Sequence Analysis: A Toolbox
III: Beyond Distance and Similarity

Cees H. Elzinga
Research Program PARIS
Department of Sociology
VU University Amsterdam

August 7, 2014

Topics

- Hypothesis on distance/similarity
  - H_0: Life courses of children are more similar to those of their parents than to the life courses of a random person from the parental cohort (intergenerational transfer)
  - H_0: Life courses of the older cohort are more similar than life courses of younger cohorts (“Second Demographic Transition”)
- Hypothesis on/Exploration of “Structure”
  - Grouping around predefined (hypothetical) patterns
  - Exploring for structure: finding k groups
- Relating Structure to Covariates
  - Discrepancy Analysis

Something happened in France!

	'45-'49	'60 -'64
average # dist. states	2.87	3.11
average length	2.96	3.42
% S M MC	48.4	14.5
% S U M MC	5.7	11.6
% Miscellaneous	18.8	38.5
% dist. trajectories	13.3	18.7

. . . like in most other countries

Apples and Trees

	full		all	
	$\overline{s}_{OM}$	$\overline{s}_{M}$	$\overline{s}_{OM}$	$\overline{s}_{M}$
Random p-p	0.52	0.45	0.46	0.24
random c-c	0.38	0.17	0.47	0.15
random p-c	0.35	0.19	0.30	0.19
p-c pairs	0.39	0.23	0.33	0.35

Intergenerational transfer of life course pattern does it exist?

LCS and family formation in Austria

- 2499 Austrian women
- born between 1945 and 1964
- LCS-based similarity, without and with (MST) durations
- average s decreases, char. sequences change

Cohort	Size	$\overline{s}_{LCS}$	sd	Char. Seq.	$\overline{s}^*_{LCS}$	sd	Char. Seq.
1945-49	539	.67	.15	S M MC	.49	.13	S/50 M/9 MC/85
1950-54	558	.64	.14	S M MC	.44	.12	S/45 M/16 MC/83
1955-59	650	.60	.12	S U M MC	.41	.11	S/53 U/6 M/8 MC/77
1960-64	752	.57	.12	S U M MC	.39	.10	S/53 U/11 M/12 MC/68

- Note that these averages are NOT from "independent observations"
  - bootstrap the distributions to determine significance of differences

Topics

- Hypothesis on distance/similarity
  - H_0: Life courses of children are more similar to those of their parents than to the life courses of a random person from the parental cohort (intergenerational transfer)
  - H_0: Life courses of the older cohort are more similar than life courses of younger cohorts ("Second Demographic Transition")
- Hypothesis on/Exploration of "Structure"
  - Grouping around predefined (hypothetical) patterns
  - Exploring for structure: finding k groups
- Relating Structure to Covariates
  - Discrepancy Analysis

VU UNIVERSITY AMSTERDAM

Structure: Partitions

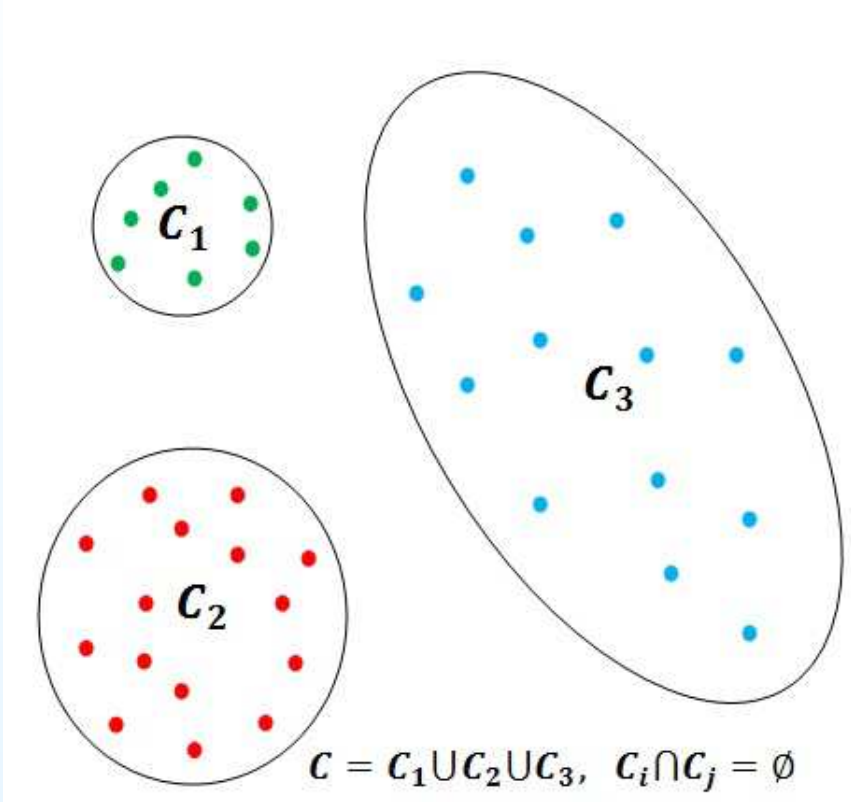
- Partition: $C = C_1 \cup \ldots \cup C_k$, Membership-function M
  - weak: $M_i(x) = \begin{cases} 1 & \text{iff } x \in C_i \\ 0 & \text{otherwise} \end{cases}$
  - strong: $M_i(x) = \begin{cases} 1 & \text{iff } x \in C_i \\ 0 & \text{otherwise} \end{cases}, C_i \cap C_j = \emptyset,$
  - fuzzy: $0 \leq M_i(x) \leq 1, \sum_i M_i(x) = 1$

Templates of Family Formation

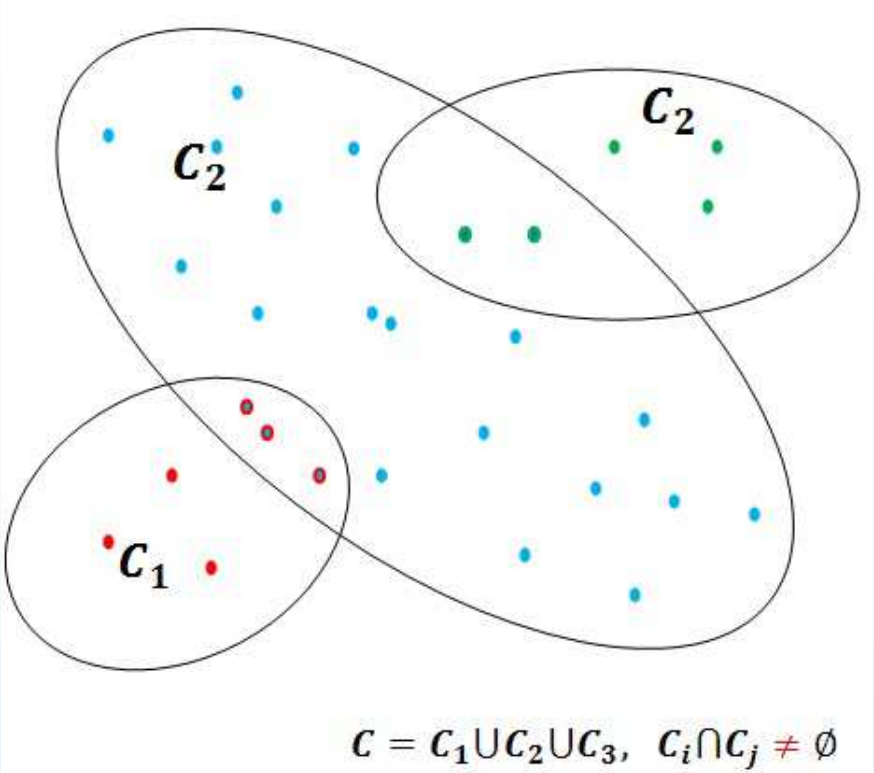
- define templates
- calculate d (ignoring duration) for all women to each template
- assign each woman to closest template

	1945-49 $N = 539$ $\overline{s}_{\mathcal{S}} = .49$		1960-64 $N = 752$ $\overline{s}_{\mathcal{S}} = .37$	
Template	%	$\overline{s}_{\mathcal{S}}$	%	$\overline{s}_{\mathcal{S}}$
S M MC	53.6	.92	23.1	.88
S U M MC	10.4	.79	25.8	.82
S U UC	8.7	.33	17.4	.59
S U S U	1.7	.59	8.2	.43
S M MC SC	16.3	.58	17.6	.44
S	9.3	.84	7.8	.82

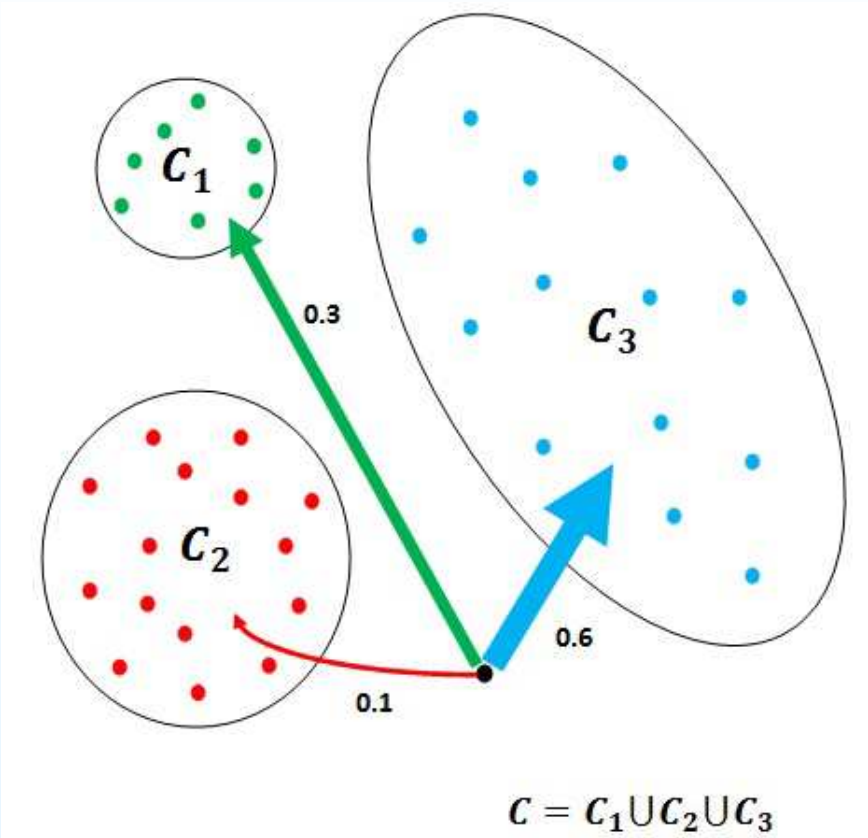
Strong Partition



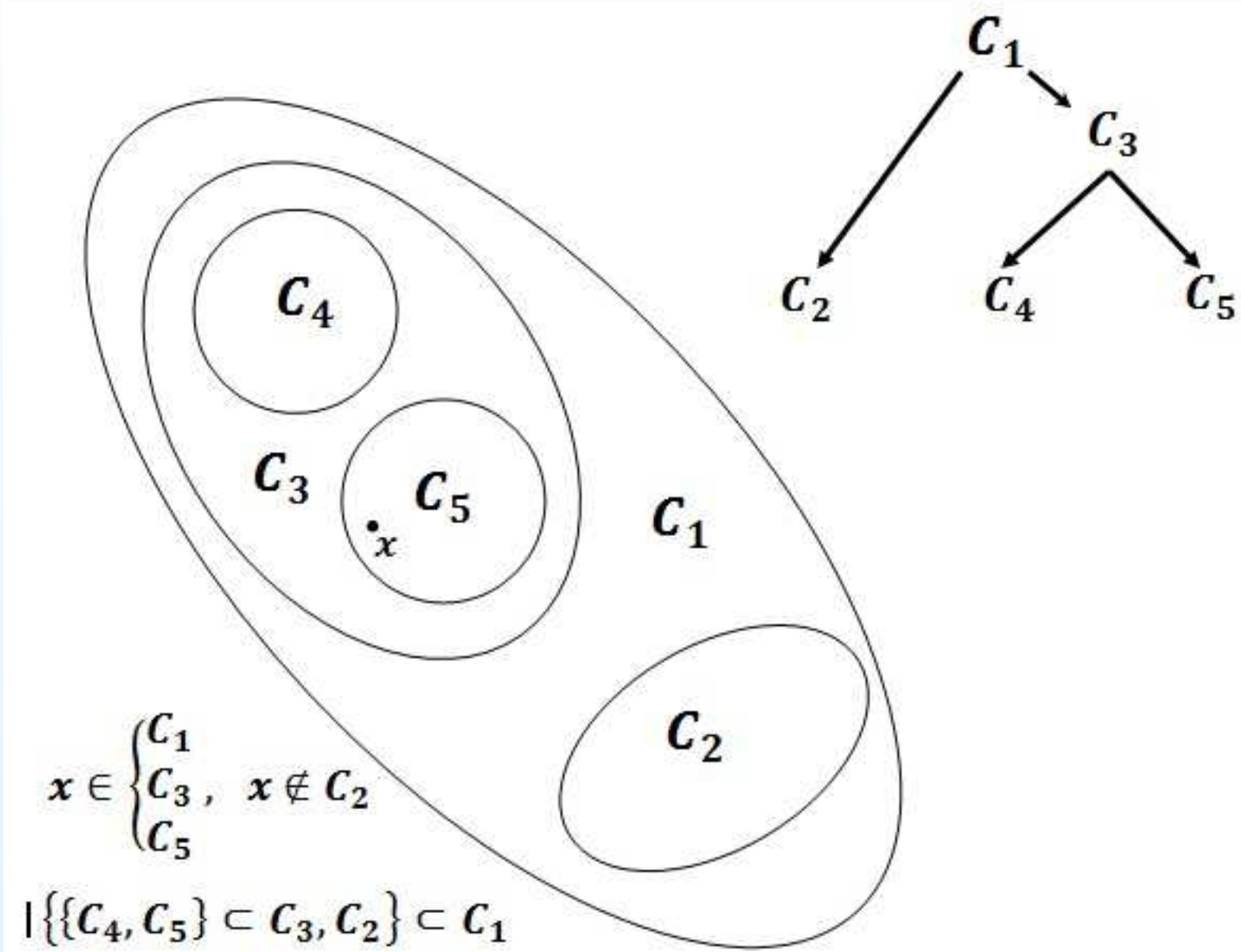
Weak Partition



Fuzzy Partition



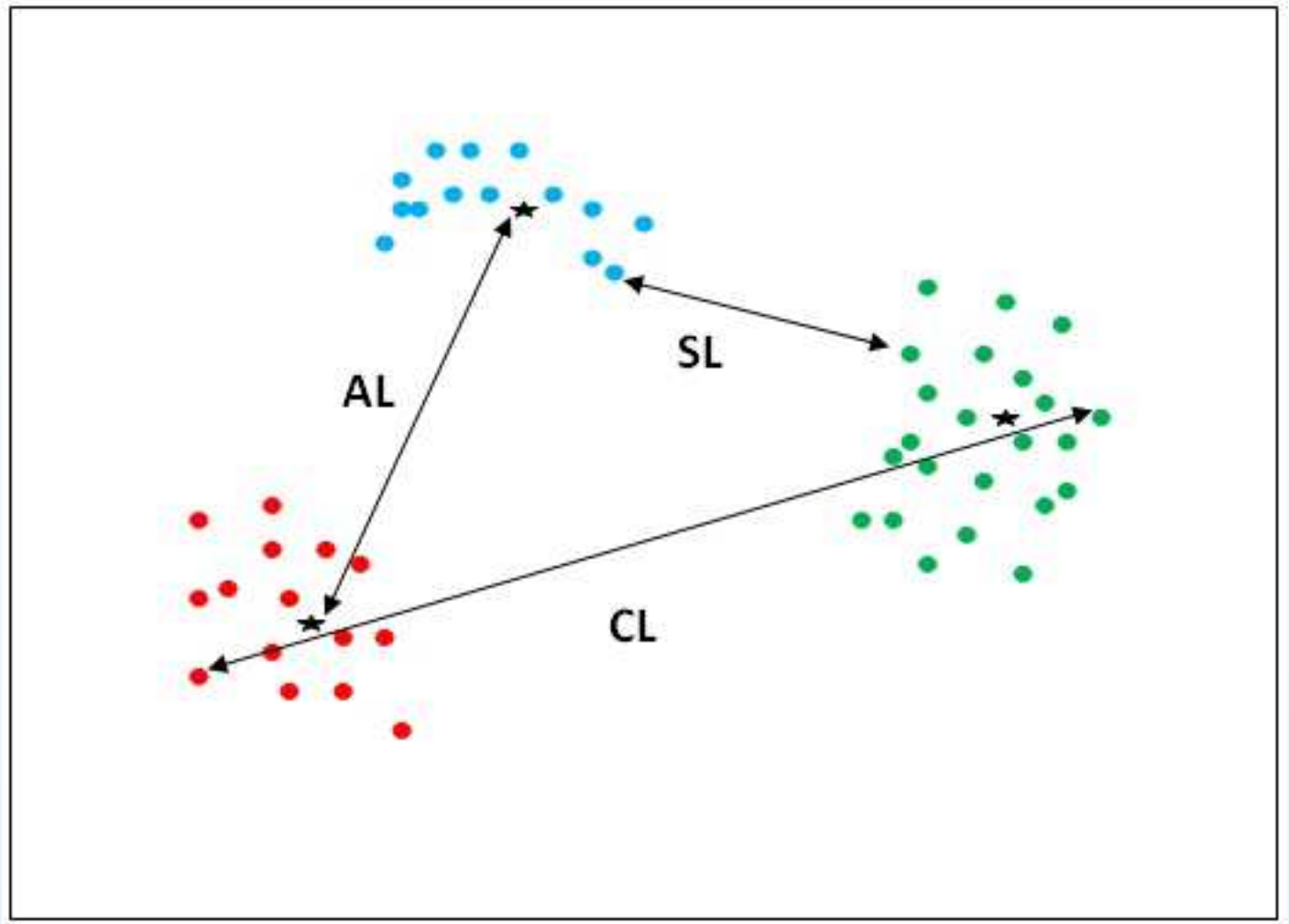
Hierarchy (Tree, Dendrogram)



Finding Structure

- Structure: groups/classes/clusters/parts that have more internal similarity than the similarity between them
  - similarity $\approx$ "inverse" distance
- define similarity between groups: distance between
  - closest members ("single linkage")
  - most remote members ("complete linkage")
  - centroids or medoids("average linkage")

Finding Structure: Linkage



Ward's Agglomerative Algorithm

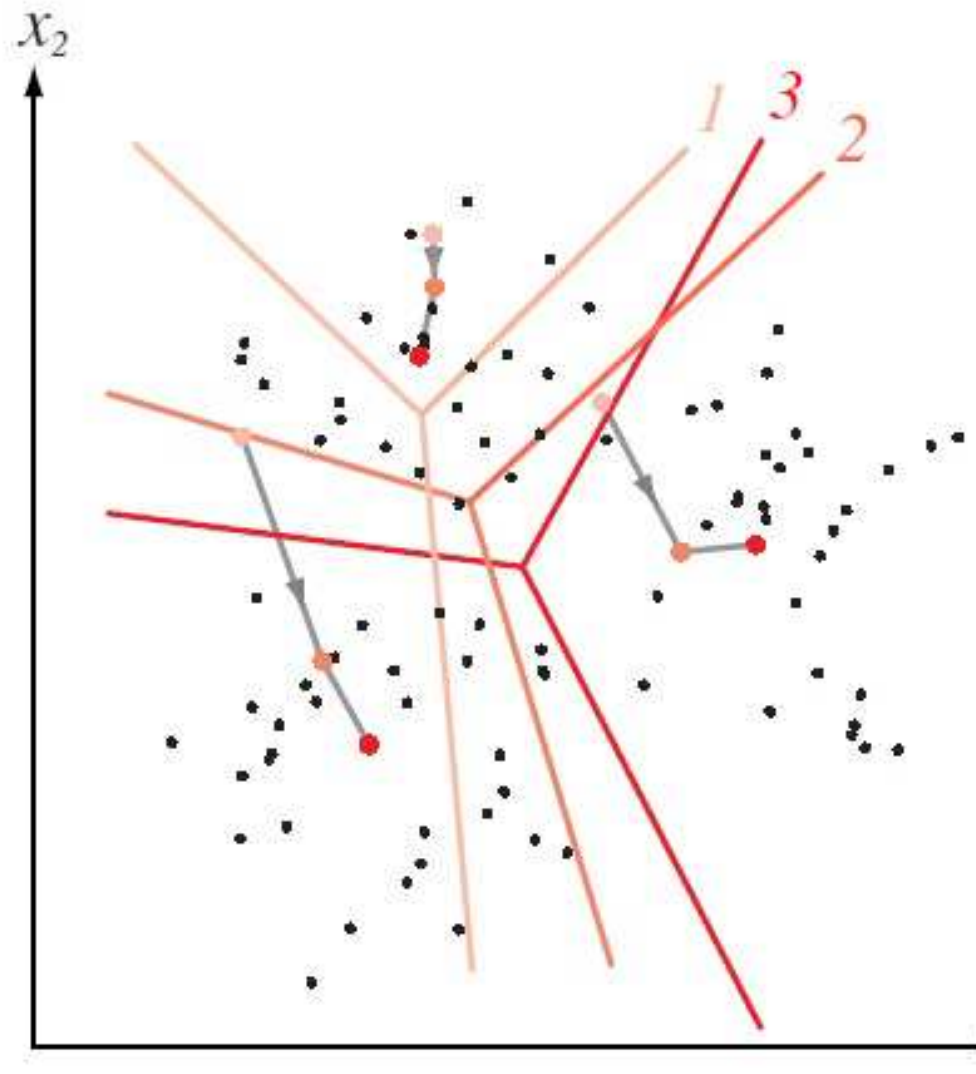
- given k clusters, calculate for all pairs of clusters i, j
- the merging cost $\delta(i,j) = \frac{n_i n_j}{n_i + n_j} d^2(\bar{c}_1, \bar{c}_2)$
  - i.e. the difference of $\sum_m d^2(x, \bar{c}_{i \cup j}) - \left(\sum_m d^2(x_{m,i}, \bar{c}_i) + \sum_m d^2(x_{m,j}, \bar{c}_j)\right)$
  - i.e. the increase in "within-cluster SS"
- merge clusters i, j for which $\delta(i,j)$ is minimal
  - tends to merge smaller cluster
  - stop merging when δ suddenly jumps
  - given k, $\sum_{j=1}^{k} \sum_{i=1}^{n_j} d^2(x_i, c_j)$ is probably **not minimal**
  - why should SS per cluster be small??

K-means Algorithm

- Step 1: Pick k objects $c_1 \ldots c_k$ from the set of objects
- Step 2: Assign each object to the nearest c_j
  - this generates k clusters
- Step 3: Calculate new centroids $c_1 \ldots c_k$ for the k clusters
- Step 4: Assign each object to the nearest c_j
- Step 5: When at least one object changes cluster, go to Step 3, else Stop.

VU UNIVERSITY AMSTERDAM

K-means at work

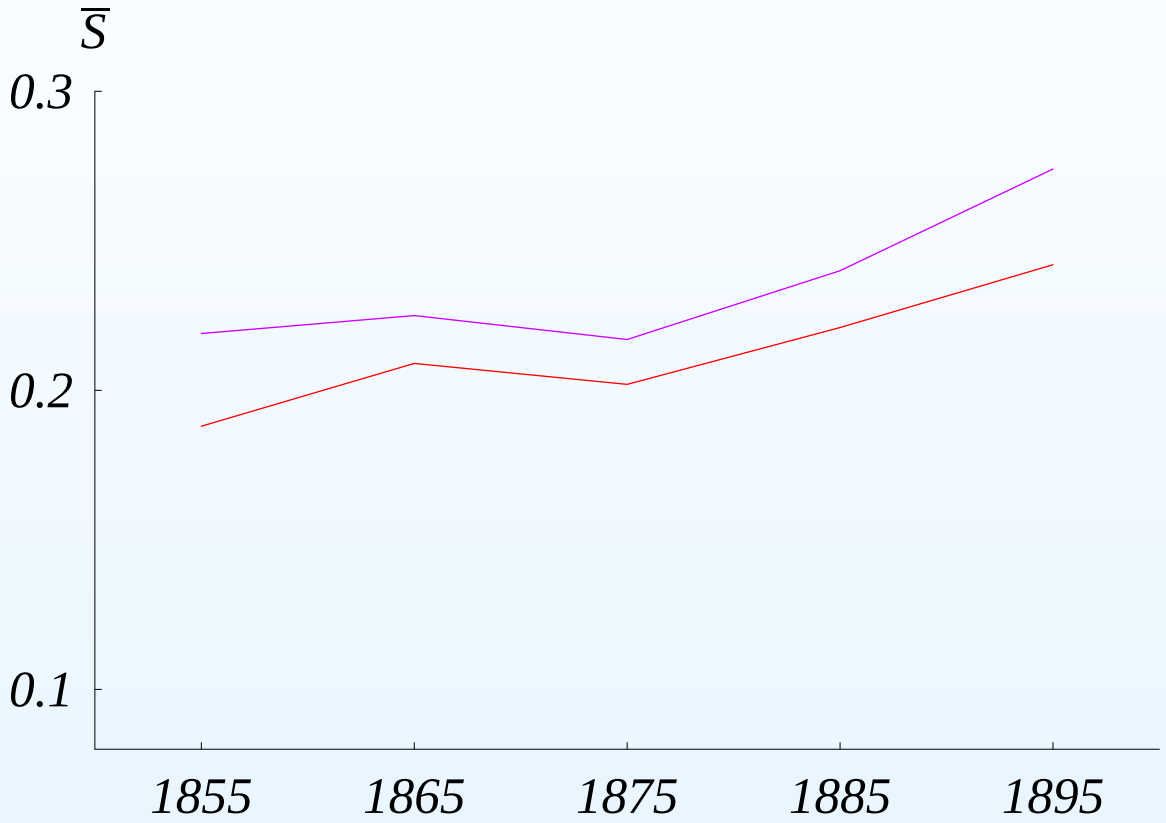


Historical Sample Netherlands (HSN)

- archive data 1850-1900, 4651 records
- age 15 - 40 years
  - monthly statuses
  - males & females
- 5 birth cohorts
  - born between 1850-59, 1860-69, 1870-79, 1880-89, 1890-99
- family formation histories: 10 distinct statuses
  -

S	Single	C	Children	N	Non-Family
P	Parents	MP	Married&Parents	D	Dead
M	Married	MPC	Mar.&Par.&Chil.		
MC	Mar.&Chil.	F	Family		

Dutch 19^{th} century Similarity



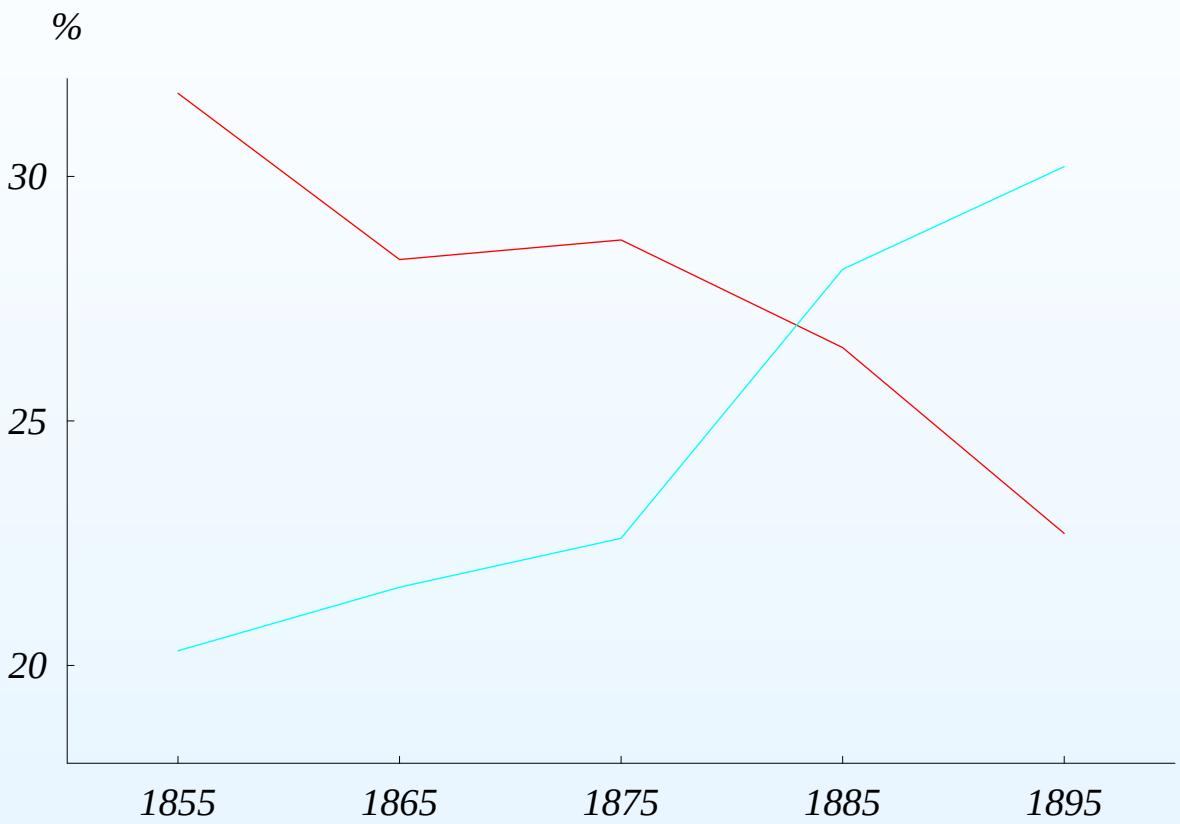
- all, unskilled workers
- Standardization!

Clustering NHS

	%	$\overline{S}$	Char. Sequence
Early Deaths	7.65	.66	P/64 D/236
Servants	14.13	.42	P/54 N/58 M/14 MC/174
TEP	26.60	.72	P/104 M/7 MC/189
TLP	11.83	.62	P/167 M/13 MC/120
Childless	5.07	.56	P/152 M/148
Non-Part	8.64	.81	P/300
Rest	26.08	.08	

- K-Means, best out of 100 initial configurations

Clusters per Cohort



- % Rest, % TEM
- Standardization!

K-means Demo

Please Start the Demo

Topics

- Hypothesis on distance/similarity
  - H_0: Life courses of children are more similar to those of their parents than to the life courses of a random person from the parental cohort (intergenerational transfer)
  - H_0: Life courses of the older cohort are more similar than life courses of younger cohorts ("Second Demographic Transition")
- Hypothesis on/Exploration of "Structure"
  - Grouping around predefined (hypothetical) patterns
  - Exploring for structure: finding k groups
- Relating Structure to Covariates
  - Discrepancy Analysis

Discrepancy Analysis I

- M. Studer et al: Discrepancy Analysis of State Sequences, Sociological Methods and Research. August 2011. Vol. 40(3), pp. 471-510
- implemented in TraMineR: downloadable from http://www.graphviz.org/

Purpose: "Explain" distances in terms of (discretized) covariates

Basic Observation:

$$SS = \sum_i (x_i - \bar{x})^2 = \frac{2}{N(N-1)} \sum_{i,j>i} d^2(i,j)$$

- the sum of squared deviations is a sum of squared distances!

Discrepancy Analysis II

Statistics are "ANOVA-like":

$$R^2 = \frac{SS_B}{SS_T}, \quad F = \frac{SS_B/(m-1)}{SS_T/(\frac{1}{2}N(N-1)-m)}$$

	F	R^2	p
gcse5eq	184.09	.206	.000
grammar	89.87	.112	.000
fmpr	23.00	.031	.000
funemp	24.05	.033	.000
sex	13.85	.019	.001
region	7.26	.039	.001
liveboth	2.61	.004	.240
religion	1.88	.003	.365

Thank You

- Topics
- Something happened in France!
- Apples and Trees
- LCS and family formation in Austria
- Topics
- Structure: Partitions
- Templates of Family Formation
- Strong Partition
- Weak Partition
- Fuzzy Partition
- Hierarchy (Tree, Dendrogram)
- Finding Structure
- Finding Structure: Linkage
- Ward's Agglomerative Algorithm
- K-means Algorithm
- K-means at work
- Historical Sample Netherlands (HSN)
- Dutch 19^{th} century Similarity
- Clustering NHS
- Clusters per Cohort
- K-means Demo
- Topics
- Discrepancy Analysis

VU UNIVERSITY AMSTERDAM