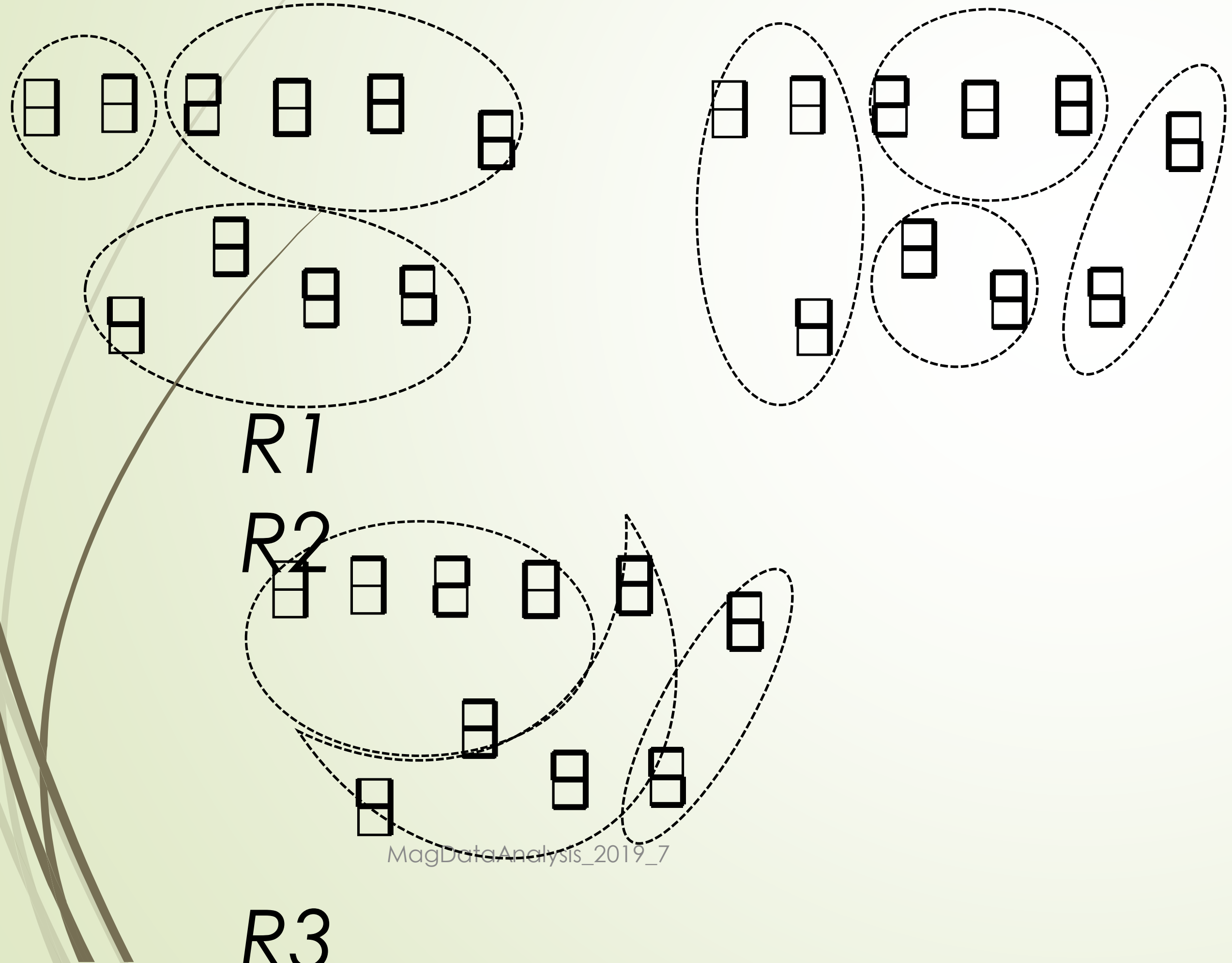
Mag 2019: Consensus Partition and similarity clustering

- The problem of consensus partition
- Mismatch (Mirkin's) distance between partitions
- Consensus matrix
- Mismatch distance consensus partition
- Muchnik test and failure of the mismatch distance consensus
- Regression distance between partitions
- Regression distance consensus partition
- Relation to k-Means clustering

Consensus partitioning

Consensus partition operationally



Given an entity set I, and *m* partitions of it, *R1, R2, …, Rm,*
Define consensus partition *S* of I:

$$\text{Min}_S \sum_{t=1}^{m} d(S, Rt)$$

with *d(R,S)*, predefined distance over partitions of I

Mismatch (Mirkin's) distance between partitions, 1

- Given partition $R=\{R_1,R_2,\ldots,R_K\}$ on I, define $N \times N$ matrix $r=(r_{ij})$, $N=\#I$:

$$r_{ij} = \begin{cases} 1, \text{if } i,j \in R_k \text{ for some } k = 1,2,\ldots,K \\ 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

- For R={1-2-3, 4-5-6},

$$\mathbf{r}= \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}$$

Mismatch distance between partitions, 2

- A binary $N \times N$ matrix to correspond to any partition:
- R={1-2-3, 4-5-6}, S={1-3, 2-4-5-6}, T={1-4, 2-5, 3-6}

	R: 1 23 456	S: 1 234 56	T: 1 234 5 6
1	111000	101000	100100
2	111000	010111	010010
3	111000	101000	001001
4	000111	010111	100100
5	000111	010111	010010
6	000111	010111	001001

Mismatch distance between partitions, 3

- Given partitions $R=\{R_1,\ldots,R_K\}$, $S=\{S_1,\ldots,S_L\}$ of I, and binary $N \times N$ matrices $r=(r_{ij})$, $s=(s_{ij})$, representing them:
- $d(R,S) = \sum_{i,j=1}^{N} |r_{ij} - s_{ij}| = \sum_{i,j=1}^{N} (r_{ij} - s_{ij})^2$ (definition)
- $d(R,S) = \sum_{i,j=1}^{N} (r_{ij} + s_{ij} - 2\, r_{ij}\, s_{ij})$ (derivation: note, no exponent 2)
- $d(R,S) = \sum_{k=1}^{K} N_k^2 + \sum_{l=1}^{L} N_l^2 - 2 \sum_{k=1}^{K} \sum_{l=1}^{L} N_{kl}^2$
- This expresses distance *d* through the contingency table between *R* and *S*

Quiz: what is the maximum distance between partitions on *I*

- Tip: $N(N-1)$ where $N=\#I$. That is the mismatch distance between trivial partition 0 consisting of N singletons, and the universal partition consisting of just one part, the set I itself.

Example of Mismatch distance

- R={1-2-3, 4-5-6}, S={1-3, 2-4-5-6}, T={1-4, 2-5, 3-6}

	1 234 56	1 234 56	1 234 5 6
1	111000	101000	100100
2	111000	010111	010010
3	111000	101000	001001
4	000111	010111	100100
5	000111	010111	010010
6	000111	010111	001001

Compute $d(R,S)=|R|+|S|-2|RS|=(3^2+3^2)+(2^2+4^2) - 2(2^2+1^2+3^2)=18+20-28=10$, **since** $RS=\{1\text{-}3, 2, 4\text{-}5\text{-}6\}$.

Relative distance is $\delta(R,S)=d(R,S)/[N(N-1)]=10/30=0.33$.

Quiz: Compute $d(R,T), d(S,T)$.

Mismatch distance between partitions, 4

- $d(R,S) = \sum_{k=1}^{K} N_k^2 + \sum_{l=1}^{L} N_l^2 - 2\sum_{k=1}^{K}\sum_{l=1}^{L} N_{kl}^2$
- Here all the ordered pairs (*i,j*) are considered in R_k, S_l, and $R_k \cap S_l$.
- Frequently, only unordered pairs {i,j} are considered only; their number is the binomial coefficient. Then
- $d(R,S) = \sum_{k=1}^{K} \binom{N_k}{2} + \sum_{l=1}^{L} \binom{N_l}{2} - 2\sum_{k=1}^{K}\sum_{l=1}^{L} \binom{N_{kl}}{2}$

  where $\binom{n}{2} = n(n-1)/2.$

This you will see in literature.

Consensus partition with mismatch distance,1

- Given partitions $R^1, R^2, \ldots, R^m$, of I, find $S=\{S_1,\ldots,S_L\}$ to minimize:

$$\sum_{t=1}^{m} d(R^t, S) = \sum_{t=1}^{m} \sum_{i,j=1}^{N} (r_{ij}^t + s_{ij} - 2r_{ij}^t s_{ij})$$

since

$$d(R^t, S) = \sum_{i,j=1}^{N} |r_{ij}^t - s_{ij}| = \sum_{i,j=1}^{N} (r_{ij}^t - s_{ij})^2$$

- Note, once again r and s with no exponent! Because they are binary: $1^2=1$ and $0^2=0$.

Consensus partition with mismatch distance,2

- Define consensus matrix: given partitions R^t with $N \times N$ matrices r^t ($t=1,\ldots,m$), **Consensus matrix** $A=(a_{ij})$:

$$A = \sum_{t=1}^{m} r^t$$

so that

$$a_{ij} = \sum_{t=1}^{m} r_{ij}^t$$

- a_{ij} is the number of partitions R^t in which i,j belong in the same part

Consensus partition with mismatch distance,3

- Given partitions R^1, R^2,…, R^m, of I, find S={S_1,…,S_L} to minimize:

$$\sum_{t=1}^{m} d(R^t,S) = \sum_{t=1}^{m}\sum_{i,j=1}^{N}(r_{ij}^t + s_{ij} - 2r_{ij}^t s_{ij}) = \sum_{i,j=1}^{N}(a_{ij} + ms_{ij} - 2a_{ij}s_{ij})$$

$$\sum_{t=1}^{m} d(R^t,S) = const - 2\sum_{i,j=1}^{N}(\,a_{ij} - \frac{m}{2})s_{ij}$$

Consensus partition with mismatch distance,4

- Given partitions R^1, R^2,…, R^m, of I, find consensus $S=\{S_1,\ldots,S_L\}$ to minimize

$$\sum_{t=1}^{m} d(R^t, S) = const - 2\sum_{i,j=1}^{N}(a_{ij} - \frac{m}{2})s_{ij}$$

That is, find $S=\{S_1, S_2, \ldots, S_K\}$ maximizing

$$f(S) = \sum_{i,j=1}^{N}(a_{ij} - \frac{m}{2})s_{ij} = \sum_{k=1}^{K}\sum_{i,j\in S_k}(a_{ij} - \frac{m}{2})$$

This is the within-cluster summary criterion with shift value $m/2$ subtracted

Example: m=11 partitions of the Ten Digits

4 parts in R6, R7; **3** parts in the 9 others

Digit	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
1	1	1	1	2	1	1	1	1	1	2	1
2	2	2	2	2	3	3	3	2	2	3	2
3	3	3	3	2	3	3	4	3	3	2	3
4	3	1	1	3	3	1	1	1	3	2	1
5	3	3	3	1	2	2	2	3	3	1	3
6	2	2	2	1	2	2	2	2	2	1	2
7	1	1	1	2	1	3	1	1	1	2	1
8	2	2	2	3	3	4	3	2	2	3	2
9											
0											

Example: Consensus matrix for m=11 partitions of the Ten D

	0	1		2	3	4	5	6	7	8	9
1	0	11	1	1	2	7	0	0	10	0
2	3	1	10	11	3	1	0	6	2	9
3	9	2	2	3	11	4	6	0	3	1
4	4	7	1	1	4	11	2	0	6	2
5	6	0	0	0	6	2	11	5	0	0
6	0	0	6	6	0	0	5	11	0	6
7	1	10	1	2	3	6	0	0	11	0
8	3	0	10	9	1	2	0	6	0	11
9	11	0	2	3	9	4	6	0	1	3
0										

Example: Consensus matrix for m=11 partitions of the Ten Digits, with $m/2=5.5$ subtracted and diagonal zeroed

	1	2	3	4	5	6	7	8	9	0
1	0	-4.5	-3.5	1.5	-5.5	-5.5	4.5	-5.5	-5.5	-4.5
2	-4.5	0	-2.5	-4.5	-5.5	0.5	-3.5	3.5	-2.5	4.5
3	-3.5	-2.5	0	-1.5	0.5	-5.5	-2.5	-4.5	3.5	-3.5
4	1.5	-4.5	-1.5	0	-3.5	-5.5	0.5	-3.5	-1.5	-4.5
5	-5.5	-5.5	0.5	-3.5	0	-0.5	-5.5	-5.5	0.5	-5.5
6	-5.5	0.5	-5.5	-5.5	-0.5	0	-5.5	0.5	-5.5	0.5
7	4.5	-3.5	-2.5	0.5	-5.5	-5.5	0	-5.5	-4.5	-4.5
8	-5.5	3.5	-4.5	-3.5	-5.5	0.5	-5.5	0	[illegible]	[illegible]
9										
0										

Example: Consensus matrix for m=11 partitions of the Ten Digits … zeroed

	1	2	3	4	5	6	7	8	9	0
1	0	-4.5	-3.5	1.5	-5.5	-5.5	4.5	-5.5	-5.5	-4.5
2	-4.5	0	-2.5	-4.5	-5.5	0.5	-3.5	3.5	-2.5	4.5
3	-3.5	-2.5	0	-1.5	0.5	-5.5	-2.5	-4.5	3.5	-3.5
4	1.5	-4.5	-1.5	0	-3.5	-5.5	0.5	-3.5	-1.5	-4.5
5	-5.5	-5.5	0.5	-3.5	0	-0.5	-5.5	-5.5	0.5	-5.5
6	-5.5	0.5	-5.5	-5.5	-0.5	0	-5.5	0.5	-5.5	0.5
7	4.5	-3.5	-2.5	0.5	-5.5	-5.5	0	-5.5	-4.5	-4.5
8	-5.5	3.5	-4.5	-3.5	-5.5	0.5	-5.5	0	-2.5	4.5
9	-5.5	-2.5	3.5	-1.5	0.5	-5.5	-4.5	-2.5	[illegible]	[illegible]
0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Clusters are obvious here:

S_1 1-4-7

S_2 2-6-8-0

S_3 3-5-9

Why obvious? All positive entries within clusters, all negative entries between clusters, Thus making

However, consensus partition for mismatch distance fails the Muchnik test.

Muchnik test

- Take a partition $R=\{R_1, R_2, \ldots, R_K\}$ of I in $K \geq 2$ parts R_k .
- Create a set of K two-part partitions $Rk=\{R_k, I-R_k\}$, $k=1,2,\ldots, K$.
- Find a consensus partition S for the K two-part partitions Rk, $k=1,2,\ldots, K$.
- If $S==R$, the test is passed; if not, the test is failed.

Muchnik test for mismatch distance

- Take a partition $R=\{R_1,R_2,\ldots,R_K\}$ of I in $K\geq 2$ parts R_k.
- Create a set of K two-part partitions $Rk=\{R_k, I\text{-}R_k\}$, $k=1,2,\ldots, K$.
- Find consensus matrix for set of $Rk=\{R_k, I\text{-}R_k\}$, $k=1,2,\ldots, K$.
- $a_{ij} = \begin{cases} K & \text{if } i,j \in R_k \text{for some } k \\ K-2, & \text{if } i \in R_k, j \in R_l \text{ and } k \neq l \end{cases}$
- Since shift K/2 <K-2 at $K\geq 5$, the maximum $f(S)$ is reached at the universal partition $S=\{I\}$: **the test fails at $K\geq 5$.**

Therefore, Mismatch distance fails Muchnik test and is not that good, overall.

Luckily, a similar distance can be defined to pass the Muchnik test: the regression distance.

Regression distance between partitions,1

- Given partitions $R=\{R_1,\ldots,R_K\}$, $S=\{S_1,\ldots,S_L\}$ of I, define binary $N\times K$ and $N\times L$ incidence matrices $X=(x_{ik})$ and $Y=(y_{il})$:
- $x_{ik} = \begin{cases} 1, & \text{if } i \in R_k \ (k = 1,2,\ldots,K) \\ 0, & \text{otherwise} \end{cases}$
- $y_{il} = \begin{cases} 1, & \text{if } i \in S_l \ (l = 1,2,\ldots,L) \\ 0, & \text{otherwise} \end{cases}$
- Define orthogonal projector $P_X=X(X^TX)^{-1}X^T$.
- In P_X: (i,j)-th entry is 0, except for the case at which both $i,j\in R_k$ for some k – in this case that is $1/N_k$ where $N_k= \#R_k$.

Regression distance between partitions,2

Given partitions $R=\{R_1,\ldots,R_K\}$, $S=\{S_1,\ldots,S_L\}$ of I, define binary $N\times K$ and $N\times L$ incidence matrices $X=(x_{ik})$ and $Y=(y_{il})$:

- $y_{il} = \begin{cases} 1, & \text{if } i \in S_l \ \ (l = 1,2,\ldots,L) \\ 0, & \text{otherwise} \end{cases}$
- Define $mean(Y)=(p_1,\ldots,p_L)$, vector of relative frequencies
- Define $Y'=Y - mean(Y)$ (subtracting $mean(Y)$ from every row of Y)
- Define orthogonal projector $P_X=X(X^TX)^{-1}X^T$.
- Regression distance: the summary squared difference between Y' and P_XY':

$$Rd(R,S)=\|Y'-P_XY'\|^2$$

- Meaning: How well is $span(X)$ for reproducing Y' ?

Regression distance Rd between partitions,3

Given partitions $R=\{R_1,\ldots,R_K\}$, $S=\{S_1,\ldots,S_L\}$ of I, define binary $N{\times}K$ and $N{\times}L$ incidence matrices $X=(x_{ik})$ and $Y=(y_{il})$:

- $y_{il} = \begin{cases} 1, & \text{if } i \in S_l \ (l = 1,2,\ldots,L) \\ 0, & \text{otherwise} \end{cases}$ Define $Y'=Y - mean(Y)$
- Regression distance: the summary squared difference between Y' and $P_X Y'$:

$$Rd(R,S)=\|Y' - P_X Y'\|^2$$

Rd consensus:

Given partitions R^1, R^2,…, R^m, of I, find $S=\{S_1,\ldots,S_L\}$ to minimize:

$$\sum_{t=1}^{m} Rd(S,R^t)$$

Regression distance consensus,1

Given partitions $R=\{R_1,\dots,R_K\}$, $S=\{S_1,\dots,S_L\}$ of I, define binary $N\times K$ and $N\times L$ incidence matrices $X=(x_{ik})$ and $Y=(y_{il})$.

- Regression distance: the summary squared difference between Y' and P_XY':

$$Rd(R,S)=\|Y'-P_XY'\|^2$$

Rd consensus: Given partitions R^1, R^2,…, R^m, of I, find $S=\{S_1,\dots,S_L\}$ to minimize $\sum_{t=1}^{m} Rd(S,R^t)$.

Theorem 1. Rd consensus S maximizes the semi-average criterion

$$G(S)=\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i,j\in S_k} a_{ij}$$

where $A=(a_{ij})$ is the consensus matrix.

Regression distance consensus,2

Given partitions $R=\{R_1,\ldots,R_K\}$, $S=\{S_1,\ldots,S_L\}$ of I, define binary $N\times K$ and $N\times L$ incidence matrices $X=(x_{ik})$ and $Y=(y_{il})$.

- Regression distance: the summary squared difference between Y' and P_XY':

$$Rd(R,S)=\|Y'-P_XY'\|^2$$

Rd consensus: Given partitions R^1, R^2,…, R^m, of I, find $S=\{S_1,\ldots,S_L\}$ to minimize $\sum_{t=1}^{m} Rd(S,R^t)$.

Theorem 2. Rd consensus passes Muchnik test.

Experimental observation

- ISSUE: At multiple runs of K-Means, the output is a set of partitions of the entity set. Which one of them to choose?
- Strategy 1. Choose the partition S1, that is the best over K-Means criterion W(S,c), that is, minimizes W(S,c).
- Strategy 2. Build a consensus partition S2.

- At synthetic data (*K* Gaussian clusters generated with various degree of noise), S2 usually is worse than S1 over W(S,c), but better than S1 in recovering the generated clusters.

Relation to K-Means clustering

- **K-Means: Maximize the anomalous clusters criterion**

$$F(S,c) = \sum_{k=1}^{K} N_k < c_k, c_k > \qquad (*)$$

Put $c_{kv} = \sum_{i \in S_k} y_{iv} / N_k$

in (*) to obtain the semi-average criterion, with no c_k

$$F(S) = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i,j \in S_k} < y_i, y_j > \qquad (**)$$

where N_k – the number of elements in S_k, and y_i, i-th entity (row of data matrix)

Criterion inherited from k-means

Kernel trick

- **Maximize semi-average criterion**

$$\boldsymbol{F(S) = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i,j \in S_k} < y_i, y_j >}$$

Kernel $K(x,y) = <\psi(x),\ \psi(y)>$ *is the inner product in* ψ*-space*

$$F_{\psi}(S) = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i,j \in S_k} < \psi(y_i), \psi(y_j) >$$

Kernel trick

- Given a semi-definite positive similarity matrix $A=(a_{ij})$, $i,j=1,\ldots,n$,
- Maximize semi-average criterion

$$G(S)=\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i,j\in S_k}a_{ij}\qquad (***)$$

Two approaches to locally maximize *G*(*S*), *out of many*

- One cluster by one
- Agglomeration

Agglomeration to maximize semi-average $G(S)$

Denote $$A(I_1, I_2) = \sum_{i \in I_1} \sum_{j \in I_2} a_{ij}$$

At merging S_k and S_l:

$\Delta(k,l) = G(S) - G(S(k,l)) =$

$= [N_k A(S_l,S_l)/N_l + N_l A(S_k,S_k)/N_k - 2A(S_k,S_l)]/(N_k + N_l)$

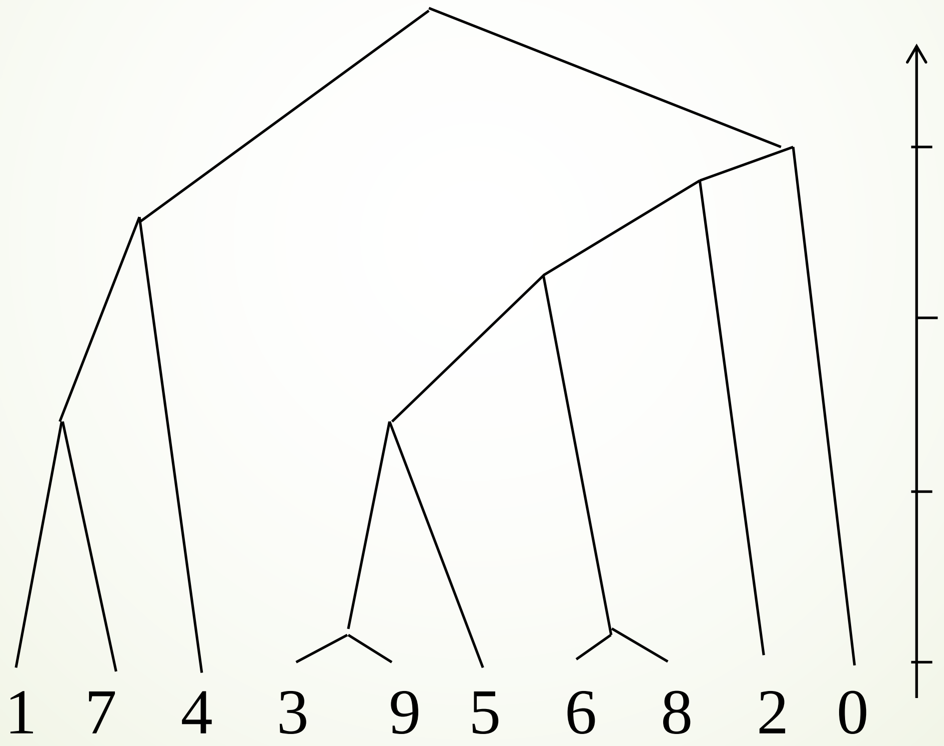
Agglomeration:

1. Start at N-singleton S={ {1}, {2}, …, {N}}
2. Find k*, l* maximizing $\Delta(k,l)$
3. If $\Delta(k^*,l^*) > 0$, merge together S_{k^*} and S_{l^*}, add row l* to row k*, then add column l* to column k* and remove both row and column l*. Go to 2. Else, stop and output S.

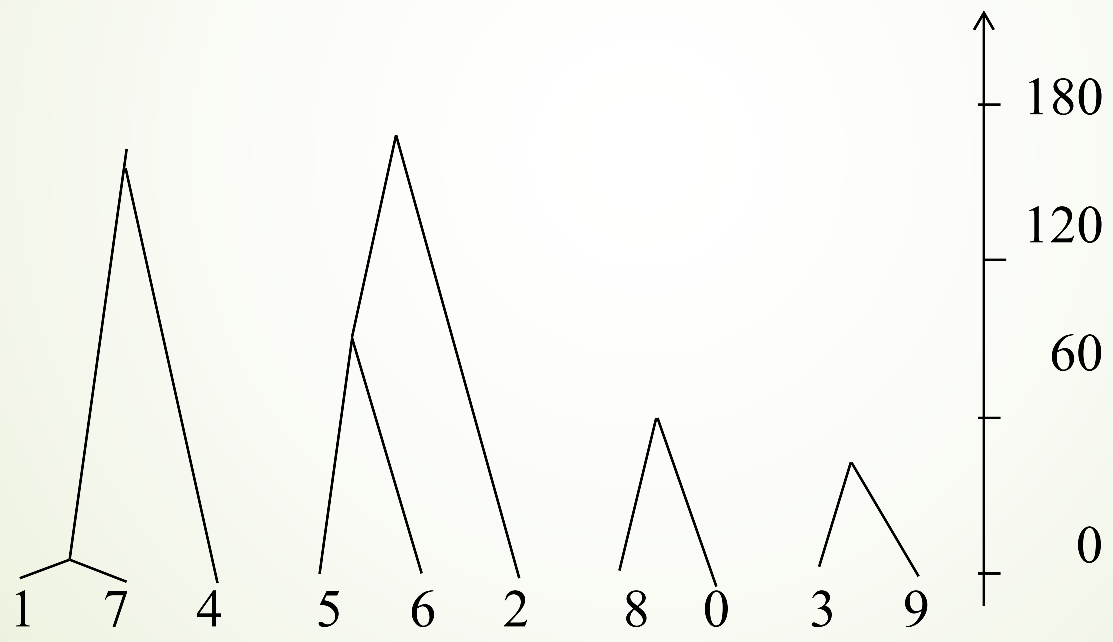
Example: Digits 0, 1, 2,…, 9 confusion data (symmetrized)

timulus	Response									
	1	2	3	4	5	6	7	8	9	0
	877	11	18	86	9	20	165	6	15	11
	11	782	38	13	31	31	9	29	18	11
	18	38	681	6	31	4	31	29	132	11
	86	13	6	732	9	11	26	13	44	6
	9	31	31	9	669	88	7	13	104	11
	20	31	4	11	88	633	2	113	11	31
	165	9	31	26	7	2	667	6	13	16
	6	29	29	13	13	113	6	577	75	122
	15	18	132	44	104	11	13	75	550	32
	11	11	11	6	11	31	16	122	32	818

Agglomerative tree for confusion between digits



Semi-average clustering results at between-digit confusion with zeroed diagonal



Consensus partition: No kernel trick

- Just represent the given partitions as a single 0/1 object-to-cluster matrix

Lecture contents

- Statement of the problem of consensus partition
- Introduction to Mismatch (Mirkin's) distance between partitions
- Consensus matrix
- Mismatch distance consensus partition as a maximizer of the within cluster summary consensus similarity
- Muchnik test and failure of the mismatch distance consensus
- Regression distance between partitions
- Regression distance consensus partition as a maximizer of the semi-average consensus similarity
- Relation to k-Means clustering