Annals of Mathematics and Artificial Intelligence (2020) 88:313–346
https://doi.org/10.1007/s10472-019-09682-2

Sufficient conditions for the existence of a sample mean of time series under dynamic time warping

Brijnesh Jain[1] · David Schultz[1]

Published online: 10 January 2020

Abstract

Time series averaging is an important subroutine for several time series data mining tasks. The most successful approaches formulate the problem of time series averaging as an optimization problem based on the dynamic time warping (DTW) distance. The existence of an optimal solution, called sample mean, is an open problem for more than four decades. Its existence is a necessary prerequisite to formulate exact algorithms, to derive complexity results, and to study statistical consistency. In this article, we propose sufficient conditions for the existence of a sample mean. A key result for deriving the proposed sufficient conditions is the Reduction Theorem that provides an upper bound for the minimum length of a sample mean.

Keywords Time series · Dynamic time warping · Fréchet function · Sample mean

Mathematics Subject Classification (2010) 49J52 · 68T99

1 Introduction

Time series such as stock prices, climate data, energy consumptions, sales, and biomedical measurements are sequences of time-dependent observations that can vary in temporal dynamics. Here, variation in temporal dynamics encompasses variations in length and speed as well as shifts in phase. For example, the same word can be uttered with different speaking speeds. Similarly, monthly temperature or precipitation extremes of certain regions can differ in duration and may occur out of phase for a period of a few weeks.

To account for such variations in proximity-based time series mining, the *dynamic time warping* (DTW) distance [27] is often the preferred choice of proximity measure [1, 3, 4]. The basic idea of the DTW distance is to warp both time series non-linearly in the

✉ Brijnesh Jain
brijnesh.jain@dai-labor.de

David Schultz
david.schultz@dai-labor.de

[1] Distributed Artificial Intelligence Laboratory, TU Berlin, Berlin, Germany

time dimension such that some cost function on the warped time series is minimized. Since different cost functions can be used, the DTW distance comprises a whole family of distances rather than a single distance function.

One important problem in DTW-based time series mining is time series averaging. The problem of time series averaging consists in finding a typical representative that summarizes a sample of time series. Time series averaging is an important subroutine to improve nearest neighbor classifiers [19, 25, 26, 30], to accelerate similarity search [32], and to generalize machine learning methods to DTW spaces such as self-organizing maps [22], learning vector quantization [19, 30], and k-means clustering [15, 25, 26, 29].

The different averaging techniques have in common that they first align the time series with respect to a DTW distance and then synthesize the aligned time series to an average. The most successful approaches pose time series averaging as an optimization problem based on the following property of the arithmetic mean in Euclidean spaces: Given a sample of points, the arithmetic mean is the unique point that minimizes the sum of squared Euclidean distances from the sample points [12, Section III.A]. As suggested by Fréchet [13], the optimization-based property of the arithmetic mean can be adopted to distance spaces for which a pairwise addition is not well-defined. In this sense, state-of-the-art approaches of time series averaging minimize the sum of squared DTW distances from the given sample time series [15, 23, 24, 28].

After more than four decades of devising and applying heuristics, the existence of a sample mean is still an unsolved problem. During this period, numerous algorithms have been proposed without any evidence whether the improved results follow a phantom or indeed approximate an existing solution [2, 15, 23, 24, 26, 28, 33]. The concept of time series average has no precise meaning. Virtually, any time series returned by an algorithm that was accepted as an averaging procedure was qualified to be an average of a sample of time series. Such an understanding of time series average lead to misconceptions (see [20] for a discussion). As discussed by Brill et al. [6], some proposed solutions have been falsely claimed as optimal and (problematic) claims on the computational complexity of the sample mean problem have been stated without having any evidence of whether the problem is solvable at all. Thus, showing the existence of a sample mean is important for several reasons: In optimization theory, existence of a solution is one of the most basic questions. It gives the notion of time series average as used in current applications a precise meaning. It is a necessary prerequisite to formulate exact algorithms [6] and to determine the computational complexity of the sample mean problem [7]. In statistical inference, existence of a sample mean is necessary to investigate whether the sample means are consistent estimators of the population means (expectations). In applications, consistency justifies the common practice to draw a finite but sufficiently large number of sample time series, because we have high confidence that nothing unexpected will happen when sampling further time series.

In this article, we generalize the concept of sample mean of time series and present sufficient conditions for its existence. Generalizations include a broader class of objective functions than the sum of squared DTW distances and impose no restrictions on the elements of the time series. Their elements can be, for example, real values, feature vectors, symbols, trees, graphs, and mixtures thereof. Sufficient conditions of existence for the generalized sample mean refer to two types of problems reported in the literature: Given a sample of time series of (possibly) varying length, find a sample mean of pre-specified length [15, 24, 28] and find a sample mean of arbitrary but finite length [6, 23].

The rest of this paper is structured as follows: We conclude the introduction with a technical presentation of the main contributions. Section 2 states the main results of this contribution and Section 3 proofs the results of Section 2. Finally, Section 4 concludes with a summary of the main findings and an outlook for further research.

Technical overview of the results The state-of-the-art in time series averaging minimizes the Fréchet function

$$F : \mathbb{R}^m \rightarrow \mathbb{R}, \quad x \mapsto \sum_{k=1}^{N} \delta\left(x, x^{(k)}\right)^2, \tag{1}$$

where $\delta(x, x')$ is the DTW distance and $x^{(1)}, \dots, x^{(N)}$ are univariate time series with real-valued elements. Though the length of the sample time series $x^{(k)}$ may vary, the search space $\mathbb{R}^m$ consists of time series of a pre-specified but fixed length m. A sample mean (if exists) is a global minimizer of the Fréchet function $F(x)$. Its existence has neither been proved nor challenged.

In this article, we derive sufficient conditions of existence for the sample mean in a more general setting than described in the previous paragraph. The first generalization concerns the domain of time series to be averaged. We assume that the elements of the time series are not restricted to real valued signals but can take values from any set. For this, we need to adopt the DTW distance accordingly. As second generalization, we consider Fréchet functions of the form

$$F : \mathcal{U} \rightarrow \mathbb{R}, \quad x \mapsto \sum_{k=1}^{N} h_k\left(\delta\left(x, x^{(k)}\right)\right),$$

where $\mathcal{U}$ is the search space and $h_k : \mathbb{R} \rightarrow \mathbb{R}$ is the loss function of the k-th sample time series $x^{(k)}$. Common examples of loss functions are the identity loss $h_k(u) = u$ and the squared loss $h_k(u) = u^2$ for all k. By using loss functions of the form $h_k(u) = w_k \cdot u^2$ with $w_k \geq 0$, we obtain weighted sample means [6, 8]. We recover the Fréchet function defined in (1) by setting $\mathcal{U} = \mathbb{R}^m$ and $h_k(u) = u^2$.

We consider two forms of sample means: (i) unrestricted form: the search space $\mathcal{U}$ is the set of all time series of finite length and (ii) restricted form: the search space $\mathcal{U}$ is the subset of time series of length m. Note that the restricted form only restricts the time series from the search space $\mathcal{U}$, whereas the sample time series $x^{(k)}$ can be of any length.

Existence of a sample mean in both forms depends on the loss functions and on the particular choice of DTW distance. We show that common loss functions and DTW distances satisfy the proposed sufficient conditions to guarantee the existence of a sample mean.

The main challenge of this contribution is the existence proof of the unrestricted sample mean. We first prove the existence of a restricted sample mean for every length $m \in \mathbb{N}$. As explained in Section 2.5, it is not self-evident that existence of restricted sample means for every length m implies existence of an unrestricted sample mean. The key result to prove the existence of an unrestricted sample mean is the Reduction Theorem (Theorem 10). This theorem presents a sample-dependent bound ρ with the following property: For every time series x whose length exceeds the bound ρ there is a shorter time series x' such that $F(x') \leq F(x)$. In other words, we can reduce a very long time series x to a shorter time series x' without increasing the Fréchet variation. We obtain the existence proof for the unrestricted problem by applying the Reduction Theorem to restricted sample means. To prove the Reduction Theorem, we cast sample means of time series into a graph-theoretic framework.

2 Existence of a sample mean via the reduction theorem

This section first introduces the DTW distance and Fréchet functions. Then, we present the Reduction Theorem and its implications. Finally, sufficient conditions of existence of a sample mean are proposed. All proofs are delegated to Section 3.

Preliminaries We write $\mathbb{R}_{\geq 0}$ for the set of non-negative reals, $\mathbb{N}$ for the set of positive integers, and $[n]$ to denote the set $\{1, \ldots, n\}$ for a given $n \in \mathbb{N}$. Let $N \in \mathbb{N}$. Then $\mathcal{S}^N = \mathcal{S} \times \cdots \times \mathcal{S}$ is the N-fold Cartesian product of the set $\mathcal{S}$.

A norm on $\mathbb{R}^q$ is a function $\|\cdot\| : \mathbb{R}^q \rightarrow \mathbb{R}_{\geq 0}$ with the following properties for all $x, y \in \mathbb{R}^q$ and all $a \in \mathbb{R}$:

1. $\|x\| = 0 \;\Rightarrow\; x = 0$ (positive definite)
2. $\|ax\| = |a| \, \|x\|$ (absolutely homogeneous)
3. $\|x + y\| \leq \|x\| + \|y\|$ (subadditive)

A metric on a set $\mathcal{S}$ is a function $d : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}_{\geq 0}$ such that the following conditions are satisfied for all $x, y, z \in \mathcal{S}$:

1. $d(x, y) \geq 0$ (non-negativity)
2. $d(x, y) = 0 \;\Leftrightarrow\; x = y$ (identity of indiscernibles)
3. $d(x, y) = d(y, x)$ (symmetry)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

Note that every norm induces a metric but not every metric is a norm.

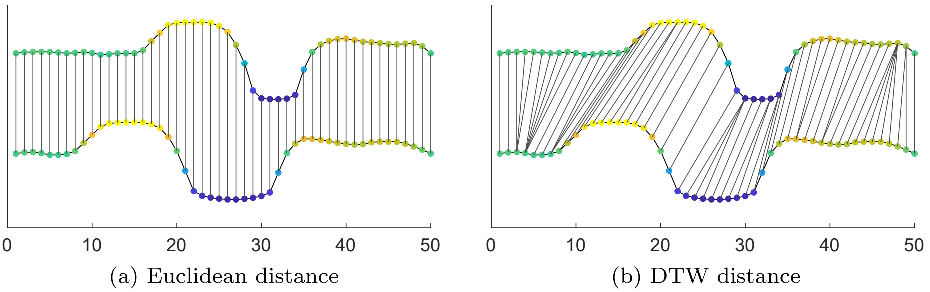
2.1 The dynamic time warping distance

Different time series representing the same concept may often vary in temporal dynamics. Such variations include variation in length and speed as well as shifts in phase. As illustrated in Fig. 1, lockstep measures such as the Euclidean distance fail to account for such variations with potential adverse effects in time series mining [4, 9] and also in time series averaging, as we will see later. To alleviate the limitations of lockstep measures, various elastic measures have been proposed [1]. The most common and widely applied elastic measure is the DTW distance.

We consider time series that can take any values. For this, we assume that $\mathcal{A}$ is a non-empty set called *attribute set*. An $\mathcal{A}$-valued *time series* x of *length* $\ell(x) = m$ is a sequence $x = (x_1, \ldots, x_m)$ consisting of *elements* $x_i \in \mathcal{A}$ for every *time point* $i \in [m]$. We denote the set of all $\mathcal{A}$-valued time series of length $n \in \mathbb{N}$ by $\mathcal{T}_n$ and the set of all $\mathcal{A}$-valued time series of finite length by $\mathcal{T} = \bigcup_{n \in \mathbb{N}} \mathcal{T}_n$. Unless otherwise stated, we assume that the attribute set $\mathcal{A}$ is fixed and briefly refer to $\mathcal{A}$-valued time series as time series.

Since we do not impose restrictions on the attribute set, the above definition of time series covers a broad range of sequential data structures. For example, to represent real-valued univariate and multivariate time series, we use the attribute set $\mathcal{A} = \mathbb{R}$ and $\mathcal{A} = \mathbb{R}^q$, respectively. For text strings and biological sequences, the set $\mathcal{A}$ is an alphabet consisting of a finite set of symbols. Further examples are time series with images or graphs as elements.

The DTW distance is based on two main components: (i) warping paths and (ii) a cost function that measures the discrepancy between two time series along a warping path. We first introduce warping paths.



(a) Euclidean distance

(b) DTW distance

Fig. 1 Euclidean distance (**a**) vs. DTW distance (**b**) of two time series. The time series are shown as sequences of colored points, where different colors represent different values. The time series have been relocated vertically for better visualization. Both time series are similar in shape but vary in speed and are shifted in phase. For example, the upper time series passes through the blue valley with higher speed than the lower one. In addition, the peak-valley phases of both time series are shifted in time. The thin gray lines indicate alignments between points. Such alignments are time preserving (the gray lines never cross). A distance is large/small if many points with large/small differences in their values are aligned. The Euclidean distance regards time series as vectors and aligns points whose time indices coincide (lockstep property). The lockstep property fails to align the peak-valley shape of both time series. Consequently, many points with different colors are aligned resulting in a large Euclidean distance. The DTW distance aims at aligning points with similar values. The peak-valley shapes are perfectly aligned. Thus, the DTW distance aligns many points of the same color and is therefore low

Definition 1 Let $m, n \in \mathbb{N}$. A *warping path* of order $m \times n$ is a sequence $p = (p_1, \dots, p_L)$ of L points (tuples) $p_l \in [m] \times [n]$ such that

1. $p_1 = (1, 1)$ and $p_L = (m, n)$ (boundary conditions)
2. $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$ for all $l \in [L - 1]$ (step condition)

We denote the set of all warping paths of order $m \times n$ by $\mathcal{P}_{m,n}$ and write $p_l \in p$ if p_l occurs as a point in warping path p. A warping path of order $m \times n$ can be thought of as a path in a $m \times n$ grid, where rows are ordered top-down and columns are ordered left-right. The boundary conditions demand that the path starts at the upper left corner and ends in the lower right corner of the grid. The step condition demands that a transition from on point to the next point moves a unit right, down, or diagonal.

Next, we define a cost function to measure the discrepancy between two time series along a warping path. A warping path $p = (p_1, \dots, p_L) \in \mathcal{P}_{m,n}$ defines an alignment between time series $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$. Every point $(i, j) \in p$ aligns element x_i to element y_j. The *cost* of aligning time series x and y along warping path p is defined by

$$c_p(x, y) = \sum_{(i,j) \in p} d\left(x_i, y_j\right),$$

where $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is a *local distance function* on $\mathcal{A}$. We demand that the local distance d is a semi-metric. A semi-metric satisfies the first three conditions of a metric but not necessarily the triangle inequality. As with the attribute set $\mathcal{A}$, we assume that the local distance d is specified without further mention. Finally, we define the DTW distance on time series.

Definition 2 Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a monotonous function. The *DTW distance* is a function $\delta : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_{\geq 0}$ defined by

$$\delta(x, y) = \min \left\{ f\left(c_p(x, y)\right) \,:\, p \in \mathcal{P}_{\ell(x),\ell(y)} \right\}.$$

for all $x, y \in \mathcal{T}$.

Of particular interest are warping paths with minimum cost. An *optimal warping path* of time series $x, y \in \mathcal{T}$ is any warping path $p \in \mathcal{P}_{\ell(x),\ell(y)}$ satisfying $\delta(x, y) = f\left(c_p(x, y)\right)$.

Even if the underlying local distance function d is a metric, the induced DTW distance only satisfies the following conditions for all $x, y \in \mathcal{T}$:

1. $\delta(x, y) \geq 0$
2. $\delta(x, x) = 0$
3. $\delta(x, y) = \delta(y, x)$

Note that possibly $\delta(x, y) = 0$ for some distinct time series $x \neq y$.

We conclude this section with an example of a common DTW distance to illustrate the concepts of Definition 2.

Example 3 The *Euclidean DTW distance* is a DTW distance specified by the attribute set $\mathcal{A} = \mathbb{R}^q$, the squared Euclidean distance $d(x, y) = \|x - y\|^2$ as local distance for all $x, y \in \mathcal{A}$, and the square root function $f(x) = \sqrt{x}$ for all $x \in \mathbb{R}_{\geq 0}$.

2.2 Fréchet Functions

All statistics are summaries [14]. The most fundamental statistic is the arithmetic mean

$$\mu = \frac{1}{N} \sum_{k=1}^{N} x^{(k)}$$

of N data points $x^{(1)}, \ldots, x^{(N)} \in \mathbb{R}^q$. The arithmetic mean is the unique global minimizer of the sum of squared Euclidean distances

$$F(x) = \sum_{k=1}^{N} \left\| x - x^{(k)} \right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. To see this, observe that the function F is convex and differentiable. Setting its gradient to zero and solving the equation yields μ. Similarly, a median minimizes the sum of Euclidean distances, that is

$$F(x) = \sum_{k=1}^{N} \left\| x - x^{(k)} \right\|.$$

The median and arithmetic mean are measures of central tendency (MCT) related to the Euclidean distance. As discussed in the previous section, the Euclidean distance fails to account for variations in temporal dynamics. This shortcoming of the Euclidean distance propagates to Euclidean MCT as indicated by Fig. 2a. To obtain MCT that can cope with variations in temporal dynamics, we relate them to the DTW distance.

We first introduce some technicalities. A *loss function* is a monotonously increasing function of the form $h : \mathbb{R}_{\geq 0} \to \mathbb{R}$. An example of a loss function is the squared loss $h(u) = u^2$ for all $u \geq 0$. The treatment rests on the following assumptions:

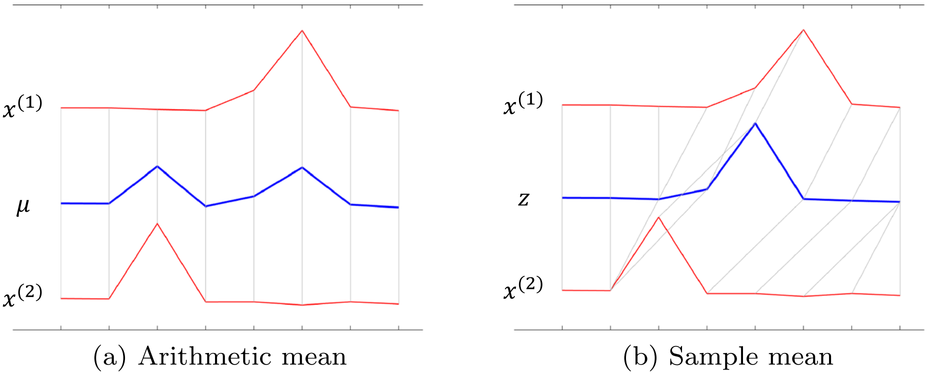


Fig. 2 Plot (a) shows two sample time series $x^{(1)}$ and $x^{(2)}$ that have a single peak but are out of phase and slightly vary in speed. We may think of $x^{(1)}$ and $x^{(2)}$ as the daily average temperature of some region during the summer at two different years. Based on this information, a typical summer of this region has a single extreme heat wave. In contrast, the arithmetic mean $\mu = (x^{(1)} + x^{(2)})/2$ has two attenuated peaks suggesting that a typical summer has two moderate heat waves. Plot (b) shows the same sample time series $x^{(1)}$ and $x^{(2)}$ as in (**a**) together with a sample mean z that minimizes the sum of squared Euclidean DTW distances (see Examples 3 and 7). In contrast to the arithmetic mean μ, the sample mean captures the characteristic properties of the sample time series $x^{(1)}$ and $x^{(2)}$ and shows a single peak as a representative summary of both sample peaks

Assumption 4 *Let $\mathcal{U} \subseteq \mathcal{T}$ be a subset, let δ be a DTW distance on $\mathcal{T}$, and let $\mathcal{X} = \left(x^{(1)}, \ldots, x^{(N)}\right) \in \mathcal{T}^N$ be a sample of N time series $x^{(k)}$ with corresponding loss function $h_k : \mathbb{R}_{\geq 0} \to \mathbb{R}$ for all $k \in [N]$.*

In line with the sum of (squared) Euclidean distances, we define Fréchet functions of a sample of time series as a sum based on DTW distances.

Definition 5 Consider Assumption 4. Then the function

$$F : \mathcal{U} \to \mathbb{R}, \quad x \mapsto \sum_{i=1}^{N} h_k\left(\delta\left(x, x^{(k)}\right)\right)$$

is the *Fréchet function* of $\mathcal{X}$ corresponding to the loss functions $h_1, \ldots, h_N$.

The domain $\mathcal{U}$ of the Fréchet function is the *search space*. Its elements are the *candidate solutions*. We omit explicitly mentioning the corresponding loss functions of a Fréchet function if no confusion can arise.

The next definition introduces the (possibly empty) set of global minimizers of a Fréchet function.

Definition 6 The *sample mean set* of $\mathcal{X}$ is the set defined by

$$\mathcal{F} = \{z \in \mathcal{U} \;:\; F(z) \leq F(x) \text{ for all } x \in \mathcal{U}\}.$$

An element of $\mathcal{F}$ (if exists) is called a *sample mean* of $\mathcal{X}$.

A sample mean of $\mathcal{X}$ (if exists) is a time series that minimizes the Fréchet function F. If the function F does not attain its infimum, the corresponding sample mean set $\mathcal{F}$ is empty. Existence of a sample mean depends on the choice of DTW distance and loss function.

Moreover, if a sample mean exists, it may not be uniquely determined. In contrast, the *total variation*

$$F^* = \inf_{x \in \mathcal{U}} F(x)$$

exists and is uniquely determined, because the DTW distance is bounded from below and the loss is monotonously increasing.

Fréchet functions have been first suggested by Fréchet in 1948 to generalize the concept of arithmetic mean to abstract metric spaces that have no well-defined addition [13]. Since then, the fields of mathematical statistics and non-Euclidean pattern recognition picked up Fréchet's idea to study properties of the sample and population mean for special classes of metric spaces such as shape [10, 21], tree [11], and graph spaces [14, 17]. All these spaces have in common that their underlying distance is a metric. Since the DTW distance is not a metric, results from mathematical statistics on the sample mean in metric spaces cannot be applied [5, 13, 31, 34].

In the remainder of this section, we present some examples. Example 7 generalizes the arithmetic mean and median from Euclidean spaces to DTW spaces.

Example 7 Let $p \geq 1$. The Fréchet function of $\mathcal{X}$ corresponding to the loss functions $h_k(u) = u^p$ takes the form

$$F(x) = \sum_{k=1}^{N} \delta^p \left(x, x^{(k)}\right).$$

For $p = 1$ ($p = 2$) the Fréchet function F generalizes the concept of sample median (mean) in Euclidean spaces.

Figure 2b shows a sample mean of two time series as a global minimizer of the Fréchet function defined in Example 2 using $p = 2$. The next example presents a formulation of the Fréchet function for weighted sample means.

Example 8 Let $w_k > 0$ for all $k \in [N]$. The Fréchet function of $\mathcal{X}$ corresponding to the loss functions $h_k(u) = w_k \cdot u^p$ is of the form

$$F(x) = \sum_{k=1}^{N} w_k \, \delta^p \left(x, x^{(k)}\right).$$

The function $F(x)$ is a weighted sum of p-distances δ^p. In the special case of $w_k = 1/N$, the function $F(x)$ averages the sum of p-distances.

Weighted sample means of two time series can be used to generalize stochastic gradient descent methods to DTW spaces. Consequently, gradient-based learning methods such as deep learning, logistic regression, and self-organizing maps can be extended to DTW spaces. To see this, observe that the update rule of stochastic gradient methods can be regarded as a weighted average between the parameter vector and the current input example. By replacing weighted averages with corresponding sample means, we obtain generalized gradient methods that can cope with variations in temporal dynamics [16].

Example 9 presents a DTW space for which a sample mean may not always exist. This example is inspired by the edit distance for sequences but drastically simplified to directly convey the main idea.

Example 9 Let $\mathcal{T}$ be the set of all time series of finite length with values from the attribute set $\mathcal{A} = \mathbb{R}$. Consider the sample $\mathcal{X} = (x^{(1)}, x^{(2)})$ consisting of the two time series

$$x^{(1)} = (1, 1) \quad \text{and} \quad x^{(2)} = (1, -1).$$

We present a Fréchet function such that the sample mean set of $\mathcal{X}$ is empty. For this, we need to specify the DTW distance and the loss functions. The DTW distance is of the form

$$\delta(x, y) = \min\left\{\sqrt{c_p(x, y)} \;:\; p \in \mathcal{P}_{\ell(x),\ell(y)}\right\}$$

for all $x, y \in \mathcal{T}$. The cost $c_p(x, y)$ of aligning x and y along warping path p are based on the local distance function

$$d(a, a') = \begin{cases} \left(a - a'\right)^2 & a \neq 0 \textit{ and } a' \neq 0 \\ 2 & a = 0 \textit{ xor } a' = 0 \\ 0 & a = 0 \textit{ and } a' = 0 \end{cases}$$

for all $a, a' \in \mathcal{A}$. We assume the squared error loss $h_1(u) = h_2(u) = u^2$ for both sample time series. Then the Fréchet function of $\mathcal{X}$ is of the form

$$F(x) = \delta\left(x, x^{(1)}\right)^2 + \delta\left(x, x^{(2)}\right)^2.$$

Under these conditions, the sample mean set of $\mathcal{X}$ is empty and the total variation is $F^* = 2$. A proof of both statements can be found in the Appendix.

2.3 Unrestricted and restricted fréchet functions

We consider two forms of search spaces $\mathcal{U}$: (i) the unrestricted form $\mathcal{U} = \mathcal{T}$ and (ii) the restricted form $\mathcal{U} = \mathcal{T}_m$ for some pre-specified $m \in \mathbb{N}$. Other forms have not been reported in the literature.

An unrestricted Fréchet function $F : \mathcal{T} \rightarrow \mathbb{R}$ imposes no restrictions on the length of the sample mean, whereas a restricted Fréchet function $F_m : \mathcal{T}_m \rightarrow \mathbb{R}$ demands that the length of a sample mean is m. Observe that we write F_m to indicate a restricted Fréchet function. In addition, we write $\mathcal{F}_m$ to denote the restricted mean set. Occasionally, we call the elements of $\mathcal{F}_m$ restricted sample means and the elements of $\mathcal{F}$ the (unrestricted) sample means.

The restricted form is often considered in practice for computational reasons [2, 8, 15, 24, 28]. The length m of a sample mean is typically chosen within the range of the lengths of the sample time series. It has been empirically shown that such a choice of m often introduces a structural error that cannot be further improved [6].

2.4 The reduction theorem

The goal of this section is to present the Reduction Theorem. This theorem is a key result to prove the sufficient conditions for existence of a sample mean in unrestricted form.

The Reduction Theorem is based on the notion of reduction bound. To introduce the reduction bound, we assume that $\mathcal{X} = \left(x^{(1)}, \ldots, x^{(N)}\right)$ is a sample of N time series $x^{(k)}$ with length $\ell(x^{(k)}) \geq 2$. Then the *reduction bound* $\rho(\mathcal{X})$ of sample $\mathcal{X}$ is of the form

$$\rho(\mathcal{X}) = \sum_{k=1}^{N} \ell\left(x^{(k)}\right) - 2(N-1). \tag{2}$$

Equation (2) shows that $\rho(\mathcal{X})$ increases linearly with the sum of the lengths of the sample time series.

Note that the reduction bound in (2) is not defined for samples that contain time series of length one. The reduction bound for arbitrary samples requires additional technicalities and is therefore delegated to Section 3.4. The subsequent Reduction Theorem assumes the general definition of a reduction bound as provided in Section 3.4.

Theorem 10 (Reduction Theorem) *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample with reduction bound $\rho(\mathcal{X})$. Suppose that $F : \mathcal{T} \to \mathbb{R}$ is the unrestricted Fréchet function of $\mathcal{X}$. Then for every time series $x \in \mathcal{T}$ of length $\ell(x) > \rho(\mathcal{X})$ there is a time series $x' \in \mathcal{T}$ of length $\ell(x') = \ell(x) - 1$ such that $F(x') \leq F(x)$.*

The Reduction Theorem deserves some explanations. To illustrate these explanations, we also refer to Fig. 3. We make the following observations:

1. From the proof of the Reduction Theorem follows that every candidate solution x whose length exceeds the reduction bound has an element that can be removed without increasing the value $F(x)$. Such elements are said to be *redundant* (see Fig. 3a).
2. Removing a redundant element can even decrease the value of the Fréchet function (see Fig. 3a).
3. The reduction bound of the sample in Fig. 3a is given by

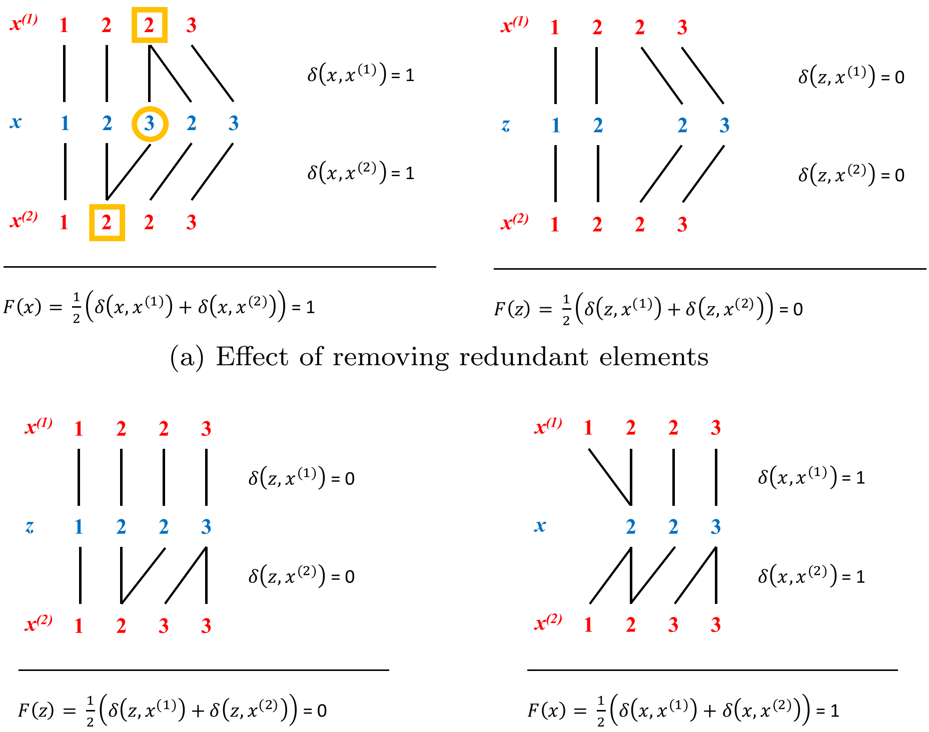
$$\rho(\mathcal{X}) = \ell\left(x^{(1)}\right) + \ell\left(x^{(2)}\right) - 2(N-1) = 4 + 4 - 2 = 6.$$

   The length of time series x is only $\ell(x) = 5 < \rho(\mathcal{X})$. This shows that short candidate solutions whose lengths do not exceed the reduction bound may also have redundant elements that can be removed without increasing the value of the Fréchet function. Existence of a redundant element depends on the choice of warping path between x and the sample time series. For short time series x, we can always find warping paths such that x has no redundant elements.
4. Removing a non-redundant element of a candidate solution can increase the value of the Fréchet function (see Fig. 3b).

The Reduction Theorem and observations 1–4 form the basis for existence proofs in unrestricted form and point to a technique to improve algorithms for approximating a sample mean (if exists). From observations 1–3 follows that a candidate solution x of any length could be improved or at least shortened by detecting and removing redundant elements of x. This observation is not further explored in this article and left for further research.

2.5 Sufficient conditions of existence

In this section, we derive sufficient conditions of existence of a sample mean in restricted and unrestricted form.



(a) Effect of removing redundant elements

(b) Effect of removing non-redundant elements

Fig. 3 Samples $\mathcal{X}$ consists of two univariate time series $x^{(1)}$ and $x^{(2)}$ shown in red. Blue time series x, z, and z' are candidate solutions. Black lines depict optimal warping paths. Fréchet functions assume the identity loss and the Euclidean DTW distance as defined in Example 3. Figure 3a: An element s_i of a candidate solution s is redundant if every element of the sample time series connected to s_i is also connected to another element of s. For example, the third element x_3 of x is redundant (enclosed by a circle). The elements of the sample time series connected to x_3 are enclosed by a square. Both squared elements are connected to two elements of x. The time series z is obtained from x by removing x_3. The DTW distances of the sample time series from x are both one and from z are both zero. Thus, we have $F(x) > F(z)$. This result suggests that removing a redundant element does not increase the value of the Fréchet function. [0.1cm] Figure **??**: Removing a non-redundant element can increase the Fréchet function. The time series x is obtained from z by removing the first element, which is not redundant. The DTW distances of the sample time series from z are both zero and from x are both one. This shows that $F(z) < F(x)$

The Reduction Theorem guarantees existence of a sample mean in unrestricted form if sample means exist in restricted forms. Thus, existence proofs in the general unrestricted form reduce to existence proofs in the simpler restricted form. This is the assertion of the following result:

Corollary 11 *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample and let $\rho \in \mathbb{N}$ be the reduction bound of $\mathcal{X}$. Suppose that $\mathcal{F}_m \neq \emptyset$ for every $m \in [\rho]$. Then $\mathcal{F} \neq \emptyset$.*

It is not self-evident that existence of restricted sample means implies existence of a sample mean. To see this, consider the restricted total variation

$$F_m^* = \inf_{x \in \mathcal{T}_m} F_m(x).$$

If F_m attains its infimum, then $\mathcal{X}$ has a restricted sample mean. Suppose that $\mathcal{X}$ has a restricted sample mean for every $m \in \mathbb{N}$. Then

$$v_m = \min_{l \leq m} F_l^*.$$

is the smallest restricted total variation over all lengths $l \in [m]$. The sequence $(v_m)_{m \in \mathbb{N}}$ is bounded from below and monotonously decreasing. Therefore, the sequence $(v_m)_{m \in \mathbb{N}}$ converges to the unrestricted total variation F_*. Then $\mathcal{X}$ has a sample mean only if the sequence $(v_m)_{m \in \mathbb{N}}$ attains its infimum F_*. Corollary 11 guarantees that the sequence $(v_m)_{m \in \mathbb{N}}$ indeed attains its infimum F^* latest at $m = \rho(\mathcal{X})$.

Also note that Corollary 11 does not exclude existence of sample means whose lengths exceed the reduction bound. Figure 4 presents an example for which a sample mean can have almost any length.

Next, we present sufficient conditions of existence of a sample mean. The first result proposes sufficient conditions of existence of a restricted and unrestricted sample mean for time series (sequences) with values from a finite attribute set.

Proposition 12 *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample. Suppose that $\mathcal{A}$ is a finite attribute set. Then the following statements hold:*

1. $\mathcal{F}_m \neq \emptyset$ *for every* $m \in \mathbb{N}$.
2. $\mathcal{F} \neq \emptyset$.

Proposition 12 imposes no other restrictions on $\mathcal{A}$ than being finite. This means that the time series to be averaged can be, for example, sequences with elements from a finite alphabet, sequences of trees, and temporal networks.

The second result proposes sufficient conditions of existence of a restricted and unrestricted sample mean of time series with elements from $\mathcal{A} = \mathbb{R}^q$.

Proposition 13 *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample. Suppose that the following assumptions hold:*

1. $(\mathcal{A}, \|\cdot\|)$ *is a normed vector space with* $\mathcal{A} = \mathbb{R}^q$.
2. *The loss functions* $h_1, \ldots, h_N$ *are continuous and strictly monotonously increasing.*

Then the following statements hold:

1. $\mathcal{F}_m \neq \emptyset$ *for every* $m \in \mathbb{N}$.
2. $\mathcal{F} \neq \emptyset$.

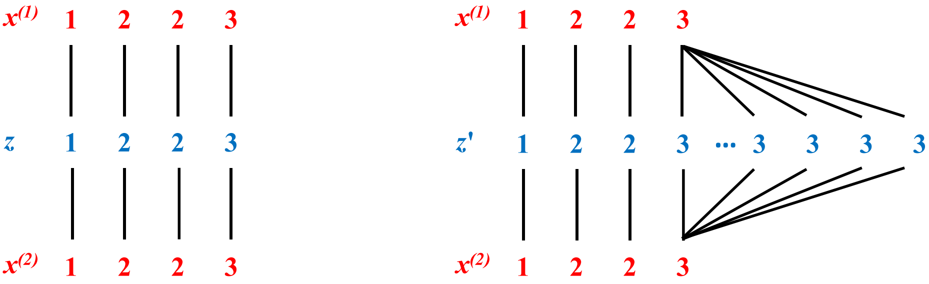


Fig. 4 The time series z is a sample mean of $\mathcal{X}$, because $F(z) = 0$ is the lowest possible value. The time series z' is obtained from the sample mean z by appending arbitrarily many time points with element 3. From $F(z') = 0$ follows that z' is also a sample mean of $\mathcal{X}$

The attribute set $\mathcal{A}$ covers the case of univariate ($q = 1$) and the case of multivariate ($q > 1$) time series. The local cost function d on $\mathcal{A}$ is a metric induced by a norm on $\mathbb{R}^q$. Loss functions $h : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ of the form $h(u) = w \cdot u^p$ are continuous and strictly monotonously increasing for $w > 0$ and $p \geq 1$. Thus, the sufficient conditions of Proposition 13cover customary DTW spaces. We conclude this section with a remark on weighted means.

Remark 14 Proposition 13 holds when we replace the loss functions h_k by the loss functions $h'_k = w_k h_k$ with $w_k \in \mathbb{R}_{\geq 0}$ for all $k \in [N]$.

Note that only loss functions $h'_k = w_k h_k$ with positive weights $w_k > 0$ are strictly monotonously increasing. Hence, assumption (2) of Proposition 13 is violated for loss functions h'_k with zero-weights $w_k = 0$. Remark 14 relaxes the condition of strictly positive weights to non-negative weights. For a proof of Remark 14 we refer to Section 3.5.

3 Theory of warping graphs

In this section, we prove the assertions of Section 2 by using a graph-theoretic approach.

3.1 Overview

The goal of this section is to outline the proposed graph-theoretic approach for proving the results of Section 2. For this, we represent Fréchet functions as graphs. Suppose that $\mathcal{X} = \left(x^{(1)}, \ldots, x^{(N)}\right)$ is a sample of time series. Consider a simple Fréchet function of $\mathcal{X}$ with respect to the identity loss

$$F(x) = \sum_{k=1}^{N} \delta\left(x, x^{(k)}\right).$$

Every candidate solution x is related to the k-th sample time series $x^{(k)}$ by an optimal warping path via the DTW distances $\delta\left(x, x^{(k)}\right)$. We represent warping paths by graphs, where nodes represent the elements of a time series and edges represent alignments between the corresponding time series elements (see Fig. 5, left plot). Then, we glue the N warping paths together by identifying the candidate solution x (see Fig. 5, right plot).

The graphs representing warping paths have an additional structure that also flows into the glued warping graph: First, there are different groups of nodes. Each group represents a time series. Second, the nodes of a group are ordered in the same sequential order as the time series they represent. Third, edges connect only nodes from a candidate solution with nodes from a sample time series. Fourth, nodes and edges are labeled. The labels of the nodes represent the attribute values of the corresponding elements of the time series and the labels of the edges represent local cost incurred when comparing the attribute values of the connected nodes.

To prove the results from Section 2, we exploit the additional structure of warping graphs and glued warping graphs. We follow a bottom-up approach. We first introduce warping chains to model the order properties of time series and warping paths. Then, we proceed to warping graphs to study the structural properties of warping paths. Thereafter, we study glued warping graphs that model the structure of Fréchet functions. Finally, we enhance the glued warping graphs with node and edge labels to derive the Reduction Theorem and the other assertions presented in Section 2.

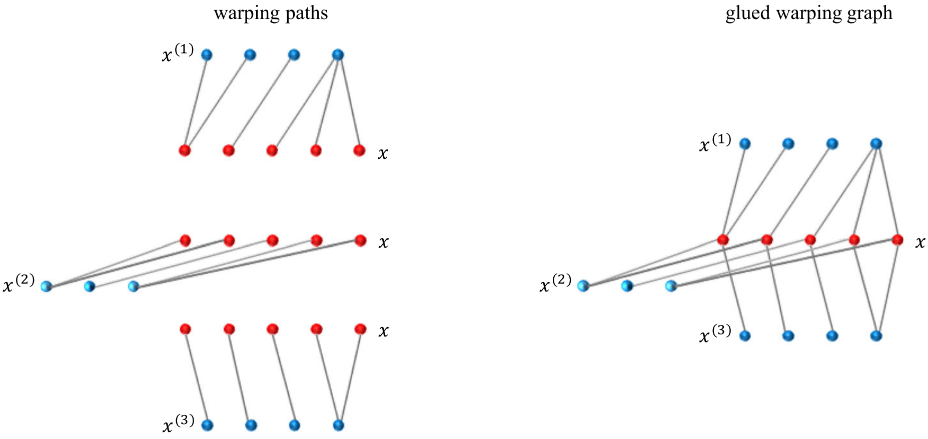


Fig. 5 The left plot shows three graphs representing warping paths between a candidate solution x and sample time series $x^{(1)}, x^{(2)}, x^{(3)}$. Balls represent elements of time series and edges represent points of warping paths. Red balls refer to elements of the candidate solution and blue balls to elements of the sample time series. The right plot shows the glued warping graph that identifies the candidate solution x of the three warping graphs

3.2 Warping Chains

In this section, we study the order properties of warping paths induced by the sequential order of time series together with the boundary and step conditions.

A binary relation $\leq_{\mathcal{V}}$ is a linear (or total) order on a set $\mathcal{V}$ if the following conditions hold for all $i, j, k \in \mathcal{V}$:

1. if $i \leq_{\mathcal{V}} j$ and $j \leq_{\mathcal{V}} i$ then $i = j$ (antisymmetríc)
2. if $i \leq_{\mathcal{V}} j$ and $j \leq_{\mathcal{V}} k$ then $i \leq_{\mathcal{V}} k$ (transitive)
3. $i \leq_{\mathcal{V}} j$ or $j \leq_{\mathcal{V}} i$ (connex relation)

The connex property implies reflexivity, that is $i \leq_{\mathcal{V}} i$ for all $i \in \mathcal{V}$. A partial order is an antisymmetric, transitive, and reflective relation. Thus, a linear order is always partial but the converse statement does not hold, in general. The strict linear order $<_{\mathcal{V}}$ on $\mathcal{V}$ associated with $\leq_{\mathcal{V}}$ is defined by $i <_{\mathcal{V}} j$ if $i \leq_{\mathcal{V}} j$ and $i \neq j$ for all $i, j \in \mathcal{V}$.

Definition 15 A *chain* $(\mathcal{V}, \leq_{\mathcal{V}})$ is a set $\mathcal{V}$ together with a linear order $\leq_{\mathcal{V}}$.

Unless otherwise stated, we assume that all chains are finite. In addition, we occasionally say that $\mathcal{V}$ is a chain without explicitly referring to its linear order $\leq_{\mathcal{V}}$. For the sake of convenience, we denote a chain by $\mathcal{V} = \{i_1, \ldots, i_m\}$ and assume that the elements appear in strict linear order $i_1 < \cdots < i_m$. We call i_1 the first and i_m the last element in $\mathcal{V}$. The first and last element of a chain are the boundary elements. Every element $i \in \mathcal{V} \setminus \{i_1, i_m\}$ is an inner element of $\mathcal{V}$.

Example 16 Consider the following examples:

1. The pair $(\mathbb{Z}, \leq)$ is a chain. Then every subset $\mathcal{V}$ together with the linear order $\leq_{\mathcal{V}}$ obtained by restricting $\leq$ to the subset $\mathcal{V}$ is also a chain.
2. Let $x = (x_1, \ldots, x_n)$ be a time series. The set $\mathcal{V} = \{1, \ldots, n\}$ of indices is a subset of $\mathbb{Z}$. Then $(\mathcal{V}, \leq_{\mathcal{V}})$ is a chain as in the first example.

3. Let $x = (x_{t_1}, \dots, x_{t_n})$ be a time series with values x_{t_i} at time stamps t_i. Suppose that t_i is the number of seconds that have elapsed since January 1, 1970 at 00:00:00 GMT. The set $\mathcal{V} = \{t_1, \dots, t_n\}$ of time stamps is a subset of $\mathbb{Z}$. Then $(\mathcal{V}, \leq_\mathcal{V})$ is a chain as in the first example.

Let $\mathcal{V}' \subseteq \mathcal{V}$ be a subset. Suppose that $\leq_{\mathcal{V}'}$ is the partial order on $\mathcal{V}'$ obtained by restricting the linear order $\leq_\mathcal{V}$ to the subset $\mathcal{V}'$. Then $\leq_{\mathcal{V}'}$ is also a linear order and $(\mathcal{V}', \leq_{\mathcal{V}'})$ is a chain, called subchain of $\mathcal{V}$. A subchain $\mathcal{V}' \subseteq \mathcal{V}$ is contiguous if there are elements $i_p, i_q \in \mathcal{V}$ with $i_p \leq_\mathcal{V} i_q$ such that $\mathcal{V}' = \left\{i_p, i_{p+1}, \dots, i_q\right\}$.

Example 17 Consider the chain $\mathcal{V} = \{2, 4, 6, 8, 10\}$ with the usual linear order on integers. Then $\mathcal{V}' = \{4, 6, 8\}$ is a contiguous subchain and $\mathcal{V}'' = \{2, 6, 8\}$ is a subchain of $\mathcal{V}$ but not a contiguous subchain.

Next, we want to model warping paths by chains. Consider a warping path $p = (p_1, \dots, p_L) \in \mathcal{P}_{m,n}$ with points $p_l = (i_l, j_l) \in [m] \times [n]$. Note that the sets $[m] = \{1, \dots, m\}$ and $[n] = \{1, \dots, n\}$ together with the linear order $\leq$ on integers are both chains. We show that the linear order on $[m]$ and $[n]$ together with the boundary and step conditions induce a linear order on warping paths. For this, we first introduce the notion of successor and predecessor.[1] The successor i_l^+ and predecessor i_l^- of element $i_l \in \mathcal{V}$ are defined by

$$i_l^+ = \begin{cases} i_{l+1} & 1 \leq l < L \\ * & l = L \end{cases} \qquad \text{and} \qquad i_l^- = \begin{cases} i_{l-1} & 1 < l \leq L \\ * & l = 1 \end{cases},$$

where $*$ is a distinguished symbol denoting the void element. Introducing the symbol $*$ is useful to avoid case distinctions. The next definition describes the step condition of warping paths by a point-to-set map.

Definition 18 Let $\mathcal{U} = \mathcal{V} \times \mathcal{W}$ be the product of two chains $\mathcal{V}$ and $\mathcal{W}$. The *successor map* on $\mathcal{U}$ is a point-to-set map

$$S_\mathcal{U} : \mathcal{U} \to 2^\mathcal{U}, \quad (i, j) \mapsto \left\{(i^+, j), (i, j^+), (i^+, j^+)\right\} \cap \left(\mathcal{V} \times \mathcal{W}\right),$$

where $2^\mathcal{U}$ denotes the set of all subsets of $\mathcal{U}$.

Intersection of the successor map with $\mathcal{V} \times \mathcal{W}$ excludes elements with $i^+ = *$ or $j^+ = *$ and thereby ensures that the map is well-defined. The successor map sends (i, j) to the empty set if $i \in \mathcal{V}$ or $j \in \mathcal{W}$ are the last elements.

We use the successor map to define warping chains that model contiguous sub-sequences of warping paths.

Definition 19 Let $\mathcal{U} = \mathcal{V} \times \mathcal{W}$ be the product of two chains $\mathcal{V}$ and $\mathcal{W}$. A subset $\mathcal{E} \subseteq \mathcal{U}$ is a *warping chain* in $\mathcal{U}$ if its elements can be ordered by $e_1, \dots, e_L$ such that $L = |\mathcal{E}|$ and $e_{l+1} \in S_\mathcal{U}(e_l)$ for all $l \in [L-1]$.

We show that warping chains are indeed chains. For this, Proposition 20 lays the foundation.

[1] For the sake of convenience, we introduce predecessors here for later usage.

Proposition 20 *Let $\mathcal{V}$ and $\mathcal{W}$ be chains and let $\mathcal{E}$ be a warping chain in $\mathcal{V} \times \mathcal{W}$. Then any pair of elements $(i, j), (r, s) \in \mathcal{E}$ satisfies*

$$(i \leq_{\mathcal{V}} r \;\wedge\; j \leq_{\mathcal{W}} s) \;\vee\; (r \leq_{\mathcal{V}} i \;\wedge\; s \leq_{\mathcal{W}} j)\,. \tag{3}$$

Proof Let $\leq_{\mathcal{V}}$ and $\leq_{\mathcal{W}}$ be the linear orders of $\mathcal{V}$ and $\mathcal{W}$, respectively. Since $\mathcal{E}$ is a warping chain, we can find an ordering $e_1, \ldots, e_L$ of the elements of $\mathcal{E}$ such that $e_{l+1} \in S_{\mathcal{U}}(e_l)$ for all $l \in [L-1]$. Let $e_l = (i_l, j_l)$ for all $l \in [L]$. From the definition of the successor map follows that

$$i_l \leq_{\mathcal{V}} i_{l+1} \;\wedge\; j_l \leq_{\mathcal{W}} j_{l+1} \tag{4}$$

for all $l \in [L-1]$. Let $e_p, e_q \in \mathcal{E}$. We distinguish between two (non-exclusive) cases: (i) $p \leq q$ and (ii) $q \leq p$. Suppose that (i) holds. Consider the contiguous subchain $\mathcal{E}' = \left\{e_p, e_{p+1}, \ldots, e_q\right\}$. By assumption, we have $e_{l+1} \in S_{\mathcal{U}}(e_l)$ for all l with $p \leq l < q$. Repeatedly applying (4) yields

$$i_p \leq_{\mathcal{V}} i_q \;\wedge\; j_p \leq_{\mathcal{W}} j_q$$

by transitivity of the linear orders $\leq_{\mathcal{V}}$ and $\leq_{\mathcal{W}}$. Similarly, case (ii) gives

$$i_q \leq_{\mathcal{V}} i_p \;\wedge\; j_q \leq_{\mathcal{W}} j_q.$$

Combining both cases (i) and (ii) yields the assertion. □

From Proposition 20 directly follows that warping chains are chains.

Corollary 21 *Let $(\mathcal{V}, \leq_{\mathcal{V}})$ and $(\mathcal{W}, \leq_{\mathcal{W}})$ be chains and let $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{W}$ be a warping chain. Consider the binary relation $\leq_{\mathcal{E}}$ on $\mathcal{E}$ defined by*

$$(i, j) \leq_{\mathcal{E}} (r, s) \;\Leftrightarrow\; i \leq_{\mathcal{V}} r \text{ and } j \leq_{\mathcal{W}} s$$

for all $(i, j), (r, s) \in \mathcal{E}$. Then $(\mathcal{E}, \leq_{\mathcal{E}})$ is a chain.

Proof It is sufficient to show that $\leq_{\mathcal{E}}$ is a linear order. The product order is always a partial order. The relation $\leq_{\mathcal{E}}$ is the restriction of the product order on $\mathcal{V} \times \mathcal{W}$ to the subset $\mathcal{E}$. The connex property of $\leq_{\mathcal{E}}$ follows from (3) of Proposition 20. This proves that $\leq_{\mathcal{E}}$ is a linear order on $\mathcal{E}$. □

Corollary 21 justifies to denote warping chains by $\mathcal{E} = \{e_1, \ldots, e_L\}$ and assume that $e_1 <_{\mathcal{E}} \cdots <_{\mathcal{E}} e_L$.

3.3 Warping graphs

In the previous section, we introduced warping chains to study order properties of warping paths. In this section, we introduce warping graphs to study graph-theoretical properties of warping paths.

We commence with a motivating example. Let $p = (p_1, \ldots, p_L) \in \mathcal{P}_{m,n}$ be a warping path with points $p_l = (i_l, j_l)$ from the grid $[m] \times [n]$. Then, we can represent p by a graph $G = (\mathcal{U}, \mathcal{E})$, called warping graph. For this, we assume that the node set

$$\mathcal{V} = [m] \sqcup [n] = \{i_1, \ldots, i_m, j_1, \ldots, j_n\}$$

is the disjoint union of $[m]$ and $[n]$, where i_l represents element $l \in [m]$ and j_k represents element $k \in [n]$. The set $\mathcal{E} = \{p_1, \ldots, p_L\}$ of edges consists of the points $p_l = (i_l, j_l)$ of warping path p that connect nodes i_l and j_l. Observe that edges only connect nodes from $[m]$ with nodes from $[n]$. Figure 6 depicts an example of a warping graph.

We begin our formal treatment with introducing some graph-theoretical definitions. A graph is a pair $G = (\mathcal{U}, \mathcal{E})$ consisting of a finite set $\mathcal{U} \neq \emptyset$ of nodes and a set $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{U}$ of edges. We say, G is an undirected graph if $(i, j) \in \mathcal{E}$ implies $(j, i) \in \mathcal{E}$. We assume that all graphs are undirected.

A node $i \in \mathcal{U}$ is incident with edge $e \in \mathcal{E}$, if there is a node $j \in \mathcal{U}$ such that $e = (i, j)$ or $e = (j, i)$. Similarly, an edge $(i, j) \in \mathcal{E}$ is said to be incident to node i and to node j. The neighborhood of node $i \in \mathcal{U}$ is the subset of nodes defined by $\mathcal{N}(i) = \{j \in \mathcal{U} : (i, j) \in \mathcal{E} \text{ or } (j, i) \in \mathcal{E}\}$. The elements of $\mathcal{N}(i)$ are the neighbors of i. The degree $\deg(i) = |\mathcal{N}(i)|$ of node i in G is the number of neighbors of i.

A subgraph of graph $G = (\mathcal{U}, \mathcal{E})$ is an undirected graph $G' = (\mathcal{U}', \mathcal{E}')$ such that $\mathcal{U}' \subseteq \mathcal{U}$ and $\mathcal{E}' \subseteq \mathcal{E}$. We write $G' \subseteq G$ to denote that G' is a subgraph of G. A graph G is connected, if for any two nodes $i, j \in \mathcal{U}$ there is a sequence $i = u_1, u_2, \ldots, u_n = j$ of nodes in G such that $u_{k+1} \in \mathcal{N}(u_k)$ for all $k \in [n-1]$. A component C of graph G is a connected subgraph $C \subseteq G$ such that $C \subseteq C'$ implies $C = C'$ for every connected subgraph $C' \subseteq G$.

A graph $G = (\mathcal{U}, \mathcal{E})$ is bipartite, if $\mathcal{U}$ can be partitioned into two disjoint and non-empty subsets $\mathcal{V}$ and $\mathcal{W}$ such that $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{W}$. We write $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ to denote a bipartite graph with node partitions $\mathcal{V}$ and $\mathcal{W}$. The size of the bipartite graph G is defined by $|\mathcal{V}| \times |\mathcal{W}|$. A bipartite chain graph is a bipartite graph whose node partitions are chains.

Definition 22 Let $\mathcal{V} = \{i_1, \ldots, i_m\}$ and $\mathcal{W} = \{j_1, \ldots, j_n\}$ be chains. A bipartite chain graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ is a *warping graph* if the following conditions are satisfied:

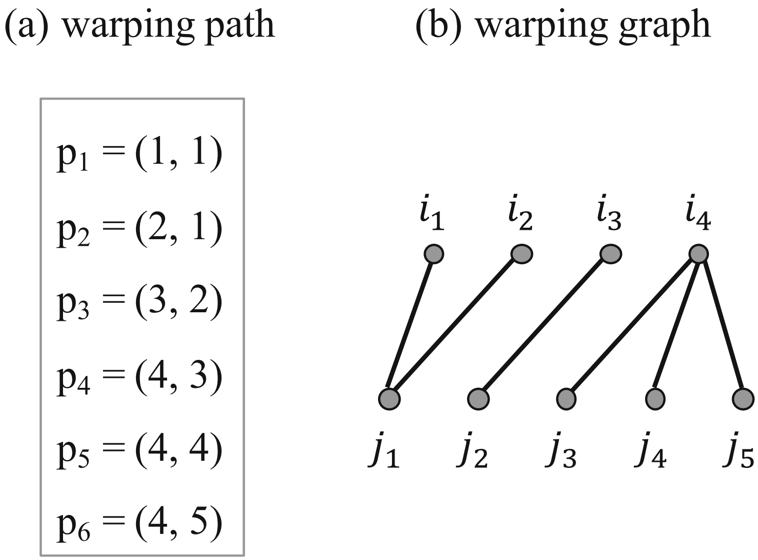


Fig. 6 Representation of a warping path by a warping graph. Table (**a**) shows a warping path $p \in \mathcal{P}_{4,5}$ with 6 points $p_1, \ldots, p_6$. Diagram (**b**) shows the corresponding warping graph that represents p. Balls represent nodes and lines represent edges. The nodes in the top (bottom) row represent the four (five) different elements occurring in the first (second) component of warping path p. Suppose that $(l, k) \in p$ is a point of warping path p. Nodes are labeled with i_l and j_k instead of l and k to ensure distinct identifiers for distinct nodes

1. $(i_1, j_1), (i_m, j_n) \in \mathcal{E}$ *(boundary condition)*
2. $\mathcal{E}$ is a warping chain in $\mathcal{V} \times \mathcal{W}$ *(step condition)*,

The set of all warping graphs of size $m \times n$ is denoted by $\mathcal{G}_{m,n}$. If $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ is a warping graph, we briefly write S_G to denote the successor map $S_{\mathcal{V}\times\mathcal{W}}$. The following result is a direct consequence of the boundary and step conditions:

Proposition 23 *Every node in a warping graph has a neighbor.*

We show that the neighborhood of a node of one partition of a warping graph is a contiguous subchain of the other partition.

Proposition 24 *Let G be a warping graph with node partitions $\mathcal{Z}$ and $\mathcal{Z}'$. The neighborhood $\mathcal{N}(i)$ of a node in $i \in \mathcal{Z}$ is a contiguous subchain of $\mathcal{Z}'$.*

Proof Suppose that the warping graph is of the form $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$. Without loss of generality, we assume that $\mathcal{Z} = \mathcal{V}$ and $\mathcal{Z}' = \mathcal{W}$. Then, we have $i \in \mathcal{V}$. The assertion trivially holds for $|\mathcal{N}(i)| = 1$. Suppose that $|\mathcal{N}(i)| > 1$. We assume that $\mathcal{N}(i)$ is not contiguous. Then there are elements $j', j'' \in \mathcal{N}(i)$ and $j \in \mathcal{W} \setminus \mathcal{N}(i)$ such that $j' \leq_{\mathcal{W}} j \leq_{\mathcal{W}} j''$. From Proposition 23 follows that there is a node $i' \in \mathcal{V} \setminus \{i\}$ such that $(i', j) \in \mathcal{E}$.

Two cases can occur: (1) $i' <_{\mathcal{V}} i$ and (2) $i <_{\mathcal{V}} i'$. It is sufficient to consider the first case $i' <_{\mathcal{V}} i$. The proof of the second case is analogue. By construction, we have edges $(i', j), (i, j') \in \mathcal{E}$ such that $i' <_{\mathcal{V}} i$ and $j' <_{\mathcal{W}} j$. Thus, the edges (i', j) and (i, j') of warping chain $\mathcal{E}$ contradict (3) of Proposition 20. Hence, $\mathcal{N}(i)$ is contiguous. □

As demonstrated by the next example, the definition of warping path admits superfluous points that can be removed without violating the boundary and step conditions.

Example 25 Consider the warping paths

$$\begin{aligned} p &= \big((1,1),(1,2),(2,2)\big) \\ q &= \big((1,1),(2,2)\big) \end{aligned}$$

from $\mathcal{P}_{2,2}$. The costs of aligning two time series $x = (x_1, x_2)$ and $y = (y_1, y_2)$ along both warping paths are given by

$$\begin{aligned} c_p(x, y) &= d(x_1, y_1) + d(x_1, y_2) + d(x_2, y_2) \\ c_q(x, y) &= d(x_1, y_1) + d(x_2, y_2) \end{aligned}$$

To determine the DTW distance, we are only interested in optimal warping paths between x and y. Since d is a distance function, all terms contributing to the costs $c_p(x, y)$ and $c_q(x, y)$ are non-negative. Therefore, we have

$$c_q(x, y) \leq c_p(x, y).$$

The second point $(1, 2)$ of warping path p can be safely removed without violating the properties of a warping path and without any effect for the DTW distance.

Figure 7 shows the two types of superfluous points that can occur in a warping path. Superfluous points are inconvenient for analytic purposes. Fortunately, we can safely remove such points, because their removal leaves the DTW distance and the Fréchet function unaffected. To get rid of superfluous points, we introduce compact warping paths. A warping path is said to be compact if none of its points can be removed without violating the boundary or step condition. The warping path in Fig. 6 is an example of a compact warping path, because removing the first or last point violates the boundary condition and removing any of the other points violates the step condition. The next definition presents a graph-theoretic formulation of compact warping paths.

Definition 26 A warping graph $G \in \mathcal{G}_{m,n}$ is *compact* if there is no warping graph $G' \in \mathcal{G}_{m,n}$ such that G' is a proper subgraph of G.

We show that the definition of a compact warping graph is indeed an equivalent formulation of a compact warping path.

Proposition 27 *Let G be a warping graph with edge set $\mathcal{E} = \{e_1, \dots, e_L\}$. Then the following statements are equivalent:*

1. *G is compact.*
2. *Let $1 < k < L$. Then $e_{l+k} \notin S_G(e_l)$ for all $l \in [L-k]$.*

Proof Suppose that $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$. For a given chain $\mathcal{C} = \{i_1, \dots, i_m\}$, we define the distance

$$\Delta_{\mathcal{C}}(i_u, i_v) = |u - v| + 1$$

for all $i_u, i_v \in \mathcal{C}$. Note that the distances $\Delta_{\mathcal{V}}$ and $\Delta_{\mathcal{W}}$ are defined, because the node sets of a warping graph are chains. Let $k, l \in \mathbb{N}$ such that $k + l \leq L$. Suppose that $e_l = (i, j)$ and $e_{l+k} = (r, s)$. For $k = 1$, we have $e_{l+1} \in S_G(e_l)$ because G is a warping graph. Then by definition of the successor map, we have $\Delta_{\mathcal{V}}(i, r), \Delta_{\mathcal{W}}(j, s) \in \{0, 1\}$ and $\Delta_{\mathcal{V}}(i, r) + \Delta_{\mathcal{W}}(j, s) \geq 1$. By induction on k, we have $\Delta_{\mathcal{V}}(i, r) + \Delta_{\mathcal{W}}(j, s) \geq k$. Then for $k \geq 3$, we have $\Delta_{\mathcal{V}}(i, r) \geq 2$ or $\Delta_{\mathcal{W}}(j, s) \geq 2$. Hence, $\Delta_{\mathcal{V}}(i, r) \notin \{0, 1\}$ or $\Delta_{\mathcal{W}}(j, s) \notin \{0, 1\}$. This shows that $e_{l+k} \notin S_G(e_l)$. Thus, for any warping graph the second statement is impossible for $k > 2$. Therefore, it is sufficient to prove the assertion for the case $k = 2$.

Let G be compact. We assume that there is an $l \in [L-2]$ such that $e_{l+2} \in S_G(e_l)$. By construction, the edge e_{l+1} is an inner element of the chain $\mathcal{E}$. From $e_{l+2} \in S_G(e_l)$ follows that removing e_{l+1} neither violates the boundary conditions nor the step condition. This contradicts compactness of G and shows that a compact warping graph G implies the second statement.

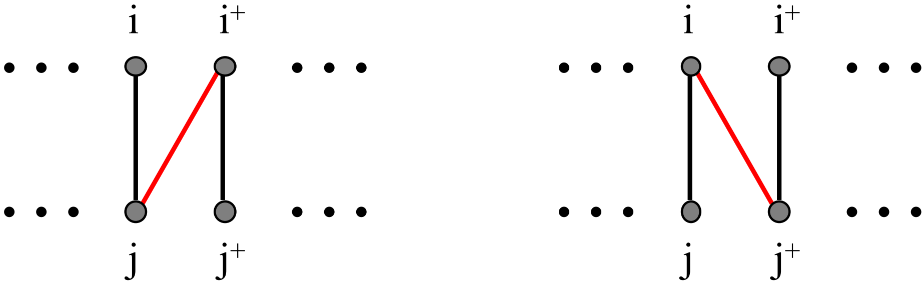


Fig. 7 Warping graphs depicting the two types of superfluous points that can occur. The superfluous points are represented by red edges

Next, we show the opposite direction. Suppose that $e_{l+2} \notin S_G(e_l)$ for all $l \in [L-2]$. We assume that G is not compact. Then there is an edge $e_k \in \mathcal{E}$ that can be removed without violating the boundary and step conditions. Not violating the boundary condition implies that $1 < k < L$. Hence, e_{k-1} and e_{k+1} are edges in $\mathcal{E}$. We set $l = k - 1$. Then, we obtain the contradiction that $e_l \in S_G(e_{l+2})$. Hence, G is compact. □

Recall that the step condition is encoded in the successor map and the boundary condition in the definition of a warping graph. The second statement of Proposition 27 indirectly states that no edge of a compact warping graph can be deleted without violating the step condition. To see this, consider the edge set $\mathcal{E} = \{e_1, \ldots, e_L\}$ of a warping graph G. Since $\mathcal{E}$ is a warping chain, every edge e_l with $l < L$ has a successor edge $e_{l+1} \in S_G(e_l)$. Suppose that we remove an inner edge e_l. Then $e_{l+1} \notin S_G(e_{l-1})$ by the second statement of Proposition 27, that is e_{l+1} is not a successor of e_{l-1}. This implies that $\mathcal{E} \setminus \{e_l\}$ violates the step condition. For our purpose, the indirect formulation of statement 2 is more convenient in later proofs than the direct formulation that no edge can be deleted without violating the step condition.

Statement 2 of Proposition 27 (indirectly) follows the characterization of compact warping paths. Later, we additionally need a second characterization of compact warping graphs that is purely graph-theoretical. To derive a graph-theoretical characterization, we first need to prove two rather technical auxiliary results (Lemma 28 and Lemma 29).

We assume that $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ is a warping graph. By $\mathcal{V} \sqcup \mathcal{W}$ we denote the disjoint union of the node partitions. If $i \in \mathcal{V} \sqcup \mathcal{W}$ is a node of one partition, then its neighborhood $\mathcal{N}(i)$ is a subset of the other node partition. From Proposition 24 follows that $\mathcal{N}(i)$ is a chain and consists of boundary and eventually inner nodes. Let $\mathcal{N}^\circ(i)$ denote the possibly empty subset of inner nodes of chain $\mathcal{N}(i)$. We show that inner nodes of $\mathcal{N}(i)$ always have degree one.

Lemma 28 *Let $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a warping graph. Suppose that $i \in \mathcal{V} \sqcup \mathcal{W}$ is a node with neighborhood $\mathcal{N}(i)$. Then* $\deg(j) = 1$ *for all $j \in \mathcal{N}^\circ(i)$.*

Proof Without loss of generality, we assume that $i \in \mathcal{V}$. Then $\mathcal{N}(i) \subseteq \mathcal{W}$ is a chain. The assertion holds for $|\mathcal{N}(i)| \leq 2$, because in this case $\mathcal{N}(i)$ has no inner node. Suppose that $|\mathcal{N}(i)| > 2$. Let $j \in \mathcal{N}(i)$ be an inner node. We assume that $\deg(j) > 1$. Then there is a node $i' \in \mathcal{V} \setminus \{i\}$ such that $(i', j) \in \mathcal{E}$. Since $\mathcal{V}$ is a chain, we find that either $i' <_{\mathcal{V}} i$ or $i <_{\mathcal{V}} i'$.

We only consider the first case $i' <_{\mathcal{V}} i$. The proof of the second case is analogue. Observe that j is an inner node of $\mathcal{N}(i)$ and $\mathcal{N}(i)$ is contiguous by Proposition 24. Then $\mathcal{N}(i)$ contains the predecessor $j' = j^-$ of node j. By construction $e = (i, j')$ and $e' = (i', j)$ are edges of $\mathcal{E}$ such that $i' <_{\mathcal{V}} i$ and $j' <_{\mathcal{W}} j$. Thus, the edges e and e' of warping chain $\mathcal{E}$ contradict (3) of Proposition 20. Hence, we have $\deg(j) = 1$. Since the inner node j was chosen arbitrarily, the assertion follows. □

Figure 8 illustrates the assertion of Lemma 28. Consider node $j_4 \in \mathcal{W}$ from the bottom row. Its neighborhood is $\mathcal{N}(j_4) = \{i_3, i_4, i_5, i_6\} \in \mathcal{V}$. The subset of inner nodes of $\mathcal{N}(j_4)$ is given by $\mathcal{N}^\circ(j_4) = \{i_4, i_5\} \subseteq \mathcal{V}$. As claimed by Lemma 28, both nodes $i_4, i_5 \in \mathcal{V}$ in the top row have degree one. Note that $j_4 \in \mathcal{W}$ is the only node of $\mathcal{V} \sqcup \mathcal{W}$ whose inner node set is non-empty.

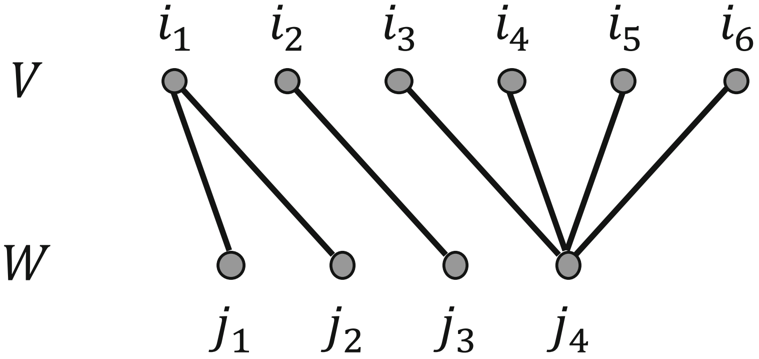


Fig. 8 Compact warping graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$. The top row represents nodes from $\mathcal{V}$ and the bottom row nodes from $\mathcal{W}$. The lines connecting the nodes from $\mathcal{V}$ and $\mathcal{W}$ is the set $\mathcal{E}$ of edges

Lemma 29 *Let $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a warping graph and let $i \in \mathcal{V} \sqcup \mathcal{W}$ be a node with neighborhood $\mathcal{N}(i)$. Suppose that $|\mathcal{N}(i)| \geq 2$ and $j \in \mathcal{N}(i)$ is a boundary node with $deg(j) \geq 2$. Then the following properties hold:*

1. *If j is the first node in $\mathcal{N}(i)$, then $i^- \in \mathcal{V}$ exists and $(i^-, j) \in \mathcal{E}$.*
2. *If j is the last node in $\mathcal{N}(i)$, then $i^+ \in \mathcal{V}$ exists and $(i^+, j) \in \mathcal{E}$.*

Proof We show the second assertion. The proof of the first assertion is analogue. Since $|\mathcal{N}(i)| \geq 2$ and j is the last node of $\mathcal{N}(i)$, we find that $j' = j^- \in \mathcal{N}(i)$ and therefore $(i, j') \in \mathcal{E}$ exists. From $\deg(j) \geq 2$ follows that there is a node $i' \in \mathcal{V}$ such that $(i', j) \in \mathcal{E}$. We assume that $i' \in \mathcal{N}(j)$ satisfies $i' <_{\mathcal{V}} i$. This implies that (i, j') and (i', j) are two edges of $\mathcal{E}$ such that $i' <_{\mathcal{V}} i$ and $j' <_{\mathcal{W}} j$. Then the edges (i, j') and (i', j) of warping chain $\mathcal{E}$ contradict (3) of Proposition 20. Thus, the assumption $i' <_{\mathcal{V}} i$ is invalid.Therefore, we have $i <_{\mathcal{V}} i'$. This in turn shows that $i^+ \in \mathcal{V}$ exists. From $i, i' \in \mathcal{N}(j)$ and $i <_{\mathcal{V}} i'$ follows $i^+ \in \mathcal{N}(j)$, because $\mathcal{N}(j)$ is a contiguous subchain of $\mathcal{V}$ by Proposition 24. This shows $(i^+, j) \in \mathcal{E}$ and completes the proof. □

A bipartite graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ is complete if $\mathcal{E} = \mathcal{V} \times \mathcal{W}$. We denote a complete bipartite graph with partitions of size $|V| = m$ and $|W| = n$ by $K_{m,n}$. Note that every two complete bipartite graphs with the same notation $K_{m,n}$ are isomorphic. Let $r \in \mathbb{N}$. Then a star graph (or star) is a complete bipartite graph of the form $K_{1,r}$ or $K_{r,1}$. By definition, a star graph has at least two nodes. A star forest is a graph whose components are star graphs.

Now, we are in the position to present a purely graph-theoretic characterization of compact warping graphs.

Proposition 30 *A compact warping graph is a star forest.*

Proof Let $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a compact warping graph and let $C = \left(\mathcal{V}', \mathcal{W}', \mathcal{E}'\right)$ be a component of G. From Proposition 23 follows that C has at least two nodes connected by an edge.

We assume that C is not a star. Then C has two nodes $i, j \in \mathcal{V}' \sqcup \mathcal{W}'$ with degree larger than one. Without loss of generality, we assume that $i \in \mathcal{V}'$. Then $\mathcal{N}(i) \subseteq \mathcal{W}'$ has at least two elements. Suppose that all nodes from $\mathcal{N}(i)$ have degree one. Since component

C is bipartite, we find that C is isomorphic to the star $K_{1,r}$, where $r = \deg(i) > 1$. This contradicts our assumption that C is not a star. Hence, there is a node $j \in \mathcal{N}(i) \subseteq \mathcal{W}'$ with $\deg(j) > 1$.

From Lemma 28 follows that node j is a boundary node of $\mathcal{N}(i)$. We show the assertion for the case that j is the last node in $\mathcal{N}(i)$. The proof for the case that j is the first node in $\mathcal{N}(i)$ is analogue. Since j is the last node in $\mathcal{N}(i)$ and $|\mathcal{N}(i)| \geq 2$, we have $j^- \in \mathcal{N}(i)$ and therefore $(i, j^-) \in \mathcal{E}$. Applying Lemma 29 yields that $i^+ \in \mathcal{V}$ exists and $(i^+, j) \in \mathcal{E}$.

By construction, we have $\left(i, j^-\right), (i, j), \left(i^+, j\right) \in \mathcal{E}$. This shows that (i, j) is not a boundary edge in $\mathcal{E}$. Since $\left(i^+, j\right) \in S_G\left(i, j^-\right)$, we can remove (i, j) without violating the step condition. Then the subgraph $G' = (\mathcal{V}, \mathcal{W}, \mathcal{E} \setminus \{(i, j)\})$ of G is a warping graph. This contradicts our assumption that G is compact. Hence, C is a star. □

For example, the compact warping graph in Fig. 8 is a star forest consisting of the three star graphs $K_{1,2}$, $K_{1,1}$, and $K_{4,1}$.

The next two results characterize the components of a compact warping graph. These results are later needed to show the existence of redundant elements.[2] The first of both results is an immediate consequence of the proof of Proposition 30.

Corollary 31 *Let $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a compact warping graph. Then every component of G is a star with at least one node in $\mathcal{V}$ and one node in $\mathcal{W}$.*

The second of both results bounds the maximum number of stars of the form $K_{1,1}$.

Proposition 32 *Let $G \in \mathcal{G}_{m,n}$ be a compact warping graph with $m > n$. Then, we have:*

1. *G has at most $n - 1$ components of the form $K_{1,1}$.*
2. *G has a component of the form $K_{r,1}$ with $r > 1$.*

Proof Let $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ with $|\mathcal{V}| = m$ and $|\mathcal{W}| = n$. To show the first assertion, we assume that G has $n' > n-1$ components of the form $K_{1,1}$. This is only possible for $n' = n$, because every component of G has at least one node in $\mathcal{W}$ by Corollary 31.

Let $C_1, \ldots, C_n$ be n components of G of the form $K_{1,1}$ with

$$\mathcal{C}_k = (\{i_k\}, \{j_k\}, \{(i_k, j_k)\})$$

for all $k \in [n]$. The union of the first node partitions over the n components C_k gives $\mathcal{V}' = \{i_1, \ldots, i_n\}$. From $m > n$ follows $\mathcal{V}' \subsetneq \mathcal{V}$. Then there is a node $i \in \mathcal{V} \setminus \mathcal{V}'$. From Proposition 30 follows that there is a component C of G is a star of the form $K_{r,s}$ that contains node i. Since $s \geq 1$ by definition of a star, component C has a node $j \in \mathcal{W}$. Then there is a $k \in [n]$ such that $j = j_k$ is a node in component $\mathcal{C}_k$. Since $i \neq i_k$ by construction, the graph $H = (\{i, i_k\}, \{j_k\}, \{(i, j_k), (i_k, j_k)\})$ is a connected subgraph of G that includes component C_k as a proper subgraph. This contradicts our assumption that C_k is a maximal connected subgraph of G. Consequently, G cannot have more than $n - 1$ components of the form $K_{1,1}$.

Next, we show the second assertion. Suppose that $C_1, \ldots, C_q$ are all components of G that are of the form $K_{1,1}$. Let $\mathcal{V}' = \left\{i_1, \ldots, i_q\right\} \subseteq \mathcal{V}$ and $\mathcal{W}' = \left\{j_1, \ldots, j_q\right\} \subseteq \mathcal{W}$ be the subsets covered by the q components C_k. From the first part of this proof follows that $q < n$

[2] We already have encountered redundant elements in Section 2 but we will formally introduce them later.

and by assumption, we have $n < m$. Then $\mathcal{V}'' = \mathcal{V} \setminus \mathcal{V}'$ and $\mathcal{W}'' = \mathcal{W} \setminus \mathcal{W}'$ are non-empty. By $m'' = \left|\mathcal{V}''\right| = m - q$ and $n'' = \left|\mathcal{W}''\right| = n - q$ we denote the respective number of nodes not contained in any of the q components C_k. From $q < n < m$ follows that $1 \leq n'' < m''$. The pigeonhole principle states that there is at least one node $j \in \mathcal{W}''$ that is connected to at least two nodes $i, i' \in \mathcal{V}''$. Let C be the component of G containing the three nodes i, i', and j. From Proposition 30 follows that C is a star of the form $K_{r,s}$. We find that $r \geq 2$, because C contains at least the two nodes i and i' from $\mathcal{V}'' \subset \mathcal{V}$. Then $s = 1$, because C is a star. This shows the second assertion. □

Observe that the compact warping graph in Fig. 8 has one star graph of the form $K_{1,1}$ and one star graph of the form $K_{r,1}$ with $r = 4 > 1$. The next definition introduces redundant nodes of a compact warping graph. Redundant nodes are nodes that can be safely deleted without violating the property of being a compact warping graph.

Definition 33 Let G be a compact warping graph. A node i of G is *redundant* if $\deg(j) \geq 2$ for all neighbors $j \in \mathcal{N}(i)$.

Let i be a node in G. Then $G - \{i\}$ is the subgraph of G obtained by deleting node i and its incident edges. We show that deleting a redundant node from a compact warping graph results again in a compact warping graph.

Proposition 34 *Let i be a redundant node of a compact warping graph G. Then $G - \{i\}$ is a compact warping graph.*

Proof Without loss of generality, we assume that $i \in \mathcal{V}$. Let $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ with warping chain $\mathcal{E} = \{e_1, \dots, e_L\}$. Suppose that $G' = G - \{i\} = \left(\mathcal{V}', \mathcal{W}, \mathcal{E}'\right)$, where $\mathcal{V}' = \mathcal{V} \setminus \{i\}$ and $\mathcal{E}'$ is the chain obtained from $\mathcal{E}$ by removing all edges incident to node i.

We first show that $\mathcal{N}(i) \subseteq \mathcal{W}$ consists of a singleton. Let C be the component of G that contains node i. Then C also includes the neighborhood $\mathcal{N}(i)$. Since G is compact, we can apply Proposition 30 and find that component C is a star of the form $K_{r,1}$ or $K_{1,r}$, where $r \geq 1$. Observe that $r = \deg(j) \geq 2$ for every neighbor $j \in \mathcal{N}(i)$, because node i is redundant. Hence, C is a star of the form $K_{r,1}$ and therefore $\mathcal{N}(i) = \{j\}$.

From $\deg(j) \geq 2$ follows that there is a node $i' \in \mathcal{V} \setminus \{i\}$ such that $(i', j) \in \mathcal{E}$. We distinguish between two cases: (1) $i' <_{\mathcal{V}} i$ and (2) $i <_{\mathcal{V}} i'$. We only consider the first case $i' <_{\mathcal{V}} i$. The proof of the second case is analogue. From Proposition 24 follows that $\mathcal{N}(j)$ is a contiguous subchain of $\mathcal{W}$ with $i', i \in \mathcal{N}(j)$. Then $i^- \in \mathcal{N}(j)$. We distinguish between two cases: (1) $e_L = (i, j)$, and (2) $e_l = (i, j)$ for some $1 < l < L$. Note that the case $e_1 = (i, j)$ cannot occur due to existence of $i^- \in \mathcal{V}$.

Case 1: From $e_L = (i, j)$ follows that the edge set of G' is of the form $\mathcal{E}' = \{e_2, \dots, e_{L-1}\}$. In addition, i is the last node in $\mathcal{V}$ and therefore i^- is the last node in $\mathcal{V}'$. Furthermore, we find that $j \in \mathcal{W}$ is the last node in $\mathcal{W}$. We show that $\mathcal{E}'$ is a warping chain. The first boundary condition is satisfied by $e_1 \in \mathcal{E}'$. From $i^- \in \mathcal{N}(j)$ follows that $(i^-, j) \in \mathcal{E}$ is an edge that satisfies the second boundary condition connecting the last nodes of $\mathcal{V}'$ and $\mathcal{W}$. Finally, from $e_{l+1} \in S_G(e_l)$ for all $l \in [L-1]$ follows that the step condition remains valid in G'. Therefore, the edge set $\mathcal{E}'$ is a warping chain of length $L' = L - 1$. It remains to show that G' is compact. For this, we assume that G' is not compact. Then from Proposition 27 follows that there is an index $l \in [L' - 2]$ such that $e_{l+2} \in S_{G'}(e_l)$. This implies that $e_{l+2} \in S_G(e_l)$ contradicting the assumption that G is compact. Hence, G' is a compact warping graph.

Case 2: There is an index $1 < l < L$ such that $e_l = (i, j)$. Hence, $\mathcal{E}$ has at least three edges and the edge set of G' is of the form $\mathcal{E}' = \{e_1, \dots, e_{l-1}, e_{l+1}, \dots, e_L\}$. To show that $\mathcal{E}'$ is a warping chain, we assume that $e_{l-1} = (i', j')$ and $e_{l+1} = (i'', j'')$. From $\mathcal{N}(i) = \{j\}$ and the step condition follows that $i' = i^-$ and $i'' = i^+$. This shows that i is neither the first nor last node in $\mathcal{V}$. Hence, e_1 and e_L satisfy the boundary conditions in $\mathcal{E}'$.

We show that $\mathcal{E}'$ satisfies the step condition. Since $\mathcal{E}$ is a warping chain, we have $e_{k+1} = S_G(e_k)$ for all $k \in [L-1]$. Since $S_{G'} = S_G$ on $\mathcal{E}'$, it is sufficient to show that $e_{l+1} \in S_{G'}(e_l)$. According to the previous parts of the proof, we have $(i^-, j) \in \mathcal{E}$. The step condition together with $i' = i^-$ imply that $e_{l-1} = (i', j') = (i^-, j') = (i^-, j)$ and therefore $j' = j$. Again, from the step condition follows that either $j'' = j$ or $j'' = j^+$. Observe that $i'^+ = i''$ in $\mathcal{V}'$. Then, we have $(i'', j) \in S_{G'}(i', j)$ and $(i'', j^+) \in S_{G'}(i', j)$. This shows that $\mathcal{E}'$ satisfies the step condition in either of both cases $j'' = j$ or $j'' = j^+$.

It remains to show that G' is compact. For this, we assume that G' is not compact. Suppose that $\mathcal{I} = [L] \setminus \{l\}$ is the index set of $\mathcal{E}'$. Then from Proposition 27 follows that there is an index $k \in \mathcal{I} \setminus \{L-1, L\}$ such that $e_{k+2} \in S_{G'}(e_k)$. This implies that $e_{k+2} \in S_G(e_k)$ contradicting the assumption that G is compact. Hence, G' is a compact warping graph. □

3.4 Glued warping graphs

In the previous section, we studied warping graphs that represent warping paths. In this section, we glue several warping graphs together to model the structure of a Fréchet function. Thereafter, we present the graph-theoretic foundation of the Reduction Theorem.

We first outline the basic structure of glued warping graphs. For further details, we refer to Section 3.1 and Fig. 5. Suppose that $\mathcal{X} = \left(x^{(1)}, \dots, x^{(N)}\right)$ is a sample of time series. Consider the same Fréchet function

$$F(x) = \sum_{k=1}^{N} \delta\left(x, x^{(k)}\right).$$

as in Section 3.1. For every $k \in [N]$, we assume that $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k)$ is the warping graph representing the optimal warping path $p^{(k)}$ between candidate solution x and sample time series $x^{(k)}$. The chain $\mathcal{V}$ represent the candidate solution x, the chain $\mathcal{W}_k$ represents the sample time series $x^{(k)}$, and the edge set $\mathcal{E}_k$ represents the points of the optimal warping path $p^{(k)}$. To obtain the structure of the Fréchet function $F(x)$, we glue the graphs $G_1, \dots, G_N$ along the warping chain $\mathcal{V}$. We formalize and study glued warping graphs as a special form of multi-partite graphs.

A graph $G = (\mathcal{U}, \mathcal{E})$ is a centered N-partite graph if the set $\mathcal{U}$ can be partitioned into $N + 1$ disjoint non-empty subsets $\mathcal{V}, \mathcal{W}_1 \dots, \mathcal{W}_N$ such that

$$\mathcal{E} \subseteq \bigcup_{k=1}^{N} \mathcal{V} \times \mathcal{W}_k.$$

Definition 35 Let $G_1, \dots, G_N$ be compact warping graphs of the form $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k)$ for all $k \in [N]$. The *glued graph* with *splice* $\mathcal{V}$ and *particles* $G_1, \dots, G_N$ is a centered N-partite graph $G = (\mathcal{V}, \mathcal{W}_1, \dots, \mathcal{W}_N, \mathcal{E})$ with edge set $\mathcal{E} = \mathcal{E}_1 \sqcup \dots \sqcup \mathcal{E}_N$.

Note that a particle of a glued graph is always a compact warping graph. The definition of a glued graph assumes that all N particles $G_1, \ldots, G_N$ share a common node partition $\mathcal{V}$ and that any two particles G_k and G_l have disjoint node partitions $\mathcal{W}_k$ and $\mathcal{W}_l$, respectively. Then the glued graph with splice $\mathcal{V}$ is obtained by taking the disjoint union of the particles $G_1, \ldots, G_N$, but by identifying the nodes from $\mathcal{V}$. A special case of a glued graph is any compact warping graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ with the first partition $\mathcal{V}$ as its splice.

The following treatment serves to prepare the graph-theoretic foundation of the Reduction Theorem. We first extend the notion of redundant nodes to glued graphs (see also Fig. 3).

Definition 36 Let G be a glued graph with splice $\mathcal{V}$ and particles $G_1, \ldots, G_N$. Node $i \in \mathcal{V}$ is *redundant* in G, if it is redundant in G_k for every $k \in [N]$.

Observe that only splice nodes can be redundant nodes. We show that removing a redundant node again results in a glued graph.

Proposition 37 *Let G be a glued graph with splice $\mathcal{V}$ and particles $G_1, \ldots, G_N$. Suppose that $i \in \mathcal{V}$ is redundant. Then $G - \{i\}$ is a glued graph with splice $\mathcal{V} \setminus \{i\}$ and particles $G_1 - \{i\}, \ldots, G_N - \{i\}$.*

Proof Let $k \in [N]$. Then particle $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k)$ is a compact warping graph by definition of a glued graph. Since splice node $i \in \mathcal{V}$ is redundant in G, it is redundant in G_k. Then from Proposition 34 follows that $G'_k = G_k - \{i\} = \left(\mathcal{V}', \mathcal{W}_k, \mathcal{E}'\right)$ is a compact warping graph with $\mathcal{V}' = \mathcal{V} \setminus \{i\}$ and $\mathcal{E}'_k = \mathcal{E}_k \setminus \mathcal{E}_k(i)$, where $\mathcal{E}_k(i) \subseteq \mathcal{E}_k$ is the subset of edges in G_k incident to node i.

The graph $G' = G - \{i\} = \left(\mathcal{V}', \mathcal{W}_1, \ldots, \mathcal{W}_N, \mathcal{E}'\right)$ has an edge set of the form $\mathcal{E}' = \mathcal{E} \setminus \mathcal{E}(i)$, where $\mathcal{E}(i) \subseteq \mathcal{E}$ is the subset of edges in G incident to node i. Since $\mathcal{E} = \mathcal{E}_1 \sqcup \cdots \sqcup \mathcal{E}_N$, we have

$$\mathcal{E}(i) = \mathcal{E}_1(i) \sqcup \cdots \sqcup \mathcal{E}_N(i) = \mathcal{E}'_1 \sqcup \cdots \sqcup \mathcal{E}'_N.$$

This shows that $G - \{i\}$ is a glued graph of particles $G'_1, \ldots, G'_N$ along $\mathcal{V}'$. □

Suppose that G is a glued graph with splice $\mathcal{V}$ and particles $G_1, \ldots, G_N$. A particle is said to be *trivial* if it is a star of the form $K_{m,1}$. By $\mathcal{I}_G \subseteq [N]$ we denote the subset of all indices $k \in [N]$ for which particle G_k is non-trivial. We call $\mathcal{I}_G$ the *core index set* (core) of G. By using the core index set, we define the reduction bound of a glued graph.

Definition 38 Let G be a glued graph consisting of splice $\mathcal{V}$ and N particles $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k)$ for all $k \in [N]$. Suppose that $\mathcal{I}_G$ is the core index set of G. Then

$$\rho(G) = \begin{cases} \sum\limits_{k \in \mathcal{I}_G} |\mathcal{W}_k| - 2\left(|\mathcal{I}_G| - 1\right) & \mathcal{I}_G \neq \emptyset \\ 1 & \mathcal{I}_G = \emptyset \end{cases}$$

is the *reduction bound* of G.

Figure 9 presents an example of the core index set and the reduction bound of a glued graph. Now, we are in the position to present the graph-theoretic foundation of the Reduction Theorem.

Theorem 39 *Let G be a glued graph with splice $\mathcal{V}$ such that $\rho(G) < |\mathcal{V}|$. Then $\mathcal{V}$ has a redundant node.*

Proof Suppose that $G_1, \ldots, G_N$ are the particles of G with $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k)$ for all $k \in [N]$. Let $m = |\mathcal{V}|$, $n_k = |\mathcal{W}_k|$, and $\mathcal{N}_k(i) = \mathcal{N}(i) \cap \mathcal{W}_k$ for every $k \in [N]$.

We first consider the special case that $\mathcal{I}_G = \emptyset$. Then all N particles are trivial, that is $n_k = 1$ for every $k \in [N]$. The reduction bound is of the form $\rho(G) = 1$. By assumption, we have $m > \rho(G)$. Hence, every particle G_k is a star of the form $K_{m,1}$, where $m > 1$. Then every splice node in $K_{m,1}$ has the same neighbor $j \in \mathcal{W}_k$ with $\deg(j) = m > 1$. Thus, every splice node is redundant.

Next, we assume that $\mathcal{I}_G \neq \emptyset$. We set $N' = |\mathcal{I}_G|$. Obviously, we have $N' \geq 1$. We say, G_k supports node $i \in \mathcal{V}$, if there is a node $j \in \mathcal{N}_k(i) \subseteq \mathcal{W}_k$ with $\deg(j) = 1$. In this case, i is not a redundant node in G_k. Thus, showing existence of a redundant node in $\mathcal{V}$ is equivalent to showing that $\mathcal{V}$ has a node not supported by any of the N particles $G_1, \ldots, G_N$. The proof proceeds in four steps.

1. We show that $n_k < m$ for any $k \in \mathcal{I}_G$. Suppose that $\mathcal{J} = \mathcal{I}_G \setminus \{k\}$. Then from $\mathcal{I}_G \neq \emptyset$ follows

$$\rho(G) = \sum_{l \in \mathcal{I}_G} n_l - 2(N' - 1) = n_k + \sum_{l \in \mathcal{J}} n_l - 2(N' - 1).$$

From $l \in \mathcal{I}_G$ follows $n_l \geq 2$. This together with $|\mathcal{J}| = N' - 1$ yields

$$\rho(G) \geq n_k + \sum_{l \in \mathcal{J}} 2 - 2(N' - 1) = n_k + 2\,|\mathcal{J}|\, - 2(N' - 1) = n_k.$$

Then from $m > \rho(G)$ follows $m > n_k \geq 2$.

2. We bound the number of splice nodes that can be supported by any non-trivial particle. For any $k \in \mathcal{I}_G$ let $\mathcal{W}'_k \subseteq \mathcal{W}_k$ be the subset of nodes in G_k that support a splice node. We define a map $\phi_k : \mathcal{W}'_k \to \mathcal{V}$ such that $(\phi_k(j), j) \in \mathcal{E}_k$. From Proposition 23

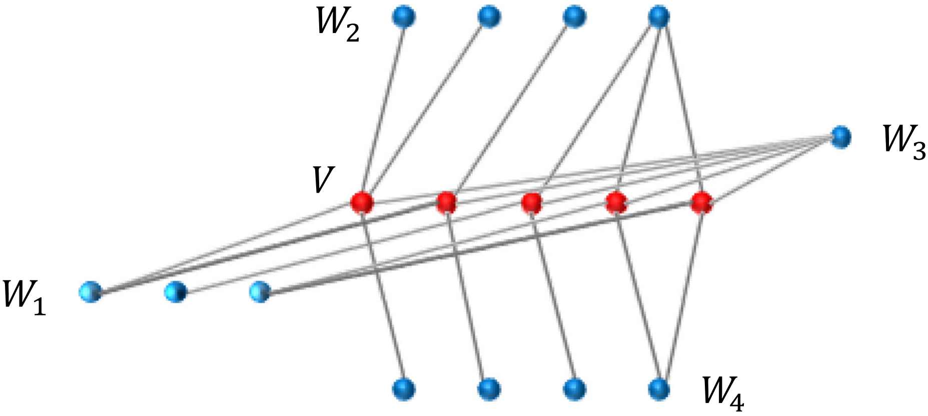


Fig. 9 Glued graph with splice $\mathcal{V}$ and four particles $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k)$. Particle G_3 is a star graph of the form $K_{5,1}$. All other particles are not star graphs. Then the core of G is the index set $\mathcal{I}_G = \{1, 2, 4\}$. The reduction bound of G is given by $\rho(G) = |\mathcal{W}_1| + |\mathcal{W}_2| + |\mathcal{W}_4| - 2\,(|\mathcal{I}_G|\, - 1) = 3 + 4 + 4 - 2(3 - 1) = 7$. The splice $\mathcal{V}$ with five nodes is below the reduction bound

follows that such a map exists due to the boundary and step conditions of warping graph G_k. Moreover, the map ϕ_k is uniquely determined, because $\deg(j) = 1$ for any node $j \in \mathcal{W}'_k$. This shows that $\mathcal{V}_k = \phi_k\big(\mathcal{W}'_k\big)$ is the set of splice nodes supported by G_k. Since ϕ_k is bijective, we have $|\mathcal{V}_k| = \big|\mathcal{W}'_k\big|$.

Recall that all G_k are compact warping graphs by definition of a glued graph. From step 1 of this proof follows that $n_k < m$. Therefore, we can apply Proposition 32 and obtain that G_k has at most $n_k - 1$ components of the form $K_{1,1}$ and at least once component of the form $K_{r,1}$ with $r > 1$. This shows that $\big|\mathcal{W}'_k\big| = |\mathcal{V}_k| \leq n_k - 1$.

3. We show that there is a splice node not supported by any non-trivial particle. For this, we define the set

$$\mathcal{U} = \bigcup_{k \in \mathcal{I}_G} \mathcal{V}_k$$

of all splice nodes that are supported by at least one non-trivial particle of G. Then it is sufficient to show that $m > |\mathcal{U}|$. We consider three cases: (1) $N' = 1$, (2) $N' = 2$, and (3) $N' > 2$.

Case 1: $N' = 1$. Suppose that $\mathcal{I}_G = \{u\}$. Since $\mathcal{I}_G \neq \emptyset$, the reduction bound is of the form

$$\rho(G) = n_u - 2(N' - 1) = n_u \geq 2.$$

According to step 2, we have $n_u - 1 \geq |\mathcal{V}_u|$. From $\mathcal{U} = \mathcal{V}_u$ follows

$$m > \rho(G) > |\mathcal{V}_u| = |\mathcal{U}|\,.$$

Case 2: $N' = 2$. Suppose that $\mathcal{I}_G = \{u, v\}$. Since $\mathcal{I}_G \neq \emptyset$, the reduction bound takes the form

$$\rho(G) = n_u + n_v - 2(N' - 1) = n_u + n_v - 2 = (n_u - 1) + (n_v - 1).$$

According to step 2, we have $n_u - 1 \geq |\mathcal{V}_u|$ and $n_v - 1 \geq |\mathcal{V}_v|$. From $\mathcal{U} = \mathcal{V}_u \cup \mathcal{V}_v$ follows

$$m > \rho(G) \geq |\mathcal{V}_u| + |\mathcal{V}_v| \geq |\mathcal{U}|\,.$$

Case 3: $N' > 2$. Suppose that the slice $\mathcal{V}$ is a chain of the form $\mathcal{V} = \{i_1, \ldots, i_m\}$ with boundary nodes $\mathrm{bd}(\mathcal{V}) = \{i_1, i_m\}$. We assume that $|\mathcal{U}| = m$. Then there are (not necessarily distinct) indices $u, v \in \mathcal{I}_G$ such that $i_1 \in \mathcal{V}_u$ and $i_m \in \mathcal{V}_v$. From the boundary and step conditions follows that the boundary nodes of any W_k ($k \in \mathcal{I}_G$) can only support the respective boundary nodes of $\mathcal{V}$ and not any other splice node. Then the first node of $\mathcal{W}_u$ only supports $i_1 \in \mathcal{V}$ and the last node of $\mathcal{W}_v$ only supports $i_m \in \mathcal{V}$.

Let $\mathcal{J} = \{u, v\}$, let $\mathcal{J}' = \mathcal{I}_G \setminus \mathcal{J}$, and let $\mathcal{V}'_k = \mathcal{V}_k \setminus (\mathcal{V}_u \cup \mathcal{V}_v)$ for all $k \in \mathcal{J}'$. The set $\mathcal{V}'_k$ consists of all splice nodes supported by G_k but not by G_u and G_v. Hence, the boundary nodes of $\mathcal{V}$ are not contained in $\mathcal{V}_k'$. Then both boundary nodes of $\mathcal{W}_k$ do not support any node in $\mathcal{V}_k'$. This implies $\left|\mathcal{V}_k'\right| \leq n_k - 2$. From the cardinality of the set union follows

$$\begin{aligned} |\mathcal{U}| &\leq \sum_{l \in \mathcal{J}} (n_l - 1) + \sum_{k \in \mathcal{J}'} (n_k - 2) \\ &= |\mathcal{J}| + \sum_{l \in \mathcal{J}} (n_l - 2) + \sum_{k \in \mathcal{J}'} (n_k - 2) \\ &= \sum_{k \in \mathcal{I}_G} n_k - 2N' + |\mathcal{J}|. \end{aligned}$$

Since $|\mathcal{J}| \leq 2$, we obtain

$$|\mathcal{U}| \leq \sum_{k \in \mathcal{I}_G} n_k - 2(N' - 2) = \rho(G) < m.$$

This contradicts the assumption $|\mathcal{U}| = m$ and shows that $|\mathcal{U}| < m$ holds.

All three cases show that $|\mathcal{U}| < m$. Hence, G has a splice node not supported by any of the non-trivial particles.

4. We show that G has a splice node not supported by any of the trivial and non-trivial particles. The non-trivial part follows from step 3. Therefore, it is sufficient to consider trivial particles only. Since $\mathcal{I}_G \neq \emptyset$ by assumption, there is a $k \in \mathcal{I}_G$. From $m > n_k$ and $n_k \geq 2$ follows $m > 2$. This implies that the trivial particles of G do not support any of the splice nodes in $\mathcal{V}$. This completes the proof.

□

3.5 Proofs of results from Section 2

The goal of this section is to prove the Reduction Theorem and the results on sufficient conditions of existence of a sample mean. For the sake of convenience, we restate all results.

In the previous sections, we studied the structure of (glued) warping graphs without considering the attributes of a time series and the local costs incurred by comparing two attributes. We first complete the graph-theoretic representation by labeling the nodes and edges of warping graphs. Node labels are the attributes of the corresponding time series elements and edge labels represent the local costs between the attributes of the connected nodes. As final step, we define the weight of a labeled warping graph that coincides with the DTW distance between the respective time series. Thereafter, we are in the position to prove the main assertions of this article.

We assume that $\mathcal{A}$ is an attribute set and $d : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ is a non-negative distance function on $\mathcal{A}$.

Definition 40 A *labeled warping graph* $H = (G, \lambda)$ consists of a warping graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{E})$ and a labeling function $\lambda : \mathcal{V} \sqcup \mathcal{W} \rightarrow \mathcal{A}$.

The labeling function λ assigns an attribute $\lambda(i) \in \mathcal{A}$ to any node $i \in \mathcal{V} \sqcup \mathcal{W}$. Thus, the nodes correspond to time points and the attributes to the elements at every time point.

The set of all labeled warping graphs of order $m \times n$ with label function λ is denoted by $\mathcal{G}_{m,n}^{\lambda}$. Since the set $\mathcal{G}_{m,n}^{\lambda}$ fixes both node partitions and the label function, the graphs in $\mathcal{G}_{m,n}^{\lambda}$ differ only in their edge sets. Thus, $\mathcal{G}_{m,n}^{\lambda}$ describes the set of all possible warping paths that align time series $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_n)$ whose elements $x_i = \lambda(i)$ and $y_j = \lambda(j)$ are specified by the labeling function λ.

Definition 41 Let $H = (G, \lambda)$ be a labeled warping graph with edge set $\mathcal{E}$. The *weight* of H is defined by

$$\omega(H) = \sum_{(i,j)\in\mathcal{E}} d(\lambda(i), \lambda(j))$$

The weight of a labeled warping graph corresponds to the cost of aligning two time series along a warping path. A DTW graph is a labeled warping graph with minimal weight.

Definition 42 A graph $H \in \mathcal{G}^{\lambda}_{m,n}$ is a *DTW graph*, if

$$\omega(H) = \min\left\{\omega(H') \,:\, H' \in \mathcal{G}^{\lambda}_{m,n}\right\}.$$

Suppose that $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a monotonous function such as the square root function. Then the value $f(\omega(H))$ is the DTW distance between the time series represented by the labeled node partitions (see Def. 2).

Now, we have developed the theory to prove the main assertions of this article.

Theorem 10 (Reduction Theorem) *Let F be the Fréchet function of a sample $\mathcal{X} \in \mathcal{T}^N$. Then for every time series $x \in \mathcal{T}$ of length $\ell(x) > \rho(\mathcal{X})$ there is a time series $x' \in \mathcal{T}$ of length $\ell(x') = \ell(x) - 1$ such that $F(x') \leq F(x)$.*

Proof Let $\mathcal{X} = \left(x^{(1)}, \dots, x^{(k)}\right)$, $m = \ell(x)$, and $n_k = \ell\left(x^{(k)}\right)$ for all $k \in [N]$. By assumption, we have $m > \rho(\mathcal{X})$.

For every $k \in [N]$ there is an optimal warping path $p^{(k)} \in \mathcal{P}_{m,n_k}$ aligning x and $x^{(k)}$. Let $H_k = (G_k, \lambda_k)$ be the DTW graph representing $p^{(k)}$. Then $\omega(H_k) = \delta(x, x^{(k)})$ and $G_k = (\mathcal{V}, \mathcal{W}_k, \mathcal{E}_k) \in \mathcal{G}_{m,n_k}$ is a warping graph with $m = |\mathcal{V}|$ and $n_k = |\mathcal{W}_k|$. Then we have

$$F(x) = \sum_{k=1}^{N} h_k(\omega(H_k)),$$

where $h_1, \dots, h_N$ are the corresponding loss functions.

Suppose that G_k is non-compact and $G'_k \subseteq G_k$ is compact. Since H_k is a DTW graph, we have $\omega(H'_k) = \omega(H_k)$, where $H'_k = (G'_k, \lambda'_k)$. Hence, without loss of generality we can assume that G_k is compact for all $k \in [N]$. Let G be the glued graph with splice $\mathcal{V}$ and particles $G_1, \dots, G_N$. Since $m > \rho(G)$, we can apply Theorem 39 and obtain that G has a redundant splice node $i \in \mathcal{V}$. Applying Proposition 37 yields that $G' = G - \{i\}$ is a glued graph with splice $\mathcal{V}' = \mathcal{V} \setminus \{i\}$ and particles $G'_1, \dots, G'_N$. The particles G'_k are of the form $G'_k = G_k - \{i\} = \left(\mathcal{V}', \mathcal{W}_k, \mathcal{E}'_k\right)$, where the edge set $\mathcal{E}'_k$ is obtained from $\mathcal{E}_k$ by removing all edges incident to splice node $i \in \mathcal{V}$.

The redundant node $i \in \mathcal{V}$ refers to element x_i of $x = (x_1, \dots, x_m)$. We denote the time series obtained from x by removing element x_i by

$$x' = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_m).$$

Let $H'_k = (G'_k, \lambda'_k)$ be the resulting labeled warping graph, where λ'_k denotes the labeling function obtained by restricting λ_k to the subset $\mathcal{V}' \sqcup \mathcal{W}_k$ for all $k \in [N]$. Then the labeled warping graphs H'_k represent warping paths $q^{(k)}$ that align time series x' with sample time series $x^{(k)}$. By construction and definition of the weight function ω, we find that $\omega(H_k') \leq \omega(H_k)$. Since the loss functions h_k are monotonously increasing, we obtain

$$F(x') = \sum_{k=1}^{N} h_k\left(\omega(H_k')\right) \leq \sum_{k=1}^{N} h_k\left(\omega(H_k)\right) = F(x).$$

By construction, we have $\ell(x') = \ell(x) - 1$. This completes the proof. □

Corollary 11 *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample and let $\rho \in \mathbb{N}$ be the reduction bound of $\mathcal{X}$. Suppose that $\mathcal{F}_m \neq \emptyset$ for every $m \in [\rho]$. Then $\mathcal{X}$ has a sample mean.*

Proof For every $m \in [\rho]$ let F_m^* denote the restricted total variation. We assume that $\mathcal{F} = \emptyset$. Then there is a time series $x \in \mathcal{T}$ of length $\ell(x) = p$ such that $F(x) = F_p(x) < F_m^*$ for all $m \in [\rho]$. This implies $p > \rho$, because otherwise we obtain the contradiction that $F_p(x) < F_p^*$. Let $q = p - \rho(\mathcal{X})$. By applying Theorem 10 exactly q-times, we obtain a time series $x' \in \mathcal{T}$ of length $\ell(x') = \rho$ such that $F_\rho^* \leq F(x') \leq F(x)$. This contradicts our assumption that $F(x) < F_\rho^*$. Hence, $\mathcal{F}$ is non-empty. □

Proposition 12 *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample. Suppose that $\mathcal{A}$ is a finite attribute set. Then the following statements hold:*

1. *$\mathcal{F}_m \neq \emptyset$ for every $m \in \mathbb{N}$.*
2. *$\mathcal{F} \neq \emptyset$.*

Proof Let $m \in \mathbb{N}$ be arbitrary. Since $\mathcal{A}$ is finite, the set subset $\mathcal{T}_m$ is also finite and consists of $m^{|\mathcal{A}|}$ time series. Then the set $F\left(\mathcal{T}^N\right)$ is a finite set. Hence, the restricted sample mean set $\mathcal{F}_m$ is non-empty and finite. Since m was chosen arbitrarily, the first assertion follows. The second assertion follows from Corollary 11. □

Proposition 13 *Let $\mathcal{X} \in \mathcal{T}^N$ be a sample. Suppose that the following assumptions hold:*

1. *$(\mathcal{A}, \|\cdot\|)$ is a normed vector space with $\mathcal{A} = \mathbb{R}^q$.*
2. *The loss functions $h_1, \ldots, h_N$ are continuous and strictly monotonously increasing.*

Then the following statements hold:

1. *$\mathcal{F}_m \neq \emptyset$ for every $m \in \mathbb{N}$.*
2. *$\mathcal{F} \neq \emptyset$.*

Proof The proof uses the notion of coercive function. Let $\alpha \in \mathbb{N}$. A continuous function $f : \mathbb{R}^\alpha \to \mathbb{R}$ is coercive if

$$\lim_{\|x\| \to \infty} f(x) = +\infty,$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^\alpha$.

We first consider the Euclidean norm $\|\cdot\|_2$ on some real-valued vector space $\mathbb{R}^\alpha$. The Euclidean norm is coercive. Since h_k is continuous and strictly monotonously increasing on $\mathbb{R}_{\geq 0}$, the composition $h_k\left(\|x\|_2\right)$ is coercive and continuous for all $k \in [N]$. Every norm $\|\cdot\|$ on $\mathbb{R}^\alpha$ is equivalent to the Euclidean norm. Therefore, we can find constants $0 < c \leq C$ for every $x \in \mathbb{R}^\alpha$ such that

$$c\,\|x\|_2 \leq \|x\| \leq C\,\|x\|_2\,.$$

Thus, $h_k(\|x\|)$ is coercive and continuous for every norm on $\mathbb{R}^\alpha$.

Suppose that $\mathcal{X} = \left(x^{(1)}, \ldots, x^{(N)}\right) \in \mathcal{T}^N$ is a proper sample of N time series $x^{(k)}$ of length $\ell\left(x^{(k)}\right) = n_k \geq 2$ for all $k \in [N]$. Let $m \in \mathbb{N}$ be arbitrary. Expanding the definition of the restricted Fréchet function F_m gives

$$F_m(x) = \sum_{k=1}^{N} h_k\left(\delta\left(x, x^{(k)}\right)\right) = \sum_{k=1}^{N} \min\left\{h_k\left(c_p\left(x, x^{(k)}\right)\right) \,:\, p \in \mathcal{P}\right\},$$

where $c_p(x, y)$ is the cost of aligning time series x and y along warping path p. Since $\mathcal{T}_m = \mathcal{A}^m = \mathbb{R}^{q \times m} \simeq \mathbb{R}^{\alpha}$, we can define the function

$$g_{p^{(k)}} : \mathbb{R}^{\alpha} \to \mathbb{R}, \quad x \mapsto c_{p^{(k)}}\left(x, x^{(k)}\right) = \sum_{l=1}^{L_k} h_k\left(\left\|x_{i_l} - x^{(k)}_{j_l}\right\|\right),$$

where $p^{(k)} \in \mathcal{P}_{m,n_k}$ is a warping path with L_k elements aligning x and $x^{(k)}$. The function $g_{p^{(k)}}$ is continuous and coercive as a sum of non-negative continuous and coercive functions. Then $g_{p^{(k)}}$ has a global minimum.

We define the set $\mathcal{P}_m = \mathcal{P}_{m,n_1} \times \cdots \times \mathcal{P}_{m,n_N}$. Then every element of $\mathcal{P}_m$ is of the form $\mathcal{C} = \left(p^{(1)}, \ldots, p^{(N)}\right)$, where $p^{(k)}$ is associated to time series $x^{(k)}$ for all $k \in [N]$. Then we can equivalently rewrite the restricted Fréchet function $F_m(x)$ as

$$F_m(x) = \min\left\{F_{\mathcal{C}}(x) \,:\, \mathcal{C} \in \mathcal{P}_m\right\},$$

where the component functions $F_{\mathcal{C}} : \mathbb{R}^{q \times m} \to \mathbb{R}$ are functions of the form

$$F_{\mathcal{C}}(x) = \sum_{k=1}^{N} g_{p^{(k)}}(x).$$

This shows that $F_{\mathcal{C}}(x)$ has a minimum. Let $F^*_{\mathcal{C}}$ denote the minimum value of $F_{\mathcal{C}}(x)$. From

$$\min_x F_m(x) = \min_x \min_{\mathcal{C}} F_{\mathcal{C}}(x) = \min_{\mathcal{C}} \min_x F_{\mathcal{C}}(x)$$

follows

$$\min_x F_m(x) = \min_{\mathcal{C} \in \mathcal{P}_m} F^*_{\mathcal{C}}.$$

Since $\mathcal{P}_m$ is a finite set, we obtain that F_m has a minimum. This shows the first assertion, because m was arbitrary. The second assertion follows from Corollary 11. □

Remark 14 *Proposition 13 holds when we replace the loss functions h_k by the loss functions $h'_k = w_k h_k$ with $w_k \in \mathbb{R}_{\geq 0}$ for all $k \in [N]$.*

Proof Suppose that all weights are zero. Then the Fréchet function corresponding to the loss functions $h'_k = 0$ is zero. Hence, every time series $z \in \mathcal{T}$ is an optimal solution and the assertion follows.

We assume that at least one weight is non-zero. Without loss of generality, let $r \in [N]$ such that $w_1, \ldots, w_r > 0$ and $w_{r+1} = \cdots w_N = 0$. Then the loss functions $h'_1, \ldots, h'_r$ are continuous and strictly monotonously increasing. Moreover, the Fréchet function $F(z)$ of sample $\mathcal{X}$ corresponding to the loss functions $h'_1, \ldots h'_N$ coincides with the Fréchet function $F'(z)$ of sample $x^{(1)}, \ldots, x^{(r)}$ corresponding to the loss functions $h'_1, \ldots h'_r$. Then the assertion follows from Proposition 13. □

4 Conclusion

This article presents sufficient conditions for the existence of a general concept of sample mean of time series in restricted and unrestricted form. The sufficient conditions hold for common loss functions and DTW distances reported in the literature. A key result is the Reduction Theorem stating that time series whose lengths exceed the reduction bound can be reduced to shorter time series without increasing the value of the Fréchet function. This result guarantees existence of a sample mean in unrestricted form if sample means exist in restricted form. The proof of the Reduction Theorem is framed into the theory of warping graphs.

After appearance of a preprint version of this article [18], two novel results based on the existence of a sample mean have been reported: a dynamic program for computing (weighted) sample means in restricted and unrestricted form [6] and a NP-hardness proof of the sample mean problem [7]. Two open theoretical problems are (i) under which conditions is a sample mean unique, and (ii) under which conditions the sample mean set is a consistent estimator of the population mean set (expectation set). Existence of weighted sample means suggests to adopt learning methods based on stochastic gradient optimization such as logistic regression and deep learning to DTW spaces. The basic idea is to replace the weighted average of the stochastic update rule in Euclidean spaces by a weighted sample mean in DTW spaces (see Example 8 and discussion). Algorithmically, it would be interesting to study to which extent existing heuristics generate superfluous points and redundant nodes and how removal of such pathological items affects the solution quality.

Acknowledgements B. Jain was funded by the DFG Sachbeihilfe `JA 2109/4-2`.

Appendix

A Proof of Example 9

Proof From the Reduction Theorem follows that it is sufficient to consider candidate solutions of length one and two. Thus, it is sufficient to consider the restricted Fréchet functions F_1 and F_2. In addition, it is sufficient to assume that all warping paths are compact (see Example 25). Then the set $\mathcal{P}_{m,n}$ with $m, n \in \{1, 2\}$ consists of exactly one warping path. Suppose that $x = (x_1)$, $y = (y_1, y_2)$ and $z = (z_1, z_2)$. Then the squared DTW distances are of the form

$$\begin{aligned}\delta(x, z)^2 &= d(x_1, z_1) + d(x_1, z_2)\\ \delta(y, z)^2 &= d(y_1, z_1) + d(y_2, z_2).\end{aligned}$$

We proceed with considering a slightly modified setting using the local distance function $d'(a, a') = (a - a')^2$ for all $a, a' \in \mathcal{A}$. Under this setting, we denote the DTW distance by δ' and the restricted Fréchet functions by F'_1 and F'_2. The function $F'_1(x)$ at time series $x = (x_1)$ is of the form

$$F'_1(x) = \underbrace{(x_1 - 1)^2 + (x_1 - 1)^2}_{=\delta'(x, x^{(1)})} + \underbrace{(x_1 - 1)^2 + (x_1 + 1)^2}_{=\delta'(x, x^{(2)})}.$$

The function F'_1 is convex and differentiable with respect to x_1. Taking the gradient, setting to zero and solving yields the unique solution $x_1 = 0.5$. Thus, $z = (0.5)$ is the restricted sample mean of $\mathcal{X}$ on $\mathcal{T}_1$ with Fréchet variation $F'_1(z) = 3$.

For a given $x = (x_1, x_2)$, a similar calculation with

$$F'_2(x) = (x_1 - 1)^2 + (x_2 - 1)^2 \; + \; (x_1 - 1)^2 + (x_2 + 1)^2$$

gives $z = (1, 0)$ as the unique restricted sample mean on $\mathcal{T}_2$ with Fréchet variation $F'_2(z) = 2$. By combining both results, we conclude that $z = (1, 0)$ is an unrestricted sample mean of $\mathcal{X}$ with total variation $F'^{*} = 2$.

Next, we assume the original local distance function d on $\mathcal{A}$ as defined in Example 9. Again, we first consider time series $x = (x_1)$ of length one. Then we have $F_1(x) = F'_1(x)$ if $x_1 \neq 0$ and $F_1(x) = 4$ if $x_1 = 0$. Thus, $z = (0.5)$ is the restricted sample mean of $\mathcal{X}$ on $\mathcal{T}_1$ with Fréchet variation $F_1(z) = 3$.

Now, we consider time series of length two. Let $x_\varepsilon = (1, \varepsilon)$ for some $\varepsilon \in \mathbb{R}$. Then

$$\lim_{\varepsilon \to 0} F(x_\varepsilon) = 2 \tag{5}$$

but $F(x_0) = 4$. Suppose there is a restricted sample mean $z = (z_1, z_2)$ on $\mathcal{T}_2$. From (5) follows that $F(z) \leq 2$. If at least one element of z is zero, we have $F(z) \geq 4$. This contradicts our assumption that z is a sample mean. Thus, the elements of z are both non-zero. Then we have $F_2(z) = F'_2(z)$. Recall that the unique minimizer of F'_2 has a zero element. This yields the contradiction $2 < F'_2(z) = F_2(z) \leq 2$. Consequently, the function F_2 has no minimizer. Thus, the unrestricted Fréchet function F never attains its infimum 2 and therefore $\mathcal{X}$ has no sample mean. □

References

1. Abanda, A., Mori, U., Lozano, J.A.: A review on distance based time series classification. Data Mining and Knowledge Discovery (2018)
2. Abdulla, W.H., Chow, D., Sin, G.: Cross-words reference template for DTW based speech recognition systems. Conference on Convergent Technologies for Asia-Pacific Region (2003)
3. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering – a decade review. Inf. Syst. **53**, 16–38 (2015)
4. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Disc. **31**(3), 606–660 (2017)
5. Bhattacharya, R., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds. Ann. Stat. **31**(1), 1–29 (2003)
6. Brill, M., Fluschnik, T., Froese, V., Jain, B., Niedermeier, R., Schultz, D.: Exact mean computation in dynamic time warping spaces. Data Mining and Knowledge Discovery (2019)
7. Bulteau, L., Froese, V., Niedermeier, R.: Hardness of Consensus Problems for Circular Strings and Time Series Averaging. arXiv:1804.02854 (2018)
8. Cuturi, M., Blondel, M.: Soft-DTW: a differentiable loss function for time-series. International Conference on Machine Learning (2017)
9. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proc. VLDB Endowment **1**(2), 1542–1552 (2008)
10. Dryden, I.L., Mardia, K.V.: Statistical shape analysis. Wiley, New York (1998)
11. Feragen, A., Lo, P., De Bruijne, M., Nielsen, M., Lauze, F.: Toward a theory of statistical tree-shape analysis. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2008–2021 (2013)
12. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging **23**(8), 995–1005 (2004)
13. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l',institut Henri Poincaré **10**, 215–310 (1948)
14. Ginestet, C.E.: Strong Consistency of Fré,chet Sample Mean Sets for Graph-Valued Random Variables. arXiv:1204.3183 (2012)

15. Hautamaki, V., Nykanen, P., Franti, P.: Time-series clustering by approximate prototypes. International Conference on Pattern Recognition (2008)
16. Jain, B.J.: Generalized gradient learning on time series. Mach. Learn. **100**(2-3), 587–608 (2016)
17. Jain, B.J.: Statistical analysis of graphs. Pattern Recogn. **60**, 802–812 (2016)
18. Jain, B.J., Schultz, D.: On the existence of a sample mean in dynamic time warping spaces. arXiv:1610.04460 (2016)
19. Jain, B.J., Schultz, D.: Asymmetric learning vector quantization for efficient nearest neighbor classification in dynamic time warping spaces. Pattern Recogn. **76**, 349–366 (2018)
20. Jain, B.: Revisiting Inaccuracies of Time Series Averaging under Dynamic Time Warping. Pattern Recogn. Lett. **125**, 418–424 (2019)
21. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bull. Lond. Math. Soc. **16**, 81–121 (1984)
22. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. Neurocomputing **21**(1-3), 19–30 (1998)
23. Liu, Y., Zhang, Y., Zeng, M.: Adaptive Global Time Sequence Averaging Method Using Dynamic Time Warping. IEEE Trans. Signal Process. **67**(8), 2129–2142 (2019)
24. Petitjean, F., Ketterlin, A., Gancarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recogn. **44**(3), 678–693 (2011)
25. Petitjean, F., Forestier, G., Webb, G.I., Nicholson, A.E., Chen, Y., Keogh, E.: Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. Knowl. Inf. Syst. **47**(1), 1–26 (2016)
26. Rabiner, L.R., Wilpon, J.G.: Considerations in applying clustering techniques to speaker-independent word recognition. J. Acoust. Soc. Am. **66**(3), 663–673 (1979)
27. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**(1), 43–49 (1978)
28. Schultz, D., Jain, B.: Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. Pattern Recogn. **74**, 340–358 (2018)
29. Soheily-Khah, S., Douzal-Chouakria, A., Gaussier, E.: Generalized k-means-based clustering for temporal data under weighted and kernel time warp. Pattern Recogn. Lett. **75**, 63–69 (2016)
30. Somervuo, P., Kohonen, T.: Self-organizing maps and learning vector quantization for feature sequences. Neural. Process. Lett. **10**(2), 151–159 (1999)
31. Sverdrup-Thygeson, H.: Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces. Ann. Stat. **9**(1), 141–145 (1981)
32. Tan, C.W., Webb, G.I., Petitjean, F.: Indexing and classifying gigabytes of time series under time warping. International Conference on Data Mining (2017)
33. Wilpon, J.G., Rabiner, L.R.: A Modified K-Means Clustering Algorithm for Use in Isolated Work Recognition
34. Ziezold, H.: On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions Random Processes and of the 1974 European Meeting of Statisticians. (1977)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.