A Flexible Metric

LaCOSA 2012

Cees H. Elzinga

PARIS/SILC Research Group Vrije Universiteit Amsterdam

A Joint Enterprise with

- Dr. Matthias Studer, Faculty of Economics and Social Sciences, University of Geneva

- Prof. Hui Wang, Computer Science Research Institute, Ulster University

- Purpose: an SMR-paper on comparing metrics for SA

VU UNIVERSITY AMSTERDAM

Sequence Comparison Methods

- Optimal Matching
  - advantages:
    - well known
    - adaptable edit cost
    - easy algorithm, readily available
  - disadvantages:
    - unequal sequence lengths problematic
    - llcs (OM with unit-cost) crude
- Feature Vectors

VU UNIVERSITY AMSTERDAM

Sequence Comparison Methods

- Optimal Matching
- Feature Vectors
  - advantages:
    - different features possible
    - handles sequences of unequal length
  - disadvantages:
    - generally not well understood
    - “no adaptable edit cost” (Hollister)

VU UNIVERSITY AMSTERDAM

Lecture's Purpose

- Discuss General methodology of feature vectors
- Explain basic algorithm: the Grid
- A flexible representation: weigh for
  - subsequence length
  - subsequence gaps
    - limit gap-size
    - penalize gap-size
  - subsequence characters
  - "edit cost": soft-matching of states
  - durations or run-lengths
- Example(s)

VU UNIVERSITY AMSTERDAM

Sequence Comparison

- To classify
  - sort into groups:
    - that are as different as possible
    - that are as homogeneous as possible
  - collect similar things
    - things that share many features
- To explain (dis)similarity

Simple Sequence Structure

- Sequences are very similar
  - same small alphabet
  - same subsequences e.g. "married - children"
  - same durations e.g. "education"
- Classification is subtle, takes many features

VU UNIVERSITY AMSTERDAM

Feature Vectors: Principles

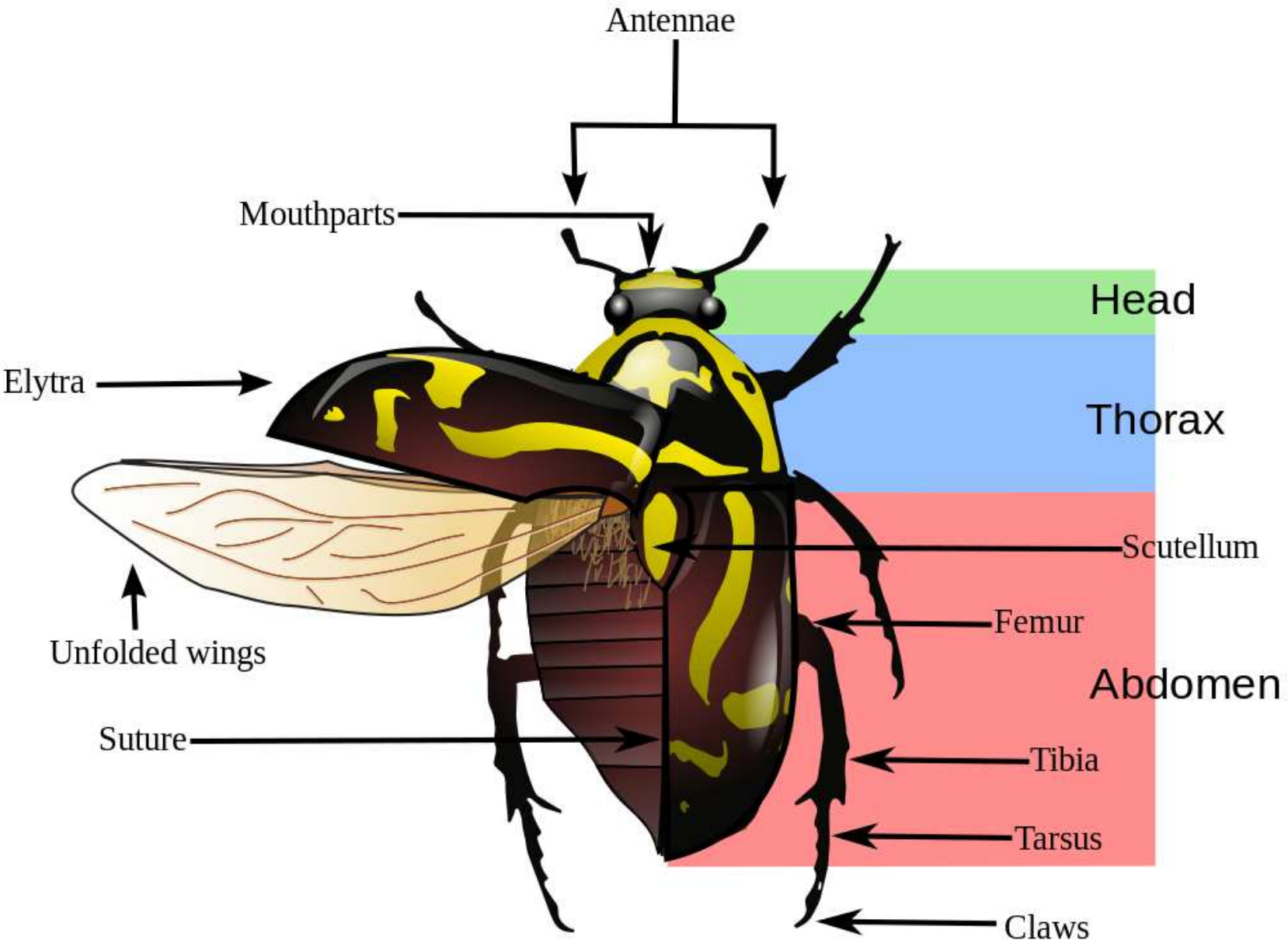
- select d features or properties $\{p_1, \ldots, p_d\}$
- map each object x to a d-vector $\mathbf{x}$
  - $\mathrm{x} \mapsto \mathbf{x} = (x_1, \ldots, x_d)$
- determine the value of the $\mathbf{x}$-coordinates x_i
  - $$x_i = \begin{cases} f(p_i) & \text{if object } \mathrm{x} \text{ has property } p_i \\ 0 & \text{otherwise} \end{cases}$$
  - simple: $f(p_i) = 1,$ all i (feature "on")

Example: Classifying Beetles

- **Step 1** : Select relevant beetle-properties as vector-coordinates
- **Step 2** : Map different beetles onto different vectors
- **Step 3** : Calculate distances in beetle-space
- **Step 4a**: If beetles are close, put them in the same class
- **Step 4b**: Else, put them in different classes
- **Step 5** : Be happy or try to “explain” the classes

VU UNIVERSITY AMSTERDAM

Simple Beetle Morphology



UNIVERSITY AMSTERDAM

Many Different Beetles

Binary Beetle Features

Feature	"1"	"0"	discriminates
Long Antennae	yes	no	yes
Compound Eyes	yes	no	yes
Functional Wings	yes	no	yes
6 legs	yes	no	no
Protruding Mouthparts	yes	no	yes
Reads Dickens	yes	no	no
Rowing Legs	yes	no	yes

- 7 binary features suffices to discern $2^7 = 524$ distinct species
- there exist $10^6 - 10^8$ distinct species
  - requires 25-30 binary features

VU UNIVERSITY AMSTERDAM

4 Beetles in Beetle Space $\{0,1\}^7$

Features	$\mathbf{a}$	$\mathbf{b}$	$\mathbf{c}$	$\mathbf{d}$
Antennae	1	0	1	1
Eyes	0	1	0	1
Wings	1	1	0	0
6 legs	1	1	1	1
Mouthparts	1	0	0	1
Reads Dickens	0	0	0	0
Rowing Legs	0	1	0	0

- inner product $\mathbf{a}'\mathbf{b} = \sum_i a_i b_i = 2$ counts common features
- inner product $\mathbf{a}'\mathbf{a} = \sum_i a_i^2 = 5$ counts features

Beetle Feature Vectors

- beetle feature space-matrix $\mathbf{X} = (\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$
- Gram-matrix $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 4 & 2 & 2 & 3 \\ 2 & 4 & 1 & 2 \\ 2 & 1 & 2 & 2 \\ 3 & 2 & 2 & 4 \end{pmatrix}$, inner products
- beetle vectors have
  - length: $\|\mathbf{a}\| = \sqrt{\mathbf{a}'\mathbf{a}} = \sqrt{\sum_i a_i^2} = \sqrt{4} = 2$ (“st. dev.”)
  - distance: $d(a, b) = \mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - 2\mathbf{a}'\mathbf{b} = 4$
  - angle: $\angle(a, b) = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} = \frac{2}{\sqrt{4 \cdot 4}} = 0.5$ (“correlation”)

Careers in Career-Space

- Alphabet $\mathcal{A} = \{a, b, c\}$ (labor market states)
- all strings $\mathcal{A}^*$: set of all possible careers
  - career $\mathrm{x} = abbcaaccbbaaaab\ldots$
  - careers are concatenations of symbols from $\mathcal{A}$
- career features: all sub-careers
  - a, ac, $abacb$, $\ldots$
- map careers onto career-feature vectors

VU UNIVERSITY AMSTERDAM

2 Careers in Career-Space

careers: $\mathrm{x} = abac \mapsto \mathbf{x}$, $\mathrm{y} = bacb \mapsto \mathbf{y}$

subcareers	$\mathbf{x}$	$\mathbf{y}$
a	1	1
$\vdots$	$\vdots$	$\vdots$
aa	1	0
ab	1	1
$\vdots$	$\vdots$	$\vdots$
aba	1	0
$\vdots$	$\vdots$	$\vdots$
acb	0	1
$\vdots$	$\vdots$	$\vdots$

- each possible subsequence is a feature

Feature Vectors: Problems I

- feature selection: relevance?
  - no beetles read Dickens (not applicable)
  - some beetles have horns (not selected)
  - all beetles have 6 legs (non-discriminating)
- feature selection: how many are necessary/acceptable?
  - $\{0,1\}^d$-vectors generate at most 2^d classes
  - dimensionality of subsequence-space is colossal: countably infinite

VU UNIVERSITY AMSTERDAM

Feature Vectors: Problems II

- feature selection: relevance?
- feature selection: how many are necessary/acceptable?
- calculating inner products
  - space/time-consuming because of size
  - “vector-avoiding” algorithms: “Kernels”
- Gram-matrix tends to be orthogonal: big diagonal
  - objects have everything in common with themselves
  - objects have little in common with other objects
  - compress sequences to shorter ones

VU UNIVERSITY AMSTERDAM

Constructing Sequence Vectors

- $\mathcal{A}$: alphabet; $\mathcal{A}^*$: set of all sequences over $\mathcal{A}$
- assign an integer, a rank number $r(u)$ to each $u \in (A)^* = \{u, v, w, \ldots\}$
- define vectors $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots)$ for each $x \in \mathcal{A}^*$ such that

$$\mathrm{x}_{r(u)} = \begin{cases} f(u,x) & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}$$

  - $f(u,x)$: "anything we like to use"
    - (as long as we can compute inner products $\mathbf{x}'\mathbf{y}$)
- distance: $d(x,y) = \mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} - 2\mathbf{x}'\mathbf{y}$
- similarity: $s(x,y) = \mathbf{x}'\mathbf{y}/(\mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} - \mathbf{x}'\mathbf{y})$

VU UNIVERSITY AMSTERDAM

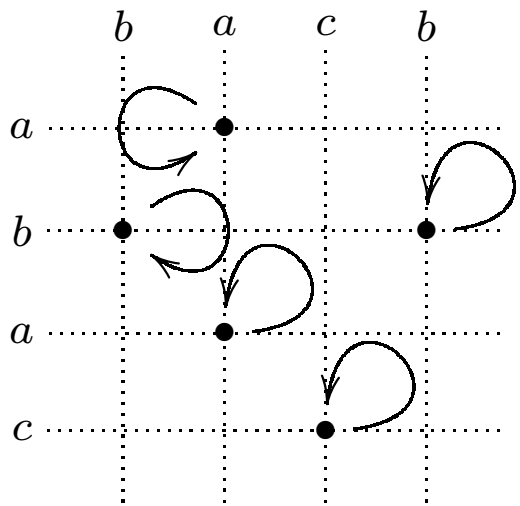
Constructing Sequence Vectors

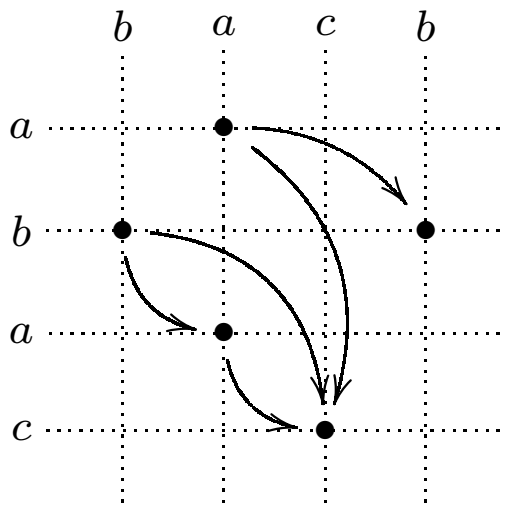
$$\mathrm{x}_{r(u)} = \begin{cases} f(u,x) & \text{if } u \sqsubseteq x \\ \\ 0 & \text{otherwise} \end{cases}$$

subsequence weighing	condition	$f(u,x)$
none (commonness)	$u \sqsubseteq x$	1
embedding frequency	$u \sqsubseteq x$	$\lvert x \rvert_u$
length	$u \sqsubseteq x$	$\ell(u)^p, \;\; p > 1$
limit gap-size	$(u \sqsubseteq x) \wedge gaps < d$	any
states	$u \sqsubseteq x$	$\prod_i w(u_i)$
duration	$u \sqsubseteq x$	$\sum_i t(u_i)$
soft-matching any subset sumultaneously	$u \sqsubseteq x$	any, with $\mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{S}\mathbf{y}$

VU UNIVERSITY AMSTERDAM

Computations in the Sequence Grid





$$\mathbf{M}^1 = \begin{pmatrix} & 1 & \\ 1 & & 1 \\ & 1 & \\ & & 1 \end{pmatrix} \quad \mathbf{M}^2 = \begin{pmatrix} & 2 & \\ 2 & & 0 \\ & 1 & \\ & & 0 \end{pmatrix} \quad \mathbf{M}^3 = \begin{pmatrix} & 0 & \\ 1 & & 0 \\ & 0 & \\ & & 0 \end{pmatrix}$$

$$\phi_k = \sum_{ij} m_{ij}^k, \quad \mathbf{x}'\mathbf{y} = \sum_k \phi_k = 5 + 5 + 1 = 11$$

VU UNIVERSITY AMSTERDAM

Embedding Frequency

- $x = abac$, $u = ac = x_1x_4 = x_3x_4$, $|x|_u = 2$
- $\mathbf{x}'\mathbf{y}$ counts "matching embeddings": $\sum_u |x|_u \cdot |y|_u$
- when repetition of patterns is important:
  - labor market careers
  - criminal careers
  - animal behavior sequences
- $$\mathrm{x}_{r(u)} = \begin{cases} |x|_u & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}$$

Length

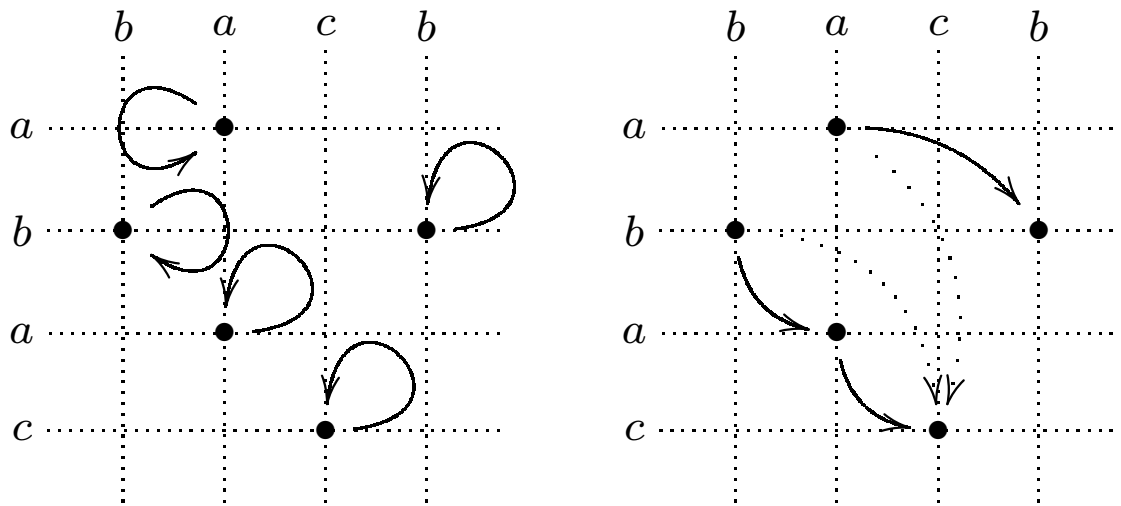
- $x = abac$, $u = ab$, $\ell(u) = 2$
- expect many short common subsequences
- focus on longer subsequences
- OM with standard costing:

$$d(x, y) = \ell(x) + \ell(y) - 2llcs(x, y)$$

- implementation $\mathbf{x}'\mathbf{y} = \sum_k k^p \phi_k$ or more sophisticated

$$\mathrm{x}_{r(u)} = \begin{cases} \ell(u)^p \cdot |x|_u & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}$$

Limiting the Gap-size



$$\mathbf{M}^1 = \begin{pmatrix} & 1 & & \\ 1 & & & 1 \\ & 1 & & \\ & & 1 & \end{pmatrix} \quad \mathbf{M}^2 = \begin{pmatrix} & 1 & & \\ 1 & & & 0 \\ & 1 & & \\ & & 0 & \end{pmatrix} \quad \mathbf{M}^3 = \begin{pmatrix} & 0 & & \\ 1 & & & 0 \\ & 0 & & \\ & & 0 & \end{pmatrix}$$

$$\phi_k = \sum_{ij} m_{ij}^k, \quad \mathbf{x}'\mathbf{y} = \sum_k \phi_k = 5{+}3{+}1 = 9$$

Weighing States

- some states may be more significant than others
  - unemployed
  - infected
- Assign a weight $w(a_i)$ to each state in $\mathcal{A} = \{a_1, a_2, \ldots\}$
- Calculate the weight of u as $\prod_i w(u_i)$
  - $w(abac) = w(a) \cdot w(b) \cdot w(a) \cdot w(c)$

VU UNIVERSITY AMSTERDAM

Weighing the States

$$x = abac, \quad y = bacb$$

$$w(a) = 2,\ w(b) = 1,\ w(c) = 3$$

$$\mathbf{M}^1 = \begin{pmatrix} & 2 & \\ 1 & & 1 \\ & 2 & \\ & & 3 \end{pmatrix} \quad \mathbf{M}^2 = \begin{pmatrix} & 8 & \\ 6 & & 0 \\ & 6 & \\ & & 0 \end{pmatrix} \quad \mathbf{M}^3 = \begin{pmatrix} & 0 & \\ 6 & & 0 \\ & 0 & \\ & & 0 \end{pmatrix}$$

$$\phi_k = \sum_{ij} m_{ij}^k, \quad \mathbf{x}'\mathbf{y} = \sum_k \phi_k = 9 + 20 + 6 = 35$$

$$\mathrm{x}_{r(u)} = \begin{cases} w(u) \cdot |x|_u & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}$$

Duration

- $x = x_1 x_2 x_3 \ldots$, $t_x = t_1 t_2 t_3 \ldots$
  - time, pitch, voltage, speed
  - *any* quantifiable state-property
- $t(x) = t_1 + t_2 + t_3 + \ldots$
- $$\mathrm{x}_{r(u)} = \begin{cases} t(u) w(u) \ell(u)^p \cdot |x|_u & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}$$
- duration is treated like any other weight

VU UNIVERSITY AMSTERDAM

Soft-Matching ("indel cost")

- "Hollister's Problem": some states are more different than others
  - (Single,Married) vs (Cohabitation,Married)
- coordinates are "hard": either 0 or >0
  - $\mathbf{x}'\mathbf{y} = \sum_i x_i y_i$, $x_i = 0$ and/or $y_i = 0 \Rightarrow x_i y_i = 0$
- Inner product compares values of equally indexed coordinates
  - never "compares" subsequences containing "Married" with subsequences containing "Single"
- In OM: substitution cost - compare different states

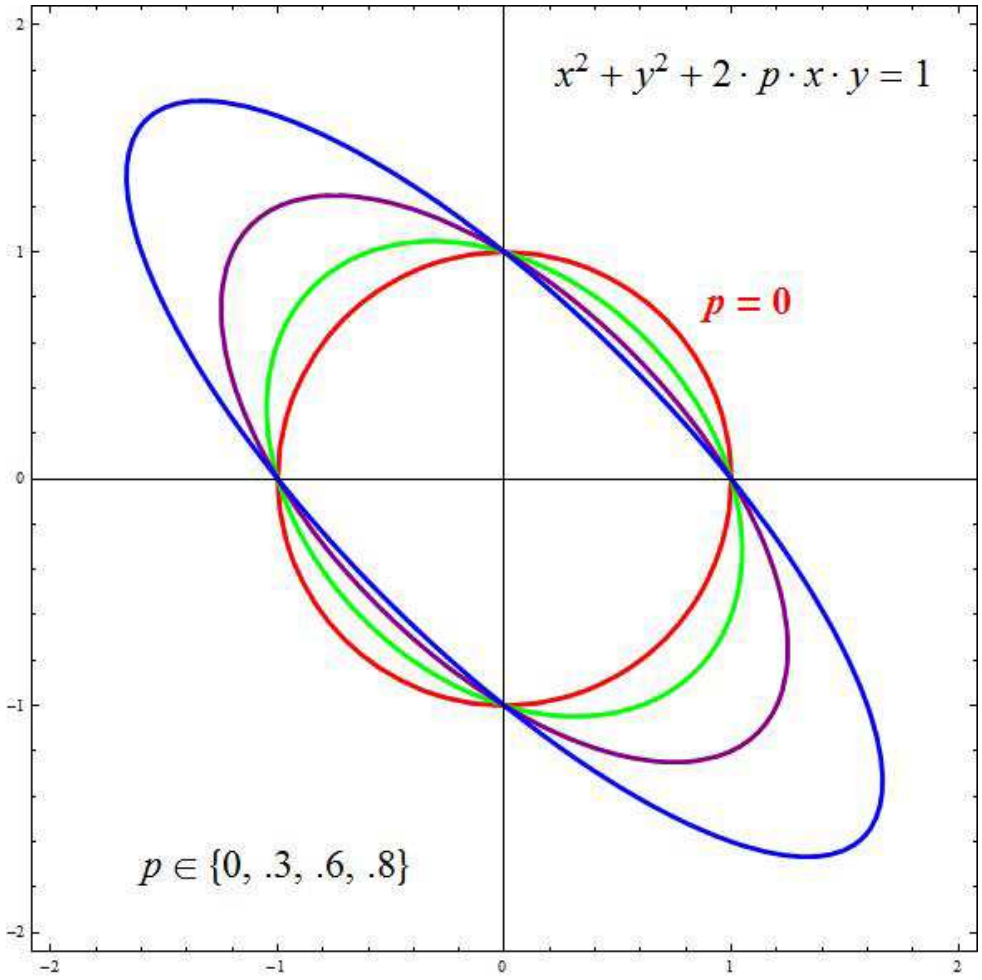
VU UNIVERSITY AMSTERDAM

Soft-Matching ("indel cost")

- Define state-similarities $\mathbf{M} = (m_{ij})$, $0 \leq m_{ij} \leq 1$, $m_{ii} = 1$, $m_{ij} = m_{ji}$
- calculate $\mathbf{x'My}$ instead of $\mathbf{x'y}$
  - $\mathbf{x'My} = \underbrace{\sum_i x_i y_i}_{\text{"}hard\text{"}} + \underbrace{2 \sum_{i \neq j} x_i \cdot m_{ij} \cdot y_j}_{\text{"}soft\text{"}}$
  - $d(x, y) = \mathbf{x'Mx} + \mathbf{y'My} - \mathbf{2x'My}$
  - Euclidean distance in "elliptical" space
  - implementation in Grid-algorithm: $\mathbf{M}^1 \leftarrow \mathbf{M}$

VU UNIVERSITY AMSTERDAM

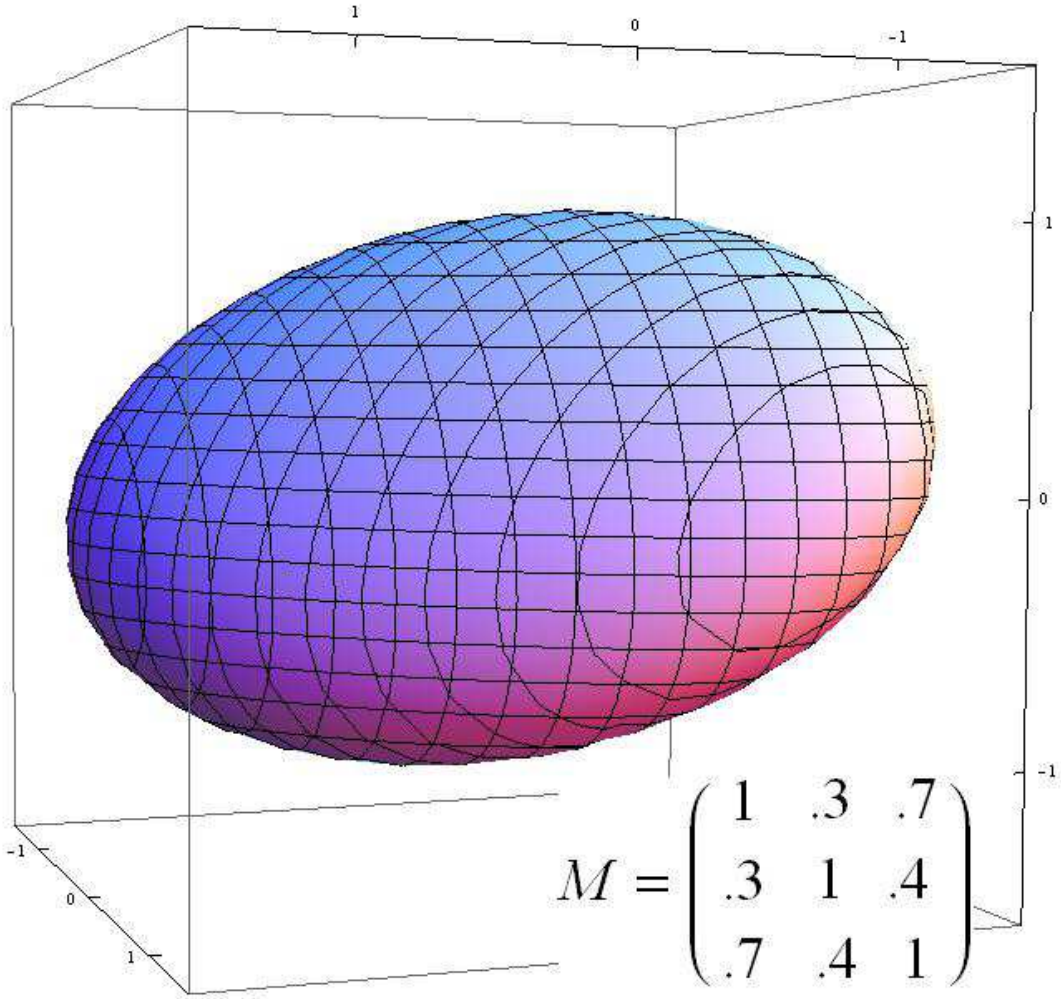
Unit-circles in Elliptical Space



M-Elliptical Circles: $\mathbf{x}'\mathbf{M}\mathbf{x} = 1, \quad \mathbf{M} = \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix}$

VU UNIVERSITY AMSTERDAM

Unit-Sphere in Elliptical Space



$$M = \begin{pmatrix} 1 & .3 & .7 \\ .3 & 1 & .4 \\ .7 & .4 & 1 \end{pmatrix}$$

M-Elliptical 1-sphere

An Application

- Data described in
  - Müller, Gabadinho, Ritschard & Studer: Extracting knowledge from life courses: clustering and visualization, *Data Warehousing and Knowledge Discovery*, Lecture Notes in Computer Science, 2008, Volume 5182/2008, 176-185

- 4318 individuals, born 1909-1972, $\geq$30 years old, retrospective data
  - 4 events: L(eft parental home), (1st) M(arriage), (1st) C(child), (1st) D(ivorce)
  - 8 states: P(no event, still with Parents), L, M, LM, C, LC, LMC, D
  - data example:

birth	L	M	C	D
1974	1992	1994	1996	-

  - sequence example:

74	. . .	91	92	93	94	95	96	97	98	. . .
P	. . .	P	L	L	LM	LM	LMC	LMC	LMC	. . .

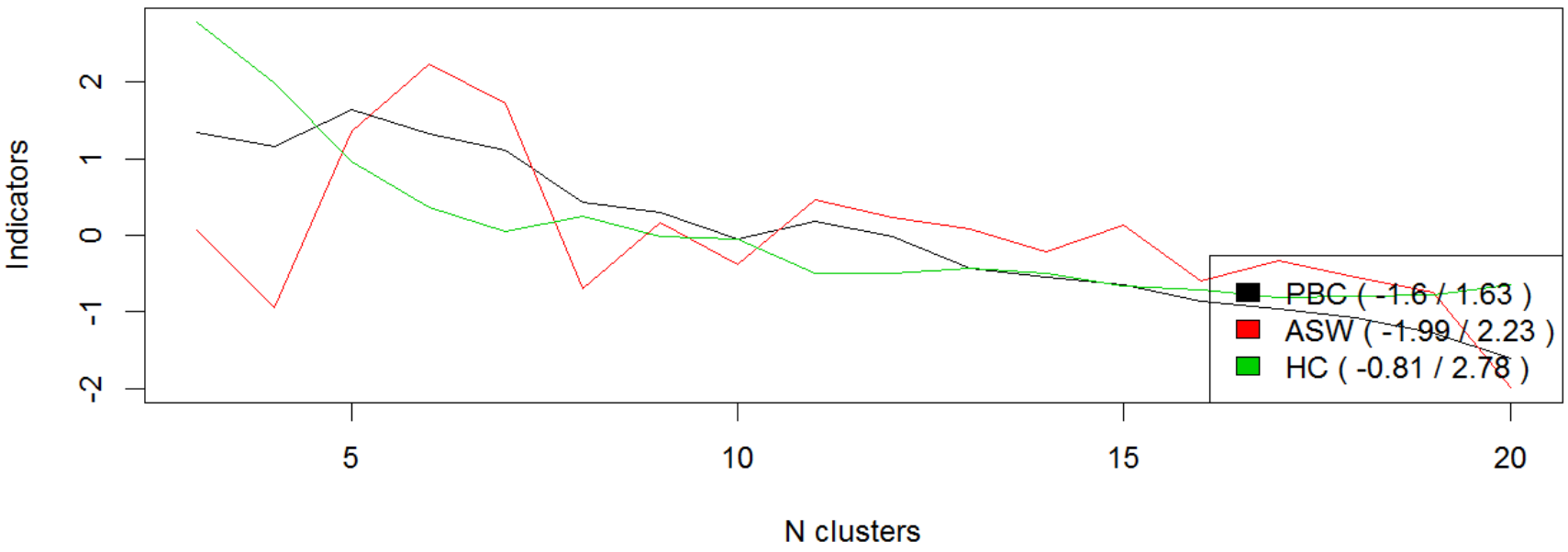
An Application

- OM-cost matrix $\mathbf{C}$:

	P	L	M	LM	C	LC	LMC	D
P	0							
L	.25	0						
M	.38	.62	0					
LM	.50	.25	.38	0				
C	.25	.50	.38	.75	0			
LC	.50	.25	.62	.50	.25	0		
LMC	.75	.50	.38	.25	.50	.25	0	
D	.75	.74	.38	.50	.75	.75	.50	0

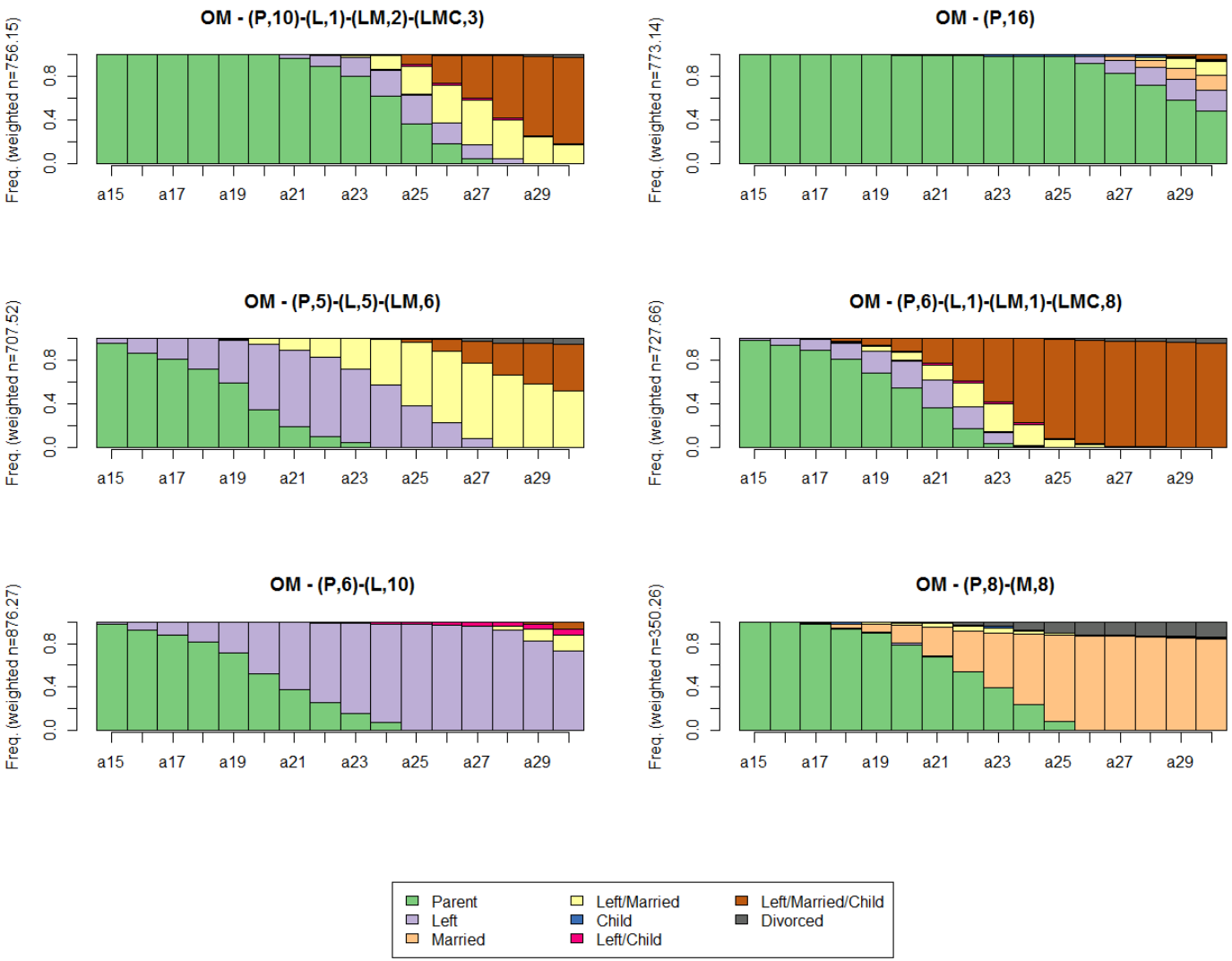
- soft-matching matrix $\mathbf{M} = \mathbf{1} - \mathbf{C}$: $m_{ij} = 1 - c_{ij}$

Finding Clusters (PAM) with d_{OM}

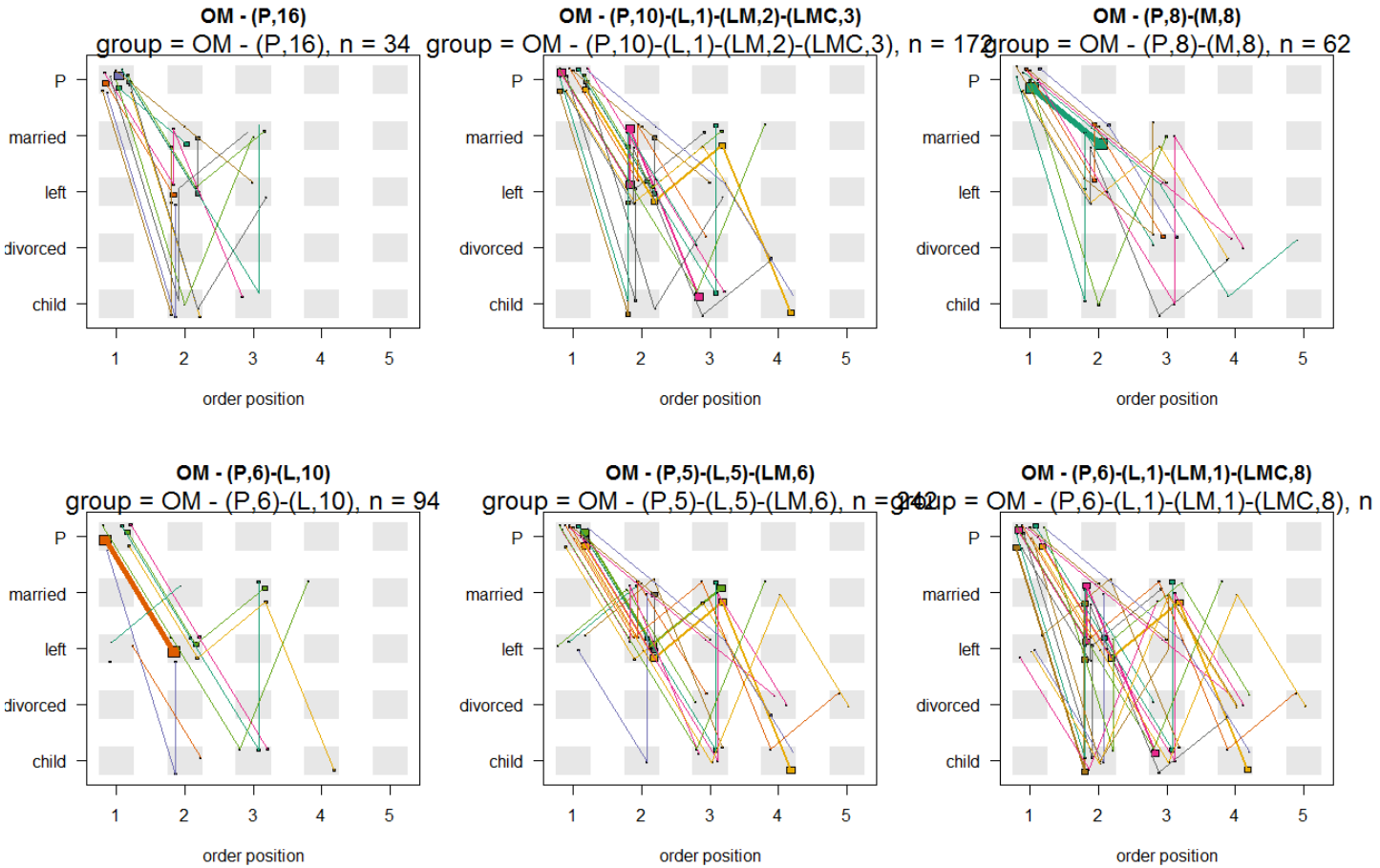


6 clusters seems optimal

OM-cluster profiles: Chronogrammes



OM-cluster profiles: order-plots

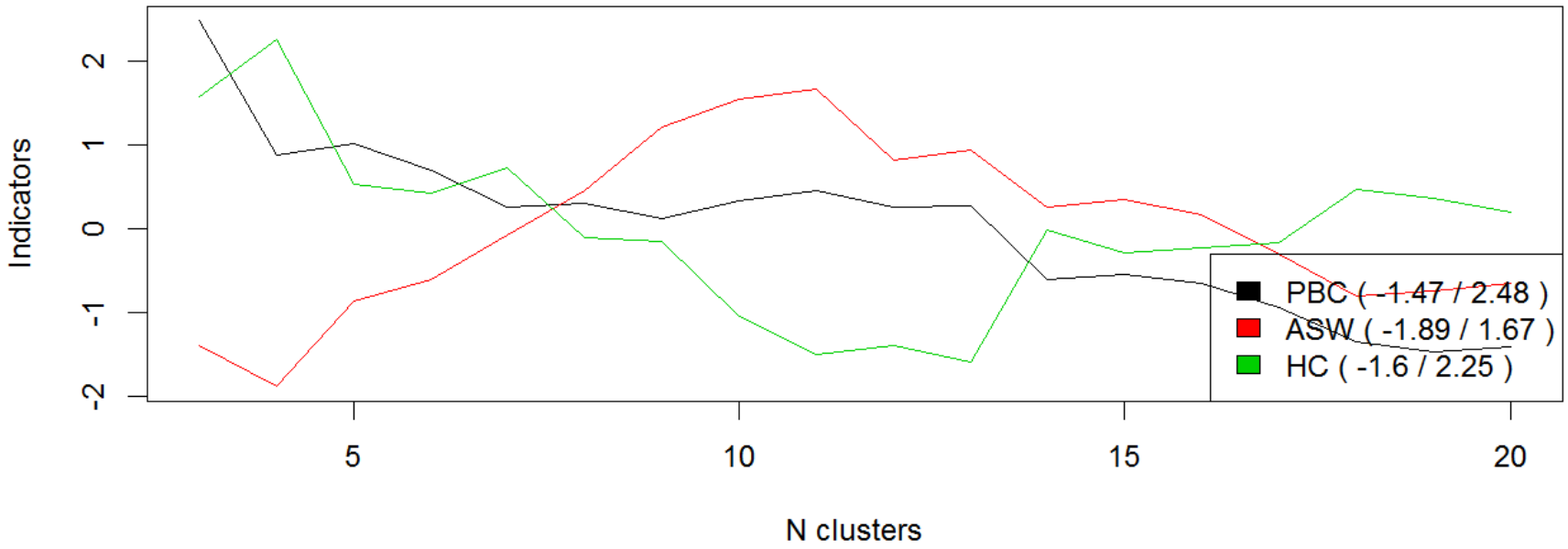


An alternative clustering

Using subsequences, weighed for

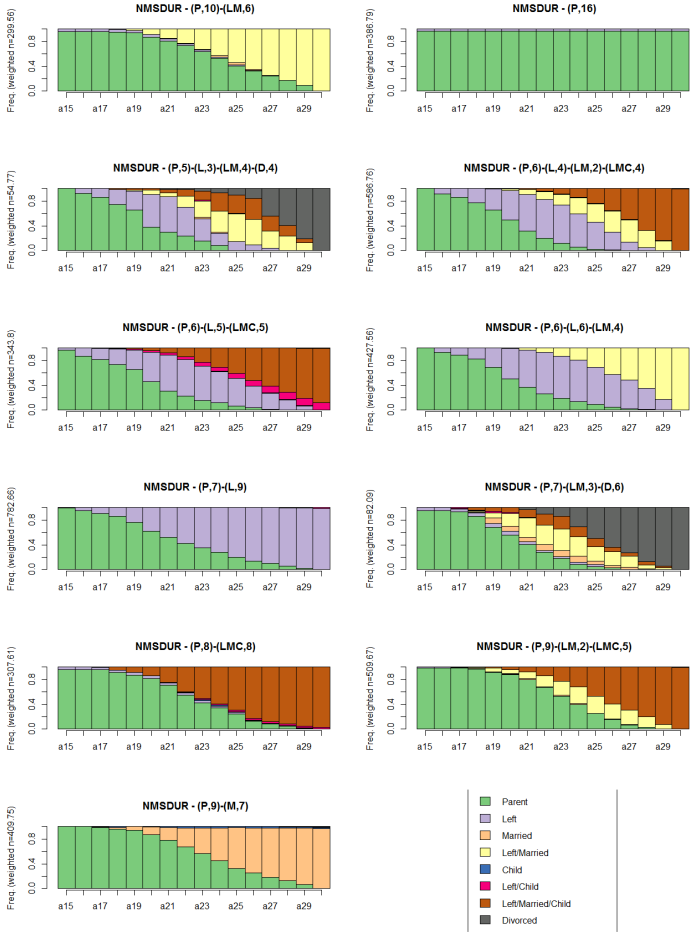
- duration
- embedding frequency

and employing soft-matching

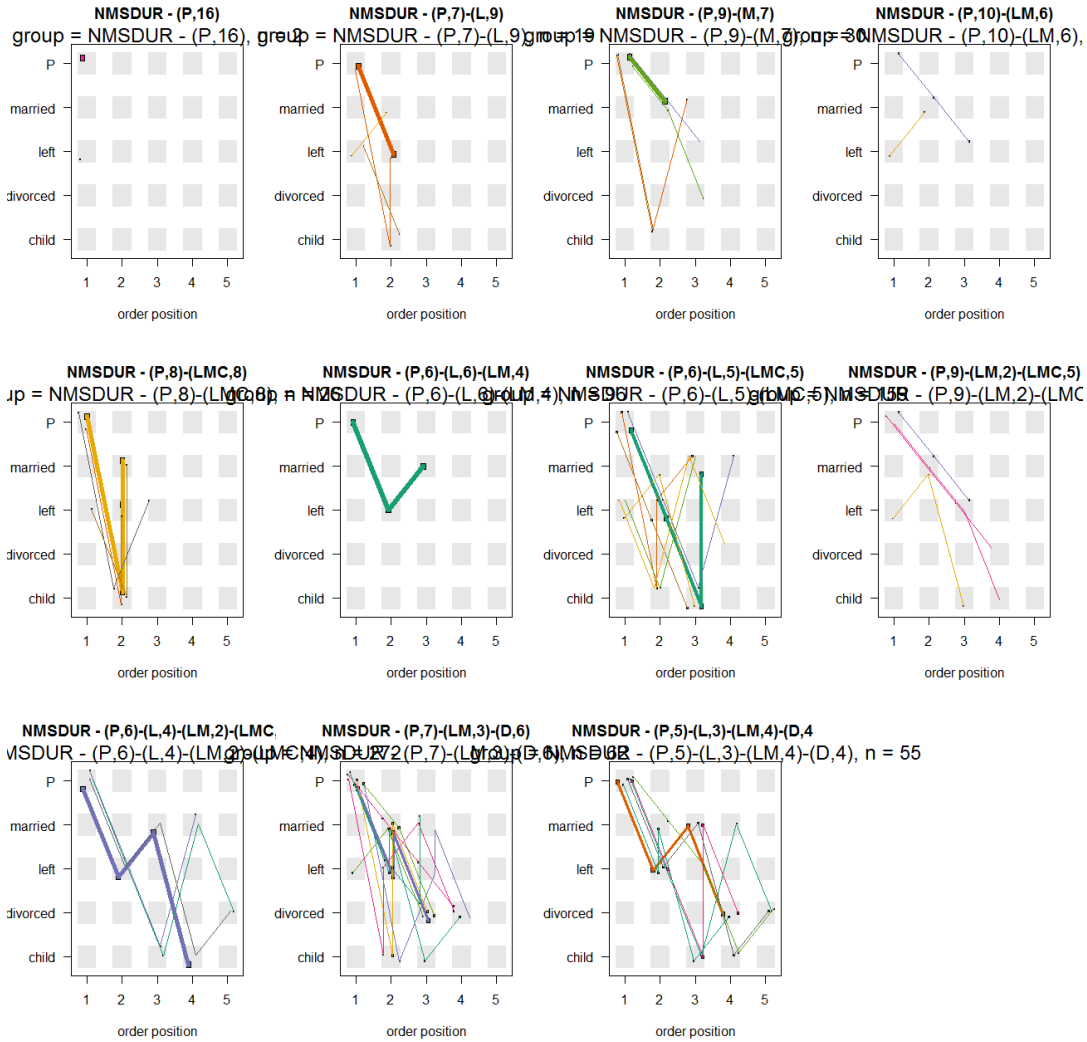


11 clusters seems optimal

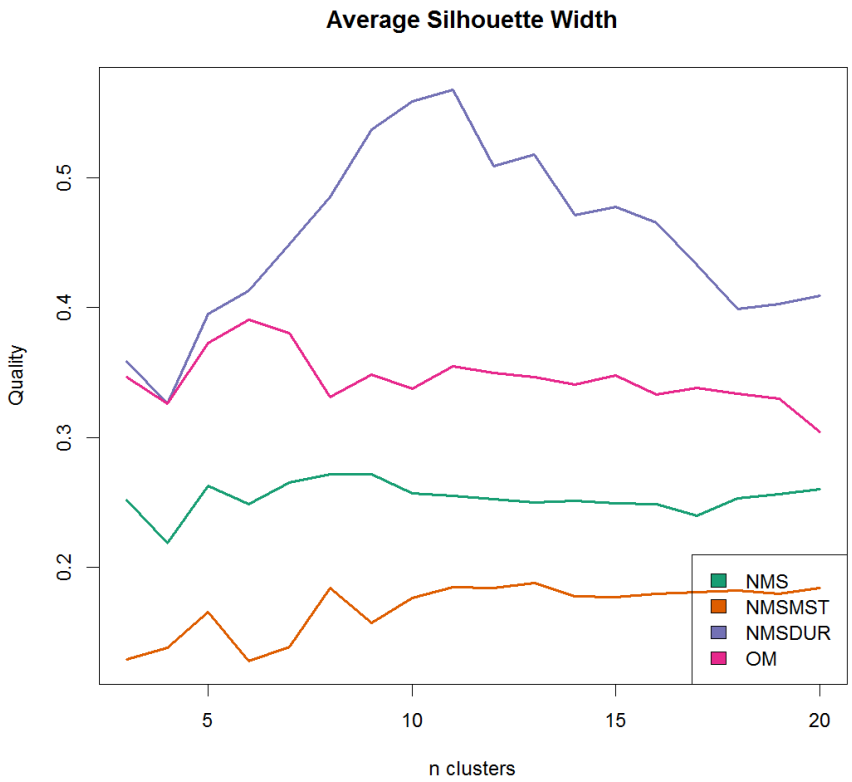
Cluster profiles: Chronogrammes

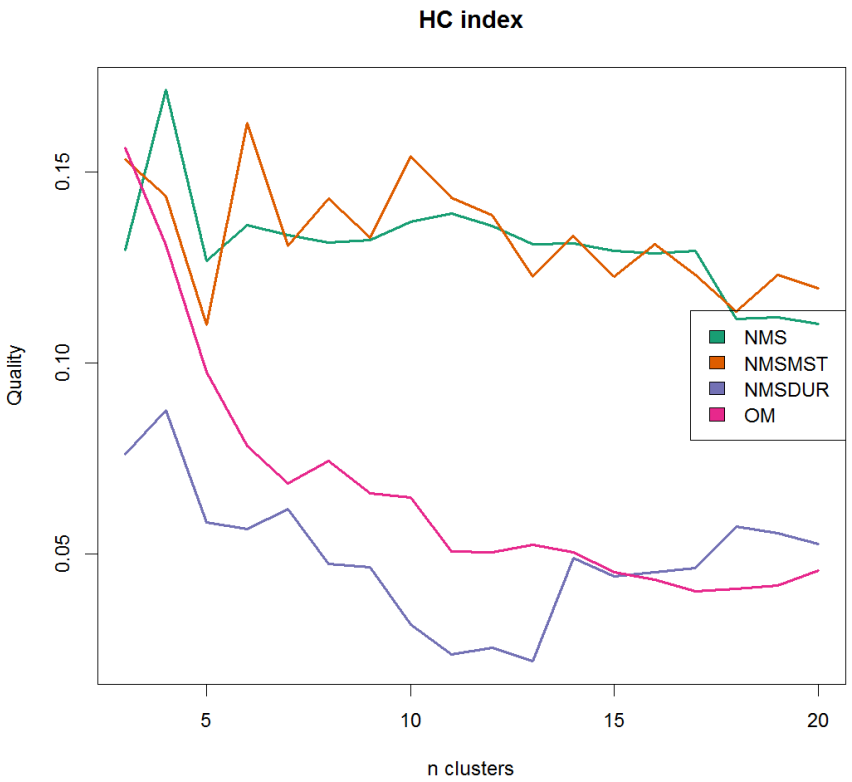


Cluster profiles: order-plots



Clustering Quality





VU UNIVERSITY AMSTERDAM

Choosing Distance is Not Trivial !

Cramer's V: Soft-Matching & OM

	emb	emb/dur	OM
emb	1.0		
emb/dur	.67	1.0	
OM	.84	.65	1.0

THANK YOU

VU UNIVERSITY AMSTERDAM