3.4, we obtain the relationship between PROUD**S** and uncertain correlation as follows:

$$Eucl(\hat{X},\hat{Y}) = \sum_{i=1}^{n}\left(E(\hat{X}_i)^2 + E(\hat{Y}_i)^2 + 2\hat{X}_i\hat{Y}_i\right) = 2(n-1) + 2\sum_{i=1}^{n}\hat{X}_i\hat{Y}_i$$

$$= 2(n-1)(1-Corr(X,Y))\blacksquare$$

This shows that PROUD**S** extends the Euclidean distance for uncertain time series in a way that there is a linear relationship between the Euclidean distance for uncertain time series and uncertain correlation, similar to the standard case [SHA04]. Moreover, if the normal form of uncertain time series in the PTC queries is defined as in Section 3.2, and uncertain Euclidean distance in the PTE queries defined as in PROUDS, similar to Lemma 5, PTC and PTE queries can be converted to each other. Next, we study how preprocessing affects the performance of deterministic similarity measures.

7.3.2. **Deterministic Similarity Measures**

As discussed in Section 7.2.2, weighted filters change the scale of data, which decreases the performance of similarity measures. However, as shown in Figure 23 and Section 7.2.2, when we normalize the filtered data, the weights are distributed among the timestamps so that the timestamps with lower error standard deviation (uncertainty) become more important than the ones with higher error standard deviation in the similarity search, but this weight distribution will not affect the scale and baseline of the data anymore. Thus, similarity measures such as the Euclidean distance and Pearson correlation coefficient will be invariant to scaling and shifting. On the other hand, weighted similarity measures like DUST [SAR10] define weights for the similarity between uncertain time series at each timestamp. For example, suppose that using DUST, we want to find the similarity between uncertain time series X and Y having the same standard deviation σ_X and σ_Y at all timestamps, respectively. Suppose that the observed value at timestamp i of X (and Y) is represented as x_i (and y_i), and the error at each timestamp has normal distribution. DUST distance between X and Y is defined as follows (see the Appendix for more details):

$$DUST(X,Y)^2 = \frac{1}{2({\sigma_X}^2 + {\sigma_Y}^2)} \sum_{i=1}^{n} (x_i - y_i)^2$$

This can be rewritten as:

$$DUST(X,Y)^2 = \frac{1}{2({\sigma_X}^2 + {\sigma_Y}^2)} Eucl(x,y)$$

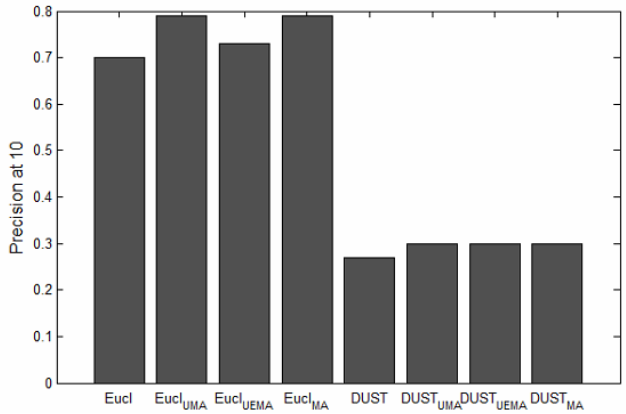


Figure 26. 10 nearest neighbor search using the Euclidean distance and DUST for normal error distribution.

The weight $1/(2({\sigma_X}^2 + {\sigma_Y}^2))$ changes the similarity between uncertain time series based on the standard deviations σ_X and σ_Y. For example, we perturbed half of the Gun-Point testing dataset [KEO] with the standard deviation 0.1 and the other half with the standard deviation 1. We then preprocessed the data and searched for the 10 nearest neighbors of a given query chosen from training set of the dataset, perturbed with the standard deviation 0.1. Figure 26 shows the precision at 10 [MAN08] for this similarity search. This precision is defined as the percentage of correct answers in 10 returned answers. As can be seen in Figure 26, DUST could not find more than 3 correct answers even when using preprocessing techniques. This shows that due to the weight given to the similarity between uncertain time series at each timestamp, preprocessing techniques could not help DUST in the similarity search. However, preprocessing is effective for DUST when the standard deviations of uncertain time series are equal at each timestamp. For example, if in the example in Figure 26, all the time series in the dataset had the same standard deviation, DUST with normal error distribution would have the same Precision at 10 as the Euclidean distance (see the Appendix for more details).

7.4. Experiments

In this section, we report the setup of our study and our findings. The setup of our experiments follows the one discussed in Chapter 4. We used 16 datasets from the UCR benchmark on real-life applications [KEO], namely, 50words, Adiac, Beef, CBF, Coffee, ECG200, FISH, FaceFour, Gun-Point, Lighting2, Lighting7, OSULeaf, OliveOil, SwedishLeaf, Synthetic-control, and Trace. In our experiments, we studied the effects of different parameters including data parameters, query parameters and filtering parameters. Similar to [ORA14, SAR10], for data parameters, we varied SDR values from 0.01 to 4, and considered normal, uniform, and exponential error distributions.

For query parameters, we studied the effect of probability threshold and similarity threshold. For the probability threshold, we considered the values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. In PTC queries, for similarity threshold, i.e., correlation threshold c, we considered 0.1 to 0.9. In PTE queries, the similarity threshold, i.e., Euclidean threshold d, was obtained from the relationship between the Euclidean distance and the Pearson correlation [SHA04], $d = 2(n-1)(1-c)$, where n is the length of the time series. For the filtering parameters, in [DAL12], the effect of different window sizes w and decreasing factors λ was studied. As in [DAL12], we chose $w = 2$ and $\lambda = 1$ as the filtering parameters in our experiments.

In the following, we report and analyze the result of our study on the performance of both deterministic and probabilistic similarity measures with and without preprocessing. Preprocessing always includes filtering and normalization.

7.4.1. Deterministic Similarity Measures

We evaluated and compared the accuracy of the deterministic similarity measures that include DUST and the Euclidean distance on non-preprocessed and preprocessed data. We began our evaluation using the 1NN-classification with K-fold cross-validation, identified as the most suitable approach for evaluating the efficiency of similarity measures [DIN08]. Similar to

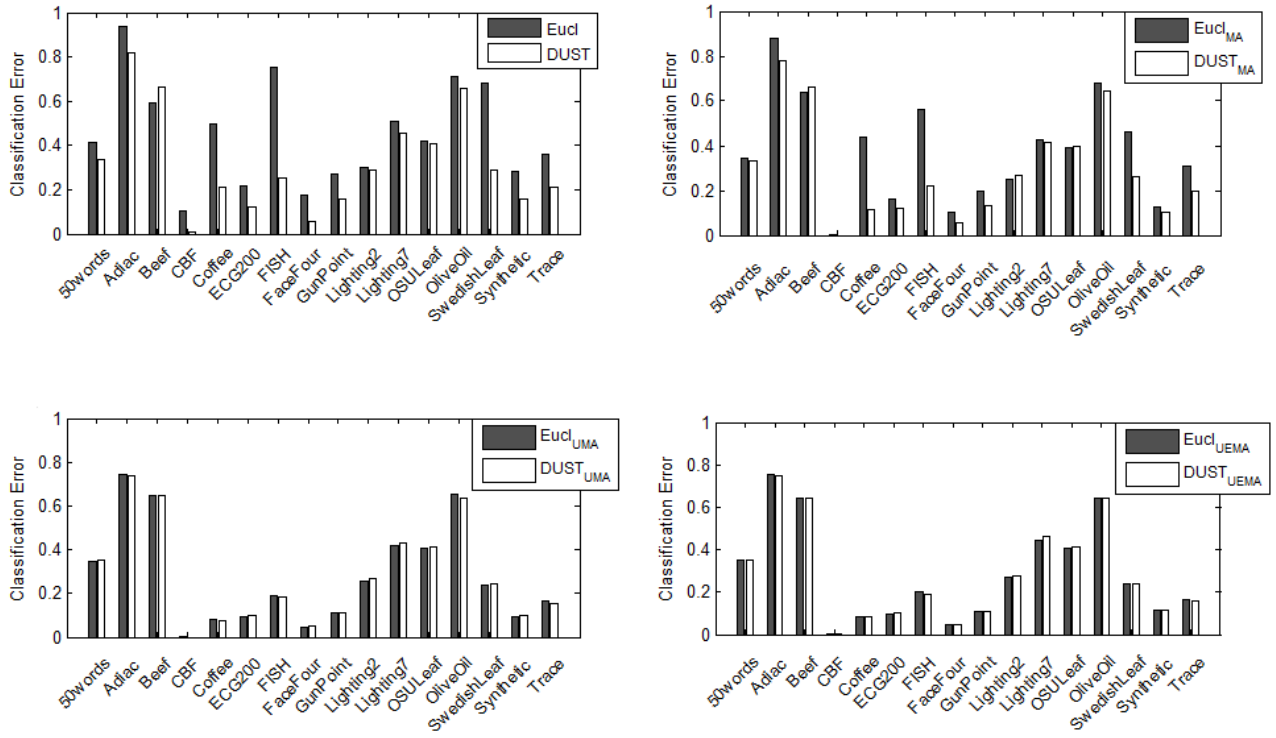


Figure 27. Effect of filtering on classification error of the Euclidean distance and DUST for *r* equal to 2 and normal distribution.

[ORA14] and [SAR10], for the first 80% of the timestamps, we used an SDR of $0.1r$, for the next 10%, we used an SDR of $0.5r$, and for the remaining 10% we used r.

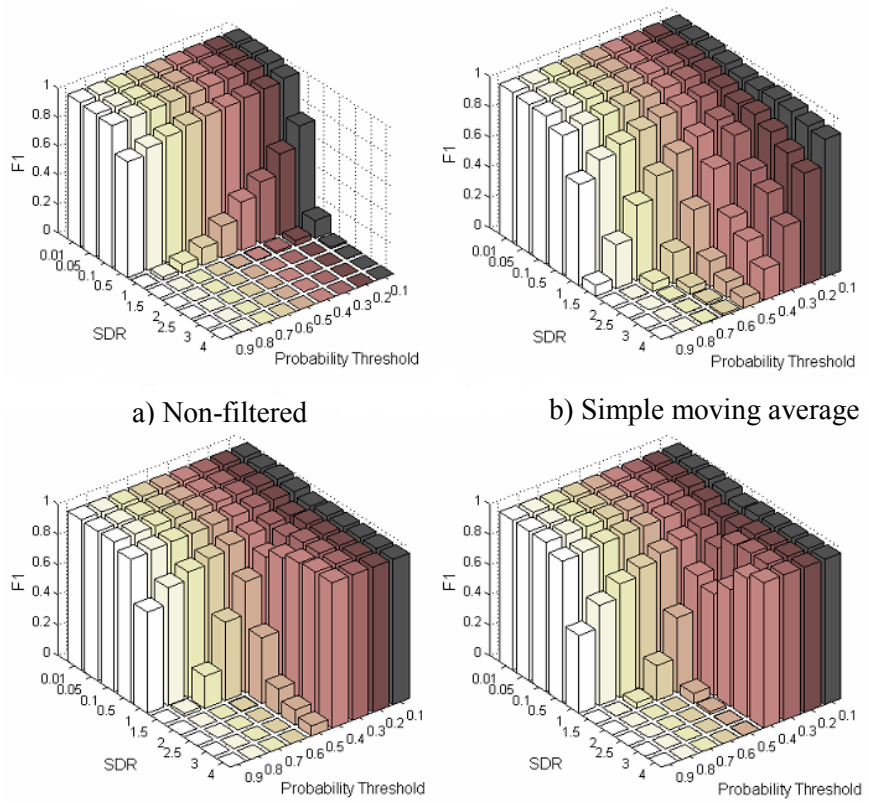
Figure 27 shows the classification error for different UCR datasets for $r = 2$ and error with normal distribution. When we applied the Euclidean distance and DUST on non-preprocessed data, in most cases, DUST had a lower classification error than the Euclidean distance. We made the same observation when the simple moving average was applied to the data. However, when uncertain time series are filtered using the uncertain moving average and uncertain exponential moving average, the difference between DUST and the Euclidean distance classification error was very small. This shows that the weighted filters helps the Euclidean distance achieve similar performance to a weighted similarity measure such as DUST.

We also studied the other *r* values and errors with exponential distribution. In all these cases, we made a similar observation as reported in this section. However, for uniform distribution, in some cases, DUST cannot be evaluated since it approaches the logarithm of zero (see the Appendix for details). This problem has also been reported in [DAL12]. As in [DAL12], we added two tails to the uniform distribution to solve the problem, however, in most cases the DUST classification error was close to 1. The complete set of experiments and results are made available to reviewers [CORY]. In these experiments, we observed that in general DUST is either better (with no preprocessing or simple moving average filters) or similar to the Euclidean distance (with uncertain filters).

7.4.2. **Probabilistic Similarity Measures**

In this section, we evaluate the performance of PTC and PTE queries with and without preprocessing. We will also compare these probabilistic queries with deterministic queries $Corr(x,q) \geq c$ and $Eucl(x,q) \leq d$. The threshold values in the deterministic queries are the same as the ones in the corresponding probabilistic queries. Moreover, the standard time series x and q would be the sequence of expected values of random variables in the corresponding uncertain time series. In all the probabilistic similarity measures, it is assumed that all random variables in the given time series are i.i.d., that is, all random variables have equal standard deviation. Note that in the experiments reported in [SAR10], it is observed that for the i.i.d. case, DUST shows similar performance to the Euclidean distance, so in this section, we only report the results for the Euclidean distance.

Similar to [DAL12] and [ORA14], we measured the performance using F_1 measure [MAN08], defined as in Section 4.3. The ground truth (i.e., correct results) is based on the result of the range query $Corr(x,q) \geq c$ for correlation, and $Eucl(x,q) \leq d$ for the Euclidean distance on the dataset without uncertainty [ORA12, YEH09]. In the following section, we present our results, beginning with uncertain correlation.



a) Non-filtered b) Simple moving average

c) Uncertain moving d) Uncertain exponential moving

Figure 28. F_1 score for probabilistic correlation queries for correlation threshold 0.5, normal distribution, and Gun-Point dataset.

7.4.2.1. Uncertain Correlation

Figure 28 shows the F_1 score for probabilistic correlation queries applied on filtered and non-filtered data. Figure 28 (a) shows the F_1 score of probabilistic queries without using any filtering method. In this case, probabilistic queries did not find any answer for the SDR values more than 2. Recall that SDR specifies uncertainty level in data; the higher the uncertainty level, the farther the observed values (expected values) from the exact values, and thus the lower the performance of probabilistic queries. Figure 28 (b) shows the F_1 score for probabilistic queries when the data is filtered using the simple moving average. Comparing Figure 28 (a) and (b), we clearly note the

positive effect of the moving average filter on the performance of probabilistic queries that can now find correlated uncertain time series for high SDRs.

Figure 28 (c) and (d) show the F_1 score of the probabilistic queries using the uncertain moving average and uncertain exponential moving average, respectively. In both cases, for low probability thresholds the F_1 score is higher than Figure 28 (b). However, for high probability thresholds, Figure 28 (b) has a higher F_1 score. For all the cases, the lower the probability threshold, the higher the F_1 score. This is due to the fact that as we decrease the probability threshold, the probabilistic query returns more candidate results, which makes it more likely to contain the correct results.

Table 1 shows the average of F_1 scores for probabilistic queries for all 16 datasets used in our experiments, with and without filtering techniques including simple moving average (MA), uncertain moving average (UMA), and uncertain exponential moving average (UEMA). As can be seen from the table, for all but the Synthetic-control dataset, filtering techniques increase the F_1 score. For each filtering technique and each dataset, the table also records the improvement in percentage (in the parenthesis) defined as the F_1 score of probabilistic queries using that filtering technique minus that of the probabilistic queries with no filtering over the F_1 score of probabilistic queries with no filtering. For all the datasets, except the Synthetic-control, the improvement observed was in the range of 18% to 48%. Our results also indicate that simple and uncertain moving average filters improve the performance of the probabilistic similarity measures more than the uncertain exponential moving average filter does.

Figure 29 shows the F_1 score of deterministic queries with and without using filtering. As expected, for the non-filtered case (Corr in the figure), the F_1 score is lower than the other cases, and as SDR increases the F_1 score decreases. Moreover, the F_1 scores of deterministic queries with uncertain moving average ($\mathrm{Corr}_{\mathrm{UMA}}$) and simple moving average ($\mathrm{Corr}_{\mathrm{MA}}$) are the same. The reason is that the standard deviations of all timestamps are equal, and after normalization the standard deviations would be canceled and have no effect on the final result, as discussed in Section 7.2.2. Furthermore, Figure 29 shows that these filters improve the performance of deterministic queries more than uncertain exponential moving average filter ($\mathrm{Corr}_{\mathrm{UEMA}}$) does.

Table 1. Average of F_1 scores and improvement percentage for different UCR datasets for normal distribution.

Dataset	Filter			
	None	MA	UMA	UEMA
50words	0.38	0.53 (+39%)	0.5 (+32%)	0.48 (+26%)
Adiac	0.51	0.72 (+41%)	0.74 (+45%)	0.71 (+39%)
Beef	0.49	0.69 (+41%)	0.69 (+41%)	0.62 (+27%)
CBF	0.38	0.48 (+26%)	0.45 (+18%)	0.45 (+18%)
Coffee	0.51	0.73 (+43%)	0.69 (+35%)	0.68 (+33%)
ECG200	0.46	0.62 (+35%)	0.62 (+35%)	0.61 (+33%)
FISH	0.5	0.71 (+42%)	0.74 (+48%)	0.69 (+38%)
FaceFour	0.39	0.53 (+36%)	0.53 (+36%)	0.5 (+28%)
Gun-Point	0.48	0.68 (+42%)	0.69 (+44%)	0.66 (+38%)
Lighting2	0.31	0.42 (+35%)	0.43 (+39%)	0.41 (+32%)
Lighting7	0.35	0.47 (+34%)	0.47 (34%)	0.45 (+29%)
OSULeaf	0.34	0.47 (+38%)	0.46 (+35%)	0.43 (+26%)
OliveOil	0.51	0.73 (+43%)	0.73 (+43%)	0.66 (+29%)
Swedish	0.46	0.62 (+35%)	0.63 (+37%)	0.60 (+30%)
Synthetic	0.33	0.33 (0%)	0.32 (-3%)	0.33 (0%)
Trace	0.47	0.66 (+40%)	0.66 (+40%)	0.62 (+32%)

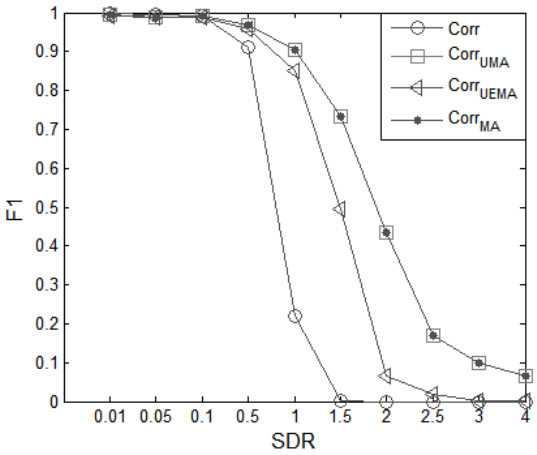


Figure 29. F_1 score of deterministic correlation queries.

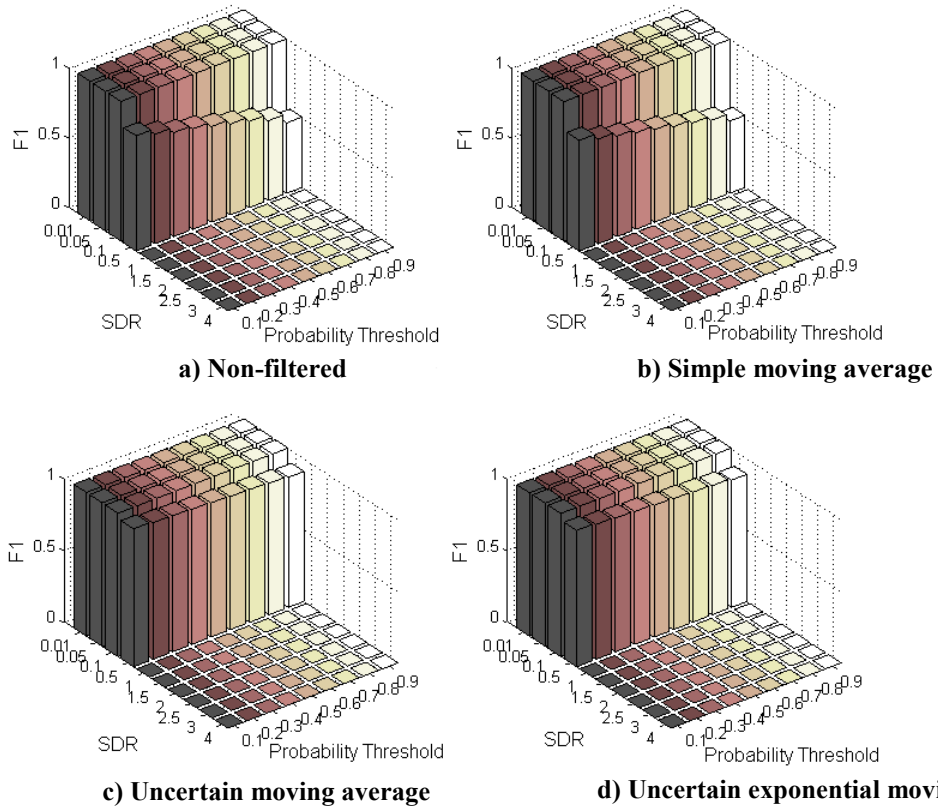
Now, let us compare probabilistic queries with deterministic ones. For the non-filtered case (Figure 28 (a) and Figure 29), we can see that both have the F_1 score equal to zero for high SDRs. However, probabilistic queries can have higher F_1 scores than deterministic ones, if a proper probability threshold is chosen. For example, for SDR=1, the F_1 score of deterministic queries is around 0.2. However, the F_1 score of probabilistic queries is near 0.7, when a small probability threshold is chosen. However, we note that using filtering methods, we can better differentiate between these two queries. We observe that probabilistic queries have higher F_1 score than deterministic ones, in particular for high SDRs. In this section, we reported our results for the Gun-point dataset, normal distribution and correlation threshold 0.5. We made the same observations using the other 15 datasets, error distribution functions, and correlation thresholds. The complete set of experiments and results are made available to reviewers [CORY].

7.4.2.2. PROUD and PROUDS

In this section, we evaluate the performance of PTE queries using PROUD and PROUDS with and without filtering. Moreover, we will compare the performance of deterministic and probabilistic queries. Figure 30 illustrates the F_1 score of PROUD for different settings. In all

these experiments, PROUD does not return any result for SDRs higher than 1. The reason for this is the way PROUD calculates the expected value of the distance, as discussed in Section 7.3.1.2. However, Figure 30 shows that uncertain and uncertain exponential moving average filters can increase the performance of PROUD.

The results of our experiments indicate that the performance of PROUD**S** is identical to that of uncertain correlation shown in Figure 28, explained in the following section. Comparing Figure 30 and Figure 28 shows that PROUD**S** performs better than PROUD for the Gun-point dataset. For all the datasets, we observed that PROUD**S** achieved on average 64% improvement compared to PROUD, and in all the cases, PROUD**S** outperforms PROUD.



a) Non-filtered **b) Simple moving average**

c) Uncertain moving average **d) Uncertain exponential moving average**

Figure 30. F_1 score for probabilistic queries using PROUD for normal distribution and the Gun-Point dataset.

Similarly, the performance of deterministic queries $Eucl(x,q) \leq d$ is exactly the same as that of $Corr(x,q) \geq c$ (Figure 29). By comparing Figure 30 and Figure 29, we can observe that deterministic queries have better performance than PROUD does. However, by comparing Figure 28 and Figure 29, it can be seen that our proposed enhancement, i.e., PROUD**S**, outperforms deterministic queries. In this section, we reported our results for the Gun-point dataset and normal distribution function. For the other 15 datasets and distribution functions, we observed similar results. The complete set of experiments and results are made available to reviewers [CORY].

Experimental Results Analysis

Here, we discuss why the performance of PROUD**S** is similar to that of uncertain correlation. Since in PROUD**S**, the squared Euclidean distance between two uncertain time series is a normal random variable, we would have:

$$P\left(Eucl(\hat{X},\hat{Y}) \leq 2(n-1)(1-c)\right) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{2(n-1)(1-c) - E\left(Eucl(\hat{X},\hat{Y})\right)}{\sqrt{2Var\left(Eucl(\hat{X},\hat{Y})\right)}}\right)\right)$$

Using Lemma 8, we obtain:

$$E\left(Eucl(\hat{X},\hat{Y})\right) = 2(n-1)\left[1 - E\left(Corr(X,Y)\right)\right], \text{and}$$

$$Var\left(Eucl(\hat{X},\hat{Y})\right) = 4(n-1)^2 Var\left(Corr(X,Y)\right)$$

Using the above information, we can compute the following probability:

$$P\left(Eucl(\hat{X},\hat{Y}) \leq 2(n-1)(1-c)\right)$$
$$= \frac{1}{2}\left(1 + \text{erf}\left(\frac{2(n-1)\left[(1-c) - \left[1 - E\left(Corr(X,Y)\right)\right]\right]}{\sqrt{8(n-1)^2 Var\left(Corr(X,Y)\right)}}\right)\right)$$

Since error function is an odd function[4], we have:

$$P\left(Eucl(\hat{X},\hat{Y}) \leq 2(n-1)(1-c)\right) = \frac{1}{2}\left(1 - \text{erf}\left(\frac{c - E\big(Corr(X,Y)\big)}{\sqrt{2Var\big(Corr(X,Y)\big)}}\right)\right)$$

$$= P(Corr(X,Y) \geq c)$$

Thus PTE and PTC queries find the same results.

7.5. **Summary**

Previous studies [DAL12, ORA14] compare the traditional similarity measures applied to filtered and normalized data to the uncertain similarity measures applied to non-preprocessed data. It is reported that the traditional measures outperform the uncertain similarity measures. In this chapter, we studied the effect of the preprocessing techniques on uncertain similarity measures and compared the measures in the same setting. We considered two settings: when data is not preprocessed and when it is filtered and normalized. In particular, we showed how preprocessing techniques improve the performance of uncertain similarity measures by using more information. We performed numerous experiments to evaluate the performance of measures with different parameters. We also improved the performance of the PROUD similarity measure and found a linear relationship between probabilistic similarity measures. We showed that probabilistic similarity measures outperform both traditional similarity measures and uncertain deterministic similarity measures with and without preprocessing. This shows the effectiveness and usefulness of probabilistic similarity measures in practice.

[4] Suppose that $f(x)$ is a real-valued function. $f(x)$ is an odd function if and only if $f(-x) = -f(x)$.

Chapter 8: Conclusions and Future Work

Due to the inherent nature of uncertain time series, a probabilistic approach has been considered key to process and analyze such data. In the probabilistic approach, probabilistic similarity measures are used to capture the similarity between uncertain time series. Unlike traditional similarity measures that consider only the expected value at each timestamp to quantify the similarity between uncertain time series, probabilistic similarity measures utilize all the available information, such as variance and probability distribution function of error. This information helps probabilistic measures to capture the similarity better than the traditional measures, in particular when there exists a high level of uncertainty.

Moreover, probabilistic similarity measures provide the users with more information about the reliability of the result. Depending on the application, this allows the users to define a confidence level, i.e., probability threshold in probabilistic queries, in order to control and/or influence the performance of similarity search tasks. In other words, the advantage of probabilistic over deterministic approach is providing a flexible trade-off between hit ratio and false alarm ratio. Besides, it provides probabilistic information on similarity, which is important in some applications [DAL12].

In this work, we considered two models for uncertain time series data, PDF-based and multiset-based models, and studied the problem of correlation analysis techniques over such data. Although these two models are different and cannot be converted to each other, the proposed techniques for multiset-based model generalize from the ones for PDF-based model. Moreover,

the proposed techniques for both models can be reduced to the corresponding techniques developed for standard time series. The major contributions of this dissertation are as follows:

- We formalized the notion of uncertain normalization which can be applied as a preprocessing step. Using this, we formulated correlation for uncertain time series data as a random variable [ORA12].

- For both PDF-based and multiset-based models, we developed probabilistic similarity search techniques that could find correlated or uncorrelated uncertain time series to the user input. The results of our numerous experiments using the UCR benchmark data illustrate that our probabilistic approach is more resilient to the uncertainty level than the deterministic one. The proposed solutions are also applicable in finding correlations between standard and standard/uncertain time series [ORA12, ORAY15].

- To speed up processing queries over the multiset-based model, we proposed two optimization techniques: probabilistic pruning and a sampling-based heuristic. Our experiments show that while our optimization techniques for multiset-based model reduce the computation time and spare utilization, they find a good approximation for uncertain correlation even when there are just a few observed values at each timestamp. This is important noting that, as discussed in [WU12], it is often impossible to have several observed values at each timestamp. The results of our experiments revealed significant improvement achieved by the proposed optimization techniques [ORAY15].

- We studied the effect of preprocessing techniques on our proposed similarity search methods as well as on the existing one. Our experimental results showed how preprocessing techniques can result in improved performance of uncertain similarity measures. Moreover, the experiments show that probabilistic similarity measures outperform traditional similarity measures with/without preprocessing techniques [ORA14,

ORAX15].

Our results shed more light on the nature and challenges of similarity search for uncertain time series. More work is required for such results find their way into the development of effective tools and software packages for high performance uncertain time series analysis and mining, similar to the tools for the standard case. The next section discusses the possible future work.

8.1. Future Work

We believe that the outcome of this dissertation provides the first step towards tools for similarity search and analysis of uncertain time series data. The following summarizes some of the problems that should be addressed as part of future work:

- Extending Boolean representation to optimize probabilistic similarity search queries when the Euclidean distance is used as the similarity measure.
- Prediction in uncertain time series.
- Developing probabilistic pruning techniques to support other probabilistic correlation queries introduced in Section 5.4. The proposed probabilistic pruning is only applicable for the PTC-1 queries, since the pruning is based on Boolean correlation that can only find positively correlated time series.
- Developing efficient correlation analysis techniques for streaming uncertain time series.
- Reinvestigating uncertain representation of dimensionality reduction techniques for standard time series. For dimensionality reduction of PDF-based uncertain time series data, uncertain representation of the Haar wavelet transform [ZHA10] has been proposed, but other dimensionality reduction techniques have not been studied yet.
- Indexing uncertain time series, which is a very challenging problem, because the uncertainty can decrease the usefulness of index structures. To the best of our knowledge,

there is no index structure designed especially for variable length queries (the length of the query reference can be different from the time series in the dataset) for uncertain time series data.

References

[AGG08] C. C. Aggarwal. On unifying privacy and uncertain data models. In Proceedings of IEEE International Conference on Data Engineering (ICDE), pp. 386-395, 2008.

[AGG09] C. C. Aggarwal. Managing and mining uncertain data. Springer-Verlag New York Inc., 2009.

[AGG13] C. C. Aggarwal. A survey of uncertain data clustering algorithms. In Data Clustering: Algorithms and Applications, pp. 457-482, 2013.

[AGG15] C. C. Aggarwal. Data Mining. Springer, pp. 457-491, 2015.

[ASF09] J. Aßfalg, H. P. Kriegel, P. Kröger, and M. Renz. Probabilistic similarity search for uncertain time series. In Proceedings of International Conference on Scientific and Statistical Database Management (SSDBM), pp. 435-443, 2009.

[BER09] T. Bernecker, H.-P. Kriegel, M. Renz, and A. Zuefle. Probabilistic ranking in uncertain vector spaces. In Proceedings of Workshop on Managing Data Quality in Collaborative Information Systems, 2009.

[BER13] T. Bernecker, R. Cheng, D. W. Cheung, H. P. Kriegel, S. D. Lee, M. Renz, F. Verhein, L. Wang, and A. Zuefle. Model-based probabilistic frequent itemset mining. Knowledge and Information Systems Journal (KAIS), 1(37):181-217, 2013.

[BOH06] C. Bohm, A. Pryakhin, and M. Schubert. The Gauss-tree: Efficient object identification of probabilistic feature vectors. In Proceedings of International Conference on Data Engineering (ICDE), 2006.

[CER11] M. Ceriotti, M. Corra, L. D'Orazio, R. Doriguzzi, D. Facchin, S. Guna, G. P. Jesi, R. L. Cigno, L. Mottola, A. L. Murphy, M. Pescalli, G. P. Picco, D. Pregnolato, and C. Torghele. Is there light at the ends of the tunnel? Wireless sensor networks for adaptive lighting in road tunnels. In Proceedings of International Conference on Information Processing in Sensor Networks (IPSN), pp. 187-198, 2011.

[CHE03] R. Cheng, D. V. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In Proceedings of ACM SIGMOD International Conference on Management of data. pp. 551-562, 2003.

[CHE04] R. Cheng, D. V. Kalashnikov, and S. Prabhakar. Querying imprecise data in moving object environments. IEEE Transactions on Knowledge and Data Engineering (TKDE), 9(16): 1112-1127, 2004.

[CHE06] R. Cheng, S. Singh, S. Prabhakar, R. Shah, J. S. Vitter, and Y. Xia. Efficient join processing over uncertain data. In Proceedings of ACM Int. Conference on Information and Knowledge Management (CIKM), pp. 738-747, 2006.

[CORX] Complete experimental results: http://tinyurl.com/qfvbauf.

[CORY] Complete experimental results for Chapter 7: http://goo.gl/Ar3kR7.

[DAL11] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas. Similarity matching for uncertain time series: analytical and experimental comparison. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data. pp. 8-15, 2011.

[DAL12] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas. Uncertain time-series similarity: return to the basics. In Proceedings of the VLDB Endowment, 5(11): 1662-1673, 2012.

[DALX14] M. Dallachiesa, T. Palpanas, and I. F. Ilyas. Top-k nearest neighbor search in uncertain data series. In Proceedings of the VLDB Endowment Journal 8(1):13-24, 2014.

[DALY14] M. Dallachiesa, G. Jacques-Silva, B. Gedik, K. L. Wu, and T. Palpanas. Sliding windows over uncertain data streams. Knowledge and Information Systems Journal (KAIS), pp. 1-32, 2014.

[DIN08] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. In Proceedings of the VLDB Endowment, 1(2): 1542-1552, 2008.

[DVO56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Annals of Mathematical Statistics, 27(3):642-669, 1956.

[EMR12] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Zufle. Querying uncertain spatio-temporal data. In Proceedings of International Conference on Data Engineering (ICDE), pp. 354-365, 2012.

[FRA15] D. Fradkin, and F. Mörchen. Mining sequential patterns for classification. Knowledge and Information Systems Journal (KAIS), pp. 1-19, 2015.

[GON15] X. Gong, Y. Xiong, W. Huang, L. Chen, Q. Lu, Y. Hu. Fast similarity search of multi-dimensional time series via segment rotation. In Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA). pp. 108-124, 2015.

[HAL00] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of International Conference on Machine Learning (ICML), pp. 359-366, 2000.

[HON13] Y. Hong. On computing the distribution function for the Poisson binomial distribution. Computational Statistics and Data Analysis, 59: 41-51, 2013.

[JIA13] B. Jiang, J. Pei, Y. Tao, and X. Lin. Clustering uncertain data based on probability distribution similarity. IEEE Transactions on Knowledge and Data Engineering (TKDE), 4(25):751-763, 2013.

[KAD08] S. Kadiyala and N. Shiri. A compact multi-resolution index for variable length queries in time series databases. Knowledge and Information Systems Journal (KAIS), 15: 131-147, 2008.

[KEO] E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei, and C. A. Ratanamahatana. The UCR Time Series Classification/Clustering Homepage: www.cs.ucr.edu/~eamonn/time_series_data/.

[KRI07] H.-P. Kriegel, P. Kunath, and M. Renz. Probabilistic nearest-neighbor query on uncertain objects. In Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA), pp. 337-348, 2007.

[LEV71] M. Levandowsky, and D. Winter, Distance between sets. Journal of Nature, 234(5): 34-35, 1971.

[LIA08] X. Lian, L. Chen, and J. W. Yu. Pattern matching over cloaked time series. In Proceedings of IEEE 24th International Conference on Data Engineering (ICDE), pp. 1462-1464, 2008.

[LIA09] X. Lian and L. Chen. Efficient join processing on uncertain data streams. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), pp. 857-866, 2009.

[LJO07] V. Ljosa, and A. K. Singh. APLA: Indexing arbitrary probability distributions. In Proceedings of International Conference on Data Engineering (ICDE), pp. 946-955, 2007.

[MA11] C. Ma, H. Lu, L. Shou, G. Chen, and S. Chen. Top-k similarity search on uncertain trajectories. In Proceedings of International Conference on Scientific and Statistical Database Management (SSDBM), pp. 589-591, 2011.

[MAN08] C. D. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. Cambridge University Press, 2008.

[MAS90] P. Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. The Annals of Probability, 18 (3):1269-1283, 1990.

[NGU08] P. Nguyen and N. Shiri. Fast correlation analysis on time series datasets. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), pp. 787-796, 2008.

[ORA12] M. Orang, and N. Shiri. A probabilistic approach to correlation queries in uncertain time series data. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), pp. 2229-2233, 2012.

[ORA14] M. Orang and N. Shiri. An experimental evaluation of similarity measures for uncertain time series. In Proceedings of International Database Engineering and Applications Symposium (IDEAS), pp. 261-264, 2014.

[ORAX15] M. Orang and N. Shiri. Improving performance of similarity measures for uncertain time series using preprocessing techniques. In Proceedings of International Conference on Scientific and Statistical Database Management (SSDBM), 2015.

[ORAY15] M. Orang and N. Shiri. Correlation analysis techniques for uncertain time series. Journal Article, Under Review.

[QIA09] A. Qian, X. Ding, and Y. Lu. Probabilistic similarity search in uncertain time series database. International Conference on Information Engineering and Computer Science (ICIECS), pp. 1-4, 2009.

[RAJ15] D. Rajalakshmi, and K. Dinakaran. A survey on effective pattern matching in uncertain time series stream data. Asian Journal of Applied Science, 2015.

[RAZ13] K. Raza, and R. Jaiswal. Reconstruction and analysis of cancer-specific gene regulatory networks from gene expression profiles. International Journal on Bioinformatics and Biosciences, 2(3): 27-36, 2013.

[ROS09] S. M. Ross. Introductory statistics. Academic Press, 2009.

[SAK15] Y. Sakurai, Y. Matsubara, and C. Faloutsos. Mining and forecasting of big time series data. In Proceedings of ACM SIGMOD, 2015.

[SAR10] S. R. Sarangi, and K. Murth. DUST: A generalized notion of similarity between uncertain time series. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 383-392, 2010.

[SHA04] D. Shasha, and Y. Zhu. High performance discovery in time series: techniques and case studies. Springer, 2004.

[SHO09] G. R. Shorack and J. A. Wellner. Empirical processes with applications to statistics. Society for Industrial and Applied Mathematics, 2009.

[TAO05] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao, and S. Prabhakar. Indexing multidimensional uncertain data with arbitrary probability density functions. In Proceedings of International Conference on Very Large Data Bases (VLDB), pp. 922-933, 2005.

[TAR14] J. Tarango, E. Keogh, and P. Brisk. Accelerating the dynamic time warping distance measure using logarithmetic arithmetic. In Proceedings of Conference on Signals, Systems and Computers, pp. 404-408, 2014.

[TRA10] G. Trajcevski, A. N. Choudhary, O. Wolfson, L. Ye, and G. Li. Uncertain range queries for necklaces. In Proceedings of International Conference on Mobile Data Management (MDM), pp. 199-208, 2010.

[WAN05] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. X. Mayer, and H. W. Mewes. Gene selection from microarray data for cancer classification: A machine learning approach. Computational Biology and Chemistry Journal, 1(29): 37-46. 2005.

[WAN13] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery Journal, 2(26): 275-309, 2013.

[WU12] W. C. H. Wu, M. Y. Yeh, and J. Pei. Random error reduction in similarity search on time series: A statistical approach. In Proceedings of IEEE International Conference on Data Engineering (ICDE), pp. 858-869, 2012.

[XU09] J. Xu, D. Yue, Y. Gu, C. Li, and G. Yu. Pk-period: A probabilistic periodic analysis approach over uncertain time series. In Proceedings of International Conference on Emerging Databases (EDB), 2009.

[YEH09] 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 Proceedings of International Conference on Extending Database Technology: Advances in Database Technology (EDBT), pp. 684-695, 2009.

[ZHA07] T. Zhang, D. Yue, G. Yu, and Y. Gu. Correlation analysis based on hierarchical Boolean representation over time series data streams. In Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2:740-744, 2007.

[ZHA10] Y. Zhao, C.C. Aggarwal, and P.S. Yu. On wavelet decomposition of uncertain time series data sets. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), pp. 129-138, 2010.

[ZHA11] L. Zhang, J. Li, and Z. Wang. Uneven two-step sampling and distance calculation for uncertain trajectory. Information and Computational Science Journal, 9(8):1505-1513, 2011.

[ZUO11] Y. Zuo, G. Liu, X. Yue, W. Wang, and H. Wu. Similarity matching over uncertain time series. In Proceedings of International Conference on Computational Intelligence and Security (CIS), pp. 1357-1361, 2011.

Appendix

To have a better understanding of DUST, this section presents our calculation of DUST distances for different error distributions. Given uncertain time series $X =< X_1, \dots, X_n >$ and $Y =< Y_1, \dots, Y_n >$, we need the distribution of the underlying (certain) standard time series and distributions of the error functions at different timestamps in order to calculate $DUST(X, Y)$ [SAR10]. Here, we consider uniform distribution for the underlying standard time series and calculate the DUST function for different error functions including normal, exponential, and uniform. DUST is defined as follows:

$$DUST(X, Y) = \sqrt{\sum_{i=1}^{n} dust(X_i, Y_i)^2}$$

where for each i ($1 \leq i \leq n$):

$$dust(X_i, Y_i) = \sqrt{-log\big(\varphi(|X_i - Y_i|)\big) + log\big(\varphi(0)\big)}$$

The constant $log\big(\varphi(0)\big)$ is added to ensure $\forall X\ DUST(X, X) = 0$. Now, we want to calculate φ at each timestamp. Let x denote the observed value at each timestamp and suppose $x = e(x) + r(x)$ were $r(x)$ is the actual/exact value and $e(x)$ denotes the error. Then, $\varphi(|x - y|)$ would be calculated as follows:

$$\varphi(|x - y|) = p(dist(0, |x - y|) = 0) = p(r(x) = r(y)|x, y)$$

$$= \int_z p(r(x) = z|x)p(r(y) = z|y)\, dz$$

Using Bayes' Theorem [ROS09], $\varphi(|x-y|)$ would be equal to:

$$\varphi(|x-y|) = \int_z \frac{p(x|r(x)=z)p(r(x)=z)}{\int_v p(x|r(x)=v)p(r(x)=v)dv} \times \frac{p(y|r(y)=z)p(r(y)=z)}{\int_v p(y|r(y)=v)p(r(y)=v)dv} dz =$$

$$= \frac{\int_z p(x|r(x)=z)p(r(x)=z)p(y|r(y)=z)p(r(y)=z)dz}{\int_v p(x|r(x)=v)p(r(x)=v)dv \int_v p(y|r(y)=v)p(r(y)=v)dv}$$

Note that $p(x|r(x)=v)$ is the PDF of the error function determined at $x-v$, and we would have:

$$\varphi(|x-y|) = \frac{\int_z p(e(x)=x-z)p(r(x)=z)p(e(y)=y-z)p(r(y)=z)dz}{\int_v p(e(x)=x-v)p(r(x)=v)dv \int_v p(e(y)=y-v)p(r(y)=v)dv}$$

Considering uniform distribution for the underlying standard time series, we would have:

$$\varphi(|x-y|)$$

$$= \frac{\int_{-\infty}^{\infty} p(e(x)=x-z)p(r(x)=z)p(e(y)=y-z)p(r(y)=z)dz}{\int_{-\infty}^{\infty} p(e(x)=x-v)p(r(x)=v)dv. \int_{-\infty}^{\infty} p(e(y)=y-z)p(r(y)=v)dv}$$

$$= \frac{\frac{1}{n^2}\int_{-\infty}^{\infty} p(e(x)=x-z)p(e(y)=y-z)dz}{\frac{1}{n^2}\int_{-\infty}^{\infty} p(e(x)=x-v)dv. \int_{-\infty}^{\infty} p(e(y)=y-z)dv}$$

$$= \int_{-\infty}^{\infty} p(e(x)=x-z)p(e(y)=y-z)dz$$

Using $\varphi(|x-y|)$, $dust(x,y)$ can be computed. In the following, we will calculate $dust(x,y)$ for different error functions including normal, exponential and uniform. Using which, we can obtain $DUST(X,Y)$.

Normal Distribution

When the error distribution is normal, $\varphi(|x-y|)$ is calculated as follows:

$$\varphi(|x-y|) = \int_{-\infty}^{\infty} p(e(x) = x - z)p(e(y) = y - z)dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi{\sigma_x}^2}} e^{-\frac{(x-z)^2}{2{\sigma_x}^2}} \frac{1}{\sqrt{2\pi{\sigma_y}^2}} e^{-\frac{(y-z)^2}{2{\sigma_y}^2}} dz$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-\frac{(x-z)^2}{2{\sigma_x}^2}} e^{-\frac{(y-z)^2}{2{\sigma_y}^2}} dz = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-(\frac{(x-z)^2}{2{\sigma_x}^2} + \frac{(y-z)^2}{2{\sigma_y}^2})} dz$$

By replacing z with $z + x$, we would have:

$$\varphi(|x-y|) = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-(\frac{z^2}{2{\sigma_x}^2} + \frac{((y-x)-z)^2}{2{\sigma_y}^2})} dz = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-(\frac{z^2}{2{\sigma_x}^2} + \frac{(y-x)^2+z^2-2z(y-x)}{2{\sigma_y}^2})} dz$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(y-x)^2}{2{\sigma_y}^2}} \int_{-\infty}^{\infty} e^{-(\frac{z^2}{2{\sigma_x}^2} + \frac{z^2-2z(y-x)}{2{\sigma_y}^2})} dz$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(y-x)^2}{2{\sigma_y}^2}} \int_{-\infty}^{\infty} e^{-(z^2(\frac{1}{2{\sigma_x}^2} + \frac{1}{2{\sigma_y}^2}) - \frac{z(y-x)}{{\sigma_y}^2})} dz$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(y-x)^2}{2{\sigma_y}^2}} \int_{-\infty}^{\infty} e^{z\frac{(y-x)}{{\sigma_y}^2} - z^2(\frac{1}{2{\sigma_x}^2} + \frac{1}{2{\sigma_y}^2})} dz$$

For $a > 0$, we know that[5]:

$$\int_{-\infty}^{\infty} e^{-az^2} e^{-2bz} dz = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{a}}, (a > 0)$$

By substituting a by $\left(\frac{1}{2{\sigma_x}^2} + \frac{1}{2{\sigma_y}^2}\right)$ and b by $-\frac{(y-x)}{2{\sigma_y}^2}$, we would have:

[5] http://en.wikipedia.org/wiki/List_of_integrals_of_exponential_functions

$$\varphi(|x-y|) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(y-x)^2}{2{\sigma_y}^2})} \int_{-\infty}^{\infty} e^{z\frac{(y-x)}{{\sigma_y}^2} - z^2\left(\frac{1}{2{\sigma_x}^2}+\frac{1}{2{\sigma_y}^2}\right)} dz$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(y-x)^2}{2{\sigma_y}^2}} \sqrt{\pi\left(\frac{2{\sigma_x}^2{\sigma_y}^2}{{\sigma_y}^2+{\sigma_x}^2}\right)} e^{\frac{\frac{(y-x)^2}{4{\sigma_y}^4}}{\frac{{\sigma_y}^2+{\sigma_x}^2}{2{\sigma_x}^2{\sigma_y}^2}}}$$

$$= \frac{\sqrt{\pi\left(\frac{2{\sigma_x}^2{\sigma_y}^2}{{\sigma_y}^2+{\sigma_x}^2}\right)}}{2\pi\sigma_x\sigma_y} e^{-\frac{(y-x)^2}{2{\sigma_y}^2}} e^{\frac{(y-x)^2}{4{\sigma_y}^4}\cdot\frac{2{\sigma_x}^2{\sigma_y}^2}{{\sigma_y}^2+{\sigma_x}^2}}$$

$$= \frac{1}{\sqrt{2\pi({\sigma_y}^2+{\sigma_x}^2)}} e^{\frac{(y-x)^2}{2{\sigma_y}^2}\left(\frac{-{\sigma_y}^2}{{\sigma_y}^2+{\sigma_x}^2}\right)} = \frac{1}{\sqrt{2\pi({\sigma_y}^2+{\sigma_x}^2)}} e^{\frac{-(y-x)^2}{2({\sigma_y}^2+{\sigma_x}^2)}}$$

Now let's calculate *dust* function:

$$dust(x,y) = \sqrt{\log\big(\varphi(0)\big) - log\big(\varphi(|x-y|)\big)}$$

$$= \sqrt{log\left(\frac{1}{\sqrt{2\pi({\sigma_y}^2+{\sigma_x}^2)}}\right) - log\left(\frac{1}{\sqrt{2\pi({\sigma_y}^2+{\sigma_x}^2)}} e^{\frac{-(y-x)^2}{2({\sigma_y}^2+{\sigma_x}^2)}}\right)}$$

$$= \sqrt{log\left(\frac{1}{\sqrt{2\pi({\sigma_y}^2+{\sigma_x}^2)}}\right) - log\left(\frac{1}{\sqrt{2\pi({\sigma_y}^2+{\sigma_x}^2)}}\right) - log\left(e^{\frac{-(y-x)^2}{2({\sigma_y}^2+{\sigma_x}^2)}}\right)} = \sqrt{-log\left(e^{\frac{-(y-x)^2}{2({\sigma_y}^2+{\sigma_x}^2)}}\right)}$$

$$= \sqrt{\frac{|x-y|^2}{2({\sigma_y}^2+{\sigma_x}^2)}} = \frac{|x-y|}{\sqrt{2({\sigma_y}^2+{\sigma_x}^2)}}$$

Thus, DUST function for error with normal distribution is calculated as follows:

$$DUST_{Normal}(X,Y) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{2({\sigma_{Xi}}^2 + {\sigma_{Yi}}^2)}}$$

Exponential Distribution

If the error follows an exponential distribution, $\varphi(|x-y|)$ is calculated as follows:

$$\varphi(|x-y|) = \int_{-\infty}^{\infty} p(e(x) = x - z)p(e(y) = y - z)dz = \int_{-\infty}^{\infty} f(x-z;\lambda_x)f(y-z;\lambda_y)dz$$

where

$$f(x-z;\lambda_x)f(y-z;\lambda_y) = \begin{cases} \lambda_x e^{-\lambda_x(x-z)}\lambda_y e^{-\lambda_y(y-z)} & x \geq z, y \geq z \\ 0 & Otherwise \end{cases}$$

Using this, $\varphi(|x-y|)$ would be calculated as follows:

$$\int_{-\infty}^{\infty} f(x-z;\lambda_x)f\big(y-z;\lambda_y\big)dz = \int_{-\infty}^{min(x,y)} \lambda_x e^{-\lambda_x(x-z)}\lambda_y e^{-\lambda_y(y-z)}dz$$

$$= \lambda_x\lambda_y e^{-(\lambda_x x+\lambda_y y)} \int_{-\infty}^{min(x,y)} e^{(\lambda_x+\lambda_y)z}dz$$

$$= \frac{\lambda_x\lambda_y}{(\lambda_x+\lambda_y)} e^{-(\lambda_x x+\lambda_y y)} e^{(\lambda_x+\lambda_y)z} \Big|_{-\infty}^{min(x,y)}$$

$$= \frac{\lambda_x\lambda_y}{(\lambda_x+\lambda_y)} e^{-(\lambda_x x+\lambda_y y)} e^{(\lambda_x+\lambda_y)min(x,y)}$$

$$= \frac{\lambda_x\lambda_y}{(\lambda_x+\lambda_y)} e^{(\lambda_x+\lambda_y)min(x,y)-(\lambda_x x+\lambda_y y)} = \frac{1}{\sigma_x+\sigma_y} e^{-|x-y|/\sigma_{\max(x,y)}}$$

Note that in exponential distribution, we have $= \frac{1}{\lambda}$. To find $dust(x,y)$, first we should calculate $log\big(\varphi(|x-y|)\big)$:

$$log\big(\varphi(|x-y|)\big) = log\left(\frac{1}{\sigma_x+\sigma_y}e^{-|x-y|/\sigma_{\max(x,y)}}\right) = log\left(\frac{1}{\sigma_x+\sigma_y}\right) - \frac{|x-y|}{\sigma_{\max(x,y)}}$$

Since we have $log\big(\varphi(0)\big) = log\left(\frac{1}{\sigma_x+\sigma_y}\right)$, we would have:

$$dust(x,y) = \sqrt{\frac{|x-y|}{\sigma_{\max(x,y)}}}$$

Finally, *DUST* would be equal to:

$$DUST_{Exp}(X,Y) = \sqrt{\sum_{i=1}^{n}\frac{|x_i-y_i|}{\sigma_{max(x_i,y_i)}}}$$

Uniform Distribution

We want to find $\varphi(|x-y|)$ defined as follows:

$$\varphi(|x-y|) = \int_{-\infty}^{\infty} p(e(x)=x-z)p(e(y)=y-z)dz$$

If we suppose that errors have uniform distribution then we would have:

$$p(e(x)=x-z) = \begin{cases}\frac{1}{b-a} & a\le x-z\le b\\ 0 & OW\end{cases} = \begin{cases}\frac{1}{b-a} & x-a\ge z\ge x-b\\ 0 & OW\end{cases}$$

$$p(e(y)=y-z) = \begin{cases}\frac{1}{d-c} & c\le y-z\le d\\ 0 & OW\end{cases} = \begin{cases}\frac{1}{d-c} & y-c\ge z\ge y-d\\ 0 & OW\end{cases}$$

Note that $\varphi(|x-y|)$is not zero if $[x-b,x-a]\cap[y-d,y-c]\neq\phi$.

If we have$[x-b,x-a]\cap[y-d,y-c]\neq\phi$, the following cases can happen:

1- If we have $[x-b,x-a]\subseteq[y-d,y-c]$

$$\varphi(|x-y|)=\int_{x-b}^{x-a}\frac{1}{(b-a)(d-c)}dz=\frac{x-a-(x-b)}{(b-a)(d-c)}=\frac{b-a}{(b-a)(d-c)}=\frac{1}{d-c}$$

2- If we have $[y-d,y-c]\subseteq[x-b,x-a]$

$$\varphi(|x-y|)=\int_{y-d}^{y-c}\frac{1}{(b-a)(d-c)}dz=\frac{y-c-(y-d)}{(b-a)(d-c)}=\frac{d-c}{(b-a)(d-c)}=\frac{1}{b-a}$$

3- If we have $y-d<x-b<y-c<x-a$ then

$$\varphi(|x-y|)=\int_{x-b}^{y-c}\frac{1}{(b-a)(d-c)}dz=\frac{y-c-(x-b)}{(b-a)(d-c)}=\frac{y-x+b-c}{(b-a)(d-c)}$$

4- If we have $x-b<y-d<x-a<y-c$ then

$$\varphi(|x-y|)=\int_{y-d}^{x-a}\frac{1}{(b-a)(d-c)}dz=\frac{x-a-(y-d)}{(b-a)(d-c)}=\frac{x-y+d-a}{(b-a)(d-c)}$$

Finally, we have:

$$\varphi(|x-y|)$$
$$=\begin{cases}\dfrac{min(x-a,y-c)-max(x-b,y-d)}{(b-a)(d-c)}, & [x-b,x-a]\cap[y-d,y-c]\neq\phi\\ 0 & OW\end{cases}$$

We know that $dust(x,y)=\sqrt{log\big(\varphi(0)\big)-log\big(\varphi(|x-y|)\big)}$. Thus, we would have:

$$dust(x,y)$$

$$= \begin{cases} \sqrt{log\big(\varphi(0)\big) - log\left(\frac{min(x-a,y-c) - max(x-b,y-d)}{(b-a)(d-c)}\right)}, [x-b,x-a] \cap [y-d,y-c] \neq \phi \\ \sqrt{log\big(\varphi(0)\big) - log(0)}\ , \qquad OW \end{cases}$$

Using this, DUST can be calculated. To simplify, we suppose that $a = c, b = d$, and we would have:

$$\varphi(|x-y|) = \begin{cases} \frac{b-a-|x-y|}{(b-a)^2}\ , & 0 \leq |x-y| < b-a \\ 0\ , & OW \end{cases}$$

dust would be calculated as follows:

$$dust(x,y) = \sqrt{log\big(\varphi(0)\big) - log\big(\varphi(|x-y|)\big)}$$

$$= \begin{cases} \sqrt{log\left(\frac{1}{b-a}\right) - log\left(\frac{b-a-|y-x|}{(b-a)^2}\right)}, & 0 \leq |y-x| < b-a \\ \sqrt{log\left(\frac{1}{b-a}\right) - log(0)}\,, & OW \end{cases}$$

$$= \begin{cases} \sqrt{-log(b-a) - (log(b-a-|y-x|) - log((b-a)^2))}\,, & 0 < |y-x| < b-a \\ \sqrt{-log(b-a) - log(0)}\,, & OW \end{cases}$$

$$= \begin{cases} \sqrt{-log(b-a) - log(b-a-|y-x|) + 2log(b-a)}\,, & 0 < |y-x| < b-a \\ \sqrt{-log(b-a) - log(0)}\,, & OW \end{cases}$$

$$= \begin{cases} \sqrt{log(b-a) - log(b-a-|y-x|)}\,, & 0 < |y-x| < b-a \\ \sqrt{-log(b-a) - log(0)}\,, & OW \end{cases}$$

Note that in uniform distribution, we have $\sigma = \frac{b-a}{2\sqrt{3}}$ so $b - a = 2\sqrt{3}\sigma$. Thus, *dust* would be calculated as follows:

$$dust(x,y) = \begin{cases} \sqrt{log\left(2\sqrt{3}\sigma\right) - log\left(2\sqrt{3}\sigma - |y-x|\right)}, & 0 < |y-x| < 2\sqrt{3}\sigma \\ \sqrt{-\left(log\left(2\sqrt{3}\sigma\right) + log(0)\right)}, & OW \end{cases}$$

Using this, we can calculate *DUST* as follows:

$$DUST(X,Y) = \sqrt{\sum_{i=1}^{n} dust(X_i, Y_i)^2}$$