Querying Time-Series Data: A Comprehensive Comparison of Distance Measures

John Paparrizos*, Chunwei Liu‡, Aaron J. Elmore†, Michael J. Franklin†
*The Ohio State University, paparrizos.1@osu.edu
†University of Chicago
‡Massachusetts Institute of Technology

Abstract

Distance measures are core building blocks in time-series analysis and the subject of active research for decades. Unfortunately, the most detailed experimental study in this area is outdated (over a decade old) and, naturally, does not reflect recent progress. Importantly, this study (i) omitted multiple distance measures, including a classic measure in the time-series literature; (ii) considered only a single time-series normalization method; and (iii) reported only raw classification error rates without statistically validating the findings, resulting in or fueling four misconceptions in the time-series literature. Motivated by the aforementioned drawbacks and our curiosity to shed some light on these misconceptions, we comprehensively evaluate 71 time-series distance measures. Specifically, our study includes (i) 8 normalization methods; (ii) 52 lock-step measures; (iii) 4 sliding measures; (iv) 7 elastic measures; (v) 4 kernel functions; and (vi) 4 embedding measures. We extensively evaluate these measures across 128 time-series datasets using rigorous statistical analysis. For the most promising measures, we present an accuracy-to-runtime analysis and summarize recent progress on a generalized lower bounding measure that accelerates all elastic distances. Our findings debunk four long-standing misconceptions that significantly alter the landscape of what is known about existing distance measures. With the new foundations in place, we discuss open challenges and promising directions.

1 Introduction

The understanding of a multitude of natural or human-made processes involves the analysis of high-dimensional observations over time. The recording of such time-varying measurements leads in an ordered sequence of data points called time series [64, 65]. In the last decades, time-series analysis has become increasingly prevalent, affecting virtually all scientific disciplines and their corresponding industries [25, 39, 47, 56, 60, 71, 72]. With sensors and devices becoming increasingly networked and with the explosion of Internet-of-Things (IoT) applications, the volume of produced time series is expected to continue to rise [41, 42, 55, 57, 74]. This growth and ubiquity of time series generates tremendous interest in the extraction of meaningful knowledge from time series [31, 67].

The basis for most analytics over time series involves the detection of similarities between time series. The measurement of similarity, through a distance or similarity measure, is the most fundamental building block in time-series data mining, fueling tasks such as querying [2, 29, 76, 78], indexing [32, 45, 112], clustering [7, 43, 68–70], classification [6, 67, 85, 107], motif discovery [53, 62, 108], and anomaly detection [9–12, 24, 75, 77, 96]. In contrast to other data types where distance measures often process observations independently, for time series, distance measures consider sequences of observations together. This characteristic complicates the definition of distance measures for time series and, therefore, it is desirable to study the factors that determine their effectiveness.

The difficulty in formalizing accurate distance measures stems from the inability to express precisely the notion of similarity. As humans we easily recognize perceptually similar time series, by ignoring a variety

of distortions, such as fluctuations, misalignments, and stretching of observations. However, it is challenging to derive definitions to reflect the similarity for mathematically non-identical time series [33]. Due to that difficulty and the need to handle the variety of distortions, dozens of distance measures have been proposed [5, 8, 16–18, 29, 34, 35, 61, 67, 68, 88, 95, 99].

Despite this abundance of time-series distance measures and their implications in the effectiveness for a multitude of time-series tasks, less attention has been given in their comprehensive experimental validation. Specifically, in the past two decades, only a single comprehensive experimental evaluation has been dedicated to studying the accuracy of 9 influential time-series distance measures over 38 datasets [29]. Unfortunately, this study suffers from three main drawbacks: (i) this study omitted multiple distance measures, including one of the most classic measures in the time-series literature, namely, the cross-correlation measure [13, 79]; (ii) this study considered only a single time-series normalization method; and (iii) this study reported raw classification error rates without performing any rigorous statistical analysis to assess the significance of the findings. Therefore, the analysis is incomplete, and, the findings might not be conclusive. Importantly, this study is now outdated (more than a decade old), and, naturally, it does not reflect recent progress. Considering the previous drawbacks as well as the remarkable interest in time-series analysis, we believe it is critical to revisit this subject.

However, our effort is not only motivated by the necessity to address the aforementioned issues or to extend the previous study with newer datasets and distance measures. Instead, the thorough experimental evaluation of time-series distance measures that we present in this paper is the byproduct of our attempt to challenge four long-standing misconceptions (see $\mathcal{M}1 - \mathcal{M}4$ in Section 2) that have appeared in the time-series literature. These misconceptions are concerned with the (i) normalization of time series; (ii) identification of the state-of-the-art distance measure in every category of measures; (iii) performance of the omitted measures against state-of-the-art measures; and (iv) detection of the most powerful category of measures. Such misconceptions originated from several influential papers [2, 8, 34, 40, 94], some of which date back a quarter of a century, and are fueled by recent inconclusive findings [29] as well as successive claims in the literature that we discuss later. Considering how widely cited and impactful these papers are, we believe it is risky not to challenge such persistent misconceptions that might disorientate newcomer researchers and practitioners.

Motivated by the aforementioned issues and our curiosity to shed some light on these misconceptions, we conduct a comprehensive experimental evaluation to validate the effectiveness of 71 time-series distance measures. These distance measures belong to five categories: (i) 52 lock-step measures, which compare the ith point of one time series with the ith point of another; (ii) 4 sliding measures, which are the sliding versions of lock-step measures when comparing one time series with all shifted versions of the other; (iii) 7 elastic measures, which create a non-linear mapping between time series by comparing one-to-many points in order to align or stretch points; (iv) 4 kernel measures, which use a function (with lock-step, sliding, or elastic properties) to implicitly map data into a high-dimensional space; and (v) 4 embedding measures, which exploit distance or kernel measures indirectly for constructing new representations for time series. In addition, we consider 8 normalization methods for time series.

We perform an extensive evaluation of these distance measures across 128 datasets [25] and compare their classification accuracy obtained from one-nearest-neighbor classifiers (1-NN) under both supervised and unsupervised settings. We conduct a rigorous statistical validation of our findings by employing two statistical tests to assess the significance of the differences in classification accuracy when comparing pairs of measures or multiple measures together. In summary, our study identifies (i) normalization methods leading to significant improvements in a number of distance measures; (ii) new lock-step measures that significantly outperform the current state of the art; (iii) an omitted baseline that most highly popular elastic measures do not outperform; and (iv) new elastic and new kernel measures that significantly outperform the current state of the art. These findings debunk the four long-standing misconceptions and alter the landscape of what is known about existing measures.

We start with the description of the four misconceptions in the literature (Section 2) and we review the relevant background (Section 3). Then, we present our contributions:

- We explore for the first time 8 normalization methods along 56 distance measures (Section 4).

- We study 52 lock-step distance measures (Section 5).
- We investigate 4 classic sliding measures omitted from every previous evaluation (Section 6).
- We compare 7 elastic measures under supervised and unsupervised settings (Section 7).
- We study for the first time 4 kernel (Section 8) and 4 embedding distance measures (Section 9).
- We present an accuracy-to-runtime analysis (Section 10).
- We summarize recent progress towards accelerating the strongest elastic distances via the use of lower bounding measures (Section 11).

Finally, we discuss new directions (Section 12) and conclude with the implications of our work (Section 13). An earlier version of the paper has been published in ACM SIGMOD 2020 [73].

2 The Four Misconceptions

In this section, we describe four misconceptions that have appeared in the time-series literature.

These misconceptions have originated in part from several influential papers [2, 8, 34, 40, 94]. Subsequently, these misconceptions were fueled by a comprehensive study of time-series distance measures [29] as well as dozens of subsequent papers in the literature trusting its findings. Even though an extension of this study appeared five years later [102], this newer version focused on elaborating on the previous results. Recent studies that have focused on time-series classification [6, 54] performed a statistical analysis of several classifiers, including the distance measures in [29, 102]. Unfortunately, these studies only considered supervised tuning of necessary parameters, which does not reflect the use of distance measures for similarity search [32]. Importantly, some results in [6] contradict other results in [54], which, in turn, validated claims that there is no significant difference between the evaluated elastic measures [29, 102]. Interestingly, the improved accuracy found for some measures was attributed to the evaluation framework used while otherwise it was claimed to be undetectable [6]. Considering such apparent difficulties in providing conclusive evidence for this important subject, it is not surprising that the following misconceptions have persisted for so long.

Before we dive into the details, we emphasize that we do not believe or imply that any of these misconceptions were created on purpose. On the contrary, we believe that they are based on evidence, trends, and resources available at the given point in time. We describe the four misconceptions in the form of answers to questions a newcomer researcher would likely identify by studying the literature.

$\mathcal{M}1$**: How to normalize time series?** The consensus is to use the z-score or z-normalization method. Starting with the work of Goldin and Kanellakis [40], a follow-up of the two seminal papers for sequence [2] and subsequence [34] search in time-series databases, that suggested first to normalize the time series to address issues with scaling and translation, z-normalization became the prevalent method to preprocess time series. Despite the proposal of alternative methods the same year [3], the z-normalization was subsequently preferred as the suggested transformations are also applicable to the widely popular Fourier representation [2, 34, 83]. Due to the ubiquity of z-normalization, a valuable resource for time series, the UCR Archive [25], offered until recently the datasets in their z-normalized form. To the best of our knowledge, no study has ever extensively evaluated normalization methods for time series. We review 8 approaches in Section 4 and study their performance in Sections 5 and 6.

$\mathcal{M}2$**: Which lock-step measure to use?** The consensus is to use the Euclidean distance (ED). ED was the method of choice in the first paper for sequence search in time series [2] due to its usefulness in many cases and its applicability over feature vectors. Considering that ED is straightforward to implement, parameter-free, efficient, as well as tightly connected with the Fourier representation and widely supported by indexing mechanisms (in contrast to other L_p-norm variants [109]), there is no surprise about its popularity. Besides, evidence that with increased dataset sizes, the classification error of ED converges to the error of more accurate measures [94], justified its use from virtually all current time-series indexing methods [32]. (Our results in Section 10 suggest that classification error of ED may not always converge to the error of more accurate measures, at least not always

with the same speed of convergence.) In Section 5, we evaluate 52 lock-step measures.

$\mathcal{M}3$: Are elastic better than sliding measures? The answer is currently unknown. Despite the wide popularity of the cross-correlation measure, also known as sliding Euclidean or dot product distance, in the signal and image processing literature [14], cross-correlation has largely been omitted from distance measure evaluations. We believe two factors are responsible for that. First, cross-correlation was considered in the seminal paper [2] as a typical similarity measure, but ED was preferred instead because (i) cross-correlation reduces to ED; and (ii) for the aforementioned reasons in $\mathcal{M}2$. Second, in the introduction of Dynamic Time Warping (DTW) [8], an elastic measure, as an alternative to ED a year later, no comparison was performed against cross-correlation, an obvious baseline. Subsequently, virtually all research on that subject focused either on lock-step or elastic measures [32, 33], with a few exceptions [52, 68, 89]. Interestingly, cross-correlation was not considered as a baseline method in any of the proposed elastic measures [16–18, 61, 95, 99], neither in any of the experimental evaluations of distance measures discussed previously [6, 29, 54, 102]. Strangely, cross-correlation was also omitted from many popular surveys [33, 84]. Therefore, it remains unknown if elastic measures outperform sliding measures. We study 4 sliding measures in Section 6 and analyse their performance against elastic measures in Section 7.

$\mathcal{M}4$: Is DTW the best elastic measure? The general consensus that has emerged is yes. Since the introduction of DTW as a distance measure for time series [8], DTW has inspired the exploration of edit-based distances and it is widely used as the baseline method for this problem [6, 16–18, 54, 61, 68, 95, 99]. It is not uncommon to identify statements even in the abstracts of papers that 1-NN with DTW is exceptionally difficult to outperform [80–82, 105]. Such statements have been backed over the years by the aforementioned extensive evaluations, which conclude that (i) the accuracy of other elastic measures is very close to that of DTW [29, 102]; (ii) there is no significant difference in the accuracy of elastic measures [54]; and (iii) that it is "a little embarrassing" that most classifiers do not outperform 1-NN with DTW [6]. Therefore, there is little space to doubt that DTW is the best elastic measure. To study that misconception, we validate 7 elastic measures in Section 7.

To complete the analysis and capture recent progress, we also include kernel measures and embedding measures in our evaluation (Sections 8 and 9). With the detailed presentation of the four misconceptions, we believe we have now convinced the reader that these misconceptions are not based on any personal biases but, instead, have originated naturally along with the evolution of this area. However, it is risky to not challenge their validity, which may result in confusion for newcomer researchers and practitioners and discourage them from tackling problems in that area. Importantly, it is surprising to consider that half a century of scientific progress has not resulted in any significant improvements over ED or the 50-year-old DTW [87].

Next, we review the relevant background required to validate the accuracy of the normalization methods and distance measures. Even though the efficiency of measures is another important factor of their effectiveness, there are many ways to accelerate each measure, ranging from hardware-aware implementations to algorithmic solutions such as the use of indexing or comparison pruning. We refer the reader to an excellent recent study of data-series similarity search [32], which shows the level of detail required to only evaluate ED. Therefore, we leave such detailed study for future work but we present an accuracy-to-runtime analysis in Section 10.

3 Preliminaries and background

In this section, we review the necessary background for our experimental evaluation.

Terminology and definitions: We consider a time-series dataset as a set of n real-valued vectors $X = [\vec{x}_1, \ldots, \vec{x}_n]^\top \in \mathbb{R}^{n \times m}$, where each time series, $\vec{x}_i \in \mathbb{R}^m$, is an m-dimensional ordered sequence of data points. From this definition, it becomes clear that we consider univariate time series of equal length, where each of these points is a scalar. Following the previous evaluations [6, 29, 102], we consider that the sampling rates of all time series are the same and omit the discrete time stamps.

Datasets: To conduct our extensive evaluation, we use one of the most valuable public resources in the time-series data mining literature, the UCR Time-Series Archive [25]. This archive contains the largest collection

of class-labeled time-series datasets. Currently, the archive consists of 128 datasets and includes time series from sensor readings, image outlines, motion capture, spectrographs, medical signals, electric devices, as well as simulated time series. Each dataset contains from 40 to $24,000$ time series, the lengths vary from 15 to $2,844$, and each time series is annotated with a single label. The majority of the datasets are already z-normalized and we apply the same normalization to all datasets.

The latest version of the archive has deliberately left a small number of datasets containing time series with varying lengths and missing values to reflect the real world. Following the recommendation of the authors of the archive, who performed similar steps to report classification accuracy numbers on the UCR archive website [25], we resample shorter time series to reach the longest time series in each dataset and we fill missing values using linear interpolation. Through these steps, we make the new datasets compatible with previous versions of the archive [66].

Evaluation framework: Following the previous studies [6, 29], we also employ the 1-NN classifier in our evaluation framework, with important differences. 1-NN classifiers are suitable methods for distance measure evaluation for several reasons [29]. Specifically, 1-NN classifiers: (i) resemble the problem solved in time-series similarity search [32]; (ii) are parameter-free and easy to implement; (iii) dependent on the choice of distance measure; and (iv) provide an easy-to-interpret (classification) accuracy measure, which captures if the query and the nearest neighbor belong to the same class.

A critical step for the effectiveness of classifiers is the splitting of a dataset into training and test sets. Previous studies [6, 29, 102] used the k-cross-validation resampling procedure, which produces k groups of time series, tunes necessary parameters on the $k-1$ groups, and evaluates the distance measures using the group of time series left. Strangely, [29, 102] tuned parameters only on a single group and evaluated the distance measures using the $k-1$ groups, which contradicts the common practice. In [6], the improved accuracy of some measures is attributed to such a resampling procedure, while otherwise, it was claimed to be undetectable. Therefore, to eliminate biases from resampling, we respect the split of training and test sets provided by the UCR archive as well as the class distribution in the datasets (i.e., some datasets contain the same number of time series in each class while other datasets contain imbalanced classes). This decision makes our evaluation framework deterministic and enables reproducibility. Refer to [73] for further details on our evaluation settings.

Statistical analysis: To assess the significance of the differences in accuracy, we employ two statistical tests to validate the pairwise comparisons of measures and the comparisons of multiple measures together. Specifically, following the highly influential [26], we use the Wilcoxon test [103] with a 95% confidence level to evaluate pairs of measures over multiple datasets, which is more appropriate than the t-test [86]. As with pairwise tests we cannot reason about multiple measures together and following [26], we also use the Friedman test [36] followed by the post-hoc Nemenyi test [63] to compare multiple measures over multiple datasets and report statistical significant results with 90% confidence level (because these tests require more evidence than Wilcoxon).

Availability of code and results: We implemented the evaluation framework in Matlab, with imported C and Java codes for several distance measures. To ensure the reproducibility of our findings, we make the code available.[1]

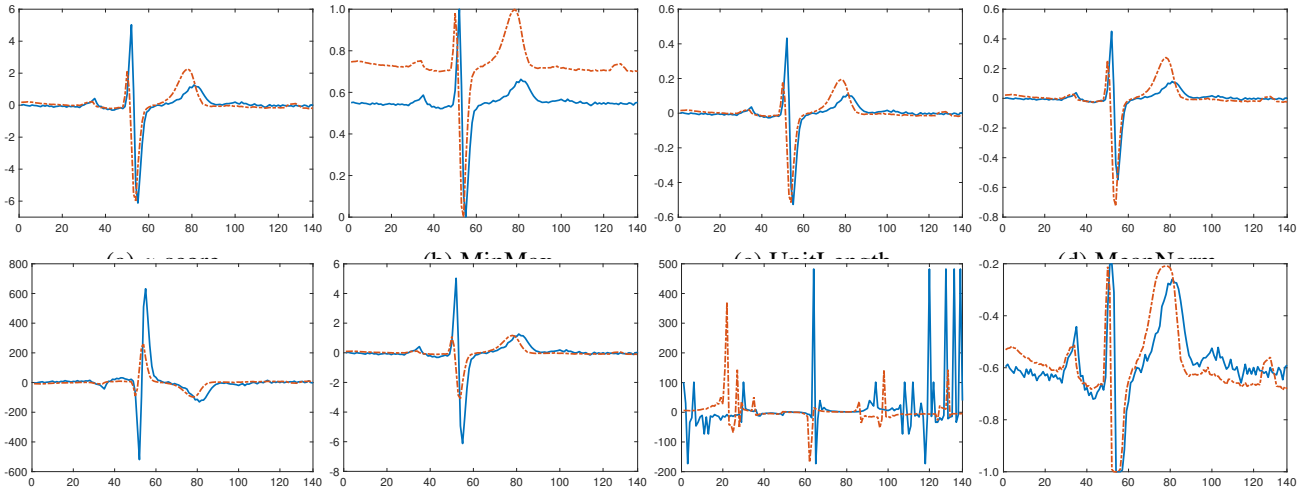
Environment: We ran our experiments on 15 identical servers: Dual Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz and 196GB RAM. Each server has 24 physical cores (12 per CPU), which provided us with 360 cores for four months.

Next, we start with the study of normalization methods.

4 Time-Series Normalizations

In this section, we review 8 normalization methods. As we discussed earlier, a critical issue when comparing time series is how to handle a number of distortions that are characteristic of the time series. For complex distortions,

[1] `https://github.com/TheDatumOrg/TSDistEval`



(e) MedianNorm (f) AdaptiveScaling (g) Logistic (h) Tanh

Figure 1: Example of how each of the 8 normalization methods transforms time series of ECGFiveDays [25].

sophisticated distance measures are required as offering invariances to such distortions is not trivial, which explains the proliferation of distance measures in the literature. However, in several cases, a simple preprocessing step is generally sufficient, as we see next.

Consider the following two examples [40]: (i) two products with similar sales patterns but different sales volume; and (ii) temperatures of two days starting at different values but exhibiting the exact same pattern. The first is an example of the difference in scale between two time series, whereas the second is an example of the difference in translation. Despite such differences, in many cases, it is useful to recognize the similarity between time series. Formally, for any constants a (scale) and b (translation), linear transformations in time series of the form $a\vec{x} + b$ should not affect their similarity.

Several methods have been proposed to handle these popular distortions. Normalization methods transform the data to become normally distributed, whereas standardization methods place different data ranges on a common scale. In the machine-learning literature, feature scaling is also used to refer to such methods. In practice, all terms are used interchangeably to refer to some data transformation.

We consider 8 popular normalization methods in our study, namely, z-score, min-max (MinMax), Mean (MeanNorm), Median (MedianNorm), Unit length (UnitLength), Adaptive scaling (AdaptiveScaling), Logistic or Sigmoid (Logistic), and Hyperbolic tangent (Tanh) normalization. (Please refer to [73] for more details and their mathematical formulas.) Figure 1, shows an example of how each one of the previously described normalization methods transforms a pair of time series from ECGFiveDays [25]. We observe that in some cases, the differences are only visible in the range of values (e.g., z-score vs. UnitLength), but, in others, the visual effect is more distinct (e.g., MinMax, MeanNorm, and AdaptiveScaling). The most unexpected visual effects come from the two non-linear transformations (i.e., Logistic and Tanh). Next, we evaluate 8 methods along with the 52 lock-step measures.

5 Time-Series Lock-Step Distances

In this section, we study 52 lock-step measures that have been proposed across different disciplines.

Distance measures provide a numerical value to quantify how distant are pairs of objects represented as points, vectors, or matrixes. Due to the difficulty in formalizing the notion of similarity, as well as the need to handle a variety of distortions and applications, hundreds of distance measures have been proposed in the literature. This proliferation of distance measures across different scientific areas has resulted in multi-year efforts to organize

this knowledge into dictionaries [27] and encyclopedias [28].

As it is understandable, not all of these measures are applicable to time-series data. Thankfully, different endeavors have already been conducted to identify appropriate measures for a variety of tasks across different fields [37, 110]. An influential study [15] identified 50 lock-step distance measures that we adapt in our evaluation of time-series distance measures. We note that a previous study [38] evaluated a subset of these measures (45) using 1-NN over 42 datasets from the UCR archive and concluded that there is no significant differences between these lock-step distance measures.

Unfortunately, we identified issues with this study. First, several of the evaluated measures are known to be equivalent to each other and, therefore, they should provide identical classification accuracy results. For example, this is the case for the Euclidean distance and the inner product (or Pearson's correlation), which under z-normalization, they should provide the same accuracy numbers. Second, several distance measures were not properly implemented, resulting in using as distance values either the real part of complex numbers or the first value of a normalized vector of the input time series. Therefore, the analysis of these lock-step measures is incomplete, and the findings are inconclusive.

In our study, we have carefully re-implemented all 50 distance measures from [15]. The distance measures belong to 7 different families of measures: (1) 4 measures belong to the L_p Minkowski family; (2) 6 measures belong to the L_1 family; (3) 7 measures belong to the Intersection family; (4) 6 measures belong to the Inner Product family; (5) 5 measures belong to the Fidelity family; (6) 8 measures belong to the L_2 family; and (7) 6 measures belong to the Entropy family. Apart from these 42 measures, we also consider the 3 measures that utilize ideas from multiple other measures (Combinations) as well as 5 measures proposed in the survey but not reported in the literature (until that point).

Besides these measures, we also include two measures that have substantial differences from the previous lock-step measures. Specifically, DISSIM [35] defines the distance as a definite integral of the function of time of the ED in order to take into consideration different sampling rates of time series. This computationally expensive operation can be approximated by a modified version of ED that considers in the distance of the ith points the $i+1$th points, which is a form of a smoothing operation. Finally, the adaptive scaling distance (ASD), embeds internally the AdaptiveScaling normalization with an inner product measure to compare time series under optimal scaling [19, 106].

Evaluation of lock-step measures: For all mathematical formulas, we refer the reader to the previous survey [15]. We evaluate 52 distance measures and their combinations with 8 normalization methods using our 1-NN classifier over 128 datasets (see Section 3). From all combinations of distance measures and normalization methods ($52 \cdot 8 = 416$ in total), we observe 14 measures with some improvement in their average accuracy in contrast to ED and overall 36 combinations with different normalization methods. However, only about half of these combinations result in statistically significant differences according to the pairwise Wilcoxon test. (Refer to [73] for raw numbers in Table 1.) To better understand the performance of lock-step measures, we also evaluate the significance of their differences in accuracy when considering several distance measures together, using the Friedman test followed by a post-hoc Nemenyi test. Specifically, we perform two analyses: (i) we evaluate different distance measures under the same normalization; and (ii) we evaluate standalone distance measures under different normalizations; Figure 2 shows the average rank across all datasets of the distance measures, which under z-score normalization, outperformed previously ED. The thick line connects measures that do not perform statistically significantly better. We observe that Lorentzian is ranked first (once we ignore the supervised Minkowski), meaning that it performed best in the majority of the datasets. All 5 measures significantly outperform ED, but we observe no difference between them. Figure 3 evaluates a standalone distance measure, the Lorentzian measure that performed the best previously, with different normalization methods against ED with z-score. We observe that the 3 out of the 4 combinations that were better than ED under the Wilcoxon test remain better under this statistical analysis, and there is no difference between them.

Debunking $\mathcal{M}1$ and $\mathcal{M}2$: Our evaluation shows clear evidence that normalization methods other than z-score can lead to significant improvements, which debunks $\mathcal{M}1$. Even though for standalone measures, we did not

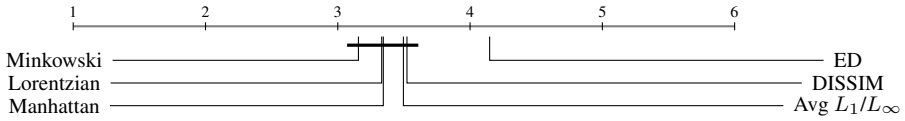


Figure 2: Ranking of lock-step measures under z-score based on the average of their ranks across datasets.

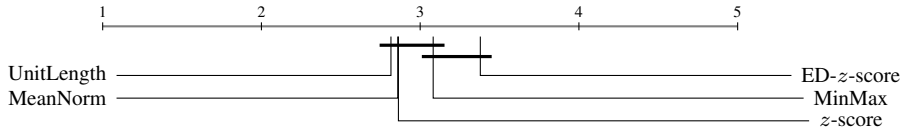


Figure 3: Ranking of normalization methods in combination with the Lorentzian distance based on the average of their ranks across datasets. ED uses z-score normalization.

observe significant improvements (e.g., ED with MeanNorm vs. ED with z-score), that does not reject our hypothesis. We note that the majority of the UCR datasets are in their z-normalized form and, therefore, for fairness, we z-normalized all datasets, which may have limited this analysis. Despite that, we identified two new distance measures, unknown until now, that only under MinMax and MeanNorm methods outperform ED with z-score and, importantly, z-score is not suitable for them. Normalizations such as MeanNorm, which combines z-score and MinMax methods, seems to perform the best for several measures. Similarly, our analysis shows that distance measures other than ED can lead to significant improvements, which debunks $\mathcal{M}2$. We identified 7 distance measures that significantly outperform ED. We emphasize that no previous study considered different normalization methods in order to challenge $\mathcal{M}1$, and our findings contradict both previous studies [29, 38], which concluded that there is no significant difference in the accuracy of lock-step measures.

Next, we focus on sliding versions of lock-step measures.

6 Time-Series Sliding Distances

We study 4 variants of cross-correlation, a measure that has largely been omitted from evaluations.

Starting with the concurrent introduction of lock-step and elastic measures for the problem of time-series similarity search [2, 8, 34], the vast majority of research focused on these two categories of measures (see $\mathcal{M}3$ in Section 2). Cross-correlation, which is similar to convolution, dates back in the 1700s [30] but received practical popularity only after the invention of Fast Fourier Transform (FFT) [20], which dramatically reduced its computational cost. Cross-correlation is one of the most fundamental operations in signal processing [14] and, lately, in deep neural networks [48, 49]. Recently, research focusing on time-series clustering used cross-correlation and achieved state-of-the-art performance for this task [68, 69]. However, this work assumed z-normalized time series and performed evaluations only against ED and DTW. (Refer to [68, 73] for the mathematical notation.)

Evaluation of sliding measures: Due to the resemblance of cross-correlation to the sliding version of Pearson's correlation, when time series are z-normalized, the majority of the literature assumes this underlying data normalization [68]. To the best of our knowledge, the performance of cross-correlation as a measure to compare time series under different normalization methods is not well explored. We measure the performance of the combinations of cross-correlation variants with normalization methods. Specifically, from 32 such combinations (i.e., 4 measures $\times$ 8 normalizations), we report only those resulted in an average accuracy higher than the one achieved by Lorentzian (with z-score followed by UnitLength), the new state-of-the-art lock-step distance measure based on our previous analysis (Section 5). (Refer to [73] for raw numbers and detailed pairwise analysis.)

In addition to these pairwise comparisons, we also evaluate the significance of the differences when considered all together. Figure 4 shows the average rank across datasets of five combinations of NCC_c with normalization methods. Similarly to the pairwise analysis, we observe that combinations with z-score, MeanNorm, and

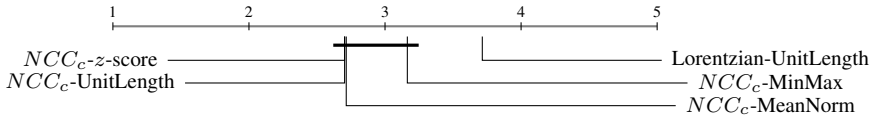


Figure 4: Ranking of different normalization methods for NCC_c based on the average of their ranks across datasets, using Lorentzian with UnitLength as the baseline method.

UnitLength normalizations lead to significant improvements according to the Friedman test followed by a post-hoc Nemenyi test to assess the significance of the differences in the ranking. Combinations of NCC_c with AdaptiveScaling or MinMax do not achieve significant improvement. We observe that both statistical evaluation approaches lead to similar conclusions.

For completeness, we report another analysis using ED as the baseline instead of the Lorentzian distance (we omit the figure due to space limitation). NCC_c in combination with z-score, UnitLength, and MeanNorm normalization methods outperform ED but, in contrast to Figure 4, now combinations with AdaptiveScaling and MinMax are also significantly better than ED. This analysis confirms our results in Section 5 that the Lorentzian distance (and other L_1 variants) are more powerful than ED. In addition, our analysis indicates that NCC_c outperforms all lock-step measures with all different normalizations, making it a strong baseline method for time-series comparison.

We now turn our focus to elastic measures and their performance against sliding measures.

7 Time-Series Elastic Measures

In this section, we study 7 elastic measures, a popular category of measures for time-series comparison.

As discussed earlier, sliding measures find a global alignment by sliding one time series against the other. In contrast, elastic measures create a non-linear mapping between time-series data points to support flexible alignment of different regions. Through this mapping, elastic measures permit time series to "stretch" or "shrink" their observations to improve time-series matching. Most elastic measures rely on dynamic programming to find this mapping efficiently by defining recursive formulas over a m-by-m matrix M that contains in each cell the ED (or some other lock-step measure) between every point of one time series against every point of another time series. In general, the goal of different elastic measures in the literature is to employ different strategies to find a warping path, $W = \{w_1, \ldots, w_k\}$, with $k \geq m$, a contiguous set of matrix cells that shows the mapping of every point of one time series to one, more, or none of the points of the other time series. To improve the efficiency and the accuracy of elastic measures, it is a common practice to introduce constraints (i.e., parameters) to guide the warping path to visit only a subset of cells in M.

The first elastic measure, DTW [87, 88], was proposed as a speech recognition tool and, later, it was introduced in the time-series literature as a suitable approach for time-series comparison [8]. DTW finds the warping path that minimizes the distances between all data points. In the original form, DTW is parameter-free, however, many approaches have been proposed to define bands (i.e., the shape of the subset cells of matrix M that the warping path is permitted to visit) and the width or window (i.e., size) of the bands. We use the Sakoe-Chiba band [88], which is the most frequently used in practice [29], and we tune the window δ using parameters shown in Table 4 of [73]. For example, a value $\delta = 10$ indicates a window size 10% of the time-series length.

The Longest Common Subsequence (LCSS) distance is another type of elastic measure that was derived from the idea of edit-distances for characters. Specifically, LCSS introduces a parameter ϵ that serves as a threshold to determine when two points of time series should match [4, 99]. Similarly to DTW, LCSS also constrains the warping window by introducing an additional parameter δ [99]. Edit Distance on Real sequence (EDR) distance [17] is another edit-distance-based measure that similarly to LCSS, uses a parameter ϵ to quantify the distance of points as 0 or 1. EDR also introduces penalties for gaps between matched subsequences. Edit Distance with Real

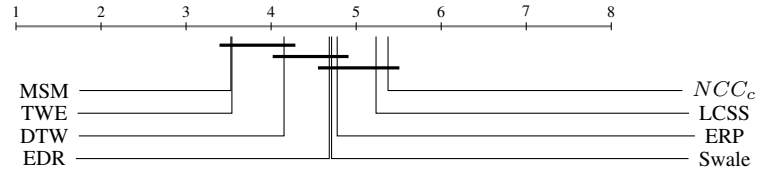


Figure 5: Ranking of elastic and sliding distance measures based on the average of their ranks across datasets, using supervised tuning for their parameters.

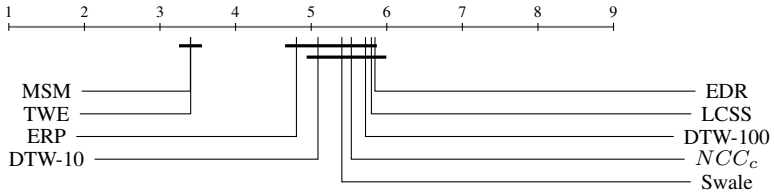


Figure 6: Ranking of elastic and sliding distance measures based on the average of their ranks across datasets, using unsupervised tuning for their parameters.

Penalty (ERP) distance [16] bridges DTW and EDR distance measures by more carefully computing the distance between gaps.

Differently than the previous approaches, the Sequence Weighted Alignment model (Swale) [61] proposes a model to compute the similarity of time series using rewards for matching points and penalties for gaps. Apart from a threshold ϵ parameter, Swale also requires parameters for the reward r and the penalty p. The Move–split–merge (MSM) distance [95] is another elastic measure based on edit-distance but in contrast to DTW, LCSS, and EDR, MSM is a metric. MSM uses a set of operations to replace, insert, or delete values in time series to improve their matching. Finally, Time Warp Edit (TWE) distance [58] is a measure that combines merits from LCSS and DTW. TWE introduces a stiffness parameter ν to control the warping but at the same point it also penalizes matched points.

Evaluation of elastic vs. sliding measures: With the introduction of the 7 elastic measures we are now in position to evaluate their performance against sliding measures, an experiment that has been omitted in all previous studies [6, 29]. Refer to [73] for detailed raw numbers and pairwise comparisons under supervised and unsupervised settings.

To understand the performance of elastic measures against NCC_c, we evaluate the significance of the differences when considered all together. Specifically, Figure 5 shows the average ranks of the elastic measures in the supervised setting and Figure 6 shows the average ranks in the unsupervised setting. We observe that even under supervised settings, 4 out of the 7 elastic measures, namely, LCSS, ERP, EDR, and Swale, do not achieve significantly better performance than NCC_c. The results for MSM, TWE, and DTW, are consistent in both statistical evaluations. For the unsupervised setting, both statistical evaluation approaches agree to an extent. In particular, Figure 6 shows clearly that MSM and TWE outperform NCC_c. However, the remaining 5 elastic measures perform similarity to NCC_c. To validate our findings, we repeat the analysis (we omit figures due to space limitation) and evaluate the significance of the differences when we consider all elastic measures together (i.e., excluding NCC_c). Specifically, we observe that Swale, ERP, EDR, and LCSS do not outperform DTW-10 with statistically significant difference. Interestingly, the supervised LCSS is slightly worse than the unsupervised DTW-10. ERP, which under pairwise evaluation appears to significantly outperform DTW-10, when all measures are considered together, both appear to achieve comparable performance. MSM, TWE, and DTW also perform similarly and all three supervised measures outperform DTW-10. However, under unsupervised settings, MSM and TWE significantly outperform all elastic measures.

Debunking $\mathcal{M}3$ and $\mathcal{M}4$: Our comprehensive evaluation shows clear evidence that sliding measures are strong baselines that most elastic measures do not manage to outperform either in supervised or unsupervised settings,

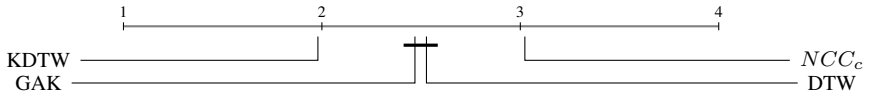


Figure 7: Ranking of kernel measures based on the average of their ranks across datasets (supervised tuning).

which debunks $\mathcal{M}3$. Specifically, from all 5 elastic measures evaluated in the decade-old study [29], namely, LCSS, Swale, EDR, ERP, and DTW, only DTW significantly outperforms cross-correlation under the supervised scenario. In the unsupervised setting, none of the 5 measures outperforms cross-correlation and, interestingly, several of them perform slightly worse. This is a remarkable finding, showing that the simplest type of alignment between time series is very effective and it should have served as a baseline method for elastic measures. Only MSM and TWE, two measures that appeared after [29] show promising results and outperform cross-correlation with statistically significant differences in both supervised and unsupervised settings. Importantly, MSM is the only method that significantly outperforms DTW under supervised settings (according to Wilcoxon) and, under unsupervised settings, both MSM and TWE significantly outperform DTW (with both statistical tests validating this result). Therefore, there is clear evidence that the widely popular DTW is no longer the best elastic distance measure, which debunks $\mathcal{M}4$.

8 Time-Series Kernel Measures

Until now, our analysis focused on three categories of distance measures, namely, lock-step, sliding, and elastic measures, with the goal to provide answers to the four-long standing misconceptions that we discussed in Section 2. Recently, kernel functions [91, 92], a different category of similarity measures, have started to receive attention due to their competitive performance [1]. In contrast to all previously described measures, kernel functions must satisfy the positive semi-definiteness property (p.s.d) [90]. The precise definition is out of the scope of this work (we refer the reader to recent papers for a detailed review [1, 67]) but in simple terms, a function is p.s.d. if the similarity matrix, which contains all pairwise similarity values, has positive eigenvalues. This important property results in convex solutions for several learning tasks involving kernels [21]. In this section, we study 4 representative kernel functions and evaluate their performance against sliding and elastic measures.

Specifically, the first kernel we consider is the Radial Basis Function (RBF) [22], a general purpose kernel function that internally exploits ED but maps data into a high-dimensional space where their separation is easier. To capture similarities between the shifted versions of time series, [100] proposed a sliding kernel to consider all possible alignments between time-series. We include a recently proposed variant of this kernel, namely, SINK, that has achieved competitive results to NCC_c and DTW [67]. Finally, we include two elastic kernel functions, the Global Alignment Kernel (GAK) [23] and Dynamic Time Warping Kernel (KDTW) [59].

Evaluation of kernel functions: Having introduced the 4 kernel functions, we are now in position to evaluate their performance against sliding and elastic measures. As before, we consider both supervised and unsupervised settings. In the supervised setting, we observe that all kernel functions significantly outperform NCC_c with the exception of RBF, which is significantly worse. In the unsupervised settings, KDTW and GAK significantly outperform NCC_c, as before, but SINK achieves comparable performance without outperforming NCC_c. To better understand the performance of KDTW and GAK, which appear to be the strongest kernel functions, we also evaluate the significance of the differences when considered together with all elastic and sliding measures. Figure 7 presents the results for supervised settings and Figure 8 for unsupervised settings. We have omitted elastic measures that based on the earlier analysis did not show competitive results. We observe that GAK achieves comparable performance to DTW under both settings. However, KDTW, significantly outperforms DTW in both unsupervised and superivsed settings. This is in contrast to TWE and MSM measures that were significantly better only under the unsupervised settings. To the best of our knowledge, this is the first time that a kernel function is reported to outperform DTW in both settings.

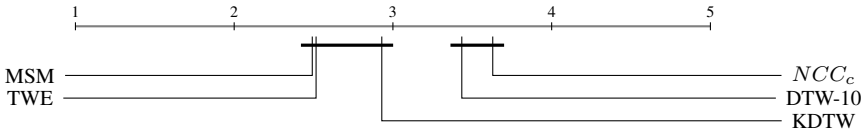


Figure 8: Ranking of kernel measures based on the average of their ranks across datasets (unsupervised tuning).

9 Time-Series Embedding Measures

Previously, we studied approaches that directly exploit a kernel function or a distance measure to compare time series. In this section, we study 4 embedding measures, which are alternative approaches that employ a similarity measure only to construct new representations [1]. These representations are similarity-preserving as the comparison of two representations with ED approximates the comparison of the corresponding original time series with the employed similarity measure.

We consider 4 approaches to construct embedding measures (i.e., ED over learned representations). Specifically, we consider the Generic RepresentAtIon Learning (GRAIL) framework, which employs the SINK kernel [67], the Shift-invariant Dictionary Learning (SIDL) method, which preserves alignment between time series [111], the Similarity Preserving Representation Learning method (SPIRAL), which employs DTW [50], and the Random Warping Series (RWS), which preserves the GAK kernel [104].

Evaluation of embedding measures: For all approaches, we follow [67] and tune required parameters using the recommended values from their corresponding papers. We construct representations of same length (100) for fairness. We observe that GRAIL, is the only framework that constructs robust representations that when ED is used for comparison (under the 1-NN settings), it achieves similar performance to NCC_c, but without significant difference. All other embedding measures perform significantly worse and none of the embedding measures outperform DTW (see detailed raw numbers in [73]. We note, however, that embedding measures (as well as kernel methods), achieve much higher accuracy under different evaluation frameworks (e.g., with SVM classifiers), as shown in [67].

10 Accuracy-to-runtime Analysis

Until now, we have extensively evaluated distance measures based on their accuracy results. However, it is also important to understand the cost associated with each one of these distance measures. In Figure 9, we summarize the accuracy-to-runtime performance of the most prominent measures. The runtime performance includes only inference time (i.e., evaluation on the testing sets). We observe that ED, and all other lock-step measures (omitted), are the fastest but achieve relatively low accuracy (all these measures have $\mathcal{O}(m)$ runtime cost). NCC_c [68] and SINK [67], two methods that rely on the classic cross-correlation measure, provide an excellent trade-off between runtime and accuracy in comparison to ED (these measures have $\mathcal{O}(m \log m)$ runtime cost). We also observe that all other elastic or kernel methods require substantially higher runtime costs to achieve comparable accuracy results to NCC_c (these measures have $\mathcal{O}(m^2)$ runtime cost). In particular, only MSM and TWE significantly outperform NCC_c (see Figure 6) but require two orders of magnitude higher runtime cost. Instead, embedding measures, such as GRAIL [67], show great promise as they can achieve high accuracy without sacrificing runtime performance.

11 Accelerating Elastic Measures

Despite their promise, elastic distance measures scale quadratically to the length of the time series, as noted earlier. Compared to ED, which has linear complexity, elastic distance measures incur an additional runtime overhead, often between one to three orders of magnitude (see Figure 9). This cost would prevent applications from using

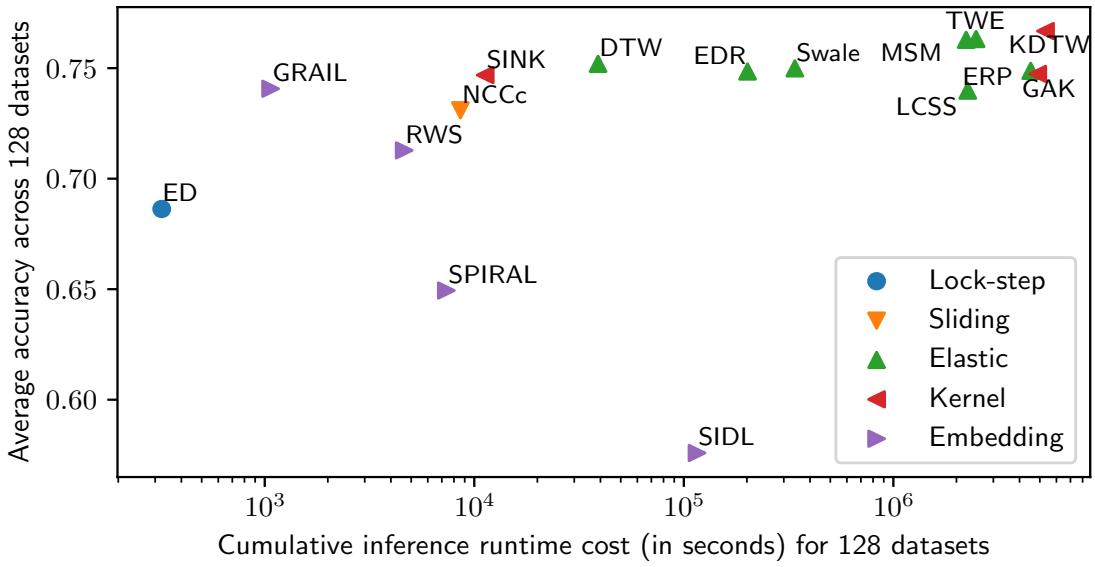


Figure 9: Accuracy-to-runtime comparison.

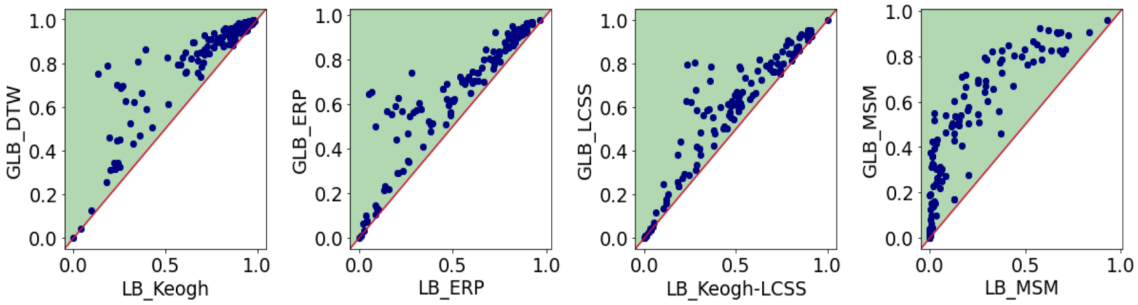


Figure 10: Comparison of the pruning power of GLB variants against state-of-the-art LBs of several popular elastic measures over 128 datasets. The blue dots above the diagonal indicate datasets over which GLB outperforms the state of the art.

elastic measures in large-scale settings. To alleviate this issue, the idea of lower bounding was developed to filter out unpromising candidates before carrying out the expensive elastic distance measure computation [34, 44, 46]. In simple terms, a lower bound (LB) is a fast distance measure that approximates an expensive elastic distance measure and is computed over some summaries of the time series instead of the actual time series.

A plethora of LBs have been developed for elastic distance measures [16, 44, 46, 51, 93, 97, 98], with the goal to improve their pruning power (i.e., *tightness* of LB). Unfortunately, the research effort on LBs has been disproportionally concentrated on Dynamic Time Warping (DTW) [87, 88], which is the oldest elastic measure with at least eight established LBs (see [78] for details). In contrast, newer and better-performing elastic distance measures, such as MSM and TWE, have received little attention, and their LBs are performing poorly. Unfortunately, developing LBs is a challenging task. It is unsustainable to expect a similar research effort for each elastic measure. For this reason, a generalized framework, namely GLB, was recently proposed [78] to accumulate the knowledge from previously developed LBs and eliminate the need for designing separate LBs for each elastic measure. Specifically, GLB outperforms all established LBs across different elastic measures. Figure 10 shows the improvement in pruning power (i.e., the percentage of the true distance computation avoided) achieved by GLB for several popular elastic measures (more details in [78]). Considering that MSM and TWE are the new state-of-the-art elastic measures, we note that GLB accelerates MSM up to 10× and TWE up to 26× in an extensive analysis we performed across 128 datasets [25].

12 Future Directions

With the new knowledge in place, several new challenges open that we hope to spark new research directions. Below, we provide three areas that we believe require more attention and can potentially lead to substantial improvements in the entire area of time-series similarity search:

- Identifying more accurate normalizations. Our work was the first to study the performance of 8 normalization methods. We identified multiple distance measures outperforming the previous SOTA measures only when combined with appropriate normalization methods. In our view, inventing a new normalization method that achieves significant accuracy improvements by preprocessing data differently and without changing existing methods and systems would be a breakthrough.
- Tuning parameters, or selecting appropriate distance measures per dataset in an unsupervised manner. Unfortunately, there are no principled methodologies currently for selecting distance measures or tuning their parameters, despite significant recent attention in AutoML for other domains.
- Improving and evaluating the performance of embedding measures. These measures show the most promise based on their runtime-to-accuracy trade-off. To the best of our knowledge and based on our comprehensive study, there are no embedding measures that significantly outperform the most vigorous elastic measures in terms of accuracy. Recent advances in deep neural networks [101] may lead to embeddings that substantially outperform elastic measures.

13 Conclusion

We presented a comprehensive evaluation to validate the performance of 71 distance measures. Our study debunked four long-standing misconceptions in the time-series literature and established new state-of-the-art results for lock-step, sliding, elastic, kernel, and embedding measures. Our findings prepare the ground for the development of distance measures with implications across time-series analytical tasks. Importantly, our work has implications for general-purpose similarity search problems over high-dimensional data. For example, several similarity search methodologies rely heavily on the concepts of lower bounding to prune unnecessary comparisons [32, 76]. Similarly to how GLB abstracted the costs of different elastic measures and generalized lower bounds for time series, we believe a similar concept can be applied in the case of lock-step measures (e.g., Euclidean distance) and the corresponding data summarization methods. In addition, our work identified lock-step measures that outperform Euclidean distance and lock-step measures performing exceptionally well only under certain normalizations. However, the literature in the similarity search area has largely focused on developing methods assuming Euclidean distance is the underlying distance measure. Our work may lead to new solutions for the new, better-performing distance measures. Finally, the methodologies presented for constructing embedding measures are sufficiently generic and can complement solutions focusing on learning embeddings from data [101] (e.g., concatenate deep embeddings with our similarity-preserving embeddings or improve deep embeddings by integrating our similarity-preserving embeddings in the loss functions).

Acknowledgments: We thank Kaize Wu for his help and useful discussions. This research was supported in part by a Google DAPA Research Award, gifts from NetApp, Cisco Systems, and Exelon Utilities, and an NSF Award CCF-1139158. Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation.

References

[1] Amaia Abanda, Usue Mori, and Jose A Lozano. A review on distance based time series classification. Data Mining and Knowledge Discovery, 33(2):378–412, 2019.

[2] Rakesh Agrawal, Christos Faloutsos, and Arun N. Swami. Efficient similarity search in sequence databases. In FODO, pages 69–84, 1993.

[3] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, and Kyuseok Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In Proceeding of the 21th International Conference on Very Large Data Bases, pages 490–501. Citeseer, 1995.

[4] Henrik André-Jönsson and Dushan Z Badal. Using signature files for querying time-series data. In European Symposium on Principles of Data Mining and Knowledge Discovery, pages 211–220. Springer, 1997.

[5] Johannes Aßfalg, Hans-Peter Kriegel, Peer Kröger, Peter Kunath, Alexey Pryakhin, and Matthias Renz. Similarity search on time series based on threshold queries. In International Conference on Extending Database Technology, pages 276–294. Springer, 2006.

[6] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 31(3):606–660, 2017.

[7] Mohini Bariya, Alexandra von Meier, John Paparrizos, and Michael J Franklin. k-shapestream: Probabilistic streaming clustering for electric grid events. In 2021 IEEE Madrid PowerTech, pages 1–6. IEEE, 2021.

[8] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In AAAI Workshop on KDD, pages 359–370, 1994.

[9] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. Sand: streaming subsequence anomaly detection. Proceedings of the VLDB Endowment, 14(10):1717–1729, 2021.

[10] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. Sand in action: subsequence anomaly detection for streams. Proceedings of the VLDB Endowment, 14(12):2867–2870, 2021.

[11] Paul Boniol, John Paparrizos, Yuhao Kang, Themis Palpanas, Ruey S Tsay, Aaron J Elmore, and Michael J Franklin. Theseus: navigating the labyrinth of time-series anomaly detection. Proceedings of the VLDB Endowment, 15(12):3702–3705, 2022.

[12] Paul Boniol, John Paparrizos, and Themis Palpanas. New trends in time series anomaly detection. In Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023, pages 847–850. OpenProceedings.org, 2023. doi: 10.48786/edbt.2023.80. URL https://doi.org/10.48786/edbt.2023.80.

[13] R Bracewell. Pentagram notation for cross correlation. the fourier transform and its applications. New York: McGraw-Hill, 46: 243, 1965.

[14] Lisa Gottesfeld Brown. A survey of image registration techniques. ACM computing surveys (CSUR), 24(4):325–376, 1992.

[15] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. City, 1(2):1, 2007.

[16] Lei Chen and Raymond Ng. On the marriage of Lp-norms and edit distance. In VLDB, pages 792–803, 2004.

[17] Lei Chen, M Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In SIGMOD, pages 491–502, 2005.

[18] Yueguo Chen, Mario A Nascimento, Beng Chin Ooi, and Anthony KH Tung. Spade: On shape-based pattern detection in streaming time series. In ICDE, pages 786–795, 2007.

[19] Kelvin Kam Wing Chu and Man Hon Wong. Fast time-series searching with scaling and shifting. In Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 237–248. Citeseer, 1999.

[20] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90):297–301, 1965.

[21] Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine learning, 20(3):273–297, 1995.

[22] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[23] Marco Cuturi. Fast global alignment kernels. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 929–936, 2011.

[24] Michele Dallachiesa, Themis Palpanas, and Ihab F Ilyas. Top-k nearest neighbor search in uncertain data series. Proceedings of the VLDB Endowment, 8(1):13–24, 2014.

[25] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[26] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 7:1–30, 2006.

[27] Michel-Marie Deza and Elena Deza. Dictionary of distances. Elsevier, 2006.

[28] Michel Marie Deza and Elena Deza. Encyclopedia of distances. In Encyclopedia of distances, pages 1–583. Springer, 2009.

[29] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 1(2):1542–1552, 2008.

[30] Alejandro Domínguez. A history of the convolution operation [retrospectroscope]. IEEE pulse, 6(1):38–49, 2015.

[31] Adam Dziedzic, John Paparrizos, Sanjay Krishnan, Aaron Elmore, and Michael Franklin. Band-limited training and inference for convolutional neural networks. In International Conference on Machine Learning, pages 1745–1754. PMLR, 2019.

[32] Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. The lernaean hydra of data series similarity search: An experimental evaluation of the state of the art. Proceedings of the VLDB Endowment, 12(2):112–127, 2018.

[33] Philippe Esling and Carlos Agon. Time-series data mining. ACM Computing Surveys (CSUR), 45(1):12, 2012.

[34] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. In SIGMOD, pages 419–429, 1994.

[35] Elias Frentzos, Kostas Gratsias, and Yannis Theodoridis. Index-based most similar trajectory search. In ICDE, pages 816–825, 2007.

[36] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32:675–701, 1937.

[37] Daniel G Gavin, W Wyatt Oswald, Eugene R Wahl, and John W Williams. A statistical approach to evaluating distance metrics and analog assignments for pollen records. Quaternary Research, 60(3):356–367, 2003.

[38] Rafael Giusti and Gustavo EAPA Batista. An empirical comparison of dissimilarity measures for time series classification. In BRACIS, pages 82–88, 2013.

[39] Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. The social dynamics of language change in online networks. In Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I 8, pages 41–57. Springer, 2016.

[40] Dina Q Goldin and Paris C Kanellakis. On similarity queries for time-series data: constraint specification and implementation. In International Conference on Principles and Practice of Constraint Programming, pages 137–153. Springer, 1995.

[41] Hao Jiang, Chunwei Liu, Qi Jin, John Paparrizos, and Aaron J Elmore. Pids: attribute decomposition for improved compression and query performance in columnar storage. Proceedings of the VLDB Endowment, 13(6):925–938, 2020.

[42] Hao Jiang, Chunwei Liu, John Paparrizos, Andrew A Chien, Jihong Ma, and Aaron J Elmore. Good to the last bit: Data-driven encoding with codecdb. In Proceedings of the 2021 International Conference on Management of Data, pages 843–856, 2021.

[43] Eamonn Keogh and Jessica Lin. Clustering of time-series subsequences is meaningless: Implications for previous and future research. Knowledge and Information Systems, 8(2):154–177, 2005.

[44] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. Knowledge and Information Systems, 7(3):358–386, 2005.

[45] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Locally adaptive dimensionality reduction for indexing large time series databases. In SIGMOD, pages 151–162, 2001.

[46] Sang-Wook Kim, Sanghyun Park, and Wesley W Chu. An index-based approach for similarity search supporting time warping in large sequence databases. In Data Engineering, 2001. Proceedings. 17th International Conference on, pages 607–614. IEEE, 2001.

[47] Sanjay Krishnan, Aaron J Elmore, Michael Franklin, John Paparrizos, Zechao Shang, Adam Dziedzic, and Rui Liu. Artificial intelligence in resource-constrained and shared environments. ACM SIGOPS Operating Systems Review, 53(1):1–6, 2019.

[48] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10):1995, 1995.

[49] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436, 2015.

[50] Qi Lei, Jinfeng Yi, Roman Vaculin, Lingfei Wu, and Inderjit S Dhillon. Similarity preserving representation learning for time series analysis. arXiv preprint arXiv:1702.03584, 2017.

[51] Daniel Lemire. Faster retrieval with a two-pass dynamic-time-warping lower bound. Pattern recognition, 42(9):2169–2180, 2009.

[52] Chung-Sheng Li, Philip S. Yu, and Vittorio Castelli. Hierarchyscan: A hierarchical similarity search algorithm for databases of long sequences. In ICDE, pages 546–553. IEEE, 1996.

[53] Michele Linardi and Themis Palpanas. Scalable, variable-length similarity search in data series: The ulisse approach. Proceedings of the VLDB Endowment, 11(13):2236–2248, 2018.

[54] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. Data Mining and Knowledge Discovery, 29(3):565–592, 2015.

[55] Chunwei Liu, Hao Jiang, John Paparrizos, and Aaron J Elmore. Decomposed bounded floats for fast compression and queries. Proceedings of the VLDB Endowment, 14(11):2586–2598, 2021.

[56] Shinan Liu, Tarun Mangla, Ted Shaowang, Jinjin Zhao, John Paparrizos, Sanjay Krishnan, and Nick Feamster. Amir: Active multimodal interaction recognition from video and network traffic in connected environments. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7(1):1–26, 2023.

[57] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. Machine learning for internet of things data analysis: A survey. Digital Communications and Networks, 2017.

[58] Pierre-François Marteau. Time warp edit distance with stiffness adjustment for time series matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):306–318, 2008.

[59] Pierre-François Marteau and Sylvie Gibet. On recursive edit distance kernels with application to time series classification. IEEE transactions on neural networks and learning systems, 26(6):1121–1133, 2014.

[60] Kathy McKeown, Hal Daume III, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R Fleischmann, et al. Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology, 67(11):2684–2696, 2016.

[61] Michael D Morse and Jignesh M Patel. An efficient and accurate method for evaluating time series similarity. In SIGMOD, pages 569–580, 2007.

[62] Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover. Exact discovery of time series motifs. In Proceedings of the 2009 SIAM international conference on data mining, pages 473–484. SIAM, 2009.

[63] Peter Nemenyi. Distribution-free Multiple Comparisons. PhD thesis, Princeton University, 1963.

[64] Themis Palpanas. Data series management: the road to big sequence analytics. ACM SIGMOD Record, 44(2):47–52, 2015.

[65] Ioannis Paparrizos. Fast, scalable, and accurate algorithms for time-series analysis. PhD thesis, Columbia University, 2018.

[66] John Paparrizos. 2018 ucr time-series archive: Backward compatibility, missing values, and varying lengths, January 2019. `https://github.com/johnpaparrizos/UCRArchiveFixes`.

[67] John Paparrizos and Michael J Franklin. Grail: efficient time-series representation learning. Proceedings of the VLDB Endowment, 12(11):1762–1777, 2019.

[68] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pages 1855–1870. ACM, 2015.

[69] John Paparrizos and Luis Gravano. Fast and accurate time-series clustering. ACM Transactions on Database Systems (TODS), 42 (2):8, 2017.

[70] John Paparrizos and Sai Prasanna Teja Reddy. Odyssey: An engine enabling the time-series clustering journey. Proceedings of the VLDB Endowment, 16(12):4066–4069, 2023.

[71] John Paparrizos, Ryen W White, and Eric Horvitz. Detecting devastating diseases in search logs. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 559–568, 2016.

[72] John Paparrizos, Ryen W White, and Eric Horvitz. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. Journal of oncology practice, 12(8):737–744, 2016.

[73] John Paparrizos, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. Debunking four long-standing misconceptions of time-series distance measures. In Proceedings of the 2020 ACM SIGMOD international conference on management of data, pages 1887–1905, 2020.

[74] John Paparrizos, Chunwei Liu, Bruno Barbarioli, Johnny Hwang, Ikraduya Edian, Aaron J Elmore, Michael J Franklin, and Sanjay Krishnan. Vergedb: A database for iot analytics on edge devices. In CIDR, 2021.

[75] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proceedings of the VLDB Endowment, 15(11):2774–2787, 2022.

[76] John Paparrizos, Ikraduya Edian, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. Fast adaptive similarity search through variance-aware quantization. In 2022 IEEE 38th International Conference on Data Engineering (ICDE), pages 2969–2983. IEEE, 2022.

[77] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S Tsay, Themis Palpanas, and Michael J Franklin. Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection. Proceedings of the VLDB Endowment, 15(8):1697–1711, 2022.

[78] John Paparrizos, Kaize Wu, Aaron Elmore, Christos Faloutsos, and Michael J Franklin. Accelerating similarity search for elastic measures: A study and new generalization of lower bounding distances. Proceedings of the VLDB Endowment, 16(8):2019–2032, 2023.

[79] Athanasios Papoulis. The Fourier integral and its applications. McGraw-Hill, 1962.

[80] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition, 44(3):678–693, 2011.

[81] François Petitjean, Germain Forestier, Geoffrey I Webb, Ann E Nicholson, Yanping Chen, and Eamonn Keogh. Dynamic time warping averaging of time series allows faster and more accurate classification. In 2014 IEEE international conference on data mining, pages 470–479. IEEE, 2014.

[82] François Petitjean, Germain Forestier, Geoffrey I Webb, Ann E Nicholson, Yanping Chen, and Eamonn Keogh. Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. Knowledge and Information Systems, 47(1):1–26, 2016.

[83] Davood Rafiei and Alberto Mendelzon. Similarity-based queries for time series data. In ACM SIGMOD Record, volume 26, pages 13–25. ACM, 1997.

[84] Chotirat Ann Ralanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das. Mining time series data. In Data mining and knowledge discovery handbook, pages 1069–1103. Springer, 2005.

[85] Chotirat Ann Ratanamahatana and Eamonn Keogh. Making time-series classification more accurate using learned constraints. In SDM, pages 11–22, 2004.

[86] John Rice. Mathematical statistics and data analysis. Cengage Learning, 2006.

[87] Hiroaki Sakoe and Seibi Chiba. A dynamic programming approach to continuous speech recognition. In ICA, pages 65–69, 1971.

[88] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1):43–49, 1978.

[89] Yasushi Sakurai, Spiros Papadimitriou, and Christos Faloutsos. Braid: Stream mining through group lag correlations. In SIGMOD, pages 599–610. ACM, 2005.

[90] Bernhard Schölkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[91] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In International Conference on Artificial Neural Networks, pages 583–588. Springer, 1997.

[92] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5):1299–1319, 1998.

[93] Yilin Shen, Yanping Chen, Eamonn Keogh, and Hongxia Jin. Accelerating time series searching with large uniform scaling. In Proceedings of the 2018 SIAM International Conference on Data Mining, pages 234–242. SIAM, 2018.

[94] Jin Shieh and Eamonn Keogh. i sax: indexing and mining terabyte sized time series. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 623–631. ACM, 2008.

[95] Alexandra Stefan, Vassilis Athitsos, and Gautam Das. The move-split-merge metric for time series. TKDE, 25(6):1425–1438, 2013.

[96] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, and Themis Palpanas. Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. Proceedings of the VLDB Endowment, 16(11):3418–3432, 2023.

[97] Chang Wei Tan, François Petitjean, and Geoffrey I Webb. Elastic bands across the path: A new framework and method to lower bound dtw. In Proceedings of the 2019 SIAM International Conference on Data Mining, pages 522–530. SIAM, 2019.

[98] Chang Wei Tan, François Petitjean, and Geoffrey I Webb. Fastee: Fast ensembles of elastic distances for time series classification. Data Mining and Knowledge Discovery, 34(1):231–272, 2020.

[99] Michail Vlachos, George Kollios, and Dimitrios Gunopulos. Discovering similar multidimensional trajectories. In Proceedings 18th international conference on data engineering, pages 673–684. IEEE, 2002.

[100] Gabriel Wachman, Roni Khardon, Pavlos Protopapas, and Charles R Alcock. Kernels for periodic time series arising in astronomy. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 489–505. Springer, 2009.

[101] Qitong Wang and Themis Palpanas. Seanet: A deep learning architecture for data series similarity search. IEEE Transactions on Knowledge and Data Engineering, 2023.

[102] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, pages 1–35, 2013.

[103] Frank Wilcoxon. Individual comparisons by ranking methods. Biometrics Bulletin, pages 80–83, 1945.

[104] Lingfei Wu, Ian En-Hsu Yen, Jinfeng Yi, Fangli Xu, Qi Lei, and Michael Witbrock. Random warping series: A random features method for time-series embedding. In AISTATS, pages 793–802, 2018.

[105] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. In Proceedings of the 23rd international conference on Machine learning, pages 1033–1040. ACM, 2006.

[106] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In WSDM, pages 177–186, 2011.

[107] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 947–956. ACM, 2009.

[108] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In 2016 IEEE 16th international conference on data mining (ICDM), pages 1317–1322. IEEE, 2016.

[109] Byoung-Kee Yi and Christos Faloutsos. Fast time sequence indexing for arbitrary lp norms. VLDB, 2000.

[110] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. Similarity search: the metric space approach, volume 32. Springer Science & Business Media, 2006.

[111] Guoqing Zheng, Yiming Yang, and Jaime Carbonell. Efficient shift-invariant dictionary learning. In SIGKDD, pages 2095–2104. ACM, 2016.

[112] Kostas Zoumpatianos, Stratos Idreos, and Themis Palpanas. Ads: the adaptive data series index. The VLDB Journal—The International Journal on Very Large Data Bases, 25(6):843–866, 2016.