STAT406- Methods of Statistical Learning Lecture 25

Matias Salibian-Barrera

UBC - Sep / Dec 2017

Multidimensional scaling - Goal

- Given observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$

  find lower dimensional points $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ (e.g. in 2 or 3 dimensions)

  that have a similar configuration (pair-wise distances) to that of the $\mathbf{X}_j$'s

- This is different from PCA

Classical (Metric) Solution

- Look at the original pairwise distances

$$\delta_{ij} \;=\; \left\| \mathbf{X}_i - \mathbf{X}_j \right\|$$

construct $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^k$ with

$$d_{ij} \;=\; \left\| \mathbf{Y}_i - \mathbf{Y}_j \right\| \;\approx\; \delta_{ij}$$

Classical (Metric) Solution

- δ_{ij} are a dissimilarity if

$$\delta_{ii} \; = \; 0 \, , \quad \delta_{ij} \; \geq \; 0 \, , \quad \delta_{ij} \; = \; \delta_{ji}$$

- It is Euclidean if there exist $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^r$ such that

$$\delta_{ij} \; = \; \left\| \mathbf{w}_i - \mathbf{w}_j \right\|$$

Multidimensional scaling

- Theorem: (Gower, 1966; Mardia, 1979) Let $a_{ij} = -\delta_{ij}^2/2$ and

$$\mathbf{B} \ = \ (\mathbf{I}_n - \mathbf{1}_n \, \mathbf{1}_n'/n) \ \mathbf{A} \ (\mathbf{I}_n - \mathbf{1}_n \, \mathbf{1}_n'/n)$$

  then δ is Euclidean iff $\mathbf{B} \geq \mathbf{0}$
- Theorem is constructive

Multidimensional scaling

- Finds $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^f$ where $f = \text{rank}(\mathbf{B})$ such that

$$\delta_{ij} \; = \; \left\| \mathbf{y}_i - \mathbf{y}_j \right\|$$

(the $\mathbf{y}$'s are the rows of the $n \times f$ matrix of eigenvectors of $\mathbf{B}$ with non-zero eigenvalues, properly standardized)

MDS

- Do not need $\mathbf{X}_j$'s
- When δ is not Euclidean some eigenvalues of $\mathbf{B}$ are negative
- We can still take the k eigenvectors (properly standardized) associated to the largest positive eigenvalues

Non-metric MDS

- Focus on the order of the pairwise distances (rather than their magnitudes)

$$\text{minimize } \mathrm{STRESS}^2 = \frac{\sum \left(f\left(\delta_{ij}\right) - d_{ij}\right)^2}{\sum d_{ij}^2}$$

where f is non-decreasing

Non-metric MDS

- Start with initial d_{ij}
- Find optimal $\hat{f}\left(\delta_{ij}\right)$ ($\hat{f}$ non-decreasing)
- Given $\hat{f}\left(\delta_{ij}\right)$ find configuration with optimal d_{ij} in STRESS
- Iterate

Languages

TABLE 12.3 NUMERALS IN 11 LANGUAGES

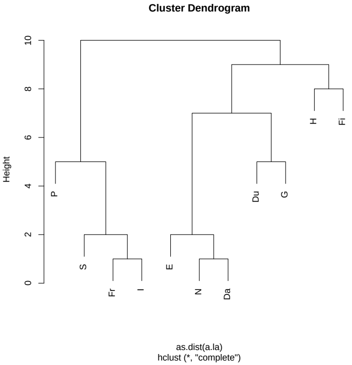
English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

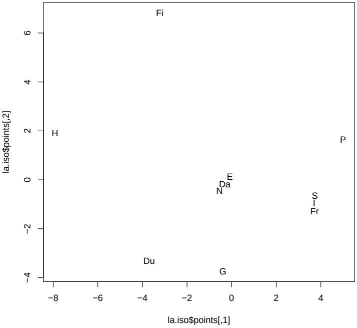
Languages

Dissimilarities

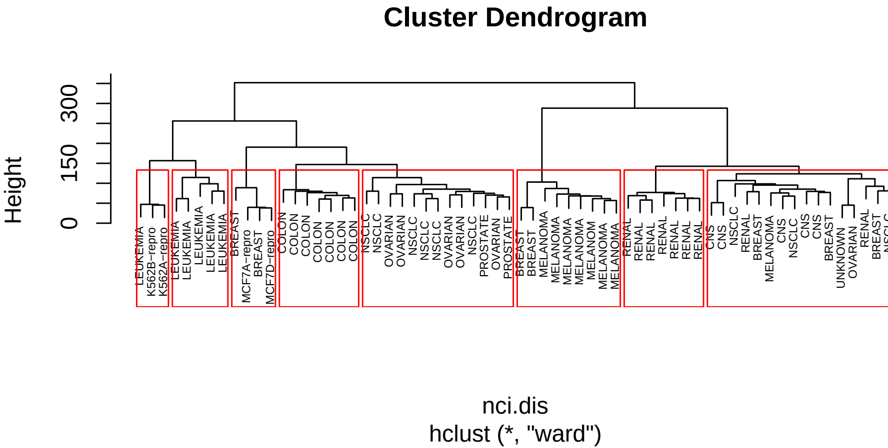
	E	N	Da	Du	G	Fr	S	I	P	H	Fi
E											
N	2										
Da	2	1									
Du	7	5	6								
G	6	4	5	5							
Fr	6	6	6	9	7						
S	6	6	5	9	7	2					
I	6	6	5	9	7	1	1				
P	7	7	6	10	8	5	3	4			
H	9	8	8	8	9	10	10	10	10		
Fi	9	9	9	9	9	9	9	9	9	9	8

Complete linkage + MDS

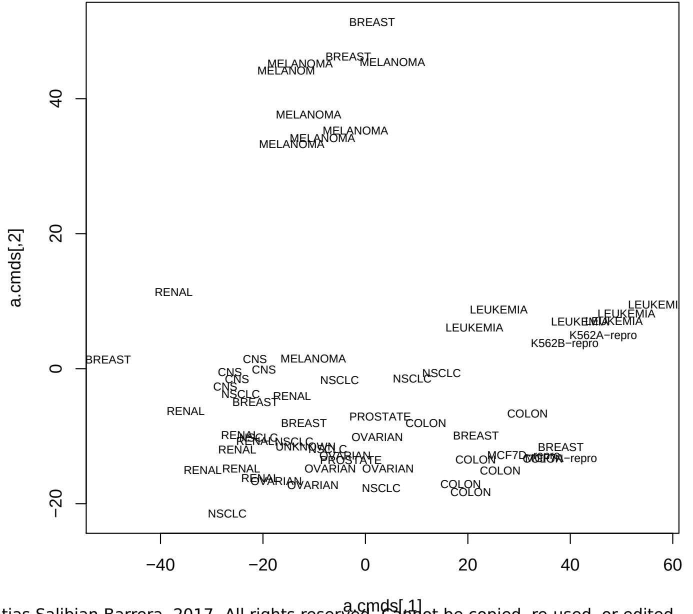




Cancer

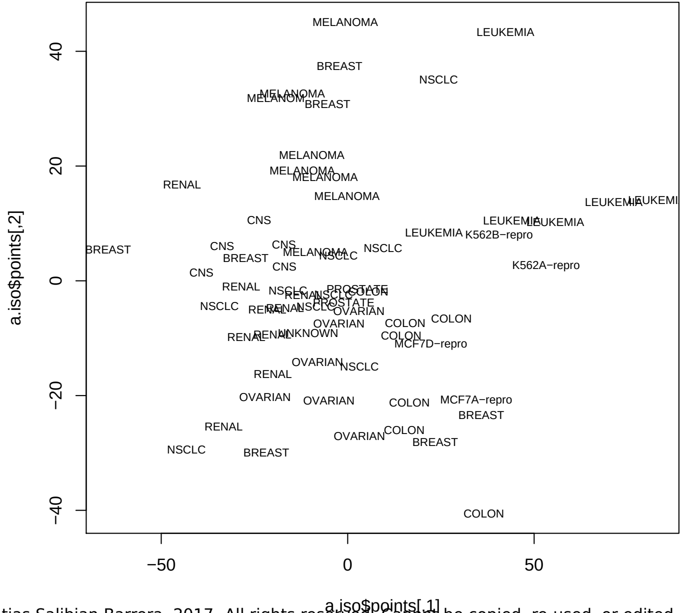


Cancer

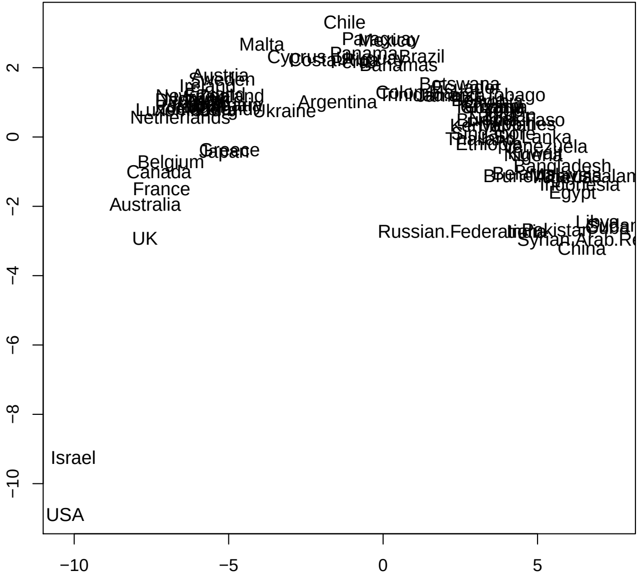


a.cmds[,1]
a.cmds[,2]
BREAST
MELANOMA
MELANOM
LEUKEMIA
K562A−repro
K562B−repro
CNS
NSCLC
RENAL
PROSTATE
COLON
OVARIAN
UNKNOWN
MCF7D−repro

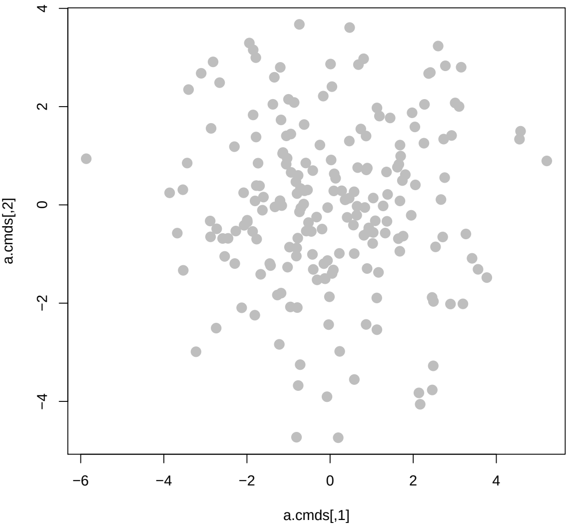
Cancer


MELANOMA
LEUKEMIA
BREAST
NSCLC
MELANOMA
MELANOMA
BREAST
MELANOMA
MELANOMA
MELANOMA
RENAL
MELANOMA
LEUKEMIA
LEUKEMI
CNS
LEUKEMIA
LEUKEMIA
LEUKEMIA
K562B−repro
BREAST
CNS
CNS
MELANOMA
NSCLC
NSCLC
BREAST
CNS
CNS
K562A−repro
RENAL
NSCLC
RENAL
NSCLC
PROSTATE
COLON
PROSTATE
NSCLC
RENAL
RENAL
NSCLC
OVARIAN
OVARIAN
COLON
COLON
RENAL
RENAL
UNKNOWN
COLON
MCF7D−repro
OVARIAN
NSCLC
RENAL
OVARIAN
OVARIAN
COLON
MCF7A−repro
BREAST
RENAL
OVARIAN
COLON
BREAST
NSCLC
BREAST
COLON
a.iso$points[,2]
a.iso$points[,1]
40
20
0
−20
−40
−50
0
50

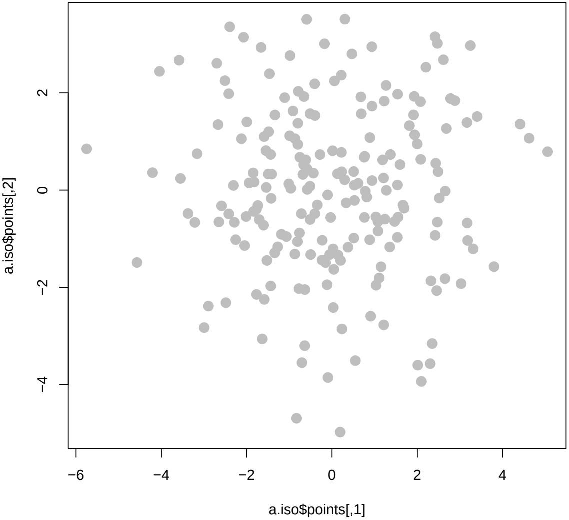
UN Votes


Chile
Malta
Brazil
Bahamas
Austria
Ukraine
Argentina
Netherlands
Botswana
Greece
Japan
Belgium
Canada
France
Australia
UK
Venezuela
Bangladesh
Indonesia
Egypt
China
Israel
USA
2
0
−2
−4
−6
−8
−10
−10
−5
0
5

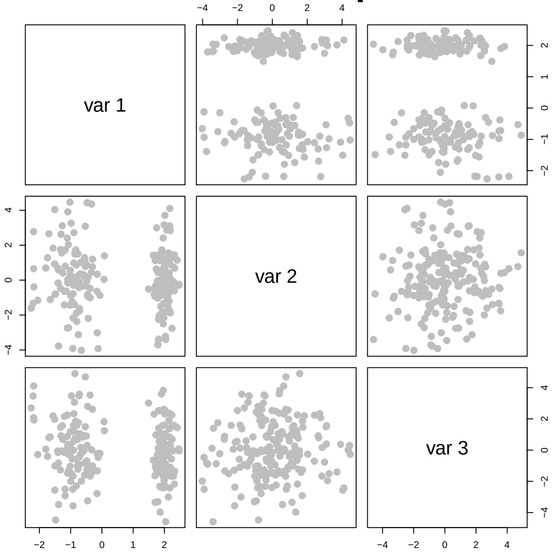
A word of caution - cmdscale



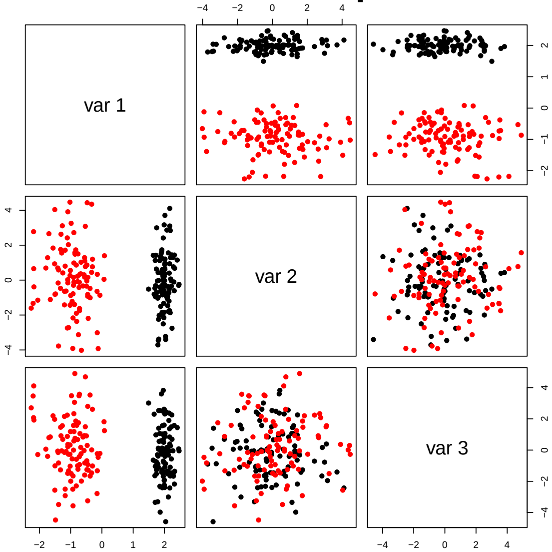
A word of caution - isoMDS



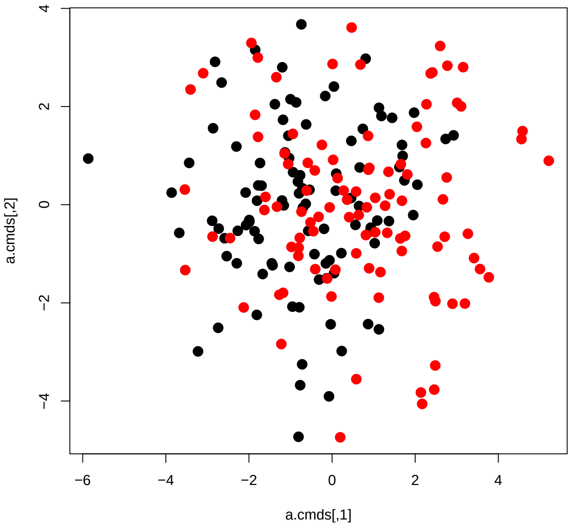
A word of caution - pairs



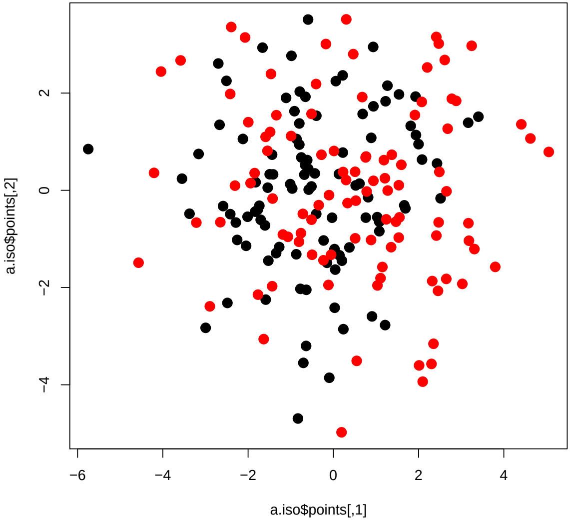
A word of caution - pairs



A word of caution - cmdscale



A word of caution - isoMDS

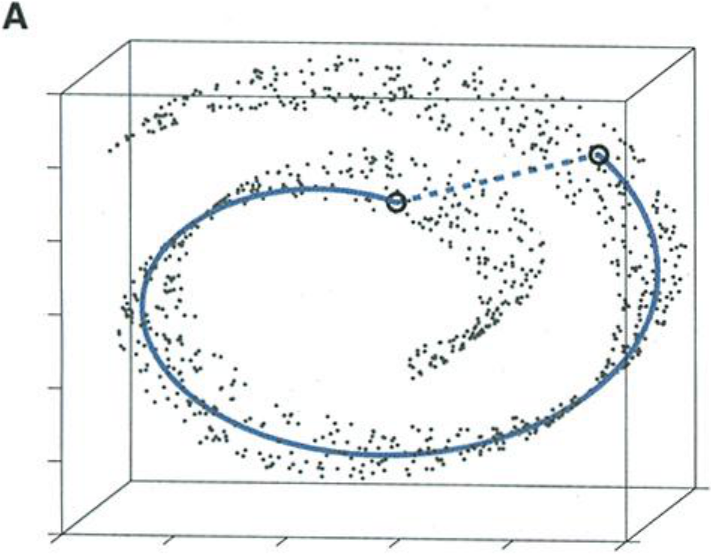


ISOMAP

- Construct a graph to approximate geodesic distances
- Use local distances (e.g. to closest neighbours)
- Distance on manifold is approximated by distance on the graph (shortest path)
- Apply classical MDS to the graph distances

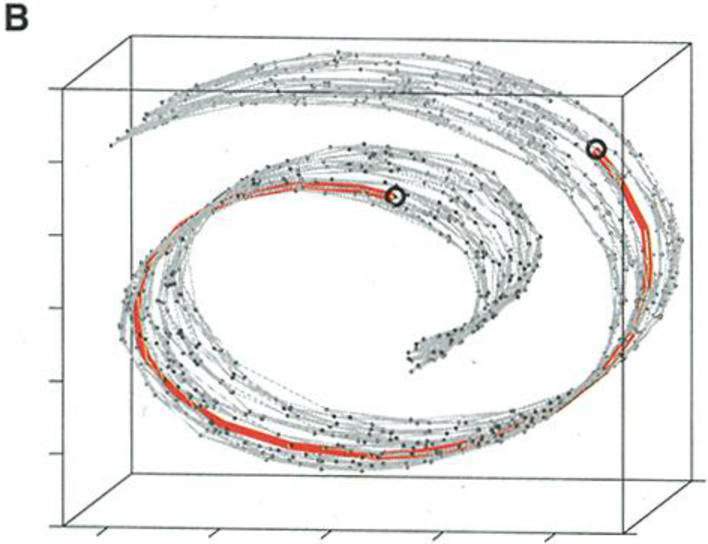
Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, Science 290: 2319-2323

ISOMAP



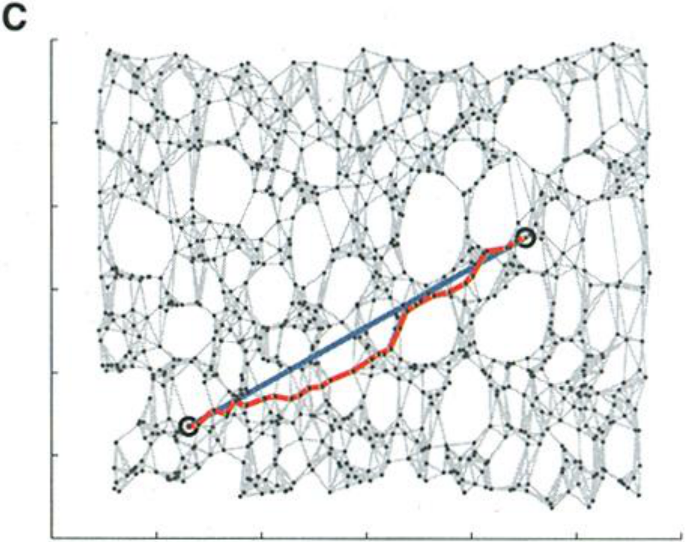
Original content from Tenenbaum et al (2000), Science

ISOMAP



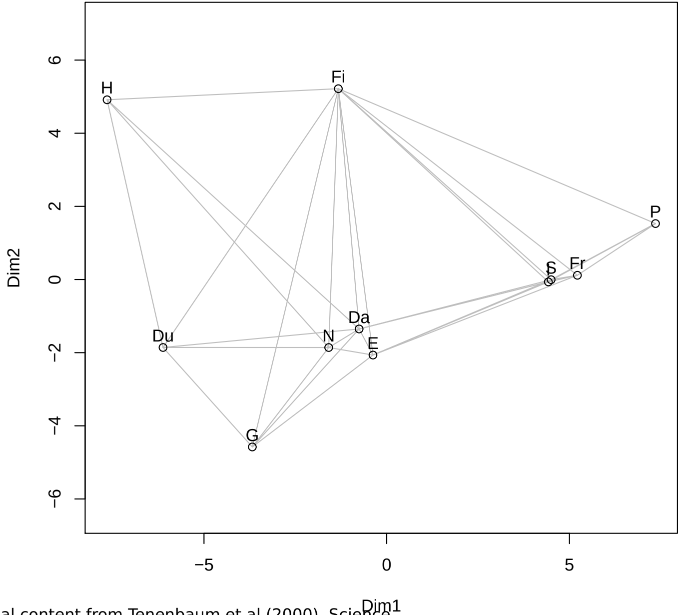
Original content from Tenenbaum et al (2000), Science

ISOMAP



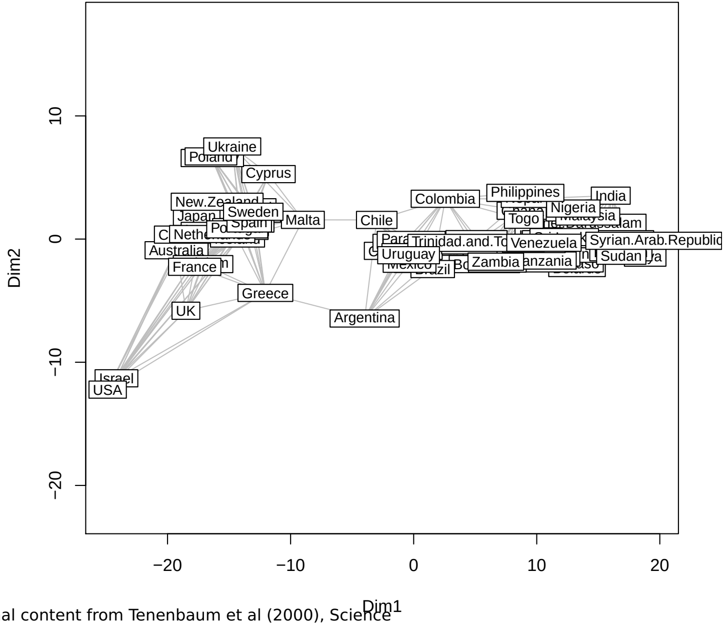
Original content from Tenenbaum et al (2000), Science

ISOMAP - Languages



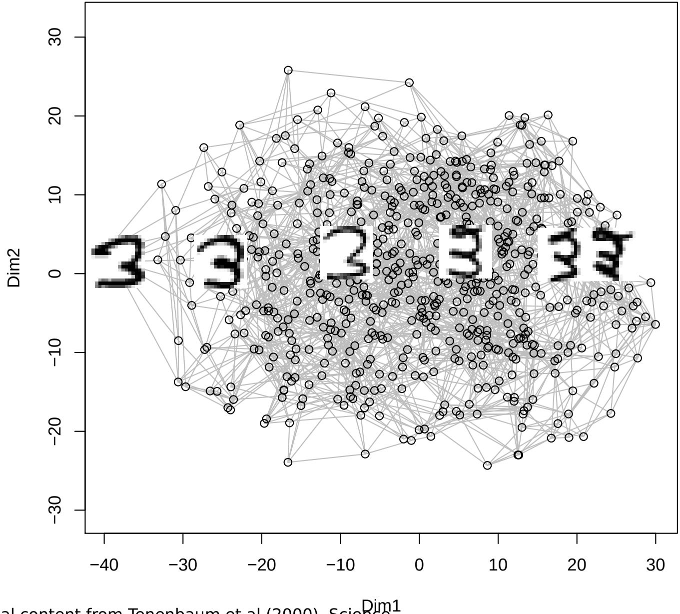
Original content from Tenenbaum et al (2000), Science

ISOMAP - UN votes



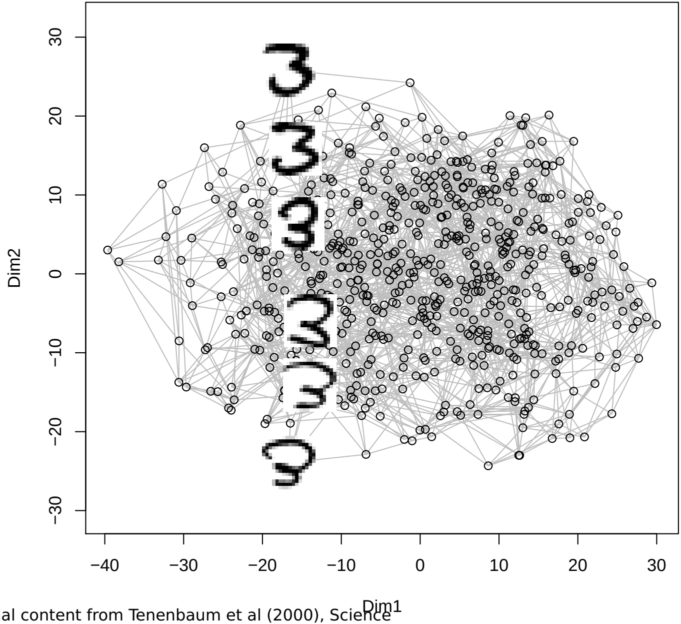
Original content from Tenenbaum et al (2000), Science

ISOMAP - Digits



Original content from Tenenbaum et al (2000), Science

ISOMAP - Digits



Original content from Tenenbaum et al (2000), Science

Multidimensional scaling

- δ_{ij} are a dissimilarity if

$$\delta_{ii} = 0\,, \quad \delta_{ij} \geq 0\,, \quad \delta_{ij} = \delta_{ji}$$

- It is Euclidean if there exist $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^r$ such that

$$\delta_{ij} = \left\| \mathbf{w}_i - \mathbf{w}_j \right\|$$

Original content from Tenenbaum et al (2000), Science