Linear-trend normalization for multivariate subsequence similarity search

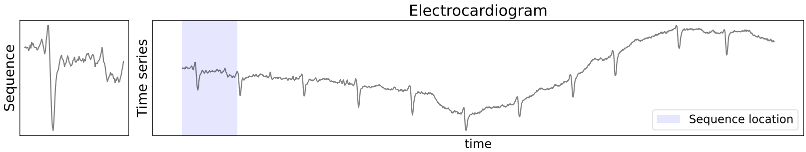
T. Germain[1,2], C. Truong[1,2], and L. Oudre[1,2]

[1]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190, Gif-sur-Yvette, France
[2]Université de Paris, CNRS, Centre Borelli, F-75005 Paris, France

MulTiSA, May 2024

Motivation

Similarity search: Searching for subsequences (S_i^w) in a large time series S similar to a query sequence Q based on a similarity/distance measure d.



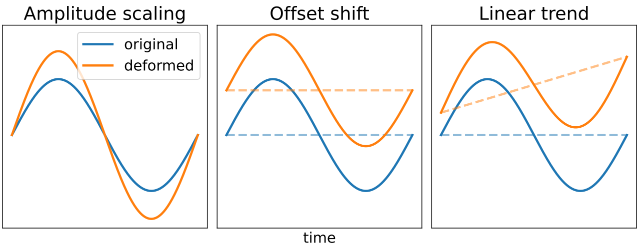
Desirable properties of the distance function d:

1. Invariance to some deformations: Let $G = \{g \mid g : \mathbb{R}^w \mapsto \mathbb{R}^w\}$ a group of deformations acting on $\mathbb{R}^w$, d is invariant to the action of G if for any $(S_1, S_2) \in \mathbb{R}^w \times \mathbb{R}^w$, $(g_1, g_2) \in G \times G$:

$$d(g_1(S_1), g_2(S_2)) = d(S_1, S_2)$$

2. Fast computation of distance profiles.

Elementary deformations



Main idea: Subsequences are normalized so that the Euclidean distance between their representation is invariant to the deformation.

Deformation	Group action	Normalization N $(d(x,y) = \|N(x) - N(y)\|)$
Amp. scaling	$g : x \mapsto \lambda x$, $\lambda \in \mathbb{R}_+^*$	$x \mapsto \frac{x}{\|x\|}$
Offset shift	$g : x \mapsto x + b1$ $b \in \mathbb{R},\ 1 = (1, \dots, 1)$	$x \mapsto x - \mu_x 1$ $s.t\ \mu_x = \text{mean}(x)$
Linear trend	$g : x \mapsto x + (at + b1)$ $(a,b) \in \mathbb{R}^2,\ t = (1, \dots, w)$	$x \mapsto x - (\alpha_x t + \beta_x 1)$ $s.t\ (\alpha_x, \beta_x) = \underset{(\alpha,\beta)}{\arg\min} \|x - (\alpha t + \beta 1)\|^2$

Z-normalization: Invariance

Definition (Euclidean Z-normalized distance)

The Euclidean Z-normalized distance between x and y is :

$$d_Z(x,y) = \left\| \frac{x - \mu_x 1}{\sigma_x} - \frac{y - \mu_y 1}{\sigma_y} \right\| = \sqrt{w} \left\| \frac{x - \mu_x 1}{\|x - \mu_x 1\|} - \frac{y - \mu_y 1}{\|y - \mu_y 1\|} \right\|$$

Proposition (Invariance)

*d_Z is invariant to **amplitude scaling** and **offset shift**.*

Z-normalization: Fast computation

Proposition

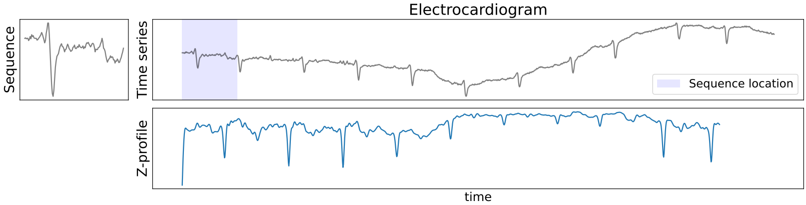
The computation of the Z-distance profile between $Q \in \mathbb{R}^w$ and $S \in \mathbb{R}^n$, $(w < n)$, is in $\mathcal{O}(n\log(n))$.

Proposition

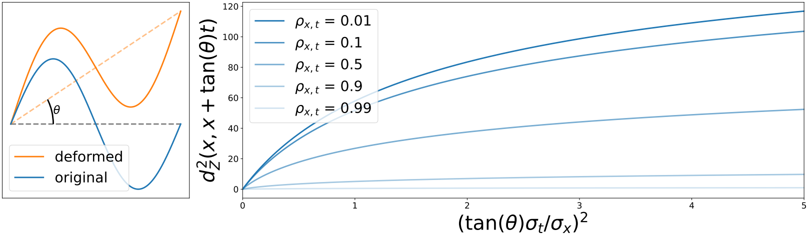
The Z-normalized distance between x and y can be written as:

$$d_Z(x,y) = \sqrt{2\left(w - \frac{\langle x,y\rangle - w\mu_x\mu_y}{\sigma_x\sigma_y}\right)}$$

Z-normalization: Running example



Influence of linear trend on the Euclidean Z-normalized distance:



LT-normalization: Invariance

Definition

The Euclidean LT-normalized distance between x and y is:

$$d_{LT}(x,y) = \left\| \frac{x - (\alpha_x t + \beta_x 1)}{\|x - (\alpha_x t + \beta_x 1)\|} - \frac{y - (\alpha_y t + \beta_y 1)}{\|y - (\alpha_y t + \beta_y 1)\|} \right\|$$

where $t = (1, \ldots, w)$ and (α_x, β_x) are solutions of the least square problem: $\arg\min_{(\alpha,\beta)} \|x - (\alpha t + \beta 1)\|^2$

Proposition

*d_{LT} is invariant to **amplitude scaling**, **offset shift**, and **linear trend**.*

LT-normalization: Fast computation

Proposition

The LT-normalized distance between x and y can be written as:

$$d_{LT}(x, y) = \sqrt{2\left(1 - \frac{\langle x, y\rangle - w(\mu_x\mu_y + \alpha_x\alpha_y\sigma_t^2)}{\eta_x\eta_y}\right)}$$

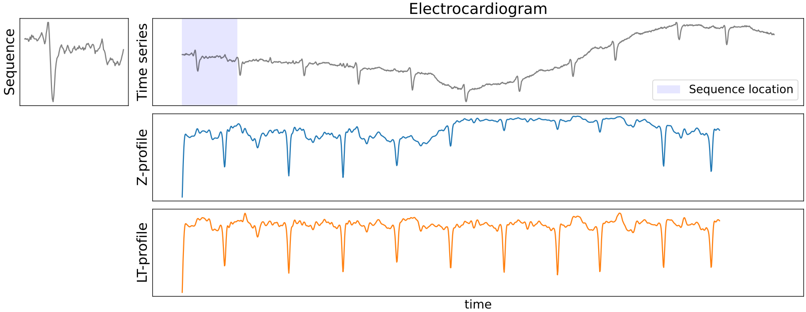
where:

$$\begin{cases} \eta_x = \|x - (\alpha_x t + \beta_x 1)\| \\ \beta_x = \mu_x - \alpha_x\mu_t \\ \alpha_x = \mathrm{cov}(x, t)/\sigma_t^2 = \frac{1}{w}(\langle x, t\rangle - \mu_x\mu_t)/\sigma_t^2 \end{cases}$$

Proposition

The computation of the LT-distance profile between $Q \in \mathbb{R}^w$ and $S \in \mathbb{R}^n$, $(w < n)$, is in $\mathcal{O}(2n\log(n))$.

LT-normalization: Running example



LT-normalization: Extension to multivariate time series

Definition

The multivariate LT-normalized distance between $x \in \mathbb{R}^{d \times w}$ and $y \in \mathbb{R}^{d \times w}$ is:

$$d_{MLT}(x, y) = \sqrt{\frac{1}{d} \sum_{k=1}^{d} d_{LT}^2(x^{(k)}, y^{(k)})}$$

where $x^{(k)}$ is the k^{th} dimension of the time series.

Remark

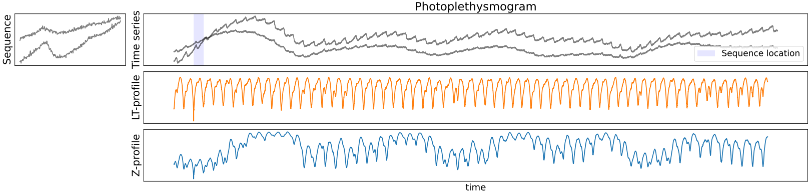
We considered an uniform averaging over dimensions. Other aggregating method can be considered[1]*.*

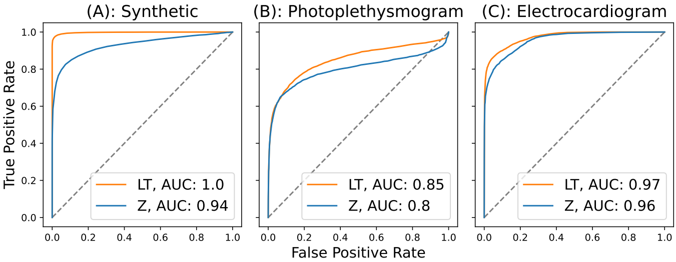
[1] Chin-Chia Michael Yeh, Nickolas Kavantzas, and Eamonn Keogh. "Matrix profile VI: Meaningful multidimensional motif discovery". In: *2017 IEEE international conference on data mining (ICDM)*. IEEE. 2017, pp. 565–574.

Experiment 1: Similarity search

Algorithm: Fast similarity search[2]

Distances: Z-normalized (Z), LT-normalized (LT)





[2] Abdullah Mueen et al. *The Fastest Similarity Search Algorithm for Time Series Subsequences under Euclidean Distance*. Aug. 2022.

Experiment 2: Motif set discovery

Motif set algorithm: STOMP[3] (matrix profile).
Distances: Euclidean (Euc), Z-normalized (Z), LT-normalized (LT), Trend removal[4] & Z-normalized (STL+Z).
Metric: f1-score.

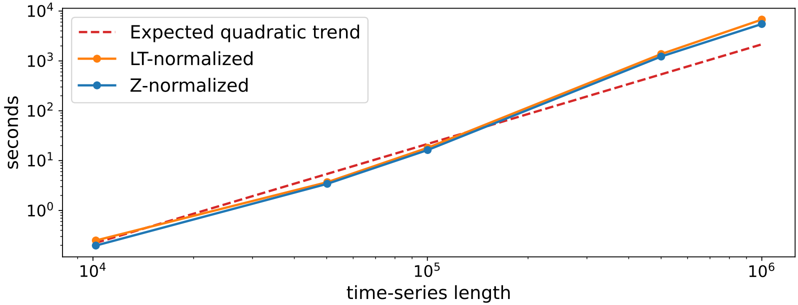
distance / dataset	Euc	STL+Z	Z	LT
s-search	0.20	0.86	**0.87**	0.86
m-set	0.25	0.62	0.62	**0.62**
mitdb1	0.42	0.54	0.50	**0.58**
mitdb2	0.16	0.44	0.43	**0.45**
ptt-ppg	0.54	**0.58**	0.53	0.57
arm-coda	0.25	**0.26**	0.25	0.25

[3] Yan Zhu et al. "Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins". In: *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE. 2016, pp. 739–748.

[4] Robert B Cleveland et al. "STL: A seasonal-trend decomposition". In: *J. Off. Stat* 6.1 (1990), pp. 3–73.

Experiment 3: Scalability

Scalability of STOMP algorithm (matrix profile) with the time series length for Z-normalized (blue) and LT-normalized (orange) distances.



Thank you