Harpoon or maggot ?

A comparison of various metrics to fish for sequence patterns

Lausanne Conference On Sequence Analysis

Wednesday 6th June, 2012

Nicolas Robette, *Printemps (UVSQ-CNRS)*

Xavier Bry, *I3M, Université de Montpellier II*

Sequence analysis

- Trajectories built as sequences of states
- Computation of pairwise dissimilarities

  *(algorithms = Optimal Matching Analysis, and **many others**)*

  → Distance matrix

  → Clustering (HCA...; or reduction by MDS)

  → Typology of trajectories

Many dissimilarity metrics

- Related to ‘sequence analysis’ tradition (oma, etc.)…
- … or to ‘geometric data analysis’ tradition

Optimal Matching Analysis (1)

- Widely used in bioinformatics (DNA)
- Introduced in social sciences by Andrew **Abbott** (80's)
- **Principle**: measuring dissimilarity between pairs of sequences by calculating the cost of the transformation of one sequence into the other

See for example Macindoe & Abbott, 2004

Optimal Matching Analysis (2)

- 3 elementary operations:
  - insertion
  - deletion
  - substitution
- each operation is assigned a **cost**
- the distance between two sequences is equal to the **minimal cost** needed to transform one sequence into the other

The choice of costs (1)

Important issue in OMA (?):

- **Substitution**:

  retains the temporal structure (**timing**)
  but distorts events (order)

- **Insertion/deletion**:

  distort time
  but retain **order** of events

The choice of costs (2)

- **substitution cost** matrix :
  - according to theoretical assumptions: hierarchy of states…
  - data driven: transition likelihoods…

- **insertion/suppression** (*indel*) costs :
  - if order prevails ⟶ low *indel* /substitution
  - if timing prevails ⟶ high *indel* /substitution

Elzinga's metrics *(2003;2008)*

- **Criticism :** OMA doesn't take order into account (substitution of A to B or B to A are equivalent)

- **Several alternatives** :
  - Longer common prefix (LCP)
  - Longer common subsequence (LCS)
  - Number of common subsequences (NCS)
  - Number of matching subsequences (NMS)
  - …

Lesnard's 'Dynamic Hamming' (2010)

- **Criticism**: Transition likelihoods are time-dependant

- **Principle**:
  - no insertion/deletion
  - substitution costs computed for each time point

- Applications to time-use diary data

Rousset *et al* (2012)

- **Principle**:
  - based on transition likelihoods
  - possibility of a delay cost

‘Geometric Data Analysis’ metrics (1)

A fictitious example of school-to-work transition:

S = studies

U = unemployment

J = job

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

‘Geometric Data Analysis’ metrics (2)

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

- **Indicator matrix**

18S	18U	18J	…	25S	25U	25J
1	0	0	…	0	0	1

PCA → Euclidean distance *(see Grelet, 2002)*

CA → χ^2 distance

→ **duration** and **timing**

‘Geometric Data Analysis’ metrics (3)

18	19	20	21	22	23	24	25
S	S	S	U	J	J	J	J

- **Summarized calendar (Qualitative Harmonic Analysis)**

18-20 S	18-20 U	18-20 J	21-25 S	21-25 U	21-25 J
1	0	0	0	0,2	0,8

CA -> χ^2 distance *(see Robette & Thibault, 2008)*

→ **duration** and **timing**

(timing less precise, but less sensitive to « shifts »)

→ allows to « **weight** » sub-periods

A few existing comparisons

- **OMA with different cost schemes**: Abbott & Hrycak 1990; Chan 1995; Anyadikes-Danes & McVicar 2002 & 2010 …

- **OMA vs other metric**: Lesnard 2010 (DHD); Robette & Thibault 2008 (QHA); Aisenbrey & Fasang 2010 (DHD,NMS) …

- **Geometric Data Analysis**: Grelet 2002

→ **broad agreement**: *"minor analytic decisions are unlikely to drastically change results"* (Abbott & Hrycak, 1990)

Limitations

- Only a few metrics at a time
- Based on one set of empirical data
- Examination of clusters

Our empirical protocol

- A “reasoned” set of simulated sequences *(+ one empirical set as “ control”)*
- Correlation b/w dissimilarity matrices
- Avg distances within / between subsets of simulated sequences

A “reasoned” sequence data set

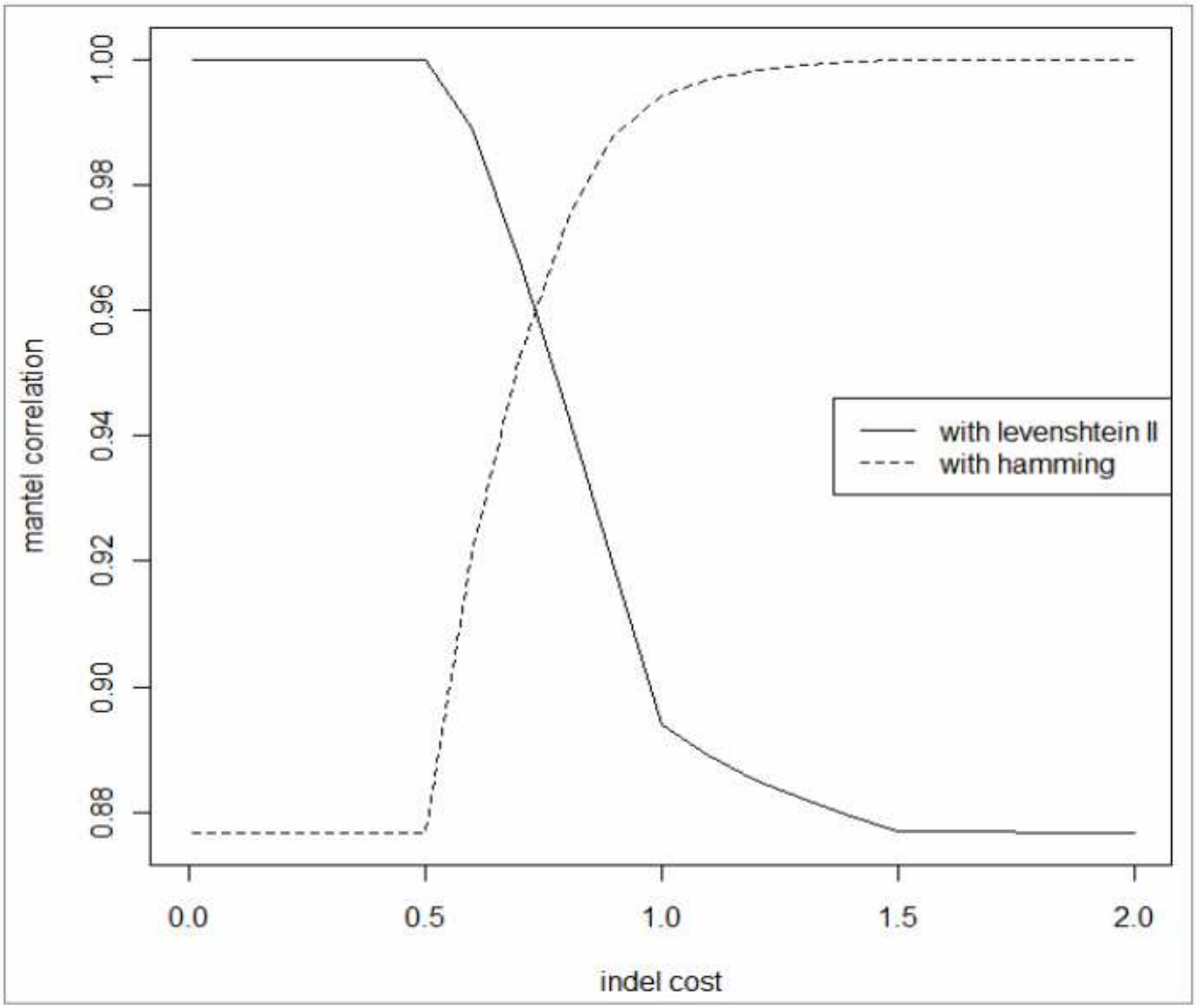
- An artificial set (N=854), designed to contain the various kinds of **regularities / differences**: shifts, swaps, insertions, deletions, replacements, repetitions of spells *(Barban & Billari, 2011)*
- Examples:

1. **Time warping**: subset of sequences A-B-C with varying durations in A, B and C
2. **Shifts**: A-B-C with B spell of fixed length equal to 6 and varying durations in A and C
3. **Reversal**: Initial sequences (subset #1) in reversed order, i.e. C-B-A
4. **Swaps**: Initial sequences (subset #1) with B and C swapped (i.e. A-C-B) or A and B swapped (i.e. B-A-C)
5. Etc…

An empirical sequence data set

- *Biographies et entourage* event-history survey (INED, 2001)
- Occupational careers of 1421 men
- 37 years, from 14 to 50
- 9 states:
  - *farmers, self-employed, higher-level intellectual occupations, intermediate occupations, clerical and sales workers, manual workers,*
  - *student,*
  - *military conscripts,*
  - *other inactivity*

Correlation b/w dissimilarity matrices
with varying *indel* (subst=1)



The set of metrics

- Hamming, ie OMA with no indel (HAM)
- Levenshtein II, ie OMA with no subst (LEVII)
- OMA with data driven subst & high *indel* (OMAtr)
- Dynamic Hamming Distance (DHD)
- Rousset's alternative (ROUS)
- Elzinga's # of matching subseq. (NMS)
- Indicator matrix with CA (CA)
- Indicator matrix with PCA (PCA)
- Summarized calendar (QHA)
- *3 "control" metrics: duration (DUR), quantum (QUA), sequence = LLCS (SEQ)*

Correlation b/w dissimilarity matrices

	DUR	QUA	SEQ	LEVII	HAM	OMAtr	DHD	ROUS	PCA	CA	QHA	NMS
DUR	100	34,9	34,3	88,9	72,6	75,1	73,4	72,1	70,1	61,5	62,7	-1,8
QUA	34,9	100	82,4	37,5	28,4	30,6	30,3	27,8	20,1	28,8	29,3	67,5
SEQ	34,3	82,4	100	53,2	46,6	49,1	49,3	45,0	36,8	45,8	45,9	52,1
LEVII	88,9	37,5	53,2	100	87,7	90,4	89,2	86,5	83,4	75,1	75,8	-1,2
HAM	72,6	28,4	46,6	87,7	100	99,3	99,7	99,2	96,9	72,4	72,0	-0,6
OMAtr	75,1	30,6	49,1	90,4	99,3	100	99,6	98,2	95,4	75,7	75,5	-1,0
DHD	73,4	30,3	49,3	89,2	99,7	99,6	100	98,6	95,8	75,5	75,2	-0,8
ROUS	72,1	27,8	45,0	86,5	99,2	98,2	98,6	100	97,9	70,7	70,2	-0,3
PCA	70,1	20,1	36,8	83,4	96,9	95,4	95,8	97,9	100	63,1	62,5	-6,2
CA	61,5	28,8	45,8	75,1	72,4	75,7	75,5	70,7	63,1	100	99,6	-3,8
QHA	62,7	29,3	45,9	75,8	72,0	75,5	75,2	70,2	62,5	99,6	100	-3,7
NMS	-1,8	67,5	52,1	-1,2	-0,6	-1,0	-0,8	-0,3	-6,2	-3,8	-3,7	100

Scaled ranked distances b/w sequences

Patterns	DUR	QUA	SEQ	LEVII	HAM	OMAtr	DHD	ROUS	PCA	CA	QHA	NMS
time warping (#1 vs #1)	20	0	0	14	12	13	13	13	18	11	11	2
shifts (#2 vs #2)	10	0	0	6	14	15	15	15	19	11	11	1
reversal, ie ABC vs CBA (#1 vs #3)	20	0	50	44	50	55	55	55	64	43	45	23
swaps, ie ABC vs ACB or BAC (#1 vs #4)	20	0	10	24	25	28	27	28	33	21	21	8
total permutation, ie ABC vs CAB or BCA (#1 vs #5)	20	0	10	31	52	51	55	56	63	39	39	15
1 insertion of a short D spell (#1 vs #6)	22	12	4	15	14	15	15	15	18	11	12	9
1 insertion of a long D spell (#1 vs #7)	44	12	4	35	35	39	37	38	41	44	43	29
2 insertions of short D spells (#1 vs #8)	24	27	10	17	16	17	17	17	18	11	12	55
2 insertions of long D spells (#1 vs #9)	61	27	10	55	52	56	54	56	61	54	54	68
2 insertions of short D and E spells (#1 vs #10)	26	27	10	18	18	19	19	19	20	14	15	52
2 insertions of long D and E spells (#1 vs #11)	61	27	10	55	52	56	55	56	60	73	72	60
1 deletion, ie ABC vs AB (#1 vs #12b)	33	12	4	26	24	26	25	26	34	19	20	5
1 deletion, ie ABC vs AC or BC (#1 vs #12a)	34	12	4	27	21	23	23	23	29	19	20	6
2 deletions, ie ABC vs A, B or C (#1 vs #13)	56	27	10	50	34	38	37	37	49	32	34	7
1 replacement, ie ABC vs ABF, AFC or FBC (#1 vs #14)	42	27	10	35	27	31	31	29	31	46	49	17
2 replacements, ie ABC vs AFG, FBG or FGC (#1 vs #15)	69	65	50	64	49	57	56	53	51	69	71	28
3 replacements, ie ABC vs FGH (#1 vs #16)	90	87	84	90	82	95	94	90	91	90	90	31
AB vs ABA (#12b vs #17)	19	12	4	19	18	18	18	20	26	20	21	5
AB vs ABAB (#12b vs #18)	17	27	10	14	11	11	11	12	15	10	11	20
many repetitions of AB spells (#12b vs #19)	11	100	100	19	11	12	12	13	6	6	8	100
slight shift of "ABABABABABABABABABAB" (#19 vs #20)	0	0	10	0	82	1	79	84	0	0	0	100
overall repetition, ie ABC vs ABCABC (#1 vs #21)	20	52	30	18	20	21	22	22	24	16	13	91

“OM-like” vs “CA-like”

- **“CA-like”** metrics more easily capture differences in the universe of states composing sequences, insofar as the states appearing in one sequence and not in the other correspond to long spells *(ie insertions of one long spell or two long different spells, one or two replacements)*

- **“OM-like”** metrics attach more importance to the way and the order in which spells unfold *(ie time warping and shifts, reversals, swaps, total permutations and repetitions)*

“OM-like” vs NMS

- **NMS** more sensitive to differences in the sequence of spells, even if the differing spells have a short duration *(ie repetitions of spells, two insertions especially short ones)*

- **NMS**'s focus on sequence of spells operates only in specific cases, in particular when “alien” spells are short *(ie NOT time warping and shifts, but above all reversals, swaps, total permutations, deletions and replacements)*

Among “OM-like”

- **PCA** is somewhat more sensitive than **Hamming** to *time warping and shifts, reversals, swaps and total permutations, deletions and long insertions*.

- **Levenshtein II** gives less importance to contemporaneousness (shifts and permutations), captures deletions and replacements better.

In a nutshell

- Social science sequence data are strongly structured

→ the main **patterns uncovered** by **most of the metrics**

- But as marginal differences may be of importance

→ **three groups** of heavily converging metrics, with small distinctions among them

References

- **webpage**: http://nicolas.robette.free.fr/Publis.htm

- Robette N., Bry X., 2012, « Harpoon or bait? A comparison of various metrics to fish for sequence patterns », forthcoming in ***Bulletin of Sociological Methodology***

- Robette N., 2010, *Explorer et décrire les parcours de vie : les typologies de trajectoires*, Paris : Ceped (série « les clefs pour »)
- Robette N., Thibault N., 2008, « Comparing qualitative harmonic analysis and optimal matching. An exploratory study of occupational trajectories », *Population-E*, 64(3), p. 533-556.