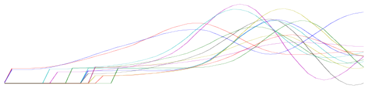
Kernelized time elastic averaging of time series

P-F. Marteau

EXPRESSION/IRISA/UBS

TS-Days/IRISA 25-26 mars 2019, Rennes



Plan

1. Time series
   - Definitions/Notations (among others)
   - Some examples
   - Time series averaging problem
   - Time series averaging problem
2. A brief history of time elastic matching
   - Maurice Fréchet
   - Richard Bellman
   - Dynamic Time Warping and some variants
   - Time elastic averaging of a set of time series with DTW
3. Kernalized Time elastic averaging of a set of time series
   - Kernel and definiteness
   - Constructing Positive Definite Time Elastic Kernels
   - Probabilistic interpretation of kernalized time elastic averaging
4. Applications
5. Conclusion

Plan

1. Time series
   - Definitions/Notations (among others)
   - Some examples
   - Time series averaging problem
   - Time series averaging problem
2. A brief history of time elastic matching
3. Kernalized Time elastic averaging of a set of time series
4. Applications
5. Conclusion

Definitions/Notations

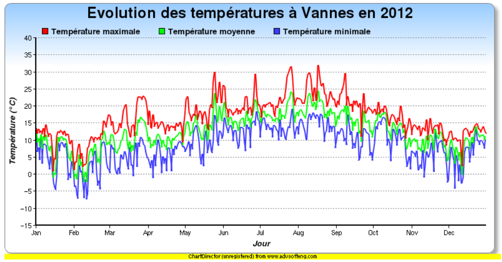
TIME SERIES are sequence of **time stamped** data,

- Let $\mathbb{U}$ be the (discrete) time series data set,
- $A_1^n = A(1)A(2)...A(n) \in \mathbb{U}$ is a finite (discrete) time series of size n
- $\forall i, A(i) = (a(i), t_{a(i)}) \in S \times T$, where S is a *space* set (set of spatial dimensions, either digital or symbolic) and T is a *time* set (an ordered set of timestamps),
- we denote by Ω the *empty* time series (as well as the empty sample),
- thus $\mathbb{U} = \bigcup_{n=1}^{\infty} (S \times T)^n \bigcup \{\Omega\}$.

Some examples

Sequential data are ubiquitous and heterogeneous

- multivariate



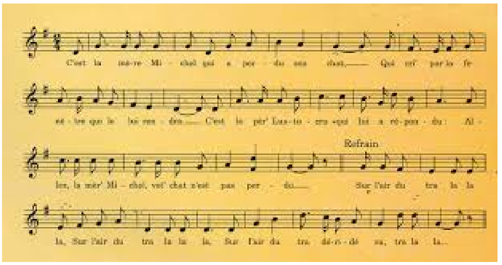
Some examples

Sequential data are ubiquitous and heterogeneous

- multivariate
- image

Some examples

Sequential data are ubiquitous and heterogeneous

- multivariate
- image
- symbolic

Some examples

Sequential data are ubiquitous and heterogeneous

- multivariate
- image
- symbolic
- text

Genome sequencing is often compared to "decoding," but a sequence is still very much in code. In a sense, a genome sequence is simply a very long string of letters in a mysterious language.
When you read a sentence, the meaning is not just in the sequence of the letters. It is also in the words those letters make and in the grammar of the language. Similarly, the human genome is more than just its sequence. [http://www.genomenewsnetwork.org]

Averaging a set of time series

Why would we consider averaging time series in the first place?

- Green computing
- Clustering
- Noise reduction
- Study the variance and the individual deviation (model temporal data)

Averaging a set of time series

The problem Let $S \subset \mathbb{U}$ be a subset of time series.

Let $\delta(.,.)$ a metric defined on $\mathbb{U}$.

The centroid time series C of S is defined as:

$$C_\delta = \underset{u \in \mathbb{U}}{argmin} \sum_{s \in S} \delta(u, s)$$

$\Rightarrow$ **What choices for δ ?**

Matching time series

Importance of time series matching $\delta(.,.)$

- Comparing time series is also an ubiquitous task in particular to detect similar patterns, predict the future from the past, cluster, classify or average temporal data, basically to extract knowledge.
- Unfortunately, in general, time series exhibit a high level of variability due to noisy measurements, noise intrinsic to the observed process, missing data, non-uniform sampling, time warp, etc.

$\Rightarrow$ Going beyond Eulidean distance while introducing ”time elasticity” is thus a long story.

Plan

1 Time series

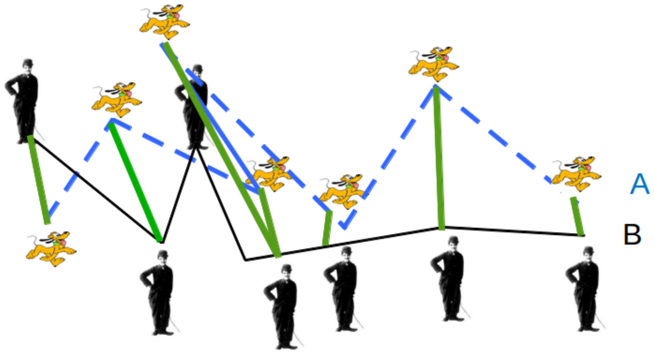
2 A brief history of time elastic matching
- Maurice Fréchet
- Richard Bellman
- Dynamic Time Warping and some variants
- Time elastic averaging of a set of time series with DTW

3 Kernalized Time elastic averaging of a set of time series

4 Applications

5 Conclusion

Maurice Fréchet



1878 – 1973

- The Fréchet's distance between the two curves is the length of the shortest possible leash. [From Wikipedia]

- $F(A,B) = \underset{\alpha,\beta}{Inf}\ \underset{i\in[0,N]}{Max} \left\{ d\Big(A(\alpha(i)),\ B(\beta(i))\Big) \right\}$
- $\forall i,\ \alpha(i) \leq \alpha(i+1)$ and $\beta(i) \leq \beta(i+1)$
- THE FRÉCHET'S DISTANCE (Fr'echet (1906)) between two curves is the minimum length of a leash required to connect a dog and its owner, constrained on two separate paths, as they walk freely but without backtracking along their respective curves from one endpoint to the other.

Combinatorics

The size of the search space $(\alpha(.), \beta(.))$ is directly related to the **Number of paths** in a $n \times m$ grid:

Delannoy's numbers $D(n,m)$

Asymptotic behavior $D(n,n) = \frac{c\gamma^n}{\sqrt{n}}(1 + O(n^{-1}))$

with $\gamma = 3 + 2\sqrt{2} \approx$ **5.828** and $c \approx 0.5727$

$D(n,n)_{n=1,2,\cdots} = 1, 3, 13, 63, 321, 1683, 8989, 48639, 265729, \cdots$ (sequence A001850 in the OEIS).

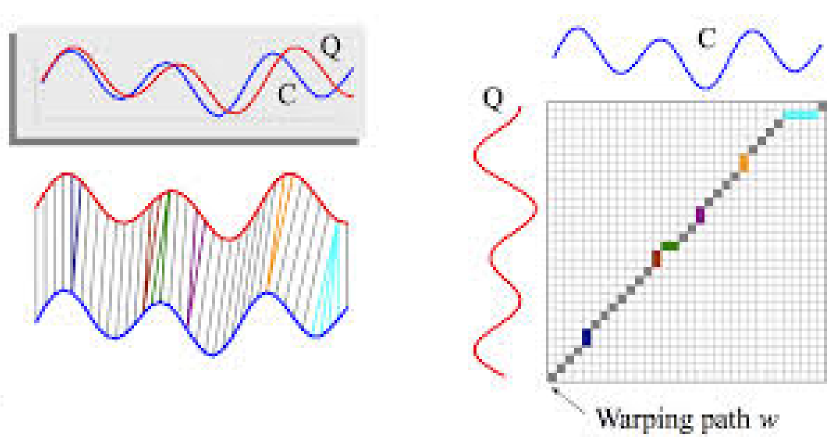
Richard Bellman

BELLMAN'S PRINCIPLE OF OPTIMALITY: An optimal policy has the property that, whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the initial decision [Wikipedia].

1920-1984

- Principle of Optimality
- Dynamic Programming $\approx$ 1949-57 (Bellman (1957))
- Application to optimal control

Dynamic time warp alignment



from andrew.cmu.edu

Evaluated in $\mathbf{O}(\mathbf{n}^2)$

Dynamic Time Warping

$$\delta_{dtw}(A_1^p, B_1^q) = \min_{\pi} \sum_i d_{LP}(a_{\pi(i)_1}, b_{\pi(i)_2}) \quad (1)$$

$$\delta_{dtw}(A_1^p, B_1^q) = d_{LP}(a_p, b_q) + Min \begin{cases} \delta_{dtw}(A_1^{p-1}, B_1^q) \text{ deletion} \\ \delta_{dtw}(A_1^{p-1}, B_1^{q-1}) \text{ substitution} \\ \delta_{dtw}(A_1^p, B_1^{q-1}) \text{ insertion} \end{cases} \quad (2)$$

where $d_{LP}(a_p, b_q)$ is the Lp norm in $\mathbb{R}^k$.

- SPEECH RECOGNITION Velichko & Zagoruyko (Velichko and Zagoruyko (1970))
- WITH CORRIDORS Sakoe & Chiba (Sakoe and Chiba (1971))
- LOWER BOUNDING Keogh & al., (Keogh et al. (2006))

Some variants

For sequential data

- GLOBAL ALIGNMENT Needleman & Wunsch (Needleman and Wunsch (1970))
- EDIT DISTANCE: Wagner & Fisher (Wagner and Fischer (1974))
- THE LONGEST COMMON SUBSEQUENCE Hirschberg (Hirschberg (1975))
- LOCAL ALIGNMENT Smith & Waterman (Smith and Waterman (1981))
- ...

For time series

- EDIT DISTANCE WITH REAL PENALTY (Chen & Ng Chen and Ng (2004))
- TIME WARP EDIT DISTANCE Marteau (Marteau (2008))
- ...

General form of an elastic distance

$$\delta_e(A_1^p, B_1^q) = \text{Min/Max} \begin{cases} \delta_e(A_1^{p-1}, B_1^q) + \Gamma(A(p) \to \Omega_B(q)) & \textit{deletion} \\ \delta_e(A_1^{p-1}, B_1^{q-1}) + \Gamma(A(p) \to B(q)) & \textit{substitution} \\ \delta_e(A_1^p, B_1^{q-1}) + \Gamma(\Omega_A(p) \to B(q)) & \textit{insertion} \end{cases}$$

where $\Gamma(.)$ is the cost/gain of an elementary editing operation and $\Omega_{X(i)}$ is the *empty* symbol at position i of sequence X

- $A(p) \to \Omega_{B(q)}$ is interpreted as a deletion operation
- $A(p) \to B(q)$ is a substitution operation
- $\Omega_{A(p)} \to B(q)$ is interpreted as an insertion operation

Time elastic averaging of a set of time series with DTW

$$C_{\delta_{dtw}} = \underset{u \in \mathbb{U}}{argmin} \sum_{s \in S} \delta_{\mathbf{dtw}}(u, s) \ ?$$

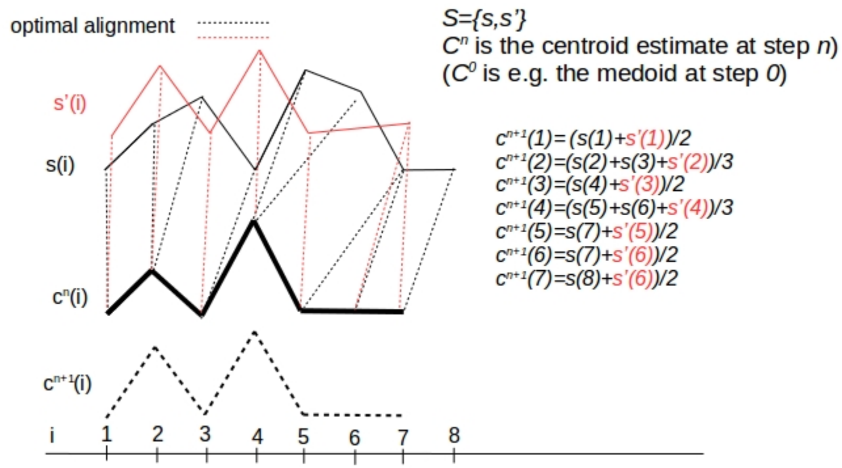
Problem complexity

- Multiple alignments have been widely studied in bioinformatics (Fasman and L. (1998)).
- determining the optimal alignment of a set of sequences under the sum of all pairs score scheme is a **NP-complete** problem (Wang and Jiang (1994), Just and Just (1999)).

$\Rightarrow$ optimal solution cannot be found in reasonable time for medium/large problems.

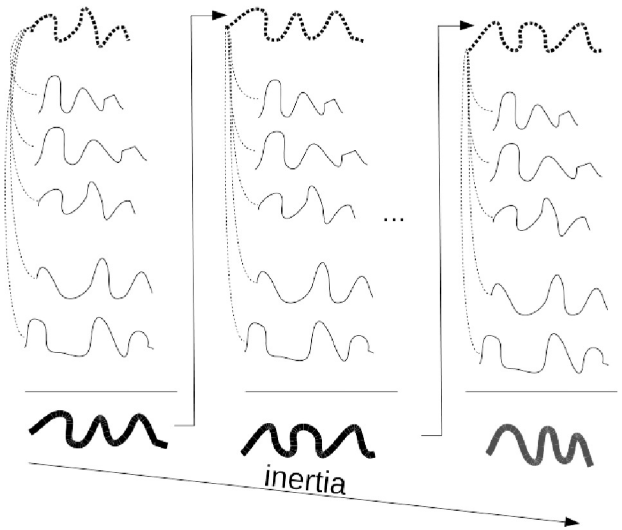
$\Rightarrow$ **heuristic solutions**.

Time elastic averaging of a set of time series with DTW



Principle of **DTW Barycenter Averaging** (DBA) (Abdulla et al. (2003), Hautamaki et al. (2008), Petitjean et al. (2014))

Time elastic averaging of a set of time series with DTW



Iterative aggregation: DTW Barycenter Averaging (**DBA**) (Abdulla et al. (2003), Hautamaki et al. (2008), Petitjean et al. (2014))

Time elastic averaging of a set of time series with DTW

Other approaches

- Hierarchical ascendant agglomerative approach (Niennattrakul and Ratanamahatana (2009)),
- Canonical Time Warp (CTW) and a Generalized version of it (GCTW) (Zhou and la Torre (2016)) that combines DTW and CCA (Canonical Correlation Analysis,
- and more ...

Time elastic averaging of a set of time series with DTW

Criticism

Inaccuracies of the proposed heuristics due to ”hardness of the problem (Niennattrakul and Ratanamahatana (2007))

- Best alignment path ⇒ lack of smoothness of the objective function (lot of local minima)
- DTW is not a metric (triangle inequality is missing). Impact?

⇒ Kernelization of time elastic distance, at least to **smooth the objective function**.

Plan

1 Time series

2 A brief history of time elastic matching

3 Kernalized Time elastic averaging of a set of time series
- Kernel and definiteness
- Constructing Positive Definite Time Elastic Kernels
- Probabilistic interpretation of kernalized time elastic averaging

4 Applications

5 Conclusion

Kernelization of elastic distance

From elastic distances to elastic kernels

Given the importance of kernel approaches in machine learning, we are led to consider the following questions:

- can we derived elastic kernels from elastic distances such as DTW?
- if not, how can we construct kernels *sufficiently closely* related to such distances such as to preserve their specific properties?

Kernel and definiteness

Definitions (Schoenberg (1938))

- Let U be a non empty set. A function $k : U \times U \rightarrow \mathbb{R}$ is called a positive (P.D.) definite kernel if and only if it is
  1. symmetric
  2. $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for all n in $\mathbb{N}$, $(x_1, x_2, ..., x_n) \in U^n$ and $(c_1, c_2, ..., c_n) \in \mathbb{R}^n$.
- For P.D. kernels, the **eigen values** of any gram matrix are all **not negative**.

Kernel and definiteness

Properties

Closure

- Sums of P.D. kernels defined on the same set are P.D. kernels.
- Product of P.D. kernels defined on the same set are P.D. kernels.

Mapping between spaces

Let U and $\tilde{U}$ two sets and define a map $\varphi(.) : U \rightarrow \tilde{U}$. If k is a P.D. kernel defined on $\tilde{U}$, then $k(\varphi(u), \varphi(u'))$ is a P.D. kernel on U.

Kernel and definiteness

Benefice of positiveness

- Allows to embed data in (high dimensional) **inner vector spaces** (Reproducing Kernel Hilbert Space)
- Gives access to a large family of **Kernel approaches** (K-PCA, K-LDA, K-ICA, Spectral Clustering, SVM, etc.)
- P.D. ensures that learning with kernel machines relates to a quadratic **convex** problem (convergence toward a single optimum)

Kernel and definiteness

Distance substitution kernels (DK) (Haasdonk and Bahlmann (2004)) are kernels composed from a distance (dissimilarity) function. Let d be a dissimilarity or distance function and O an origin element in set U, then the following quantities are DK:

- $k_l(x,y) = \langle x, y \rangle_d^O = -\frac{1}{2}(d(x,y)^2 - d(x,O)^2 - d(y,O)^2)$
- $k_P(x,y) = (1 + \gamma \langle x, y \rangle_d^O)^p, \forall p \in \mathbb{N}, \forall \gamma \in \mathbb{R}^+$
- $k_{nd}(x,y) = -d(x,y)^\beta, \beta \in [0,2]$
- $k_{rbf}(x,y) = exp(-\gamma d(x,y)^2), \forall \gamma \in \mathbb{R}^+$

If k_l is P.D., then k_p, k_{rbf} are P.D. and k_{nd} is C.P.D.

Unfortunately DK constructed from an elastic distance (EDK) are not PD.

Regularizing the Gram matrix

Spectral methods attempt to directly modify the *Gram* matrix $K(i,j)$ obtained from non P.D. kernels (Wu et al. (2005), Chen et al. (2009)) by:

- **changing the sign** of the negative eigen values (flipping)
- or **shifting** the set of eigen values by a minimal offset to make it D.P.
- Then the Gram matrix is **reconstructed** from the initial eigen vectors and the new set of eigen values to get a D.P. matrix.

Other approaches: replace the Gram matrix by the **closest** (Froebonius norm) P.D. matrix (Higham2002).

$\Rightarrow$ These spectral approaches are difficult to interpret and do not show significant benefits (to my experience).

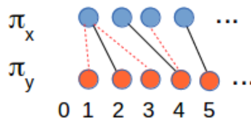
A more direct and constructive approach

A conjecture is that the presence of the *Min* or *Max* operators prevents the definiteness.

An approach to regularize EDK is to replace these *Min* or *Max* operators by a summation **(soft-min/max)** operator leading to cope with all the possible alignment paths instead of a single best one.

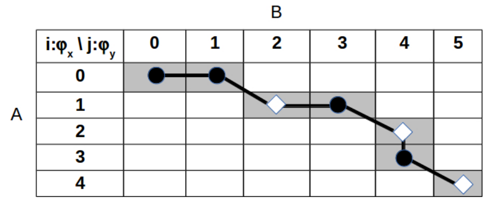
- **String alignment kernel** for protein homology Saigo et al. (2004)
- **Global alignment kernel** (Cuturi et al. (2007))
- Regularized Edit Distance Kernels (REDK) (Marteau and Gibet (2014)) ⇒ **Kernel induced by an alignment map**.

Alignment map



$\boldsymbol{\pi}$ = (0,0) (0,1) **(1,2)** (1,3) **(2,4)** (3,4) **(4,5)** ...

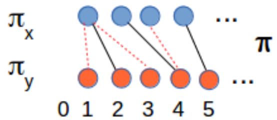
— match
...... deletion/insertion



An ordered alignment map, π, is a finite **sequence of ordered pairs** of integers $\pi(l) = (i_l, j_l)$ satisfying

- $i_l \leq i_{l-1} + 1$ and $j_l \leq j_{l-1} + 1, \forall l \in 1, .., |\pi| - 1$
- $i_{l-1} < i_l$ or $j_{l-1} < j_l, \forall l \in \{1, .., |\pi| - 1\}$

Kernel induced by an alignment map



$\boldsymbol{\pi}$ = (0,0) (0,1) **(1,2)** (1,3) **(2,4)** (3,4) **(4,5)** ...

— match
...... deletion/insertion

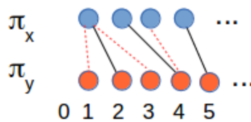
Any mapping π uniquely induces two projections in $\mathbb{U}_\pi = ((S \times T) \cup \{\Omega\})^{|\pi|}$, basically two vectorized (fixed length) representations for any $A, B \in \mathbb{U}$:

$\varphi_{\pi_x} : \mathbb{U} \to \mathbb{U}_\pi$ and $\varphi_{\pi_y} : \mathbb{U} \to \mathbb{U}_\pi$

$\varphi_{\pi_x}(A) = \Omega_{A(1)}$ **A(1)** $\Omega_{A(2)}$ **A(2)** A(3) **A(4)** $\cdots$
$\varphi_{\pi_y}(A) =$ A(1) **A(2)** A(3) **A(4)** $\Omega_{A(4)}$ **A(5)** $\cdots$
$\varphi_{\pi_x}(B) = \Omega_{B(1)}$ **B(1)** $\Omega_{B(2)}$ **B(2)** B(3) **B(4)** $\cdots$
$\varphi_{\pi_y}(B) =$ B(1) **B(2)** B(3) **B(4)** $\Omega_{B(4)}$ **B(5)** $\cdots$

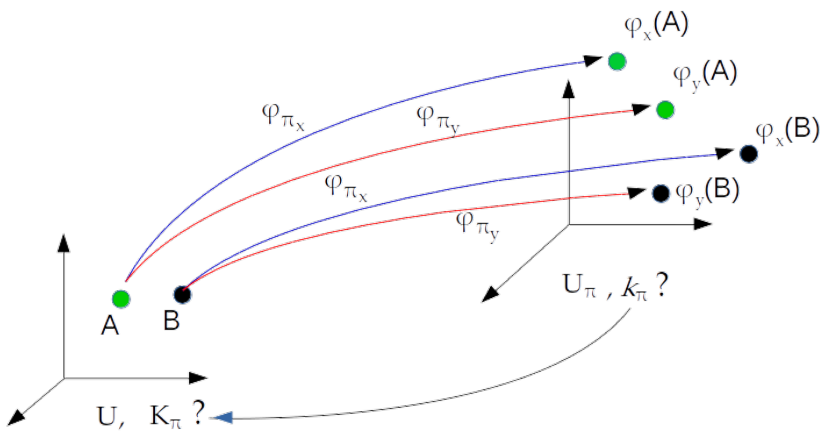
where $\Omega_{X(i)}$ is the *empty* replacement symbol in X at position i

Kernel induced by an alignment map

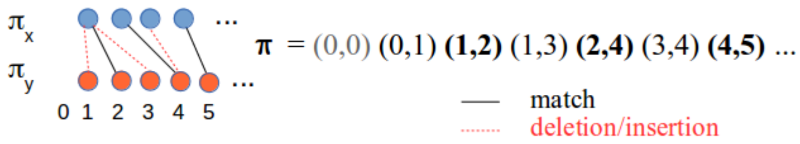


$\boldsymbol{\pi}$ = (0,0) (0,1) **(1,2)** (1,3) **(2,4)** (3,4) **(4,5)** ...

— match
...... deletion/insertion



Kernel induced by an alignment map



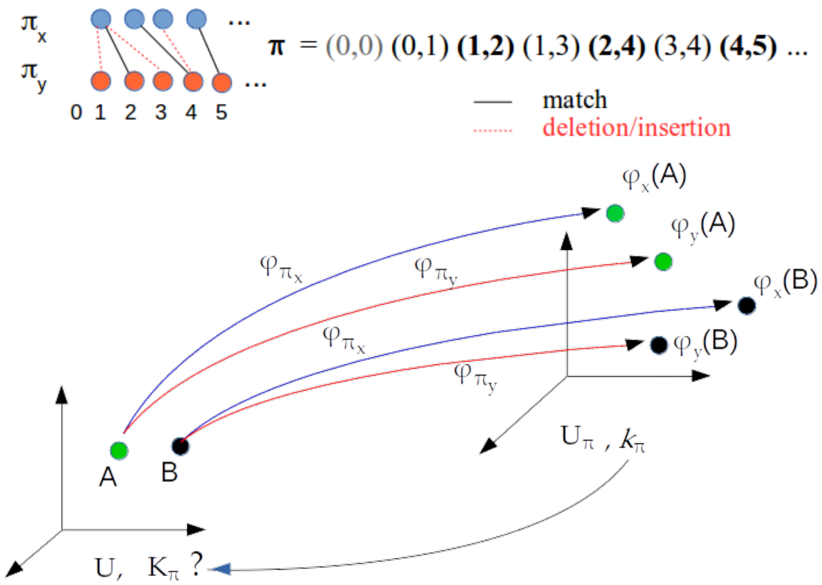
- $\varphi_{\pi_x}(A) = \Omega_{A(1)}$ **A(1)** $\Omega_{A(2)}$ **A(2)** A(3) **A(4)** $\cdots$
- $\varphi_{\pi_y}(B) =$ B(1) **B(2)** B(3) **B(4)** $\Omega_{B(4)}$ **B(5)** $\cdots$

Suppose that a **local alignment P.D. kernel** κ exists (for all editing operation) then, the following kernel $k_\pi(.,.)$ is P.D. on $\mathbb{U}_\pi$,
$k_\pi(\varphi_{\pi_x}(A), \varphi_{\pi_y}(B)) = \kappa(\Omega_{A(1)} \to B(1))\kappa(\mathbf{A(1)} \to \mathbf{B(2)})$

$$\kappa(\Omega_{A(2)} \to B(3))\ \kappa(\mathbf{A(2)} \to \mathbf{B(4)})\ \kappa(A(3) \to \Omega_{B(4)})$$
$$\kappa(\mathbf{A(4)} \to \mathbf{B(5)}) \cdots$$

(Idem for $k_\pi(\varphi_{\pi_x}(A), \varphi_{\pi_x}(B))$, $k_\pi(\varphi_{\pi_y}(A), \varphi_{\pi_y}(B))$, and $k_\pi(\varphi_{\pi_y}(A), \varphi_{\pi_x}(B))$).

Kernel induced by an alignment map

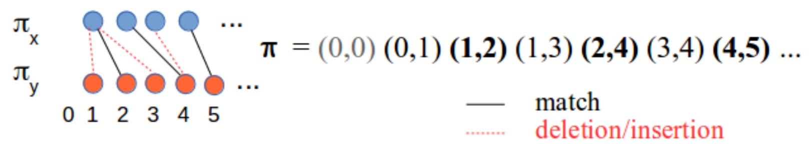


Regularizing EDK: R-Convolution theorem

David Haussler - (Haussler (1999))

- $K(x,y) = \sum_{\vec{x} \in R^{-1}(x)} \sum_{\vec{y} \in R^{-1}(y)} k(\vec{x}, \vec{y})$
- $R^{-1}(x)$ and $R^{-1}(y)$ are respectively the set of parts of x and y.
- $K(x,y)$ is P.D. iff $k(\vec{x}, \vec{y})$ is P.D. (Haussler (1999))

Kernel induced by an alignment map



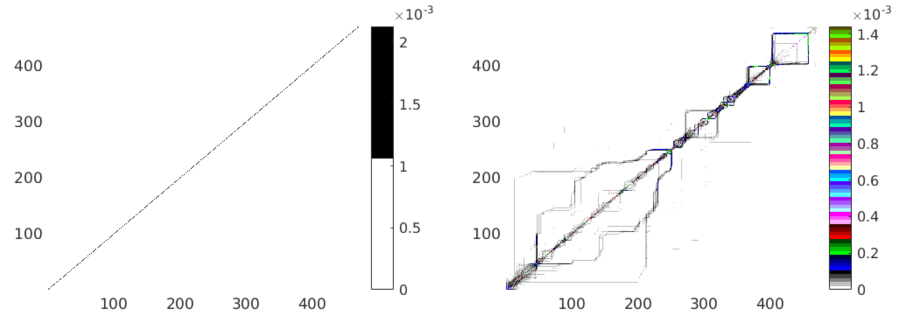
Given any mapping π and any kernel k_π defined on $\mathbb{E}_\pi$, one can defines a symmetric kernel $K_\pi(A,B)$ on $\mathbb{U}^2$ as:

- $K_\pi(A,B) = \sum_{\varphi(A) \in \mathcal{P}_\pi(A)} \sum_{\varphi(B) \in \mathcal{P}_\pi(B)} k_\pi(\varphi(A), \varphi(B))$
  where $\mathcal{P}_\pi(X) = \{\varphi_{\pi_x}(X), \varphi_{\pi_y}(X)\}$ are the set of parts of sequence X.
- $K_\pi(A,B)$ is a R-convolution kernel (Haussler (1999)).

Summary of the results

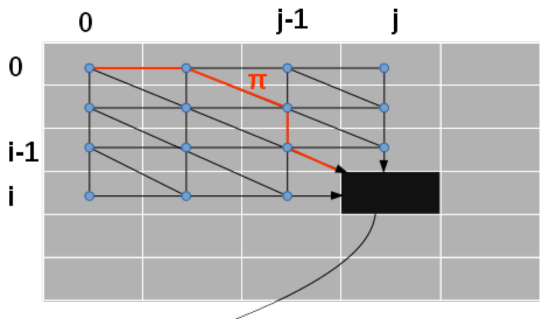
1. If **local kernel** $\kappa(x, y) = f(x \to y)$ is P.D. on $(S \times T) \cup \{\Omega\}$ then $k_\pi(X, Y) = \prod_i \kappa(X_i \to Y_i)$ is P.D. on $\mathbb{U}_\pi$, $\forall \pi \in \Pi$
2. If $k_\pi(.,.)$ is P.D. on $\mathbb{U}_\pi$, then the **R-convolution kernel** $K_\pi(.,.)$ is P.D. on $\mathbb{U}$.
3. If $K_\pi(.,.)$ is P.D. on $\mathbb{U}$ for all $\pi \in C$, then
$$\mathcal{K}_{\mathbf{C}}(.,.) = \sum_{\pi \in C \subseteq \Pi} K_\pi(.,.) \text{ is P.D. on } \mathbb{U}.$$

Property : for $\mathcal{K}_C(.,.)$, the "corridor" C does not need to be dense



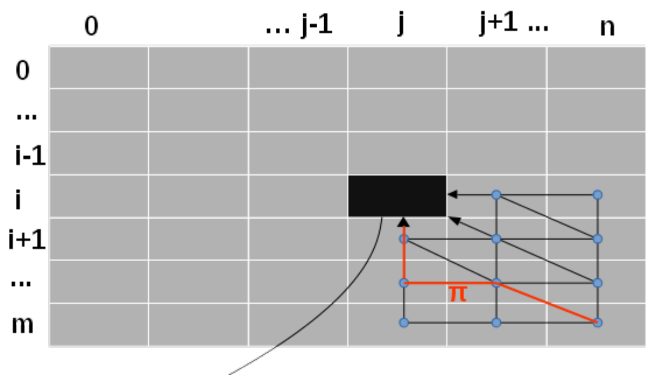
UCR Beef dataset: left Sakoe-Chiba 'optimal' corridor, right all the best DTW alignment paths (symmetrized), (Soheily-Khah and Marteau (2019)).

Exponential local kernel and probabilistic interpretation of global kernel (Forward probability)



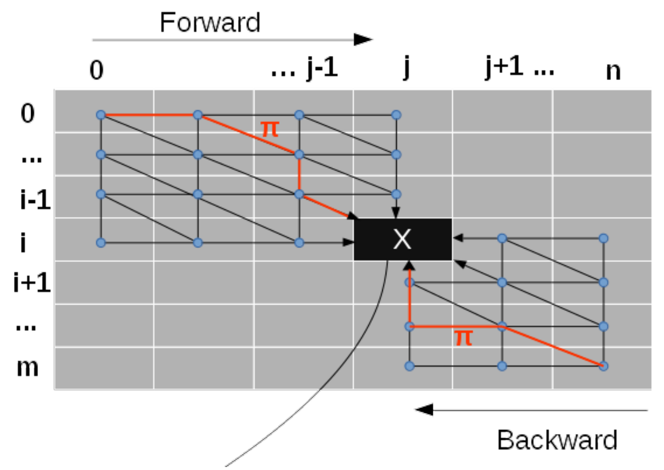
$$\mathcal{K}_C(X_0^i, Y_0^j) \approx \sum_{\pi \in C \subseteq \Pi} P_\pi(X_0^i, Y_0^j)$$
$$K_\pi(X_0^i, Y_0^j) \approx \sum_{\varphi, \varphi'} \prod_u p(\varphi(X)_u, \varphi'(Y)_u))) \approx P_\pi(X_0^i, Y_0^j)$$
$$\kappa(x \to y) = exp(-\nu \cdot (d(x, y))) \approx p(x, y)$$

Exponential local kernel and probabilistic interpretation of global kernel (Backward probability)



$$\mathcal{K}_C(X_n^i, Y_m^j) \approx \sum_{\pi \in C \subseteq \Pi} P_\pi(X_n^i, Y_m^j)$$
$$K_\pi(X_n^i, Y_m^j) \approx \sum_{\varphi, \varphi'} \prod_u p(\varphi(X)_u, \varphi'(Y)_u))) \approx P_\pi(X_n^i, Y_m^j)$$
$$\kappa(x, y) = exp(-\nu \cdot (d(x, y))) \approx p(x, y)$$

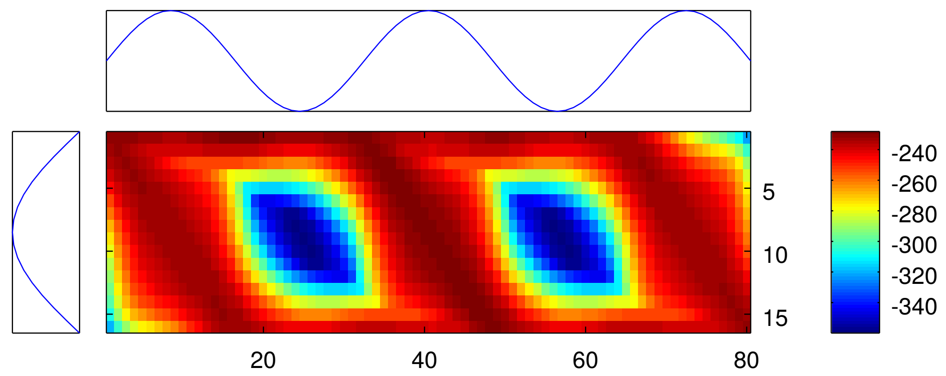
Exponential local kernel and probabilistic interpretation of global kernel (Forward-Backward probability)



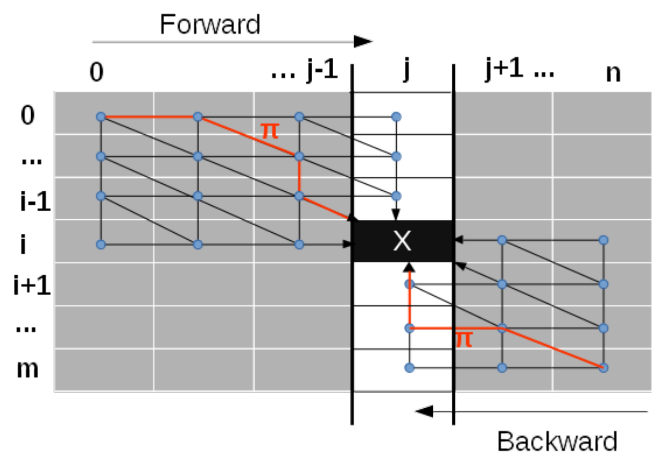
$$\mathcal{FB}_{X,Y}(i,j) \approx \sum_{\pi \in \mathcal{C} \subseteq \Pi} P_\pi(X_0^i, Y_0^j) \times P_\pi(X_n^i, Y_m^j)$$

Sum of the probabilities of the global alignment paths that cross cell (i,j).

Forward-Backward alignment matrix example



The utility of the Forward-Backward



$P_{i|j}$: Probability to align sample *i* given sample *j*

$$P_{i|j} \approx \frac{\mathcal{FB}_{X,Y}(i,j)}{\sum_t \mathcal{FB}_{X,Y}(t,j)}$$

The utility of the Forward-Backward

- Expectation and Standard deviation of the **samples of X that are aligned** with sample **Y(j)**:
  $Ex(x|Y(j)) \approx \sum_{t=1}^{m} X(t) \cdot P_{t|j}$
  $St(x|Y(j)) \approx \sqrt{\frac{\sum_{t=1}^{m}(X(t) - Ex(x|Y(j))^2 \cdot P_{t|j}}{m-1}}$
- Expectation and Standard deviation of the **time (index)** of occurrences of the **samples of X that are aligned with sample Y(j)**:
  $Ex(t'|Y(j)) \approx \sum_{t=1}^{m} t \cdot P_{t|j}$
  $St(t'|Y(j)) \approx \sqrt{\frac{\sum_{t=1}^{m}(t - Ex(t'|Y(j))^2 \cdot P_{t|j}}{m-1}}$

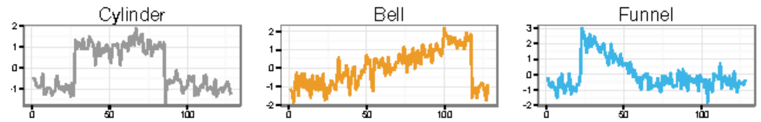
Averaging a set of time series (iterative approach)

TEKA ALGORITHM

Let $\mathcal{Y} = \{Y_1, Y_2, \cdots, Y_N\}$ a set of time series and C_0 **an initial centroid** (the medoid of $\mathcal{Y}$ for instance).

- Centroid value at index t after q iterations: $C_q(t) = 1/N \sum_{k=1}^{N} Ex(y_k|C_{q-1}(t))$
- Time of occurrence of centroid value at time index t after q iterations: $\mathcal{T}_q(t) = 1/N \sum_{k=1}^{N} Ex(t_k|\mathcal{T}_{q-1}(t))$
- Similar derivations for the standard deviations.
- Needs interpolation (resampling) to get a uniform sampling.

Cylinder/Bell/Funnel example

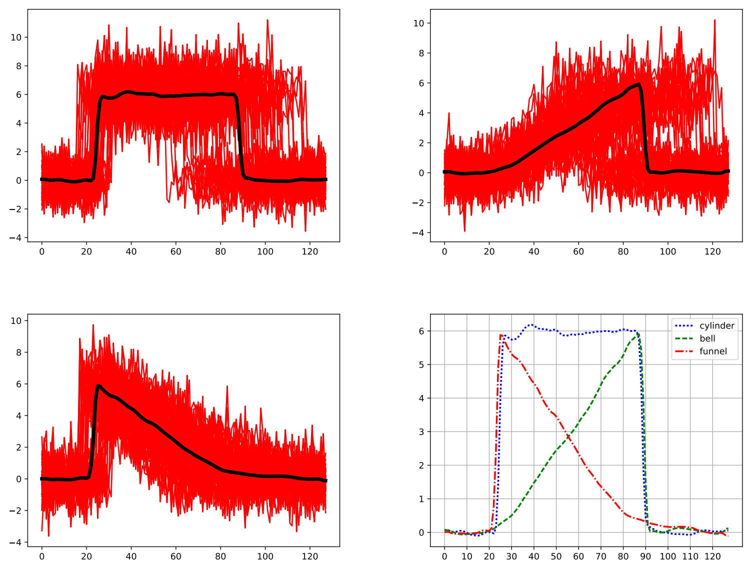


$c(t) = (6+\eta)\cdot\chi_{[a,b]}(t) + \epsilon(t)$
$b(t) = (6+\eta)\cdot\chi_{[a,b]}(t)\cdot(t-a)/(b-a) + \epsilon(t)$
$f(t) = (6+\eta)\cdot\chi_{[a,b]}(t)\cdot(b-t)/(b-a) + \epsilon(t)$

$\chi_{[a,b]} = 0$ if $t < a \vee t > b$, 1 if $a \leq t \leq b$,
η and $\epsilon(t)$ are $N(0,1)$,
a is uniformly drawn from $[16, 32]$,
and $(b-a)$ is uniformly drawn from $[32, 96]$.

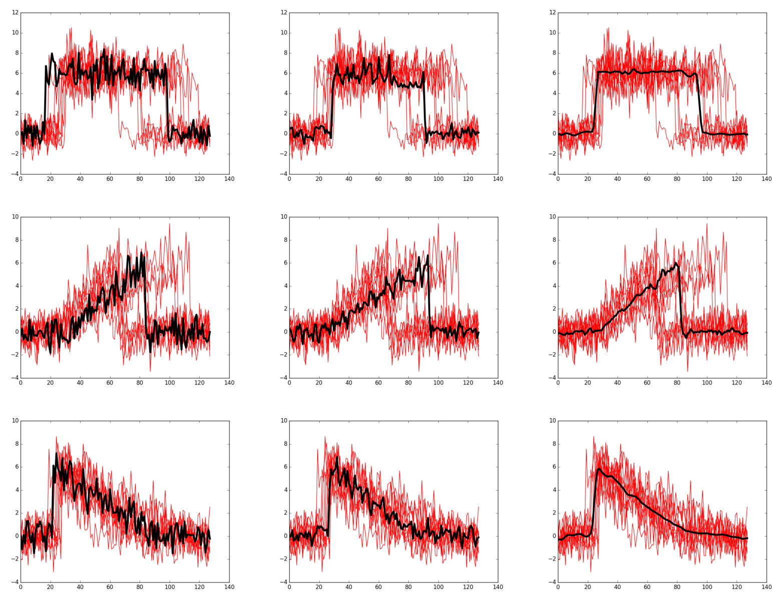
$\rightarrow$ expected start and end time stamps are respectively 24 and 88,
$\rightarrow$ expected shape duration is 64 samples.

Cylinder/Bell/Funnel example



Cylinder/Bell/Funnel example

DBA **CTW** **TEKA**



Plan

1 Time series
2 A brief history of time elastic matching
3 Kernalized Time elastic averaging of a set of time series
4 Applications
5 Conclusion

Applications

- Reducing the instance set
- Noise reduction
- Augmenting (Boostrapping) the instance set

Reducing the instance set

Motivation: In a big data context, for lazy and costly classification or regression models (e.g. k-NN), one can clusterize the training dataset to represent it using a small set of centroids.

Reducing the instance set

Comparative study using 45 data sets (UCR or UCI): classification error rates evaluated on the TEST data set (in %) obtained using the **1-NN classification rule**

DATASET	**DTW-M**	**DBA**	**CTW1**	**CTW2**	**KRDTW-M**	**TEKA**
# Best Scores	1	7	0	9	6	**27**
# Uniquely Best Scores	1	5	0	7	5	**23**
Average rank	4.56	2.87	4.62	2.97	3.22	**1.6**

DTW-M, KRDTW-M, (medoids),
DBA, CTW1, CTW2 and TEKA (centroids).
A **single medoid/centroid** extracted from the training data set represents each category.

Noise reduction: synthetic dataset

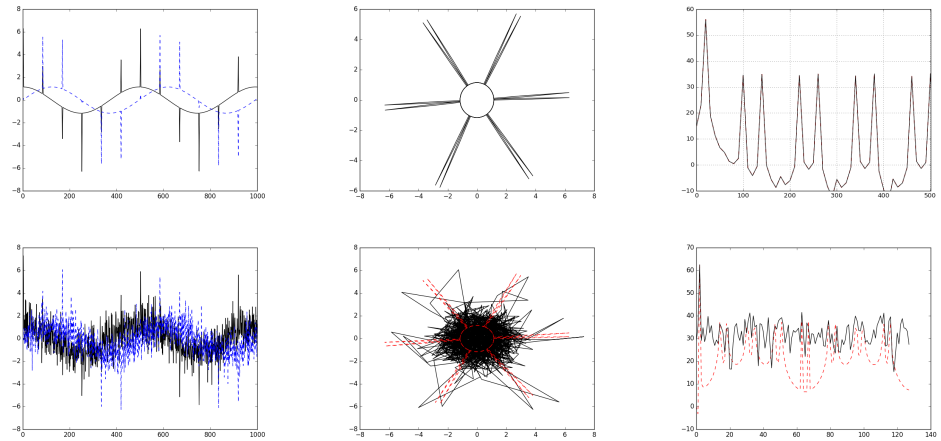
The "blinking star" signal

$$X_k(t) = \left(A_k + B_k \sum_{i=1}^{\infty} \delta\left(t - \frac{2\pi i}{6\omega_k}\right) \right) cos(\omega_k t + \phi_k) \qquad (3)$$

$$Y_k(t) = \left(A_k + B_k \sum_{i=1}^{\infty} \delta\left(t - \frac{2\pi i}{6\omega_k}\right) \right) sin(\omega_k t + \phi_k)$$

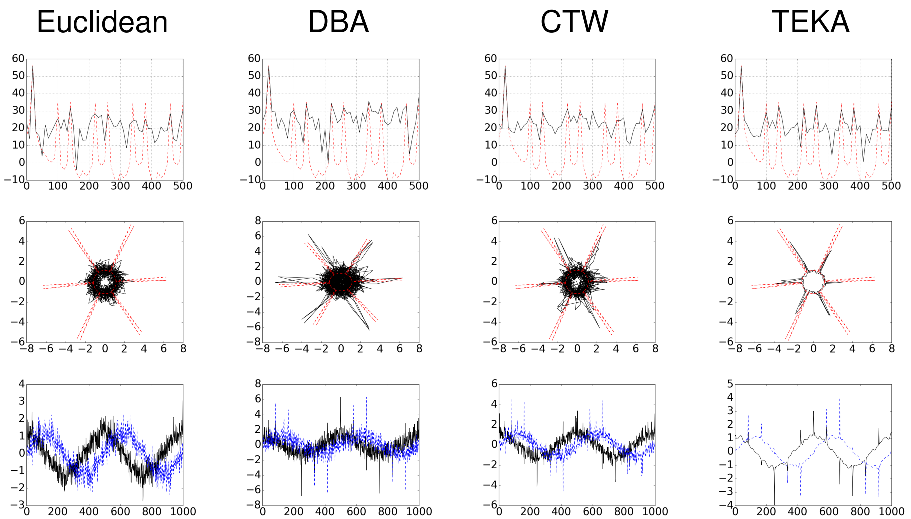
where $A_k = A_0 + a_k$, $B_k = (A_0 + 5) + b_k$ and $\omega_k = \omega_0 + w_k$, A_0 and ω_0 are constant and a_k, b_k, ω_k, ϕ_k are small perturbation in amplitude, frequency and phase respectively and randomly drawn from $a_k \in [0, A_0/10]$, $b_k \in [0, A_0/10]$, $\omega_k \in [-\omega_0/6.67, \omega_0/6.67]$, $\phi_k \in [-\omega_0/10, \omega_0/10]$.

Noise reduction: synthetic dataset



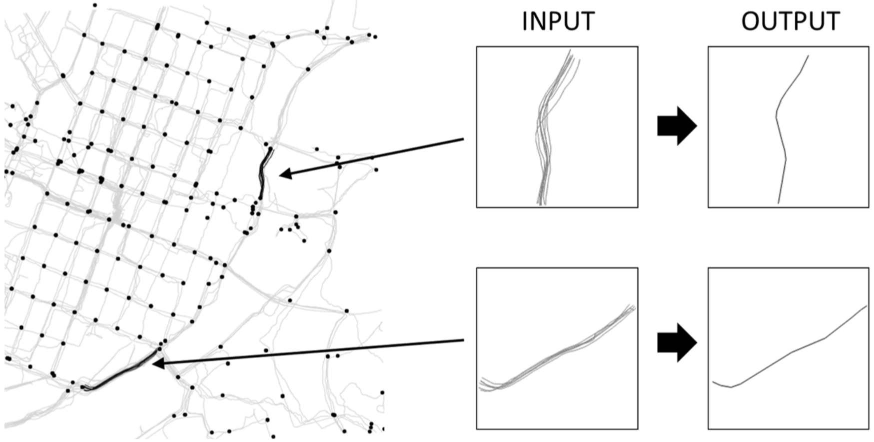
Top: clean signal
Bottom: a Gaussian noise with zero mean and variance one is added to each instances of the 2D signal.

Noise reduction: synthetic dataset



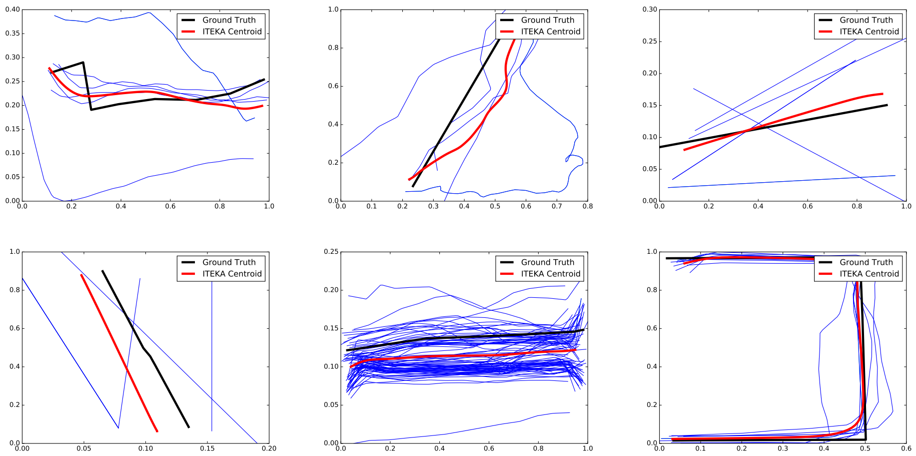
Noise reduction: GPS trajectories data set
Pasi Fränti and Radu Mariescu-Istodor [first.last@uef.fi]

Averaging GPS segments (OpenStreetMap) `https://www.mdpi.com/journal/applsci/special_issues/GPS_segment`

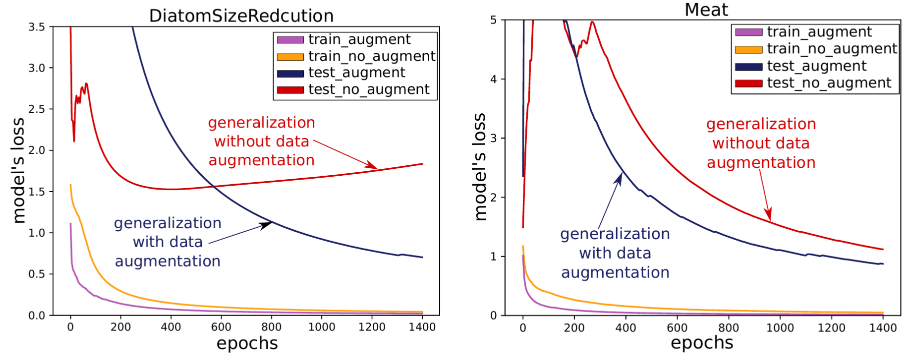


Noise reduction: GPS trajectories data set
Pasi Fränti and Radu Mariescu-Istodor [first.last@uef.fi]

Averaging GPS segments (OpenStreetMap) `https://www.mdpi.com/journal/applsci/special_issues/GPS_segment`



Boostrapping the instance set



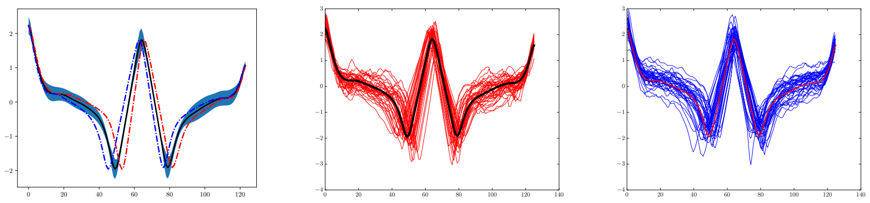
(a) DiatomSizeReduction (b) Meat

Training data augmentation by interpolating a curve improves the accuracy of 1NN DTW and Deep Neural Nets (Forestier2017), (Fawaz et al. (2018))

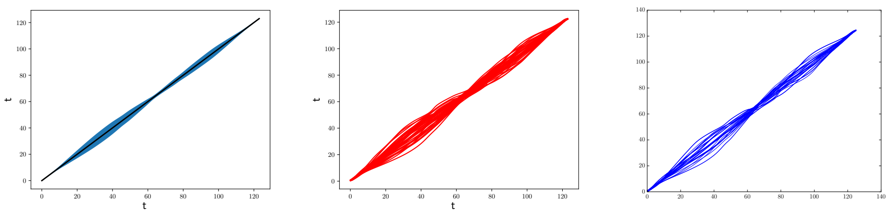
Boostrapping the instance set: UCR SwedishLeaf-12

ANALYSIS TRAIN SYNTHESIS

SHAPES



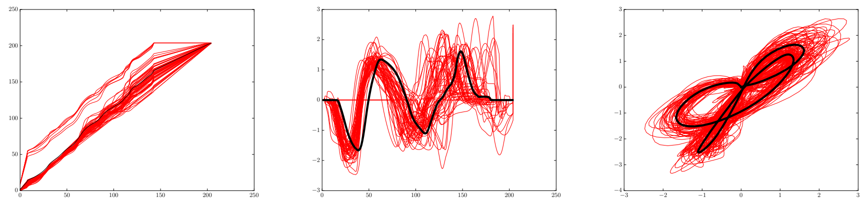
TEMPORAL PATTERNS



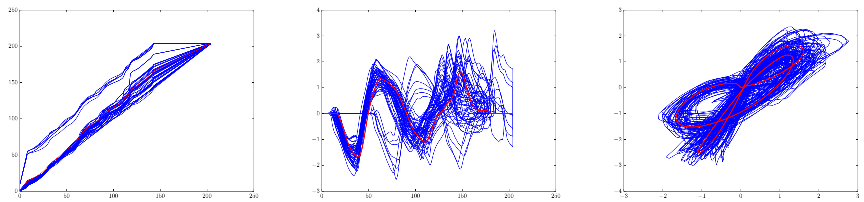
Boostrapping the instance set: UCI CharTraj_3D

TEMP. PATTERNS X 3D-SHAPE

TRAIN



SYNTHESIS



Plan

1. Time series
2. A brief history of time elastic matching
3. Kernalized Time elastic averaging of a set of time series
4. Applications
5. Conclusion

Conclusion/perspective

Time elastic averaging of set of time series

- DTW is not necessarily the *grail* elastic distance: ”soft-max” kernels exist too.
- TEKA achieves a spatio-temporal decomposition, separating the ”shape” and the temporal patterns.
- Corridor/Sparsification of the alignment path search space may reduce the quadratic complexity.

Conclusion/perspective

Hot topics

- Global alignment $\Rightarrow$ local alignment kernels: avoiding aligning what obviously cannot (or should not) be aligned ($\Rightarrow$ coping with gaps).
- K-kmean (aggregating ”distant” data leads to stability problems).
- Learn the variance model from the data in boostrap applications
- Evaluate kernelized data augmentation in classification tasks (DNN).

Thank you

Bibliography I

W. Abdulla, D. Chow, and G. Sin. Cross-words reference template for dtw-based speech recognition systems. In *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, volume 4, pages 1576–1579 Vol.4, Oct 2003. doi: 10.1109/TENCON.2003.1273186.

R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

L. Chen and R. Ng. On the marriage of lp-norm and edit distance. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pages 792–801, 2004.

Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *J. Mach. Learn. Res.*, 10:747–776, June 2009. ISSN 1532-4435.

Bibliography II

M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui. A Kernel for Time Series Based on Global Alignments. In *Proceedings of ICASSP'07*, Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, pages II–413 – II–416, Honolulu, HI, April 2007. IEEE.

K. H. Fasman and S. S. L. An introduction to biological sequence analysis. In *Comp. Methods in Mol. Biology,*, pages 21–42. In Salzberg, S.L., Searls, D.B., and Kasif, S., eds., Elsevier, 1998.

H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller. Data augmentation using synthetic data for time series classification with deep residual networks. *CoRR*, abs/1808.02455, 2018. URL http://arxiv.org/abs/1808.02455.

M. Fr'echet. *Sur quelques points du calcul fonctionnel*. Th'ese, Faculte des sciences de Paris., 1906.

Bibliography III

B. Haasdonk and C. Bahlmann. Learning with distance substitution kernels. In C. Rasmussen, H. Bülthoff, B. Schölkopf, and M. Giese, editors, *Pattern Recognition*, volume 3175 of *Lecture Notes in Computer Science*, pages 220–227. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-22945-2. doi: 10.1007/978-3-540-28649-3_27. URL http://dx.doi.org/10.1007/978-3-540-28649-3_27.

D. Haussler. Convolution kernels on discrete structures. Technical report, University of California, Santa Cruz, 1999. Technical Report.

V. Hautamaki, P. Nykanen, and P. Franti. Time-series clustering by approximate prototypes. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, Dec 2008. doi: 10.1109/ICPR.2008.4761105.

Bibliography IV

D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM 18 (6)*, pages 341–343, 1975.

W. Just and W. Just. Computational complexity of multiple sequence alignment with sp-score. *Journal of Computational Biology*, 8:615–623, 1999.

E. J. Keogh, L. Wei, X. Xi, S.-H.Lee, and M. Vlachos. Lb_keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In *ACM International Conference on Very Large Databases*, pages 882–893, 2006.

P. F. Marteau. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31 (2):306–318, 2008.

Bibliography V

P.-F. Marteau and S. Gibet. Constructing positive elastic kernels with application to time series classification. *IEEE Trans. on Neural Networks and Learning Systems*, http://arxiv.org/abs/1005.5141:1–14, 2014. doi: http://dx.doi.org/10.1109/TNNLS.2014.2333876.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

V. Niennattrakul and C. Ratanamahatana. Shape averaging under time warping. In *Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th Int. Conf. on*, volume 02, pages 626–629, May 2009. doi: 10.1109/ECTICON.2009.5137128.

Bibliography VI

V. Niennattrakul and C. A. Ratanamahatana. Inaccuracies of shape averaging method using dynamic time warping for time series data. In Y. Shi, G. D. van Albada, J. Dongarra, and P. M. A. Sloot, editors, *Computational Science – ICCS 2007*, pages 513–520, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-72584-8.

F. Petitjean, G. Forestier, G. Webb, A. Nicholson, Y. Chen, and E. Keogh. Dynamic time warping averaging of time series allows faster and more accurate classification. In *Proceedings of the 14th IEEE International Conference on Data Mining*, pages 470–479, 2014.

Bibliography VII

H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11): 1682–1689, 2004. URL `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics20.html#SaigoVUA04`.

H. Sakoe and S. Chiba. A dynamic programming approach to continuous speech recognition. In *Proceedings of the 7th International Congress of Acoustic*, pages 65–68, 1971.

I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3): 522–536, nov 1938. URL `http://links.jstor.org/sici?sici=0002-9947%28193811%2944%3A3%3C522%3AMSAPDF%3E2.0.CO%3B2-Z`.

Bibliography VIII

T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

S. Soheily-Khah and P.-F. Marteau. Sparsification of the alignment path search space in dynamic time warping. *Applied Soft Computing*, 78:630 – 640, 2019. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2019.03.009.

V. M. Velichko and N. G. Zagoruyko. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2: 223–234, 1970.

R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, Jan. 1974. ISSN 0004-5411. doi: 10.1145/321796.321811. URL `http://doi.acm.org/10.1145/321796.321811`.

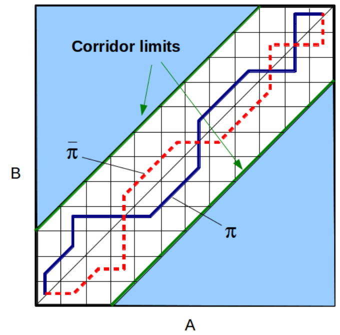
Bibliography IX

L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Jour. of Comp. Biology*, 1(4):337–348, 1994.

G. Wu, E. Y. Chang, and Z. Zhang. Learning with non-metric proximity matrices. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 411–414, New York, NY, USA, Nov. 2005. ACM. ISBN 1-59593-044-2. doi: http://doi.acm.org/10.1145/1101149.1101239.

F. Zhou and F. D. la Torre. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):279–294, Feb 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2414429.

Kernel induced by an alignment map (Marteau and Gibet (2014))



Properties

$\varphi_{\pi_x}(.) = \varphi_{\bar{\pi}_y}(.)$
$\varphi_{\pi_y}(.) = \varphi_{\bar{\pi}_x}(.)$
$K_\pi(A, B) = K_\pi(B, A)$
$K_\pi(A, B) = K_{\bar{\pi}}(A, B)$

- $K_\pi^{xy}(A, B) = k_\pi(\varphi_x(A), \varphi_y(B)) + k_\pi(\varphi_y(A), \varphi_x(B))$
- $K_\pi^{xx}(A, B) = k_\pi(\varphi_x(A), \varphi_x(B)) + k_\pi(\varphi_y(A), \varphi_y(B))$
- $K_\pi(A, B) = K_\pi^{xy}(A, B) + K_\pi^{xx}(A, B)$