Expert Systems With Applications 55 (2016) 452–460

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Ranking and significance of variable-length similarity-based time series motifs

Joan Serrà[a,b,*], Isabel Serra[c], Álvaro Corral[c], Josep Lluis Arcos[b]

[a] Telefónica Research, Barcelona, Spain
[b] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain
[c] Centre de Recerca Matemàtica, Bellaterra, Barcelona, Spain

ARTICLE INFO

Keywords:
Time series
Motif ranking
Distance modeling
Beta distribution

ABSTRACT

The detection of very similar patterns in a time series, commonly called motifs, has received continuous and increasing attention from diverse scientific communities. In particular, recent approaches for discovering similar motifs of different lengths have been proposed. In this work, we show that such variable-length similarity-based motifs cannot be directly compared, and hence ranked, by their normalized dissimilarities. Specifically, we find that length-normalized motif dissimilarities still have intrinsic dependencies on the motif length, and that lowest dissimilarities are particularly affected by this dependency. Moreover, we find that such dependencies are generally non-linear and change with the considered data set and dissimilarity measure. Based on these findings, we propose a solution to rank (previously obtained) motifs of different lengths and measure their significance. This solution relies on a compact but accurate model of the dissimilarity space, using a beta distribution with three parameters that depend on the motif length in a non-linear way. We believe the incomparability of variable-length dissimilarities could have an impact beyond the field of time series, and that similar modeling strategies as the one used here could be of help in a more broad context and in diverse application scenarios.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

With the generalized use of smartphones and the increasing adoption of wearable devices, the information sources available for decision support systems and expert systems has changed drastically. In particular, the application of such information sources to healthcare (Dolgin, 2014; Menshawya, Benharrefb, & Serhani, 2015) or living assistance to elder people (Chernbumroong, Cang, Atkins, & Yu, 2013) is becoming a fertile research area. One of the primary data types generated by the aforementioned devices are time series, and one of the first challenges to effectively process the huge amount of information they provide is the detection of repetitive patterns, commonly called motifs. Because expert systems have to deal with a diversity of activities to be monitored (e.g., the different living activities considered by Chernbumroong et al. (2013) or the detection of a variety of time-spanning events (Guralnik & Srivastana, 1999)), they have to consider and compare repeated patterns or motifs of different lengths. In this article, we uncover some problems that arise when comparing motifs of different lengths and propose a principled methodology to do so.

In the literature, two formal definitions of a time series motif coexist. The first one is based on the notion of frequency (Lin, Keogh, Lonardi, & Patel, 2002): a pattern is interesting if it has a significant amount of repetitions. The second one is based on the notion of similarity (Mueen, Keogh, Zhu, Cash, & Westover, 2009): a pattern is interesting if its occurrences are identical or too similar to happen at random. Both definitions are complementary, as a strikingly similar pattern does not necessarily need to be frequent, nor a frequent pattern does necessarily need to include extremely similar ones. Hence, algorithms exploiting both notions independently have received continuous and increasing attention (Bankó & Abonyi, 2015; Chiu, Keogh, & Lonardi, 2003; Mueen, 2013; Mueen et al., 2009; Tanaka, Iwamoto, & Uehara, 2005; Tang & Liao, 2008; Yingchareonthawornchai, Sivaraks, Rakthanmanon, & Ratanamahatana, 2013). Notice, however, that a notion of frequency necessarily implies a notion of similarity and vice versa, although these relationships may not be explicit nor straightforward to devise.

Under a frequency-based definition, the ranking of the motifs found in a time series is trivial. The most important motif is the

* Corresponding author at: Telefónica Research, Barcelona, Spain. Tel.: +34933653010.
E-mail addresses: joan.serra@telefonica.com (J. Serrà), iserra@crm.cat (I. Serra), acorral@crm.cat (Á. Corral), arcos@iiia.csic.es (J.L. Arcos).

http://dx.doi.org/10.1016/j.eswa.2016.02.026
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

one with the highest count, the second most important motif is the one with the second highest count, and so on. Moreover, we can assess the statistical significance of frequency-based motifs by comparing observed and expected counts under a null model reflecting some basic characteristics of the time series. This has been exploited by Castro and Azevedo (2011), who leverage work from the bioinformatics community to derive a motif's significance.

Using a similarity-based definition, motif ranking also looks straightforward. Given a single (usually pre-specified) motif length, the most important motif pair is the one with the lowest dissimilarity, the second most important pair is the one with the second lowest dissimilarity, and so on (equivalently for highest similarity). However, if we have motif pairs of different lengths, we cannot directly compare dissimilarities or distances, as these typically depend on the length of the given segments. In these cases, researchers rely on two different strategies. On the one hand, there is the option to compute a ranking for every motif length of interest, possibly removing covering motifs (e.g., Mueen, 2013). Consequently, we have as many orderings as lengths being considered, and the choice for the most important motif depends on the user. On the other hand, there is the possibility to normalize the dissimilarity measure by the length of the motif, or to use a measure that already incorporates some notion of normalization[1] (e.g., Yingchareonthawornchai et al., 2013). For instance, one can divide the Euclidean distance by the square root of the length, or consider the Pearson's correlation measure. In terms of motif significance, similarity-based approaches are much less developed than frequency-based ones. In fact, to the best of our knowledge, this topic has not been considered yet.

In this work, we show an important and overlooked aspect of variable-length similarity-based motifs: that they cannot be directly compared, and hence ranked, using common motif dissimilarity measures and their corresponding length normalization. Using a variety of statistical tools, we illustrate that normalized motif dissimilarities exhibit intrinsic dependencies with respect to the motif length, and that these particularly affect the lowest dissimilarities of each length. Moreover, we find that such dependencies are generally non-linear (they do not have a linear relationship with the length of the motif), and that they change with the considered measure and data set. These aspects are quantified using a combination of 8 common dissimilarity measures and 9 different publicly-available time series data sets. To further facilitate the assessment and reproducibility of our work, we make all results and code available online.

Given the aforementioned problems, and as a further contribution, we propose a solution to compare motifs of different lengths and, at the same time, derive a measure of their significance. The proposed solution consists of a compact model of the motif dissimilarity space, using a beta distribution whose parameters non-linearly depend on the length of the motif. We find this model leads to a reasonable fit for the majority of the considered lengths, measures, and data sets. Importantly, the cumulative distribution function (CDF) of the proposed model can wrap the motif dissimilarity function, hence directly yielding a p-value for each motif pair that can be used for ranking and significance assessment inside a given motif discovery algorithm.

The remainder of the article is structured as follows. Before delving into the description of our findings and the proposed methodology, we first present the considered data sets and dissimilarity measures, as well as our motif sampling strategy and all available reproducibility resources (Section 2). We then start by analyzing the problem of comparing motifs of different lengths (Section 3). Next, we introduce the proposed modeling strategy (Section 4). Finally, we conclude by summarizing our work and highlighting some future directions (Section 5).

2. Materials and methods

2.1. Time series data sets

To demonstrate that our results are not biased with regard to the data source, we consider 9 different publicly-available time series (Serrà & Arcos, 2016) of varying length, coming from distinct domains: (1) DowJones – the daily closing values of the Dow Jones industrial average (Williamson, 2012); (2) CarCount – the number of cars measured for the Glendale on ramp for the 101 North freeway in Los Angeles, CA, USA (Ihler, Hutchins, & Smyth, 2006); (3) Insect – the electrical penetration graph of a beet leafhopper[2] (Mueen et al., 2009); (4) EEG – a one hour electroencephalogram from a single channel in a sleeping patient[3] (Mueen et al., 2009); (5) FieldRecording – the spectral centroid of a field recording[4] (we used the mean of the stereo channels and the spectral centroid linear frequency plugin from Sonic Visualizer[5]); (6) Wind – the wind speed registered in the buoy of Rincon del San Jose[6], TX, USA. (7) Power – the electric power consumption of an individual household[7] (Bache & Lichman, 2013); (8) EOG – an electrooculogram tracking the eye movements of a sleeping patient[8] (Goldberger et al., 2000); and (9) RandomWalk – a random walk time series, artificially generated using $z_{i+1} = z_i + \eta$ and $z_1 = 0$, where η is a Gaussian random number with zero mean and unit variance.

2.2. Dissimilarity measurement

To demonstrate that our results are not biased with regard to the similarity measurement, we consider 8 different and commonly-used time series dissimilarity measures (see Serrà and Arcos, 2014, and references therein): (1) Euc – Euclidean distance normalized by the square root of the number of time series segment samples; (2) sqEuc – squared Euclidean distance normalized by the number of time series segment samples; (3) Corr – Pearson's correlation between time series segments; (4) Cos – cosine dissimilarity between segments; (5) DTW – dynamic time warping with path-accumulated normalization weights and a ±5% corridor window; (6) EDR – edit distance with real penalty, normalized by the alignment path length; (7) TWED – time-warped edit distance normalized by the alignment path length; and (8) MDL – minimum description length as in Rakthanmanon, Keogh, Lonardi, and Evans (2011), with an added constant to force $d \geq 0$. All dissimilarities were computed between z-normalized non-overlapping time series segments.

2.3. Motif sampling

We here employ the formal definition of similarity-based time series motifs by Mueen et al. (2009) and consider a random selection of possible motif candidates. A motif candidate is defined by

[1] All dissimilarities considered in this paper are normalized by the length of the motif (sometimes we will additionally employ the terms "normalized" or "length-normalized" to further clarify this aspect). The reader should not confuse these terms with the typical z-normalization between time series or other possible normalization strategies (see also Section 2.2).

[2] http://www.cs.ucr.edu/~mueen/MK.

[3] http://www.cs.ucr.edu/~mueen/OnlineMotif.

[4] http://www.freesound.org/people/JeffWojo/sounds/121250.

[5] http://www.sonicvisualiser.org.

[6] http://lighthouse.tamucc.edu/pq.

[7] http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption.

[8] http://www.cs.ucr.edu/~mueen/DAME.

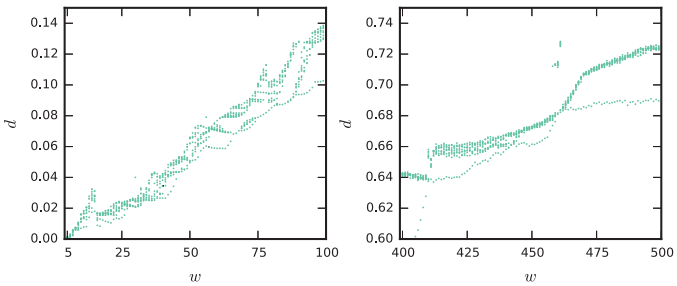

Fig. 1. Lowest 10 dissimilarities d for each segment length w considering all possible non-overlapping segments from a section of the EEG data set: $w \in [5, 100]$ (left) and $w \in [400, 500]$ (right). Importantly, notice that length-normalized Euclidean distance is used (see main text and also Section 2.2).

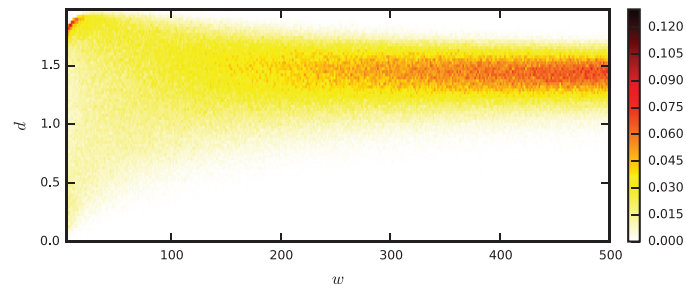

Fig. 2. Normalized histogram of length-normalized Euclidean distances using $n = 2000$ dissimilarity samples for each $w \in [5, 500]$ and the full EEG data set.

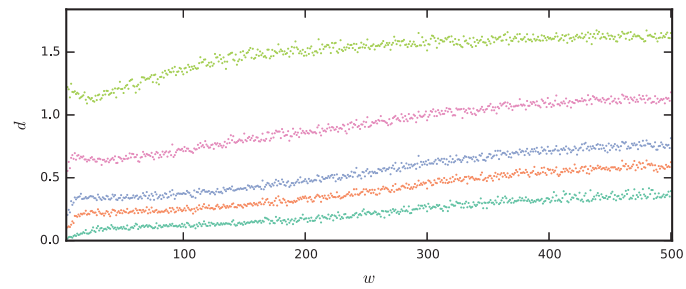

Fig. 3. Quantiles for a sample of length-normalized dissimilarities using dynamic time warping (DTW) and the WIND data set. From top to bottom, the quantiles correspond to 0.5, 0.25, 0.1, 0.05, and 0.01.

two starting points i and j and a motif length w. To obtain motif candidates, we sample the motif space. We take $n = 2000$ motif samples for each possible w uniformly at random, and explicitly avoid trivial matches corresponding to short-time delayed versions of the same pattern (Chiu et al., 2003). That is, given a motif length w, we randomly generate the start of the segments that will form the motif, $i, j \in [1, N - w]$, N being the time series length, such that $|i - j| > w$. If not stated otherwise, in the following experiments we consider $w_{\min} = 5$ and $w_{\max} = 500$, i.e., $w \in [5, 500]$. Thus, in total, every experiment is based on 2000×496 samples.

2.4. Additional results, code, and data availability

Apart from the main results presented here (for instance in Figs. 4 and 9), we make all raw results available at our web page[9]. Additional summary tables reporting specific values for each data set and measure combination and the code used to run the experiments are also available at the same web page. The data sets are all available from the original sources or mirrored[10] by Serrà and Arcos (2016).

3. Comparing motif dissimilarities

3.1. Motivating examples

To understand the issues that arise when comparing variable-length motif dissimilarities, we first take a look at some examples. Let's consider a randomly-chosen contiguous segment of 10,000 samples from the EEG data set (Section 2.1). We then compute the length-normalized Euclidean distance d (Section 2.2) between all possible non-overlapping pairs of segments of length $w \in [5, 100]$, and take the lowest 10 dissimilarities d for each w. What we observe is a clear trend of increasing d with w (Fig. 1, left). Given this trend, how can we automatically determine the most important motif using a similarity-based definition? To make it more explicit, let's assume that the best motif at length $w_1 = 30$ scores a length-normalized distance of $d_1 = 0.202$ and that the best motif at length $w_2 = 40$ scores a length-normalized distance $d_2 = 0.219$. Based on what we have seen (Fig. 1, left), which one should we prefer? Notice that, furthermore, both motifs could overlap. Would we prefer motif 2, an extension of motif 1, even if the length-normalized dissimilarity is not as low as the one of motif 1? How can we choose in an objective and informed way? This are the kind of situations this work deals with. However, we first need to demonstrate that such situation is systematically occurring, independently of the data source and the dissimilarity measurement.

Re-taking our motivating example (Fig. 1), we could argue that the observed trend is due to the short length of the segments ($w \in [5, 100]$). However, if we repeat the calculations for $w \in [400, 500]$, another trend appears (Fig. 1, right). Notice the change in the dissimilarity values, which is more than 4 times larger (Fig. 1, vertical axes). Such a difference is difficult to attribute to the effect of some characteristic timescales. Instead, it looks as a property resulting of the combination of both time series and dissimilarity measurement.

The aforementioned trends are clearly observable by a simple uniform random sampling of the motif dissimilarity space. If, for each $w \in [w_{\min}, w_{\max}]$, we select n non-overlapping segment pairs at random and compute their length-normalized Euclidean distance, we can reproduce the same phenomenon (Fig. 2). The plotted histogram gives us an indication that the empirical distribution of d changes with w. As w increases, the mode of the distribution seems to be more or less stable, but the tails (i.e., the non-central parts of the distribution) are visibly different, specially the lower one.

With further analysis, we confirm that the observed phenomenon is not unique of the EEG data set nor of the normalized Euclidean distance. In fact, if we consider other data sources and dissimilarities with their corresponding length normalization (Section 2.2), we can easily obtain more radical examples of the same phenomenon (see, for instance, Fig. 3). In this example, we can compute the statistical significance of the slopes of the plotted quantiles, obtained via ordinary least squares, for $w \in [300, 500]$. The highest p-value we obtain is $p = 1.93 \cdot 10^{-15}$, which corresponds to the slope of the median. Thus, we see that even the median can show a statistically significant trend for relatively large w. Fig. 3 also depicts a non-linear dependency of the computed quantiles with respect to w. We can also observe that such dependency is different than the one seen in Fig. 2. Clear differences are observable even if we fix the data source and change the

[9] http://www.iiia.csic.es/~jserra/motifranking.
[10] http://www.iiia.csic.es/~jserra/swarmmotif.

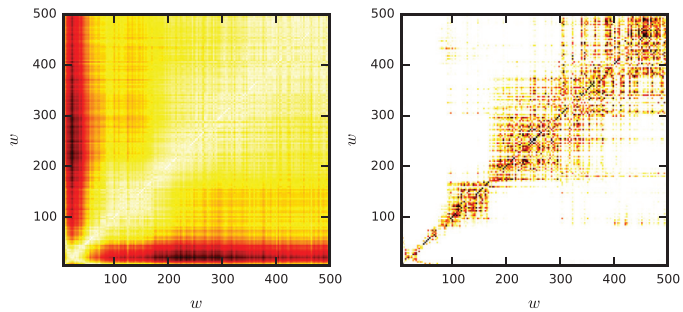

Fig. 4. Matrices comparing distribution differences between all possible pairwise comparisons in w: ϵ (left) and p_{KS} (right). The samples come from using correlation and the CarCount data set. The color code goes from 0 (white) to 0.1 and 1 (black), respectively.

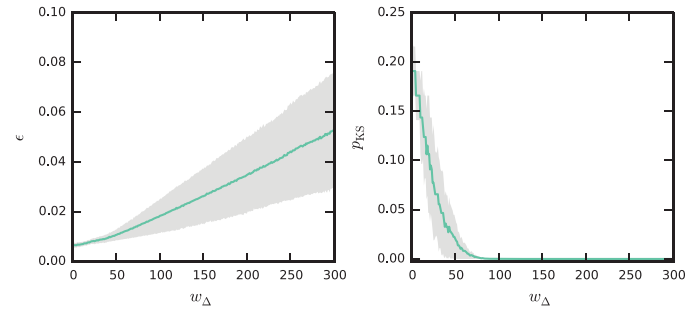

Fig. 5. Median (line) and median absolute deviation (patches) for ϵ (left) and p_{KS} (right), displayed as a function of w_Δ. Results computed from aggregating all data sets and measures (see text).

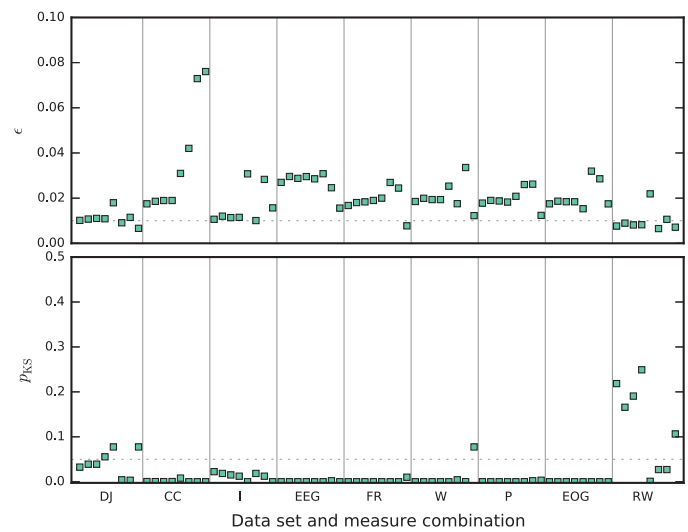

Fig. 6. Incomparability of distributions: median values for ϵ (top) and p_{KS} (bottom) for $w_\Delta = 100$ and every studied combination of data set and dissimilarity measure. There are a total of $9 \times 8 = 72$ such combinations (Sections 2.1 and 2.2). Vertical lines separate data set blocks: DowJones (DJ), CarCount (CC), Insect (I), EEG (EEG), FieldRecording (FR), Wind (W), Power (P), EOG (EOG), and RandomWalk (RW).

dissimilarity measurement. This also suggests that the observed behaviors are not due, to a large extent, to the effect of some characteristic timescales of the time series.

3.2. Quantitative evaluation

To quantify the incomparability of d with respect to w in a more formal and rigorous way, we employ two basic measures of the difference between distributions. First, we consider the global disagreement between empirical CDFs. We quantify it using

$$\epsilon = \frac{1}{k}\sum_{i=1}^{k} |F_i - G_i|,$$

where k is an arbitrarily chosen bin resolution and F and G correspond to the two empirical CDFs being compared. Notice that ϵ is conceptually similar to the total variation distance between probability distributions (Levin, Peres, & Wilmer, 2009). Nonetheless, since we use CDFs and take the average, $\epsilon \in [0, 1]$ gives a rapid and intuitive idea of the average difference between distributions. The bin resolution for all experiments reported here corresponds to $k = 100$ equally-spaced bins between the minimum and the maximum of each sample for each w.

As we are interested in the best motifs, we need to pay special attention to the lower tails of the dissimilarity distributions (i.e., the lowest sample values of each w). Hence, we consider a second measure based on just the lowest quartile of the empirical sample. Specifically, we consider the well-known Kolmogorov-Smirnov (KS) test (Massey, 1951) and its associated p-value, which we denote by p_{KS}. The KS test is a non-parametric test of the equality of continuous, one-dimensional probability distributions. It can be used to compare a sample with a reference probability distribution or to compare two samples. In our case, we compare the first quartile of the two samples to assess whether they significantly differ or not. The p_{KS} value thus denotes the probability of observing a difference equal to or more extreme that the actually observed among the 25% smallest sampled dissimilarities of the two distributions.

Computing ϵ and p_{KS} for all possible pairwise comparisons of samples in w yields two matrices that can be post-processed in order to aggregate the information for each data set and dissimilarity measure (Fig. 4). If, for a given data set and measure, we take statistics of the diagonals of these matrices, we obtain an assessment of the distribution differences as a function of $w_\Delta = |w_i - w_j|$, the absolute difference between two motif lengths w_i and w_j. Specifically, for a given w_Δ, we compute the median and the median absolute deviation of ϵ and p_{KS}. Aggregating these results for all possible combinations of data set and dissimilarity measure[11] gives us an idea of the expected differences when comparing two distributions separated by w_Δ (Fig. 5). For instance, if we compare a motif pair of length $w_i = 150$ with a motif pair of length $w_j = 190$ ($w_\Delta = 40$), we can expect an average CDF error $\epsilon \approx 0.01$ and a $p_{KS} \approx 0.03$ (Fig. 5). The former tells us that, on average, there will be a difference between CDFs of one per cent. The latter tells us that the tails of the distributions are hardly comparable, given that p_{KS} is systematically lower than the significance threshold of 0.05. Thus, in general, we see that comparing motifs whose lengths differ by more than 40 samples is hardly justifiable.

Let's take $w_\Delta = 100$ and analyze the results for individual combinations of data set and measure (Fig. 6). We observe that nearly all distribution differences ϵ are above 0.01 and that almost no combination passes the KS test at a significance level of $p_{KS} > 0.05$. However, there is one notable exception: the last 8 combinations, which correspond to the RandomWalk data (Fig. 6, right). Several dissimilarity measures on this data set achieve acceptable p_{KS} values while keeping $\epsilon \approx 0.01$. This is to be expected, and tells us that, for the case of artificially generated random Gaussian data (Section 2.1), length-normalized motif dissimilarities tend to comparable, even across very different lengths. Apart from the RandomWalk data, the first 8 combinations, which correspond to the

[11] The raw results can be found in the online results document (see Section 2.4). Please refer to it for further detail.

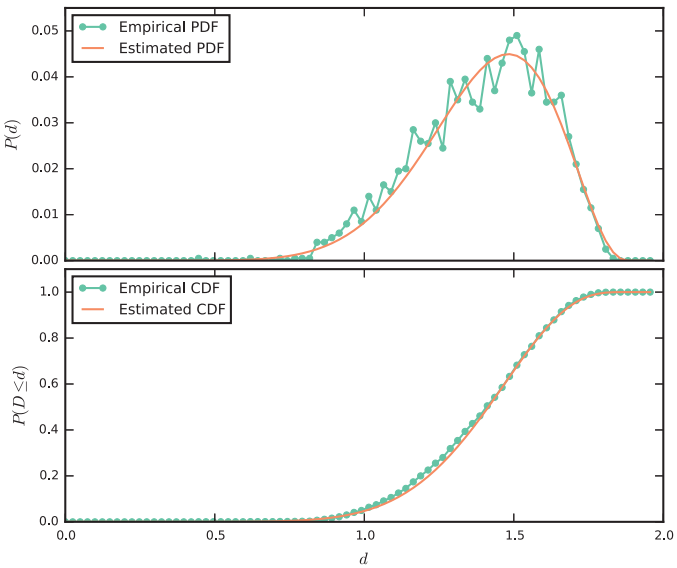

Fig. 7. Examples of an empirical PDF (top) and CDF (bottom) and their estimated fits. The sample comes from taking the length-normalized Euclidean distance and the INSECT data set with $w = 460$. Here, our procedure estimates $\alpha = 9.37$, $\beta = 3.03$, and $m = 1.93$.

DOWJONES data, seem to achieve larger p_{KS} values than the rest (Fig. 6, left). This is interesting, as in economics the random walk hypothesis has been used to model share prices and other factors for a long time (Malkiel, 1973).

4. Modeling motif dissimilarities

4.1. Main idea

To overcome the drawbacks described in the previous section, we now propose a procedure to model the dissimilarity space. Our aim is to produce a good and compact model of the empirical dissimilarity distributions for each w from a given combination of data set and dissimilarity measure. The main idea behind our modeling strategy is to achieve a 'normalization' of the dissimilarity space. We want to transform the dissimilarity space into a uniform probability space in which given motifs of different lengths can be compared in a meaningful way[12].

In Section 3, we have seen that, for two length-normalized dissimilarities d_i and d_j obtained from w_i and w_j, respectively, the relation $d_i < d_j$ does not necessarily imply that d_j should be ranked after d_i. Our observation is that, by considering an estimated CDF for each w, we can mix motifs of different lengths and meaningfully compare them. For instance, if $P_w(D \leq d)$ denotes the estimated CDF of the dissimilarities for a fixed w, then $P_{w_i}(D \leq d_i) < P_{w_j}(D \leq d_j)$ implies that d_j should indeed be ranked after d_i. We will further develop this idea, and specially the way to estimate P_w, in the next sections. In the end, we plan to substitute a given dissimilarity d by $d' = P_w(D \leq d)$.

4.2. Preliminary analysis

An illustration of the empirical probability distribution function (PDF) for dissimilarities with fixed w is shown in Fig. 7. Observe that a Gaussian model could initially appear as a reasonable model. However, this is not so. The Gaussian model is a good model for the central part of an empirical distribution, but it has the limitation that the kurtosis is always zero. Hence, in general, it does not correctly model the observed tails. Contrastingly, the similarity-based motif discovery task requires to get accurate estimations at the tails of such distributions. In fact, we are only interested in the smallest existing dissimilarities (the most relevant motifs). Thus, our modeling task requires a good model for the tails. In particular, it requires a model with a good fit in the left, lowest dissimilarity tail.

Extreme value theory (EVT) is focused on accurately modeling the tail of an empirical distribution (Beirlant, Goegebeur, Teugels, & Segers, 2004). In EVT, such tails are classified by a real number, called the tail index. In summary, there are two approaches to estimate the tail index: analyzing the empirical distribution of block minimums, and analyzing the empirical tail distribution. In any case, using models for tails requires the existence of an optimal threshold defining the starting point of the tail (Coles, 2001). In practice, one must verify that the sample size is large enough to accommodate a sub-sample of the tail of the distribution. In pre-analysis, we considered the Euclidean and DTW measures and confirmed that this property holds for all data sets and w. For each combination of measure, data, and w we tried, the estimation of the optimal threshold provided an estimation of the tail index inside the confidence interval for the tail index obtained with the analysis of block minimums (del Castillo & Serra, 2015). Thus, we found the considered data fulfilled the aforementioned requirement.

Obviously, the left tail distribution of the computed dissimilarities has a bounded range, since $d \geq 0$. That is called a short tail, and it corresponds to a negative value of the tail index. Therefore, the distributions considered as models for the lowest dissimilarities have to contain short tails. Since both tail distributions showed this behavior, we consider the simplest model to fit two-side short tails (Beirlant et al., 2004): the beta distribution. Besides the tails, we also observed that the behavior in the central part of the beta distribution was very similar to the behavior in the central part of most of the empirical distributions obtained for the considered cases. Thus, in addition to being a theoretically plausible model, the beta distribution was found to visually correspond to the empirical data.

4.3. Model fit

The beta distribution typically depends on two shape parameters, each of them corresponding to the tail index of each side. The extreme value for close-to-zero dissimilarity is zero, but in the case of the maximum, we have seen it depends on the original data set and w. Therefore, we consider the three-parameter beta distribution

$$P(d) = \frac{1}{mB(\alpha,\beta)}\left(\frac{d}{m}\right)^{\alpha-1}\left(1-\frac{d}{m}\right)^{\beta-1}, \tag{1}$$

where α, $\beta > 0$ are the so-called shape parameters, m is a scale parameter, and $B(\alpha, \beta)$ is the beta function. Eq. 1 is defined for $0 \leq d \leq m$. For values of d outside this range, $P(d) = 0$.

We start by fitting one beta distribution for each w. We do so by employing the maximum likelihood. Given n normalized dissimilarities $d = d_1, \ldots d_n$ computed from a uniform random sampling of all possible non-overlapping segments of length w (Section 2.3), we can calculate the log-likelihood

$$\begin{aligned}\ln(\mathcal{L}(\alpha,\beta,m|d)) = (\alpha-1)\sum_{i=1}^{n}\ln(d_i) \\ +(\beta-1)\sum_{i=1}^{n}\ln(m-d_i) - n\ln(B(\alpha,\beta)) \\ -n(\alpha+\beta-1)\ln(m).\end{aligned} \tag{2}$$

[12] Note that we do not claim that the combination of data set and dissimilarity measure yields a particular expression nor that the proposed model corresponds to such expression.

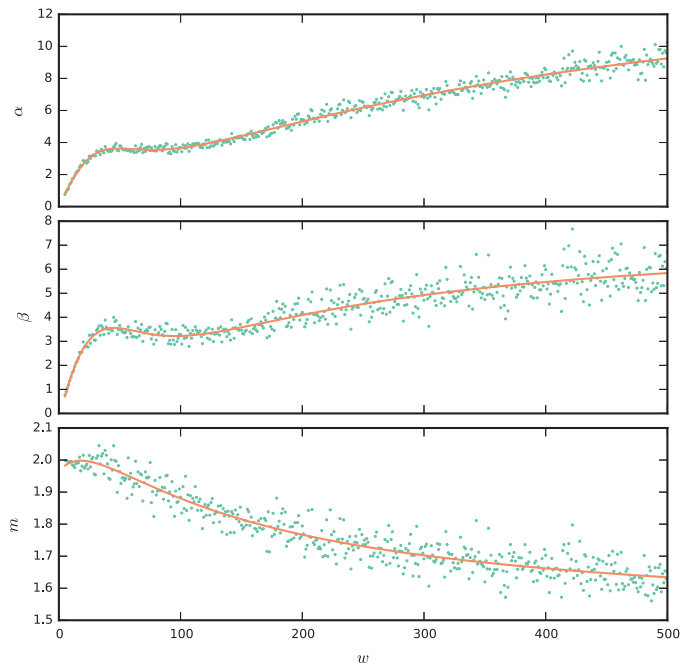

Fig. 8. Example of the estimated parameters α (top), β (middle), and m (bottom) for each w. The fitted rational functions are also displayed. The samples come from using the cosine dissimilarity and the Power data set.

From here, we have to find the values of α, β, and m that maximize Eq. 2. To do so, we choose a particle swarm optimizer (Poli, Kennedy, & Blackwell, 2007).

Particle swarm optimization (PSO) is a well-known population-based stochastic approach for solving continuous and discrete optimization problems. PSO makes few or no assumptions about the problem being optimized, does not require it to be differentiable, can search very large spaces of candidate solutions, and can be applied to problems that are irregular, incomplete, noisy, dynamic, etc. (see Poli et al., 2007; Parsopoulos and Vrahatis, 2010, and references therein). We here use the canonical PSO algorithm (Poli et al., 2007), with 25 particles and a local best configuration, and run 300 iterations. Further details can be found in the provided code (Section 2.4). The motivation for using PSO comes from our experience in optimization problems. However, we believe that more classical optimization algorithms would yield comparable, if not identical results. Essentially, any suitable optimization procedure available in typical scientific programming environments could do. The only constrains it needs to handle are α, $\beta > 0$ and $m > \max(d)$. To facilitate the search, we additionally force $m < 2.1\max(d)$.

If we repeat the previous procedure for all $w \in [w_{\min}, w_{\max}]$, we end with three series of parameters: one for α, one for β, and one for m (Fig. 8). This can represent a huge number of parameters for our model $(3 \times (w_{\max} - w_{\min} + 1))$. However, as we have seen in Section 3.2, close distributions with $w_\Delta < 40$ are objectively similar, and this similarity increases as w_Δ decreases. Because of this, the estimated parameters exhibit a continuity in w (Fig. 8). We can exploit this continuity to fulfill two desirable objectives at the same time: reducing the number of parameters of our model, and removing some of the potential noise introduced in the sampling and/or the fitting procedure. This brings us to the next important step.

Given the three parameter series for α, β, and m, we fit a curve to each of them by using rational functions, i.e., the ratio of two polynomial functions (Ghosh & Rao, 1996). A rational function model is a generalization of the polynomial model, as the former contains the latter as a subset. Rational function models provide several advantages over polynomial models while still having a moderately simple form (Ghosh & Rao, 1996). In particular, they are relatively easy to fit, take on an extremely wide range of shapes, and have very good interpolatory and extrapolatory properties. Thus, the three parameter beta distribution accounting for the full range of w becomes

$$P_w(d) = \frac{1}{m_w B(\alpha_w, \beta_w)} \left(\frac{d}{m_w}\right)^{\alpha_w - 1} \left(1 - \frac{d}{m_w}\right)^{\beta_w - 1},$$

where

$$\alpha_w = Q_\alpha(w)/R_\alpha(w), \tag{3}$$

$$\beta_w = Q_\beta(w)/R_\beta(w), \tag{4}$$

$$m_w = Q_m(w)/R_m(w), \tag{5}$$

and Q_z and R_z correspond to polynomials of degrees u_z and v_z, respectively, such that

$$Q_z(w) = \sum_{i=0}^{u_z} q_{z_i} w^i \tag{6}$$

and

$$R_z(w) = 1 + \sum_{i=1}^{v_z} r_{z_i} w^i. \tag{7}$$

To fit the rational functions, we employ the default implementation of the Levenberg-Marquardt algorithm (LMA; Gill & Murray, 1978) available in the Matlab's curve fitting toolbox[13]. However, as with the case of PSO, we believe that any other suitable curve fitting or optimization algorithm could be used with very similar or identical results. The motivation for using the LMA is its improved robustness over the typical Gauss–Newton algorithm (Gill & Murray, 1978). We recursively compute the fits for all pairwise combinations of $u_z = 1, 2, 3$ and $v_z = 0, 1, 2, 3$, and take the one that yields the lowest Akaike information criterion (Burnham & Anderson, 2002). For further details about this fitting procedure, we refer the interested reader to the provided code (Section 2.4).

The final model P_w is parameterized by the rational functions α_w, β_w, and m_w. Hence, it consists of $u_\alpha + 1$, v_α, $u_\beta + 1$, v_β, $u_m + 1$, and v_m coefficients. From the values of u_z and v_z considered above, we see that the total number of model coefficients ranges from 6 $(3 \times (2 + 0))$ to 21 $(3 \times (4 + 3))$. A model with 6 to 21 coefficients can be considered a compact model given the size and complexity of the dissimilarity spaces we are dealing with (Section 3), which comprise $w_{\max} - w_{\min} + 1$ different lengths or individual empirical distributions.

4.4. Model usage

As mentioned, our end goal is to 'normalize' the dissimilarity space with respect to variations in w. To do so, we just need to compute α_w, β_w, and m_w following Eqs. 3–7 and consider the CDF of the proposed model,

$$P_w(D \le d) = \frac{B\left(\frac{d}{m_w}; \alpha_w, \beta_w\right)}{B(\alpha_w, \beta_w)}, \tag{8}$$

where $B(x; \alpha, \beta)$ is the incomplete beta function, a generalization of the beta function. The incomplete beta function can be efficiently calculated using functions that are commonly included in

[13] http://www.mathworks.com/products/curvefitting.

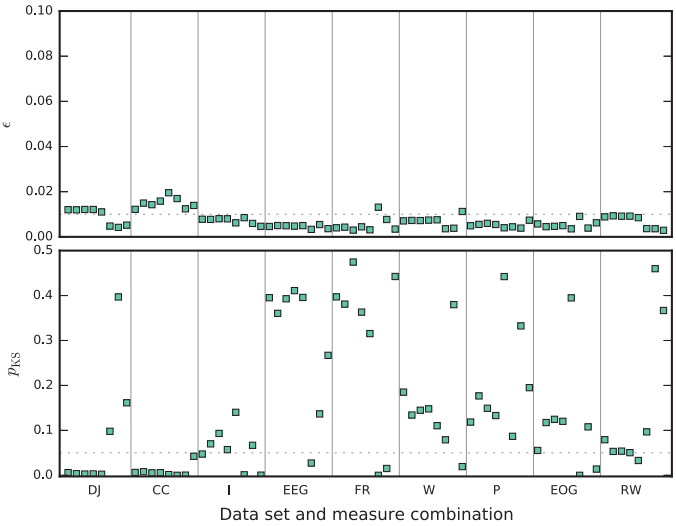

Fig. 9. Model accuracy: median values for ϵ (top) and p_{KS} (bottom) for $w_\Delta = 100$ and every studied combination of data set and dissimilarity measure. There are a total of $9 \times 8 = 72$ such combinations (Sections 2.1 and 2.2). Vertical lines separate data set blocks: DowJones (DJ), CarCount (CC), Insect (I), EEG (EEG), FieldRecording (FR), Wind (W), Power (P), EOG (EOG), and RandomWalk (RW).

spreadsheet or programming systems (Press, Teukolsky, Vetterling, & Flannery, 2007).

Because $P_w(D \leq d)$ is only defined for $0 \leq d \leq m_w$, we propose the new dissimilarity measure d' for ranking and comparing motifs of different lengths under the same conditions:

$$d' = \begin{cases} 0 & \text{if } d < 0, \\ P_w(D \leq d) & \text{if } 0 \leq d \leq m_w, \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

Note that the case of $d < 0$ is impossible for most dissimilarity measures since typically $d \geq 0$. Moreover, if d took negative values, we could always apply any suitable transformation to make it strictly positive (e.g., e^d). The case of $d > m_w$ might happen in practice, as our estimation of the maximum d for each w could be inaccurate or underestimating the true maximum (if this exists). However, this latter case is of little interest in motif discovery, as it corresponds to extremely dissimilar segment pairs. Thus, without compromising the accuracy of the task, we can tolerate some error and consider these motifs to form a tie in the last positions of the ranking ($d' = 1$ for all of them).

The new dissimilarity measure d' is a wrapper of d, and can be inserted in any motif discovery algorithm once P_w has been estimated (offline or prior to the execution of the algorithm). Furthermore, d' is easily interpretable, as it corresponds to the probability of seeing a dissimilarity equal to or smaller than d. This gives us a raw idea of the significance of the motif with respect to the dissimilarity space. In the next two paragraphs, we exemplify and give more detail on the method's usage.

Suppose that, as practitioners, we are interested in finding similar patterns or motifs in the EEG time series of Figs. 2 and 10 and that we are interested in the temporal range that goes from $w \in [100, 200]$ samples. When we consider the Euclidean distance and extract motifs at the considered lengths, we can easily find repetitions at $w = 100$ whose distance is around 0.5 and repetitions at $w = 200$ whose distance is around 1.2, and we do not know which motif to prefer (this is analogous to our motivating example of Section 3.1). Therefore we run the our algorithm, sample the Euclidean dissimilarity space of $w \in [100, 200]$, and fit a beta distribution (Eq. 1) for every w following the proposed methodology.

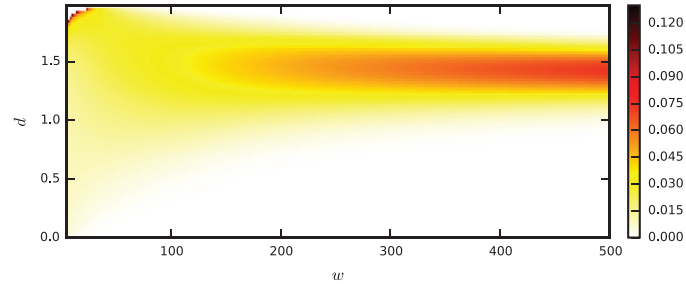

Fig. 10. Fitted model for length-normalized Euclidean distances sampled from the EEG data set (compare with Fig. 2).

We end up with a model of the dissimilarity space that depends only on w (Eqs. 8 and 9). All parameters m_w, α_w, and β_w have been estimated. In the case we had motif candidates for all lengths, we only need to recalculate the new dissimilarity measure d' using the previously obtained dissimilarities d at every w. In the case we did not have a list of candidate motifs for each length, we may include Eqs. 8 and 9 into the dissimilarity computation function and re-run the motif finding algorithm. In both cases, since the value of d' is bounded between 0 and 1, we as practitioners have a clear idea of the probability of observing a dissimilarity equal to or smaller than d in that w.

4.5. Model validation

To measure the quality of the model fit P_w, we resort to the measures introduced in Section 3.2: ϵ, the global disagreement between empirical CDFs, and p_{KS}, the p-value of the KS test on the lowest quartile of the samples. The only difference is that, here, the p_{KS} value is not the result of a two-sample test, but the result of a goodness of fit test for the plausibility of our proposed model given the available samples (we adapt the bootstrap generative procedure described by Clauset, Shalizi, and Newman (2009) for power-law models to the current model). If we compute ϵ and p_{KS} for all considered measures and data sets, we see that the fitted models generally provide a good agreement with the data (Fig. 9). In general, ϵ is never above 0.02 and rarely above 0.01. The p_{KS} value is often above 0.05, what indicates that we cannot reject the null hypothesis of the tail samples coming from the fitted distribution tail. The DowJones and the CarCount data sets achieve relatively low p_{KS} values, but ϵ is always below 0.02. The median and median absolute deviation for the aggregation of all combinations[14] are $\epsilon = 0.006 \pm 0.002$ and $p_{KS} = 0.10 \pm 0.09$. Overall, we can consider a reasonably good fit is reached for the majority of cases. We can visually confirm the agreement of our model and the empirical data by comparing the resultant PDFs against the empirical histograms obtained for each combination (compare, for instance, the obtained model in Fig. 10 with our motivating example of Fig. 2).

5. Conclusion

In recent years, expert systems built around time series-based methods have been enthusiastically adopted in engineering applications, thanks to their ease of use and effectiveness (Bankó & Abonyi, 2015). One of the first challenges of an expert system dealing with time series is recognizing repeated events or sequences, here called motifs. And a great part of this challenge implies working at different temporal resolutions or motif lengths, which arise naturally because of different interest scales (e.g., Guralnik & Srivastana, 1999; Chernbumroong et al., 2013) or because of some

[14] The raw results are available online (see Section 2.4).

concept drift (e.g., Abad, Gomes, & Menasalvas, 2016). Hence, comparing and ranking motifs of different lengths is becoming a primary task to be solved by current and future expert systems.

The main contribution of the present work is to show that time series motif dissimilarities of different lengths are not directly comparable, and thus cannot be ranked. Through both motivating examples and formal quantitative analysis, we have shown (1) that length-normalized motif dissimilarities have non-linear dependencies with the motif length, (2) that these dependencies change with the data set and the dissimilarity measure, and (3) that they particularly affect the lowest dissimilarities, which are precisely the focus of interest of any similarity-based motif discovery algorithm. Another contribution of the present work is a solution to tackle the aforementioned problems. This consists of a compact model of the dissimilarity space that allows comparing motifs of different lengths and assessing their significance with respect to the overall dissimilarity distribution. Such model is motivated by extreme value theory, and is based on a three-parameter beta distribution. We propose a procedure to fit those three parameters while taking into account the local continuity and the non-linearity of the motif dissimilarity space.

In this article, we have not explicitly dealt with motif pairs consisting of segments of different length. Instead, we have assumed the same length for the pair of segments forming a motif pair. This assumption is well motivated, as practically all existing motif discovery algorithms operate under such constraint (e.g., Lin et al., 2002; Chiu et al., 2003; Tanaka et al., 2005; Mueen et al., 2009; Castro and Azevedo, 2011; Mueen, 2013; Yingchareonthawornchai et al., 2013). It is also motivated for the case where we are interested in pairs of segments of different length, as the most common way to compute the dissimilarity between such segments is by resampling them to have the same length. That is extensively used for Euclidean distance or correlation (Yankov, Keogh, Medina, Chiu, & Zordan, 2007). For measures explicitly handling segments of different length, this is also one of the most recommended practices. For instance, it has been shown that a brute-force up-sampling to the largest segment length yields equivalent or slightly better results for classification tasks using DTW (Ratanamahatana & Keogh, 2004). Nonetheless, we acknowledge that the proposed approach could have some limitations in certain domains or applications where segment up-sampling was not appropriate. In preliminary analysis, we performed a number of experiments with DTW, ERD, and TWED dissimilarities while considering segments of different lengths. The results showed that a similar situation as with same length segments was taking place. We believe that a solution coming from the techniques exposed in the present paper could also deal with the extended case of different lengths. However, we leave the rigorous study of such solution for future work.

It is difficult to assess the potential impact of the present findings in other contexts. However, we have the impression that a similar phenomenon could happen when comparing feature vectors or quantitative descriptions of different sizes, even if these are not time series or segments. It would be interesting to analyze what happens with clustering or classification tasks with variable-length instances, and in particular with clustering or classification approaches based on dissimilarity measurements. The scarce literature on the topic we have found typically relies domain-specific knowledge (e.g., McHardy, García Martín, Tsirigos, Hugenholtz, and Rigoutsos, 2007) or makes a number of assumptions on the nature of the data (e.g., Porikli, 2004). The model and the methodology proposed here are domain-agnostic and make very few assumptions. Thus, we believe they could be good candidates to be considered in situations or applications where variable-length instance similarities need to be compared.

Acknowledgments

We would like to thank all the people who contributed the data sets used in this study. This research has been partially funded by Generalitat de Catalunya, the Spanish Government, CSIC, and the Collaborative Mathematics Project from La Caixa Foundation: 2014-SGR-118 (JS, JLA), 2014-SGR-1307 (AC, IS), FIS2012-31324 (AC), MTM2012-31118 (IS), NASAID CSIC Intramural 201550E022 (JLA), and TIN2012-38450-C03-03 (JS, JLA).

References

Abad, M. A., Gomes, J. B., & Menasalvas, E. (2016). Predicting recurring concepts on data-streams by means of a meta-model and a fuzzy similarity function. *Expert Systems with Applications, 46*, 87–105.

Bache, K., & Lichman, M. (2013). The UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

Bankó, Z., & Abonyi, J. (2015). Mixed dissimilarity measure for piecewise linear approximation based time series applications. *Expert Systems with Applications, 42*(21), 7664–7675.

Beirlant, J., Goegebeur, Y., Teugels, J., & Segers, J. (2004). *Statistics of extremes. Theory and applications*. Hoboken, USA: John Wiley & Sons.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: aA practical information-theoretic approach* (2nd ed.). Berlin, Germany: Springer.

Castro, N., & Azevedo, P. (2011). Time series motifs statistical significance. In *Proceedings of the SIAM international conference on data mining (SDM)* (pp. 687–698).

Chernbumroong, S., Cang, S., Atkins, A., & Yu, H. (2013). Elderly activities recognition and classification for applications in assisted living. *Expert Systems with Applications, 40*(5), 1662–1674.

Chiu, B., Keogh, E., & Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 493–498).

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*(4), 661–703.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Berlin, Germany: Springer.

del Castillo, J., & Serra, I. (2015). Likelihood inference for generalized Pareto distribution. *Computational Statistics and Data Analysis, 83*, 116–128.

Dolgin, E. (2014). Technology: dressed to detect. *Nature, 7508*, S16–S17.

Ghosh, S., & Rao, C. R. (1996). *Handbook of statistics 13: Design and analysis of experiments*. Amsterdam, The Netherlands: Elsevier.

Gill, P. E., & Murray, W. (1978). Algorithms for the solution of the nonlinear least-squares problem. *SIAM Journal on Numerical Analysis, 15*(5), 977–992.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220.

Guralnik, V., & Srivastana, J. (1999). Event detection from time series data. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 33–42).

Ihler, A., Hutchins, J., & Smyth, P. (2006). Adaptive event detection with time-varying Poisson processes. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 207–216).

Levin, D. A., Peres, Y., & Wilmer, E. L. (2009). *Markov chains and mixing times*. Providence, USA: American Mathematical Society.

Lin, J., Keogh, E., Lonardi, S., & Patel, P. (2002). Finding motifs in time series. In *Proceedings of the workshop on temporal data mining* (pp. 53–56).

Malkiel, B. G. (1973). *A random walk down Wall Street* (9th ed.). New York, USA: W. W. Norton & Company.

Massey, F. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46*(253), 68–78.

McHardy, A. C., García Martín, H., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods, 4*, 63–72.

Menshawya, M., Benharrefb, A., & Serhani, M. (2015). An automatic mobile-health based approach for EEG epileptic seizures detection. *Expert Systems with Applications, 42*(20), 7157–7174.

Mueen, A. (2013). Enumeration of time series motifs of all lengths. In *Proceedings of the IEEE international conference on data mining (ICDM)* (pp. 547–556).

Mueen, A., Keogh, E., Zhu, Q., Cash, S., & Westover, B. (2009). Exact discovery of time series motifs. In *Proceedings of the SIAM international conference on data mining (SDM)* (pp. 473–484).

Parsopoulos, K. E., & Vrahatis, M. N. (2010). *Particle swarm optimization and intelligence: Advances and applications*. Hershey, USA: IGI Global.

Poli, R., Kennedy, J., & Blackwell, T. M. (2007). Particle swarm optimization. *Swarm Intelligence, 1*(1), 33–57.

Porikli, F. (2004). Clustering variable length sequences by eigenvector decomposition using HMM. *Technical Report*. Mitsubishi Electric Research Laboratories.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (2nd ed.). New York, USA: Cambridge University Press.

Rakthanmanon, T., Keogh, E., Lonardi, S., & Evans, S. (2011). Time series epenthesis: clustering time series streams requires ignoring some data. In *Proceedings of the IEEE international conference on data mining (ICDM)* (pp. 547–556).

Ratanamahatana, C. A., & Keogh, E. (2004). Everything you know about dynamic time warping is wrong. In *Proceedings of ACM SIGKDD workshop on mining temporal and sequential data* (pp. 22–25).

Serrà, J., & Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems, 67*, 305–314.

Serrà, J., & Arcos, J. L. (2016). Particle swarm optimization for time series motif discovery. *Knowledge-Based Systems, 92*, 127–137.

Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time-series motif from multi-dimensional data based on MDL principle. *Machine Learning, 58*, 269–300.

Tang, H., & Liao, S. S. (2008). Discovering original motifs with different lengths from time series. *Knowledge-Based Systems, 21*, 666–671.

Williamson, S. H. (2012). Daily closing value of the Dow Jones average, 1885 to present. URL http://www.measuringworth.com/datasets/DJA/index.php.

Yankov, D., Keogh, E., Medina, J., Chiu, B., & Zordan, V. (2007). Detecting time series motifs under uniform scaling. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 844–853).

Yingchareonthawornchai, S., Sivaraks, H., Rakthanmanon, T., & Ratanamahatana, C. A. (2013). Efficient proper length time series motif discovery. In *Proceedings of the IEEE international conference on data mining (ICDM)* (pp. 1265–1270).