From ‘time’ elastic distances to ‘time’ elastic inner vector spaces

-Prospective applications in ‘time’ series, sequence or text classification-

P.-F. Marteau, SEASIDE
Challenging issue

• Embedding complex data in an inner vector space (or normed space) gives access to a wide range of applications relying on main linear algebra results
  – PCA, SVD (LSI, LSA), ‘Optimal’ Dimension Reduction, Matching pursuit, data compression, etc.

• For sequential or 'time-stamped' data, elastic distances (such as Levenshtein Dist., DTW, LCSS, ERP, TWED) have proved to be superior in general to Euclidean or Minkowski distances in particular for discrimination or regression tasks
  – Unfortunately, elastic distances do not derive from a norm (an inner product)

• **Existence of ‘time-elastic’ inner products?**
  – Preserving the access to main linear algebra results
  – Keeping benefiting from ‘time-elasticity’ capability
Content

- Challenging issue
- ‘Time’-stamped data
- ‘Time elastic’ distance or pseudo distance
- ‘Time elastic inner products’ (TEIP)
- Prospective applications
- Conclusion, perspectives
“times-tamped” data

- \( U \) : sequence set, \( A \in U \) is a finite sequence
- \( A_i^n = A(1)A(2)\ldots A(n) \) with \( A(i) \in S \times T \)
- \( S \) : “spatial” set (numeric or symbolic) : spatial dimensions
- \( T \) : “time” set : timestamp dimension
- \( A(i) = (a(i); t_{a(i)}) \)

Examples

By itself, not a whole lot. 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.
‘Time’-elastic distances or pseudo distances

\[
DTW(A^p_i, B^q_i) = \delta(a(p), b(q)) + \min \begin{cases}
    DTW(A^p_i, B^{q-1}_i) \\
    DTW(A^{p-1}_i, B^q_i) \\
    DTW(A^{p-1}_i, B^{q-1}_i)
\end{cases}
\]

- Recursive definition
- O(N^2)
- Nonlinearity (Min, Max, \(\delta(\cdot;\cdot)\)),
  - Impossibility to derive from elastic distances inner (dot) products
  - Impossibility to derive a norm
  - Existence of ‘elastic’ inner vector space?
Dot/Inner Products

• Euclidean dot product

\[ \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \]

• Inner product

• **Conjugate** symmetry: \[ \langle x, y \rangle = \overline{\langle y, x \rangle} \]. Note that in \( \mathbb{R} \), it is symmetric.

• **Linearity** in the first argument:

\[ \langle ax, y \rangle = a \langle x, y \rangle. \]
\[ \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle. \]

• **Positive-definiteness**: \[ \langle x, x \rangle \geq 0 \]
  with equality only for \( x = 0 \).
\((U^*, \oplus, \otimes)\) provided \((S, \oplus_S, \otimes_S)\)

Definition 2.3. For all \(A \in U^*\) and all \(\lambda \in \mathbb{R}\), \(C = \lambda \otimes A \in U^*\) is such that for all \(i \in \mathbb{N}\) such that \(0 \leq i \leq |A|\), \(C(i) = (\lambda.a(i), t_{a(i)})\) and thus \(|C| = |A|\).

Definition 2.4. For all \((A, B) \in (U^*)^2\), the addition of \(A\) and \(B\)

\(U^*\): set of sequences defined on \((S-\{0_s\}) \times T\)
From elastic distance to ‘time’
elastic products

Provided the existence of the addition $\oplus$ and scalar multiplication $\otimes$ on the sequence set $U$, what are the conditions on $f$, $g$, $\alpha$, $\beta$ and $\xi$ to make it an inner product?

\[
\begin{align*}
< A^p_1, B^q_1 >_{tep} = & \sum \left\{ \begin{array}{l} 
\alpha \cdot < A^{p-1}_1, B^q_1 >_{tep} \\
\beta \cdot < A^{p-1}_1, B^{q-1}_1 >_{tep} + f(a(p), b(q)) \cdot g(t_{a(p)}, t_{b(q)}) \\
\alpha \cdot < A^p_1, B^{q-1}_1 >_{tep} \end{array} \right. \\
< A, \Omega >_{tep} = & < \Omega, A >_{tep} = < \Omega, \Omega >_{tep} = \xi
\end{align*}
\]

where $\Omega$ is the null sequence, $\xi$, $\alpha$, $\beta$ are real

and $g$ is strictly positive
‘Time’ elastic inner products

\( (U, \oplus, \otimes) \) provided \( (S, \oplus_S, \otimes_S) \)

\( U^* \) is the set of sequences whose elements are in \( S \setminus \{0_S\} \)
\( (0_S, t) \) identifies the null sequence element for all \( t \)

Theorem 2.1. \( <\cdot,\cdot>_{tep} \) is an inner product on \( (U^*, \oplus, \otimes) \) iff:

i) \( \xi = 0 \).

ii) \( g : (T \times T) \to \mathbb{R} \) is symmetric and strictly positive,

iii) \( f \) is an inner product on \( (S, \oplus_S, \otimes_S) \), if we extend the domain of \( f \) on \( S \) while setting \( f(0_S, 0_S) = 0 \).

iv) \( \alpha = 1 \) and \( \beta = -1 \),

Note that according to this result, it is possible to embed sequences of various lengths or times series not-uniformly sampled and/or of various lengths in a unique ‘elastic’ vector space structure.
‘Time’ elastic **inner products**

**Example** of a ‘time elastic’ inner product

\[
\langle A_1^p, B_1^q \rangle_{teip} = \\
\sum \left\{ 
\begin{array}{l}
\langle A_1^p, B_1^{q-1} \rangle_{teip} \\
- \langle A_1^{p-1}, B_1^{q-1} \rangle_{teip} + a(p)b(q) \cdot e^{-\nu \cdot |t_{ap} - t_{bq}|} \\
\langle A_1^{p-1}, B_1^q \rangle_{teip}
\end{array}
\right.
\]

\( \nu: \text{stiffness parameter} \)
\( \nu = 0, \text{infinite elasticity} \)
\( \nu = \infty, \text{null elasticity} \)
Recursively embedded elastic dimensions

\[ < A_1^p, B_1^q >_{teip} = \]
\[ \sum \left\{ \begin{array}{l}
< A_1^{p-1}, B_1^q >_{tep} \\
- < A_1^{p-1}, B_1^{q-1} >_{tep} + g(t_{a(p)}, t_{b(q)}) \cdot < a(p), b(q) >_{teip(S)} \\
< A_1^p, B_1^{q-1} >_{tep}
\end{array} \right\} \]
Recursively embedded elastic dimensions

Time-stamped data
Time-elastic distance
Time-elastic inner prod
Prospective applications
Elastic measures

**Elastic norm:**
\[ \| A_1^p \|_e = \sqrt{< A_1^p, A_1^p>_e} \]

**Elastic distance**
\[ eDist( A_1^p, B_1^q ) = \| A_1^p \oplus ( -B_1^q ) \|_e \]

**Elastic cosine**
\[ eCos( A_1^p, B_1^q ) = \frac{< A_1^p, B_1^q>_e}{\| A_1^p \|_e \cdot \| B_1^q \|_e} \]
Sanity check: Gram-Schmidt orthogonalization of a sine basis

\[ \nu = 0.01 \]
Prospective applications in sequence or text matching

\[
< A_1^p, B_1^q >_{teip_{tm}} = \sum \left\{ \begin{array}{l}
< A_1^{p-1}, B_1^q >_{teip_{tm}} \\
- < A_1^{p-1}, B_1^{q-1} >_{teip_{tm}} + e^{-\nu|t_{a(p)}-t_{b(q)}|}\delta(a(p), b(q)) \\
< A_1^p, B_1^{q-1} >_{teip_{tm}}
\end{array} \right.
\]

(7)

where \(a(p)\) and \(b(q)\) are vectors whose coordinates identify words with weightings, \(\delta(a, b) = < a, b >\) is the Euclidan inner product, and \(\nu\) a time stiffness parameter.

**Proposition 3.3.** For \(\nu = 0\) and \(\delta\) redefined as \(\delta(a, b) = 1\) if \(a = b\), 0 otherwise, the elastic inner product defined in Eq. 7 coincides with the euclidean inner product between two vectors whose coordinates correspond to term frequencies observed into the \(A_1^p\) and \(B_1^q\) text sequences.

**Proposition 3.4:** if \(\nu \to \infty\), and if the two sequences in argument are uniformly sampled and have the same length, then the TEIP tends toward the Euclidean inner product.
Toy experiment

\[ e\text{Cos}( A_i^p , B_i^q ) = \frac{< A_i^p , B_i^q >_e}{ \| A_i^p \|_e \cdot \| B_i^q \|_e } \]

\[ \delta( x, y ) = 1 \text{ if } x = y, \ 0 \text{ otherwise} \]
Toy experiment

\[ e\cos(A,D) = \frac{A \cdot B}{||A|| \cdot ||B||} \]

\[ \delta(?,x) = \delta(x,?) = 1 \]

<table>
<thead>
<tr>
<th>A</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>C</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>D</td>
<td>b</td>
<td>b</td>
<td>?</td>
<td>b</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
</tbody>
</table>
Toy experiment

Prospective applications

Time-stamped data
Time-elastic distance
Time-elastic inner prod

$$e\text{Cos}(x,y)$$

$$e\text{Cos}(A,E)$$
$$e\text{Cos}(B,E)$$
$$e\text{Cos}(C,E)$$
$$e\text{Cos}(D,E)$$

$$e\text{Cos}(A_1^p, B_1^q) = \frac{<A_1^p, B_1^q>}{\|A_1^p\|_e \|B_1^q\|_e}$$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>C</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>D</td>
<td>b</td>
<td>b</td>
<td>?</td>
<td>b</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>E</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
</tbody>
</table>

PFM, 8 April 2011
Toy experiment

Prospective applications

Time-stamped data
Time-elastic distance
Time-elastic inner prod

Prospective applications

eCOS(x,y)

<table>
<thead>
<tr>
<th>A</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>b</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>C</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>D</td>
<td>b</td>
<td>b</td>
<td>?</td>
<td>b</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CPER: support Invent’IST
IRISA, IETR, VALORIA, IRMAR, LMAM

Platform for “Massive” Data Processing: 250 k€

Dr Nicolas Bonnel
Implementation on GPU and cluster (today ~20 nodes, perspective 32 nodes)
Early results in Text Mining

Text classification experiment: Reuters 21578, WebKB datasets
Early results in Text Mining

Text classification experiment: WebKB classes

Classif. error rate

Time-stamped data
Time-elastic distance
Time-elastic inner prod
Prospective applications
Early results in sequence classification

- Protein Classification Benchmark Collection (ICGEB/EMBNet)
  - 3-phosphoglycerate kinase (3PGK) protein sequences

<table>
<thead>
<tr>
<th>ID</th>
<th>POSITIVE TEST</th>
<th>POSITIVE TRAIN</th>
<th>NEGATIVE TEST</th>
<th>NEGATIVE TRAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea_Crenarchaeota</td>
<td>4</td>
<td>11</td>
<td>53</td>
<td>63</td>
</tr>
<tr>
<td>Archaea_Euryarchaeota</td>
<td>11</td>
<td>4</td>
<td>53</td>
<td>63</td>
</tr>
<tr>
<td>Bacteria_Actinobacteridae</td>
<td>5</td>
<td>68</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Bacteria_Firmicutes</td>
<td>35</td>
<td>38</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Bacteria_Proteobacteria</td>
<td>30</td>
<td>43</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Eukaryota_Alveolata</td>
<td>4</td>
<td>39</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Eukaryota_Euglenozoa</td>
<td>5</td>
<td>38</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Eukaryota_Fungi</td>
<td>10</td>
<td>33</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Eukaryota_Metazoa</td>
<td>12</td>
<td>31</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Eukaryota_Viridaeplanta</td>
<td>8</td>
<td>35</td>
<td>44</td>
<td>44</td>
</tr>
</tbody>
</table>
Early results in sequence classification

–3-phosphoglycerate kinase (3PGK) protein sequences

![Graph showing error rate versus time-elastic distance]
Early results in sequence classification

3-phosphoglycerate kinase (3PGK) protein sequences

Error Rate

Time-stamped data
Time-elastic distance
Time-elastic inner prod
Prospective applications

eCOS

Archaea_Crenarchaeota
Archaea_Euryarchaeota
Bacteria_Actinobacteridae
Bacteria_Firmicutes
Bacteria_Proteobacteria
Eukaryota_Alveolata
Eukaryota_Euglenozoa
Eukaryota_Fungi
Eukaryota_Metazoa
Eukaryota_Viridaeplantae
### Average AUC values for the 10 classification tasks

**Prospective applications**

<table>
<thead>
<tr>
<th>1NN</th>
<th>BLAST</th>
<th>SW</th>
<th>NW</th>
<th>LA</th>
<th>PRIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCOP95_supfam_fam_1</td>
<td>0.70</td>
<td>0.84</td>
<td>0.84</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>SCOP95_supfam_fam_kfold_2</td>
<td>0.95</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>SCOP95_fold_supfam_3</td>
<td>0.57</td>
<td>0.65</td>
<td>0.67</td>
<td>0.58</td>
<td>0.85</td>
</tr>
<tr>
<td>SCOP95_fold_supfam_kfold_4</td>
<td>0.90</td>
<td>0.94</td>
<td>0.95</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td>SCOP95_class_fold_5</td>
<td>0.60</td>
<td>0.61</td>
<td>0.63</td>
<td>0.57</td>
<td>0.78</td>
</tr>
<tr>
<td>SCOP95_class_fold_kfold_6</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
<td>0.88</td>
<td>0.93</td>
</tr>
<tr>
<td>CATH95_H_S_7</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>CATH95_H_S_kfold_8</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>CATH95_T_H_9</td>
<td>0.57</td>
<td>0.65</td>
<td>0.67</td>
<td>0.61</td>
<td>0.77</td>
</tr>
<tr>
<td>CATH95_T_H_kfold_10</td>
<td>0.90</td>
<td>0.93</td>
<td>0.93</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>CATH95_A_T_11</td>
<td>0.55</td>
<td>0.57</td>
<td>0.58</td>
<td>0.54</td>
<td>0.67</td>
</tr>
<tr>
<td>CATH95_A_T_kfold_12</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>CATH95_C_A_13</td>
<td>0.65</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td>CATH95_C_A_kfold_14</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.86</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.79</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.85</strong></td>
</tr>
</tbody>
</table>

**Time-stamped data**

**Time-elastic distance**

**Time-elastic inner prod**

**SW**: Smith-Waterman  
**NW**: Needleman-Wunsch  
**LA**: Local Alignment kernel  
**PRIDE**: Probability of similarity (Histogram of distances between motifs)

**AUC**: Area under the Curve (ROC curve)
Conclusion, perspectives

• The existence of ‘time-elastic’ products seems to be proved given the existence of $\oplus$ and $\otimes$
  – Elastic inner products generalize somehow the Euclidean inner product and the classical vector model defined for textual information retrieval
  – Several ‘time-elastic’ dimensions can be managed recursively
  – Leads to positive definite kernels (not the case for DTW-like distances)
• Early results on time series or sequential data classification show some potentiality that need to be consolidated
• Generalization to simultaneously elastic dimension? (can $T$ be multidimensional?)
• ‘time-elastic’ Linear Algebra has to be further developed