Course Schedule

- Introduction
- 1. Data visualization: PDPs, KDEs, and CDFs
- 2. detritalPy
    - Break
- 3. Statistical metrics & MDS
- 4. DZmds & DZstats
    - Break
- 5. Mixture modelling introduction & theory
- 6. DZmix application
- 7. DZnmf application
- Wrap-up

Module 3 Learning goals

- Understand how statistical metrics are calculated
  - What are the strengths and limitations of each metric
- Understand how metric and non-metric multi-dimensional scaling (MDS) proceeds.
- Understand the difference between metric and non-metric MDS
- Be able interpret MDS plots and evaluate their quality.

Module 3 Outline

- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multi-dimensional scaling (MDS)

Module 3 Outline

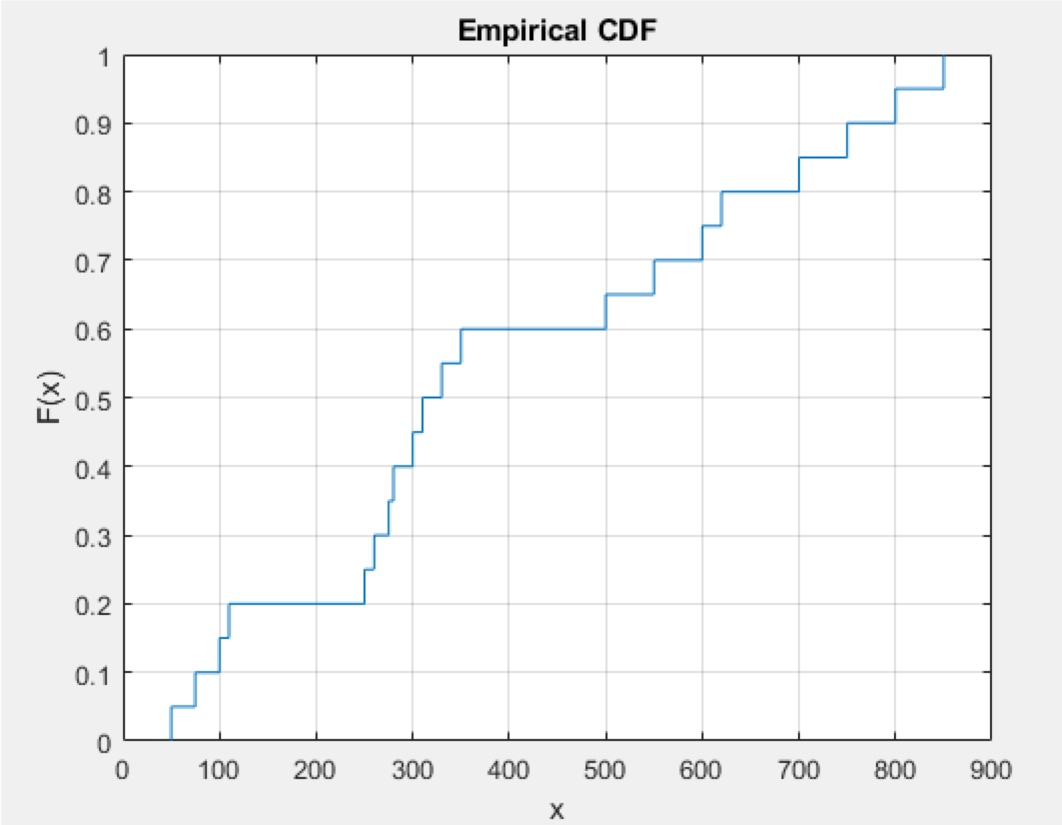
- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multi-dimensional scaling (MDS)

Kolmogorov-Smirnov distance (D value)

- EDF and CDF
  - The empirical distribution function (EDF, ECDF, sometimes CDF) is a non-parametric estimator of the underlying cumulative distribution function (CDF)
    - EDF = CDF as $n \Rightarrow \infty$
  - Calculating ECDF

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

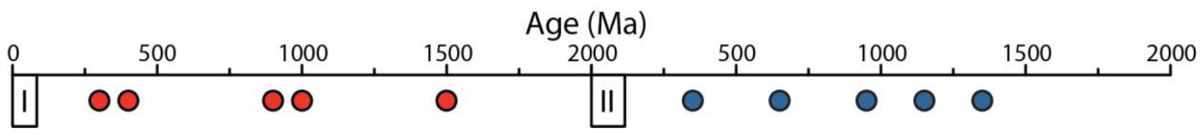
    - Where I = 1 if Xi ≤ x or 0 otherwise
    - For all real numbers x
  - The ECDF ranges from 0 to 1 with step heights of 1/n located at the values Xi.



Kolmogorov-Smirnov distance (D value)

- 2 samples, 5 ages each

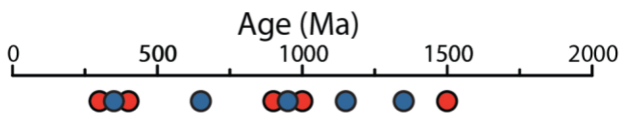
Sample 1 ages (Ma)	Sample 2 ages (Ma)
300	350
400	650
900	950
1000	1150
1500	1350



Kolmogorov-Smirnov distance (D value)

- 2 samples, 5 ages each
- Merged ages

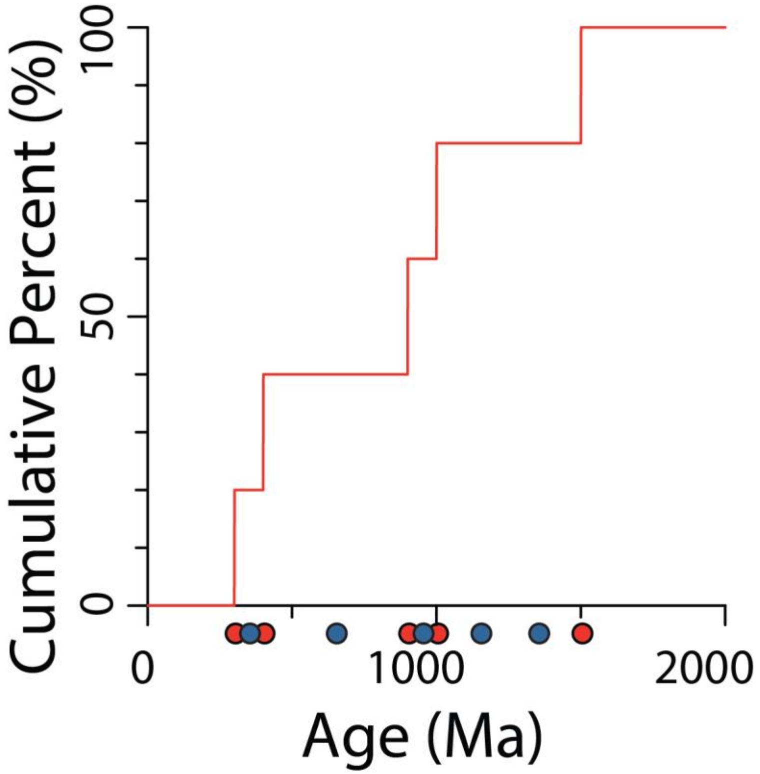
Merged ages (Ma)	CDF Sample 1	CDF Sample 2	CDF1-CDF2	CDF2-CDF1
300				
350				
400				
650				
900				
950				
1000				
1150				
1350				
1500				



Kolmogorov-Smirnov distance (D value)

- 2 samples, 5 ages each
- Merged ages
- Cumulative Distribution Function 1
- Because CDF1 is a function,
  - F(350)=0.2, F(650)=0.4, F(926.5)=0.6, etc

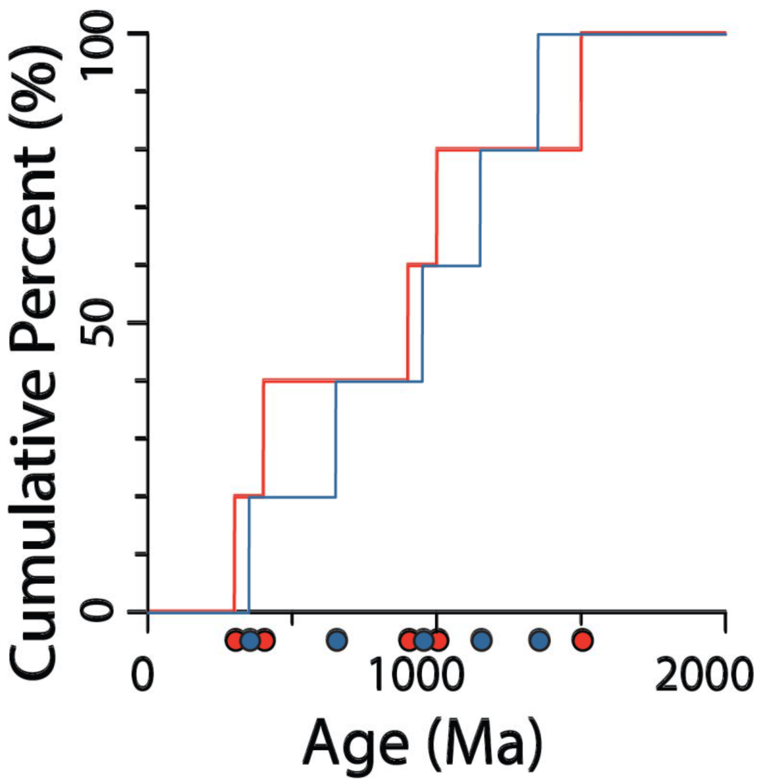
Merged ages (Ma)	CDF Sample 1	CDF Sample 2	CDF1-CDF2	CDF2-CDF1
300	**0.2**			
350	0.2			
400	**0.4**			
650	0.4			
900	**0.6**			
950	0.6			
1000	**0.8**			
1150	0.8			
1350	0.8			
1500	**1**			



Kolmogorov-Smirnov distance (D value)

- 2 samples, 5 ages each
- Merged ages
- Cumulative Distribution Function 1
- Cumulative Distribution Function 2

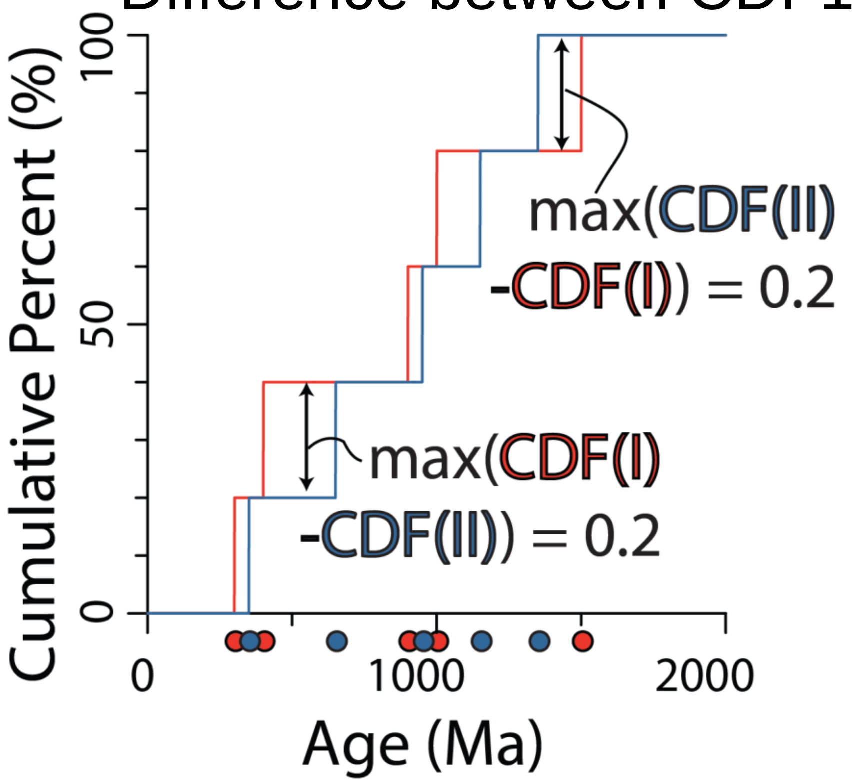
Merged ages (Ma)	CDF Sample 1	CDF Sample 2	CDF1-CDF2	CDF2-CDF1
300	**0.2**	0		
350	0.2	**0.2**		
400	**0.4**	0.2		
650	0.4	**0.4**		
900	**0.6**	0.4		
950	0.6	**0.6**		
1000	**0.8**	0.6		
1150	0.8	**0.8**		
1350	0.8	**1**		
1500	**1**	1		



Kolmogorov-Smirnov distance (D value)

- 2 samples, 5 ages each
- Merged ages
- Cumulative Distribution Function 1
- Cumulative Distribution Function 2
- Difference between CDF1 and CDF2

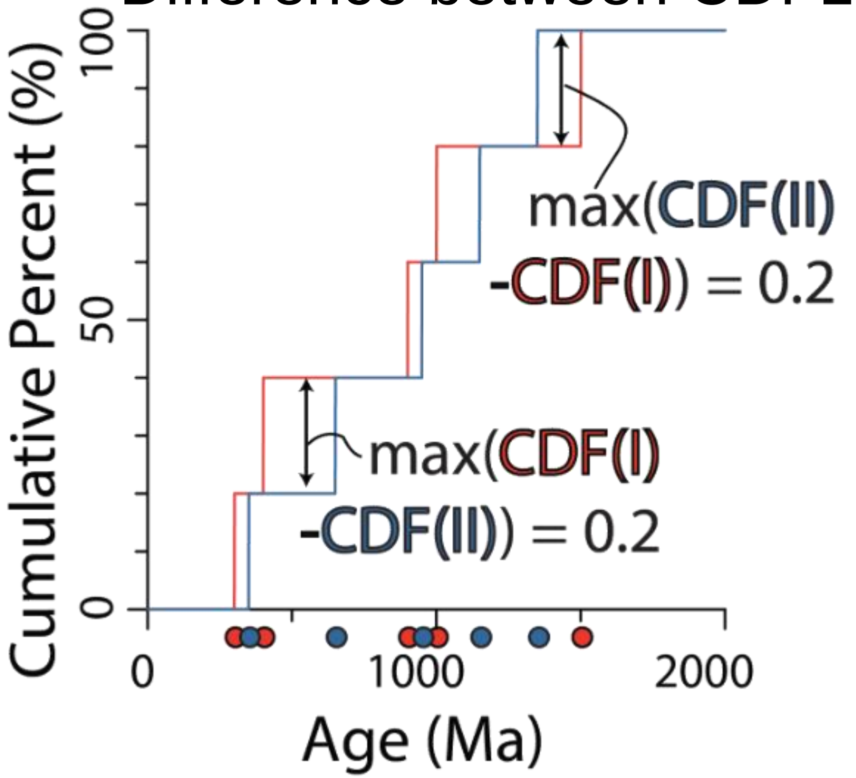
Merged ages (Ma)	CDF Sample 1	CDF Sample 2	CDF1-CDF2	CDF2-CDF1
300	**0.2**	0	**0.2**	-0.2
350	0.2	**0.2**	0	0
400	**0.4**	0.2	**0.2**	-0.2
650	0.4	**0.4**	0	0
900	**0.6**	0.4	**0.2**	0
950	0.6	**0.6**	0	0
1000	**0.8**	0.6	0.2	-0.2
1150	0.8	**0.8**	0	0
1350	0.8	**1**	-0.2	**0.2**
1500	**1**	1	0	0



Kolmogorov-Smirnov distance (D value)

- 2 samples, 5 ages each
- Merged ages
- Cumulative Distribution Function 1
- Cumulative Distribution Function 2
- Difference between CDF1 and CDF2

Merged ages (Ma)	CDF Sample 1	CDF Sample 2	CDF1-CDF2	CDF2-CDF1
300	**0.2**	0	**0.2**	-0.2
350	0.2	**0.2**	0	0
400	**0.4**	0.2	**0.2**	-0.2
650	0.4	**0.4**	0	0
900	**0.4**	0.4	**0.2**	0
950	0.6	**0.6**	0	0
1000	**0.8**	0.6	**0.2**	-0.2
1150	0.8	**0.8**	0	0
1350	0.8	**1**	-0.2	**0.2**
1500	**1**	1	0	0



Kolmogorov-Smirnov test D-value

$\max|CDF(II)-CDF(I)| = 0.2$

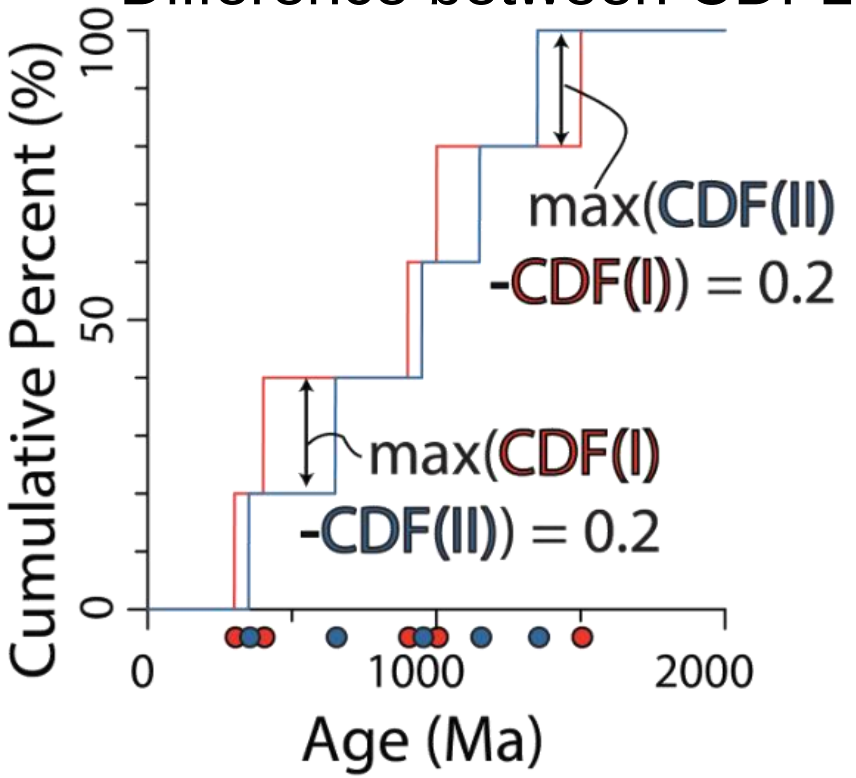
Module 3 Outline

- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multi-dimensional scaling (MDS)

Kuiper distance (V value)

- 2 samples, 5 ages each
- Merged ages
- Cumulative Distribution Function 1
- Cumulative Distribution Function 2
- Difference between CDF1 and CDF2

Merged ages (Ma)	CDF Sample 1	CDF Sample 2	CDF1-CDF2	CDF2-CDF1
300	**0.2**	0	**0.2**	-0.2
350	0.2	**0.2**	0	0
400	**0.4**	0.2	**0.2**	-0.2
650	0.4	**0.4**	0	0
900	**0.4**	0.4	**0.2**	0
950	0.6	**0.6**	0	0
1000	**0.8**	0.6	**0.2**	-0.2
1150	0.8	**0.8**	0	0
1350	0.8	**1**	-0.2	**0.2**
1500	**1**	1	0	0



Kolmogorov-Smirnov test D-value

$$\max|CDF(II)-CDF(I)| = 0.2$$

Kuiper test V-value

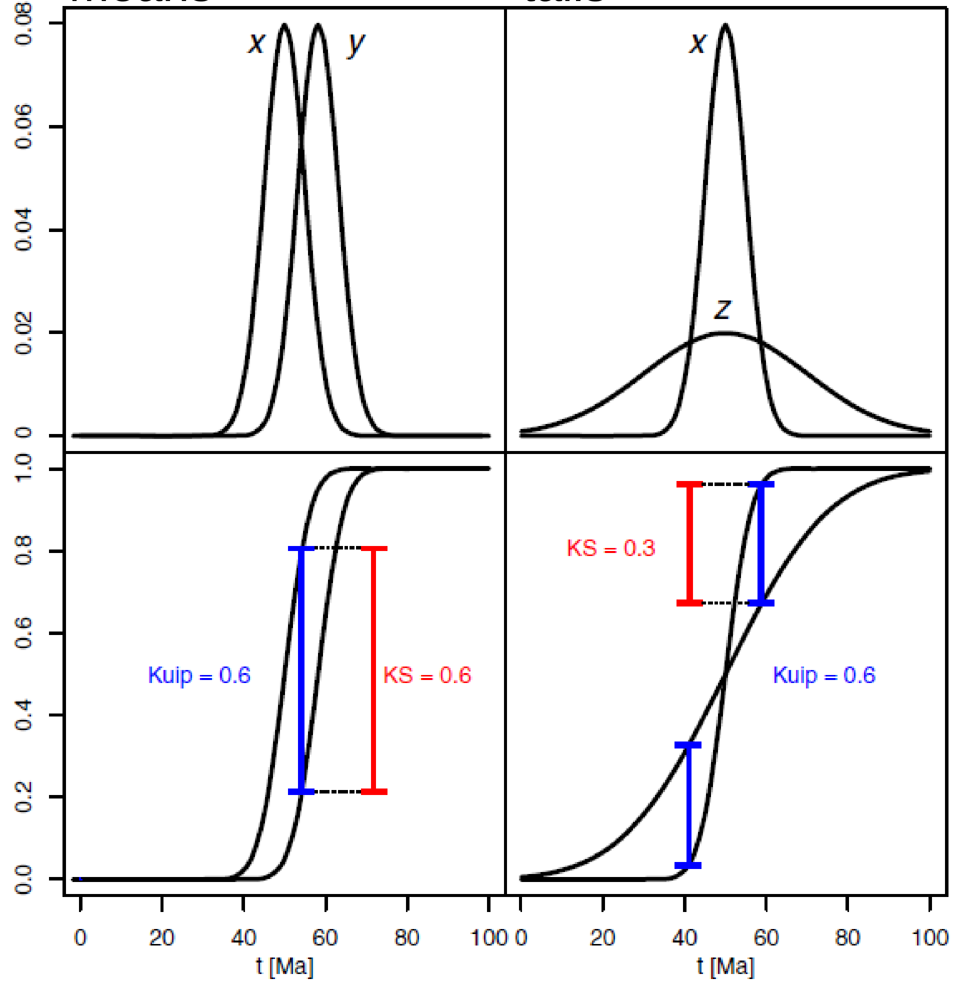
$$\max(CDF(II)-CDF(I)) + \max(CDF(I)-CDF(II)) = 0.4$$

Limitation of K-S distances

- 1) More sensitive at the center of the distribution than at the tails
  - Due to monotonically increasing nature of CDF
  - As the CDF approaches 1 or 0, the variance goes to 0

Equally sensitive to differences in the means

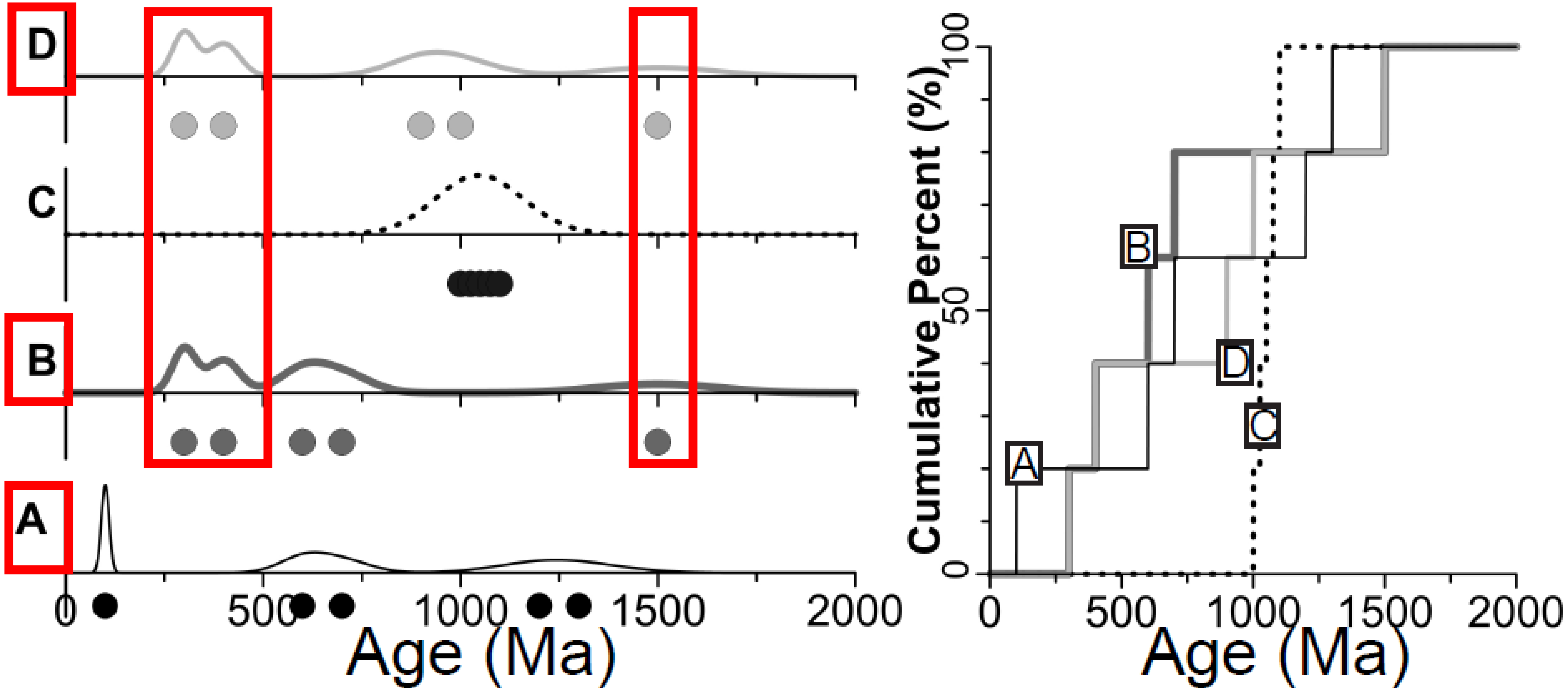
Less sensitive to differences in the tails



Vermeesch (2018)

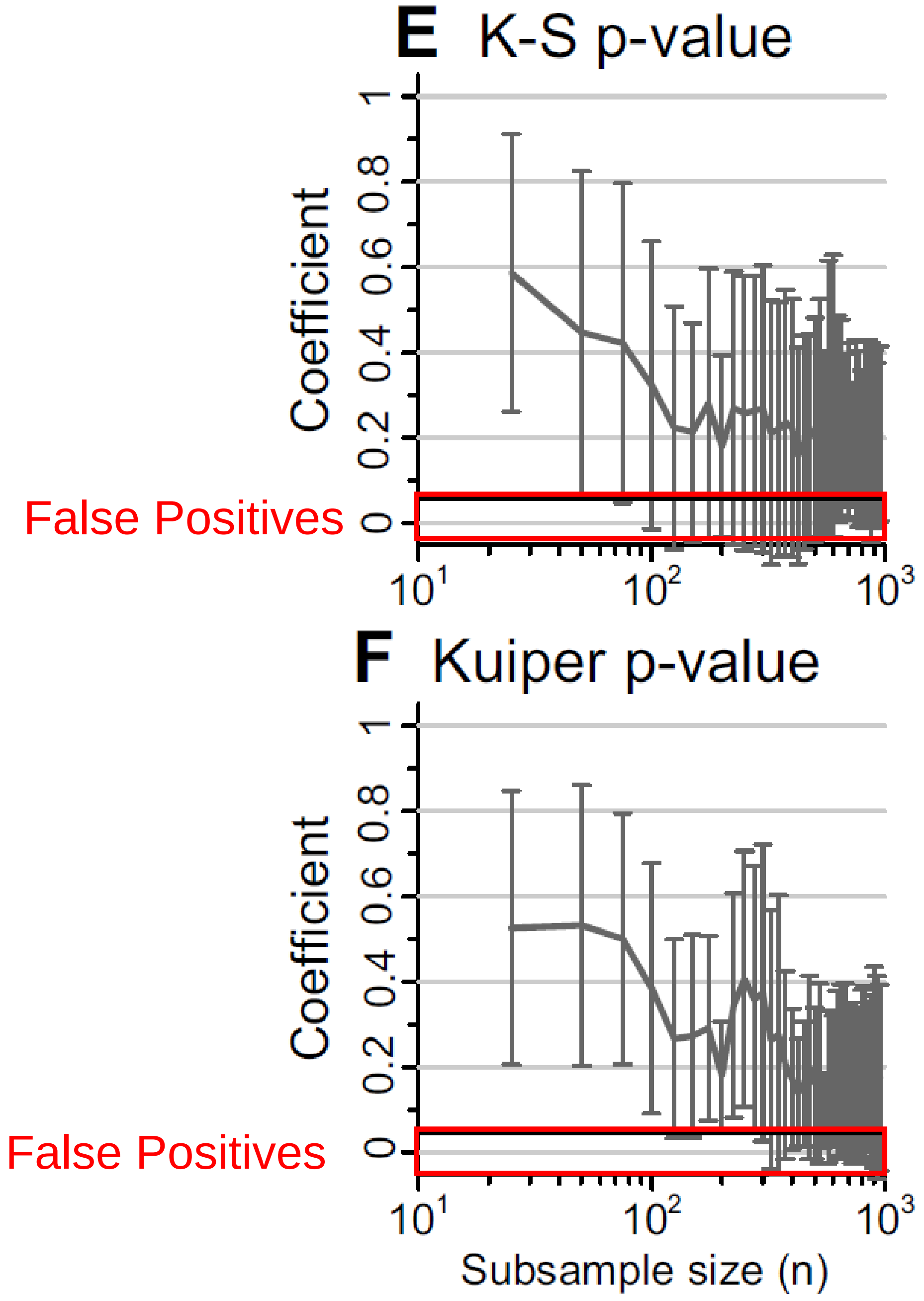
Limitations of Kuiper and K-S distances

- 2) Sensitive to age proportions and distribution
  - D value for AD (0) and BD (3) = 0.2
  - V value for AD (0) and BD (3) = 0.4



A note on p values

- Typically if the p-value is less than our confidence level, the hypothesis of common derivation is rejected.
  - For example a p value <0.05 indicates that the **null hypothesis of common derivation** can be rejected at the 95% confidence level.
- PROBLEM: over-occurrence of Type 1 errors
  - false positive (i.e., incorrectly rejecting the null hypothesis, Saylor and Sundell, 2016)
- There is always a sample size at which differences between samples are observable
  - Vermeesch (2013, 2018)



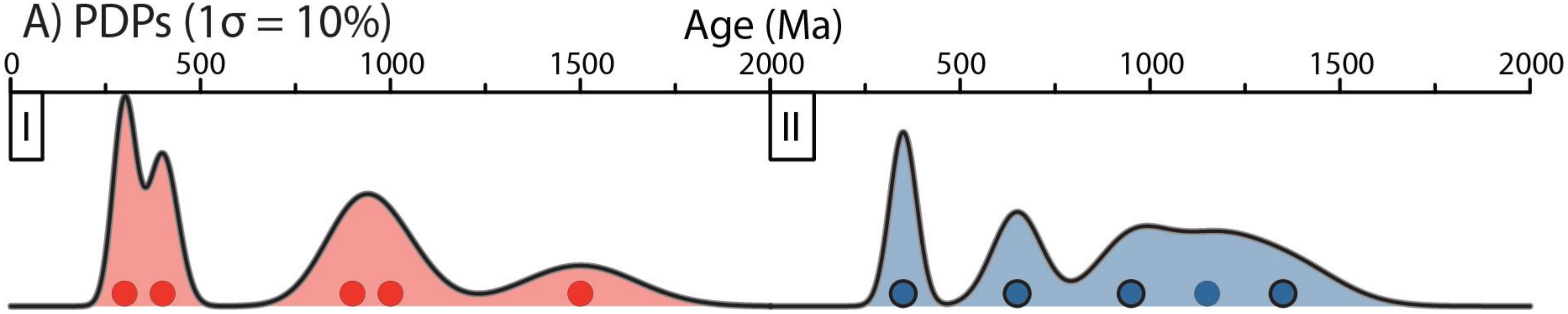
Saylor and Sundell (2016)

Module 3 Outline

- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multi-dimensional scaling (MDS)

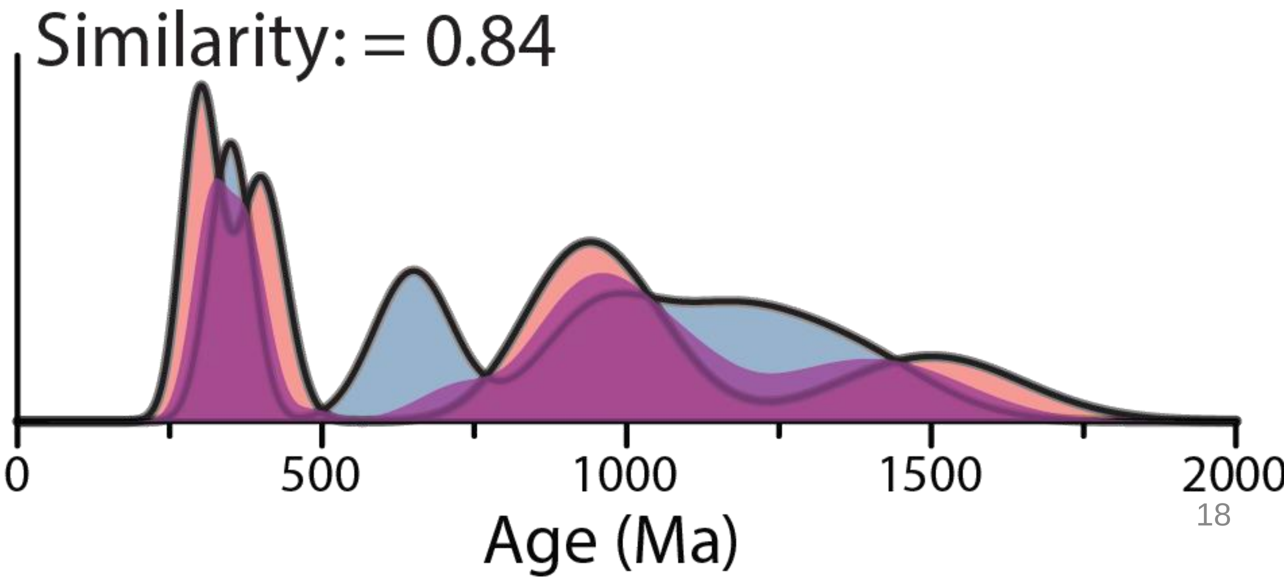
Similarity

- Bhattacharya distance (Bhattacharya, 1943; 1946)
- Introduced to detrital geochronology by Gehrels (2000)



$$S(f,g) = \sum_{i=0}^{x} \sqrt{f(x)g(x)}$$

- Recall that for this data set
  - K-S D value = 0.2
  - Kuiper V value = 0.4



Module 3 Outline

- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multi-dimensional scaling (MDS)

Mismatch/Likeness

- Mismatch introduced by Amidon et al. (2005)

$$M(f,g) = \frac{1}{2}\sum_{i=0}^{x} |f(x) - g(x)|$$

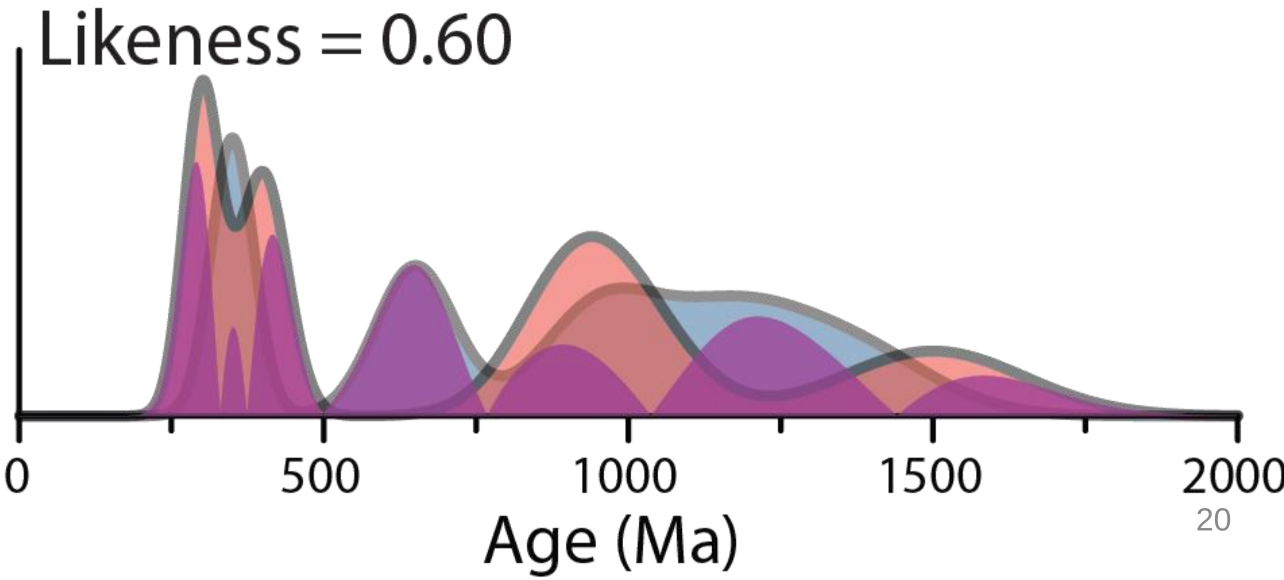
  - Ranges from 1 (no overlap) to 0 (identical)

- Modified by Satkoski et al. (2013) to Likeness

$$L\,(f,g) = 1\, - M(f,g)$$

  - Range: 0 (no overlap) to 1 (identical)

- Recall that for this data set
  - D = 0.2
  - V = 0.4
  - S = 0.84

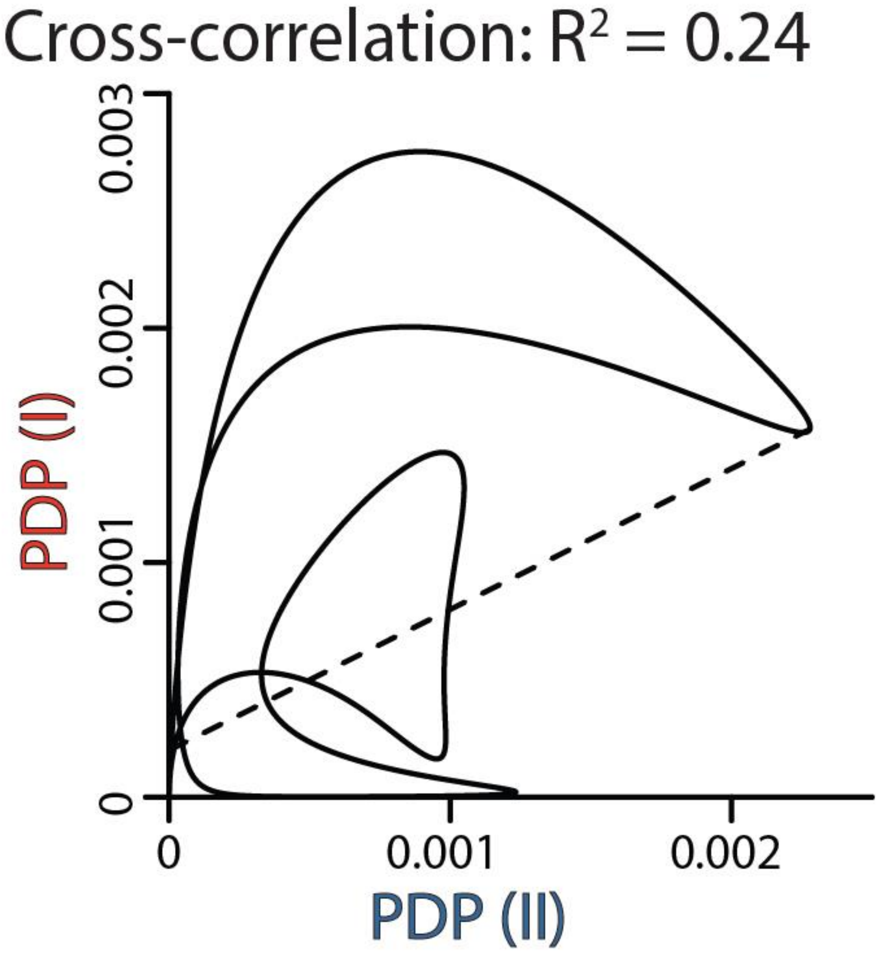


Module 3 Outline

- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multi-dimensional scaling (MDS)

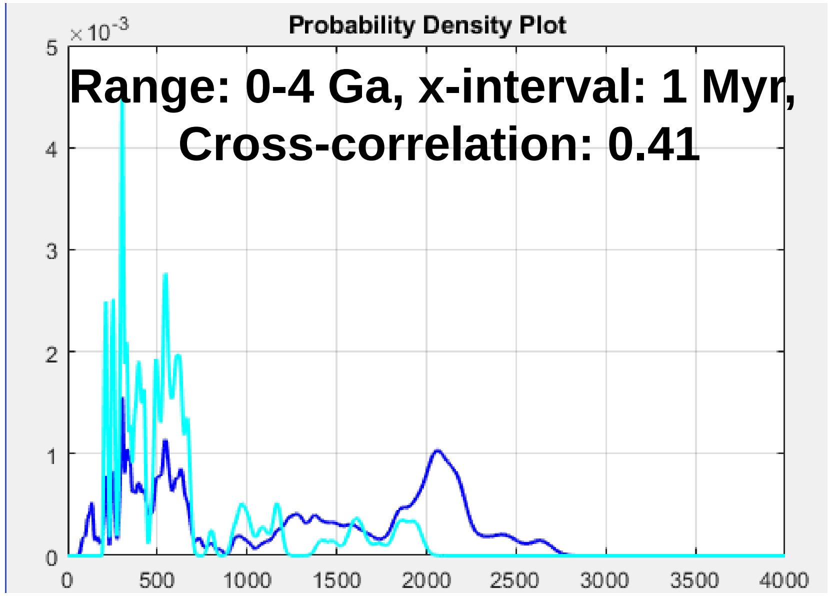
Cross-correlation

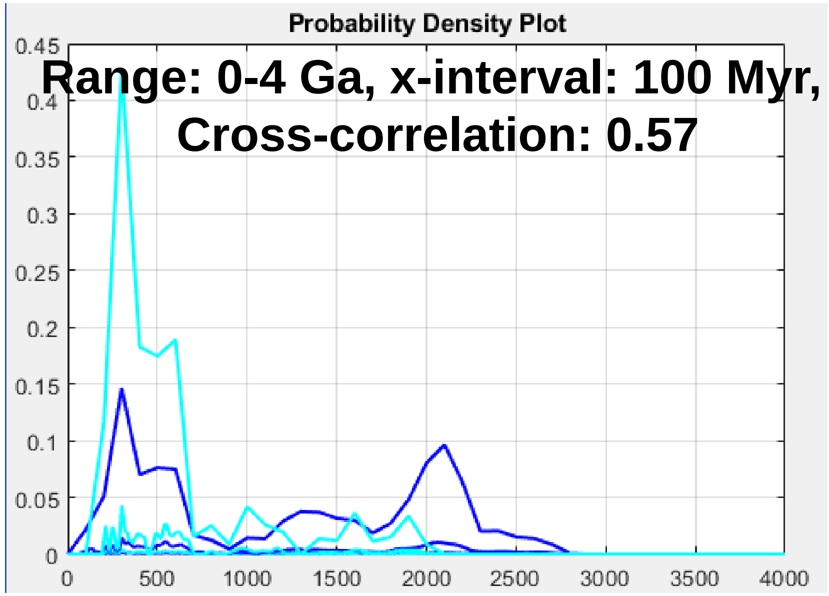
- Widely used in signal processing, template matching, image matching, and geophysics
- Pearson's correlation coefficient for co-located PDPs or KDEs
  - Squared to ensure range of 0-1
- $R(f,g)^2 = \left(\frac{\sum_{i=0}^{x}(f_i-\bar{f})(g_i-g)}{\sqrt{\sum_{i=0}^{x}(f_i-\bar{f})^2}\sqrt{\sum_{i=0}^{x}(g_i-g)^2}}\right)^2$
  - Ranges from 0 (no correlation) to 1 (perfectly correlated)
- Sensitive to the location and distribution of modes



A note on intervals & resolution

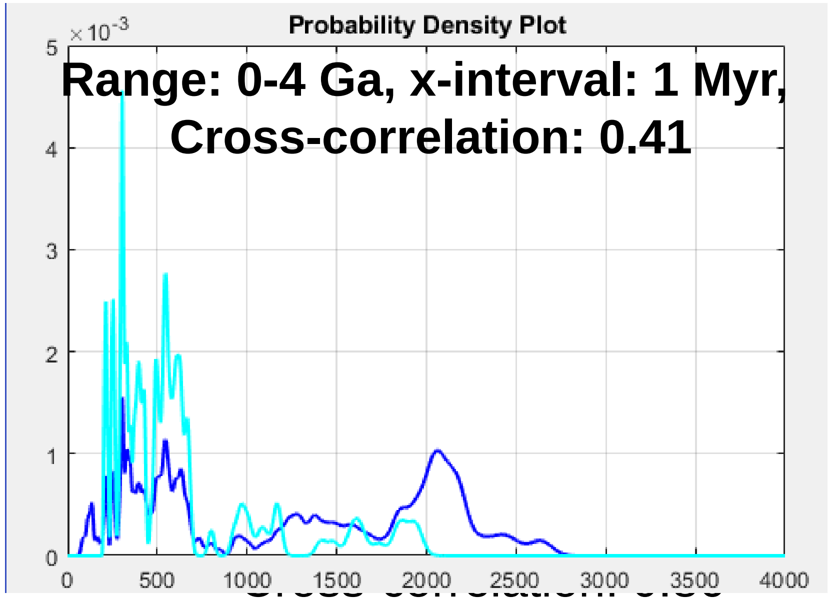
- When applied to discretized functions comparison metrics depend on
  - Coarseness of discretization (1 Myr? 0.5 Myr? 10 Myr?)
  - Applies to PDPs, KDEs, or CDFs produced from summation of them.
- Comparison metrics always depend on range
  - What are the min and max ages in the comparison?

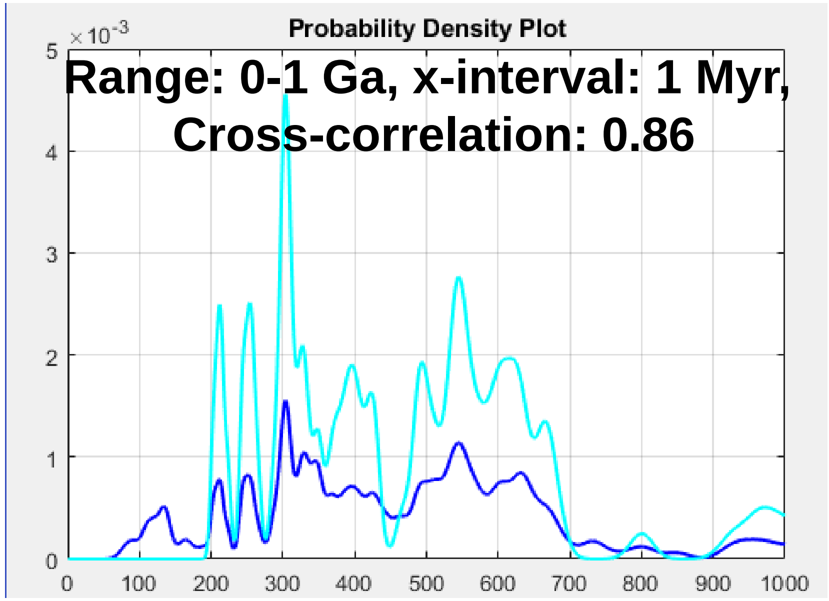




A note on intervals & resolution

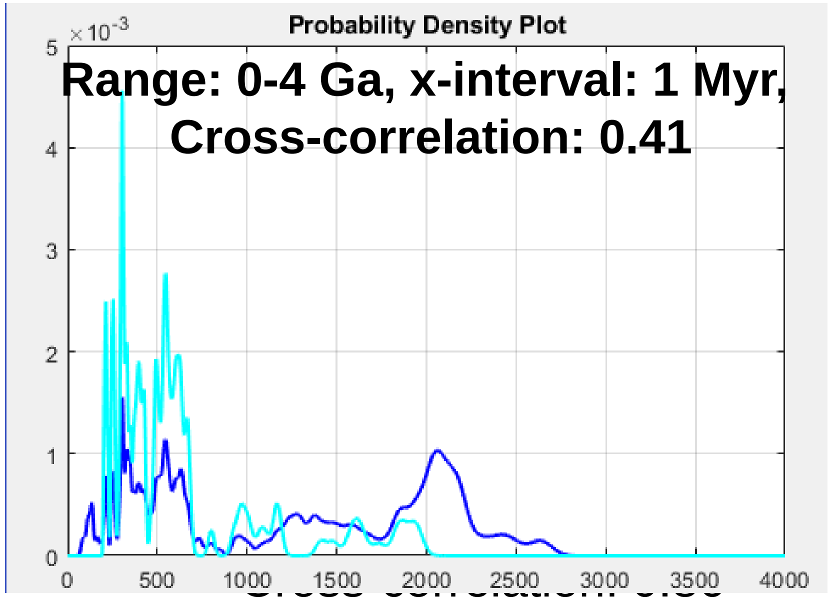
- When applied to discretized functions comparison metrics depend on
  - Coarseness of discretization (1 Myr? 0.5 Myr? 10 Myr?)
  - Applies to PDPs, KDEs, or CDFs produced from summation of them.
- Comparison metrics always depend on range
  - What are the min and max ages in the comparison?

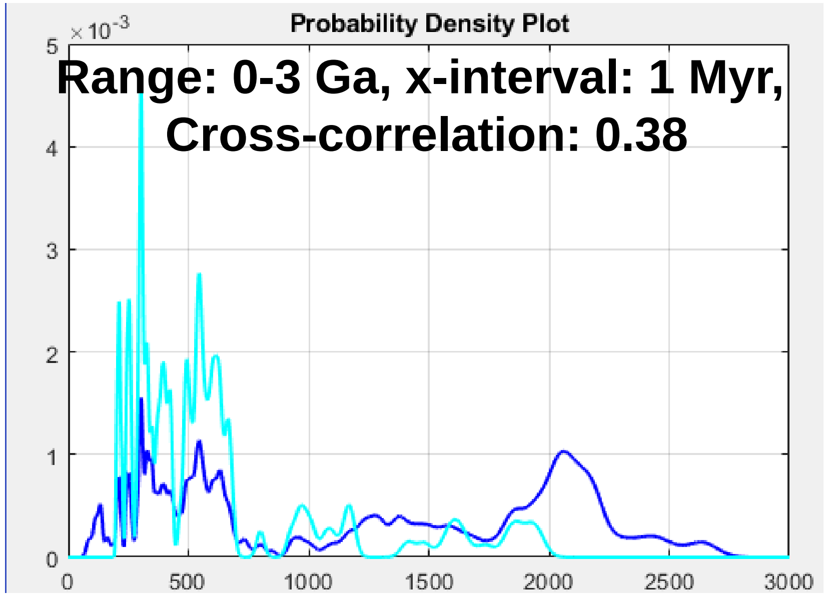




A note on intervals & resolution

- When applied to discretized functions comparison metrics depend on
  - Coarseness of discretization (1 Myr? 0.5 Myr? 10 Myr?)
  - Applies to PDPs, KDEs, or CDFs produced from summation of them.
- Comparison metrics always depend on range
  - What are the min and max ages in the comparison?
  - For Cross-correlation even zeros matter!





Module 3 Outline

- Some metrics applicable to detrital geochronology
  - Metrics based on CDF
    - Kolmogorov-Smirnov distance (D value)
    - Kuiper distance (V value)
  - Metrics based on PDPs/KDEs
    - Similarity
    - Mismatch/Likeness
    - Cross-correlation
- Application to multidimensional scaling (MDS)

Application to Multidimensional scaling (MDS)

- Converts dissimilarity to distance
  - By iterative rearrangement of the samples in Cartesian space
  - $\hat{d}(i,j) = f[p(i,j)]$
    - *p(i,j)* = (dis)similarity between samples *i* and *j*
    - $\hat{d}(i,j)$ = distance between samples *i* and *j* in Cartesian space (transformation of *p(i,j)*)
      - Referred to as “disparity” or "approximated distances“ to distinguish it from the final plotted distance.
    - $d(i,j)$ = final plotted distance between samples *i* and *j* in Cartesian space
  - Goal to minimize stress function $\left|\hat{d}(i,j) - d(i,j)\right|$
- Types
  - Nonmetric (qualitative)
  - Metric (quantitative)

Metric MDS

- “MDS [is] a method that represents (dis)similarity data as distances in a low dimensional space in order to make these data accessible to visual inspection and exploration” Borg and Groenen (1997)

TABLE 2.1. Distances between ten cities.

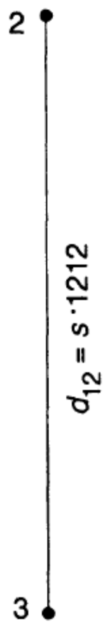
	1	2	3	4	5	6	7	8	9	10
1	0	569	667	530	141	140	357	396	570	190
2	569	0	1212	1043	617	446	325	423	787	648
3	667	1212	0	201	596	768	923	882	714	714
4	530	1043	201	0	431	608	740	690	516	622
5	141	617	596	431	0	177	340	337	436	320
6	140	446	768	608	177	0	218	272	519	302
7	357	325	923	740	340	218	0	114	472	514
8	396	423	882	690	337	272	114	0	364	573
9	569	787	714	516	436	519	472	364	0	755
10	190	648	714	622	320	302	514	573	755	0

Borg, I., and P. Groenen (1997), *Modern Multidimensional Scaling: Theory and Applications*, Springer New York.

Multidimensional scaling (MDS)

TABLE 2.1. Distances between ten cities.

	1	2	3	4	5	6	7	8	9	10
1	0	569	667	530	141	140	357	396	570	190
2	569	0	1212	1043	617	446	325	423	787	648
3	667	1212	0	201	596	768	923	882	714	714
4	530	1043	201	0	431	608	740	690	516	622
5	141	617	596	431	0	177	340	337	436	320
6	140	446	768	608	177	0	218	272	519	302
7	357	325	923	740	340	218	0	114	472	514
8	396	423	882	690	337	272	114	0	364	573
9	569	787	714	516	436	519	472	364	0	755
10	190	648	714	622	320	302	514	573	755	0



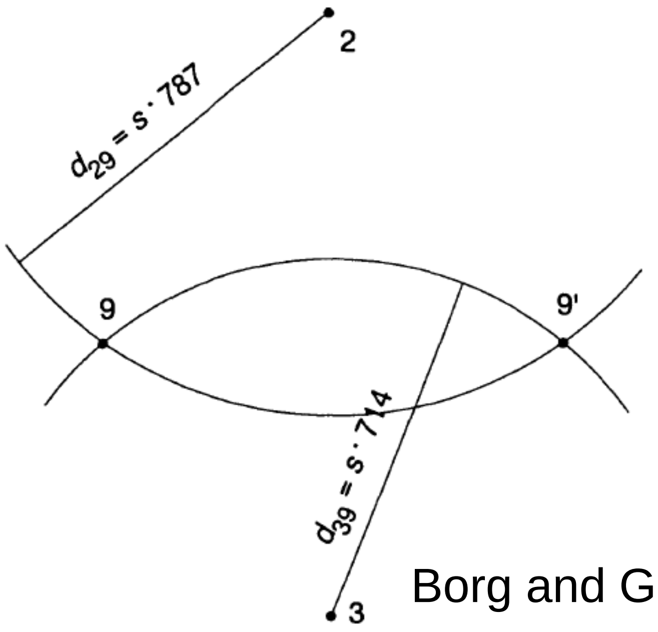
- Example from Borg and Groenen (1997) of disances between European cities
- Plot maximum distance

Multidimensional scaling (MDS)

TABLE 2.1. Distances between ten cities.

	1	2	3	4	5	6	7	8	9	10
1	0	569	667	530	141	140	357	396	570	190
2	569	0	1212	1043	617	446	325	423	787	648
3	667	1212	0	201	596	768	923	882	714	714
4	530	1043	201	0	431	608	740	690	516	622
5	141	617	596	431	0	177	340	337	436	320
6	140	446	768	608	177	0	218	272	519	302
7	357	325	923	740	340	218	0	114	472	514
8	396	423	882	690	337	272	114	0	364	573
9	569	787	714	516	436	519	472	364	0	755
10	190	648	714	622	320	302	514	573	755	0

- Triangulate intermediate distances
- 9 or 9’?
  - It doesn’t matter
  - Just a reflection (see next slides)



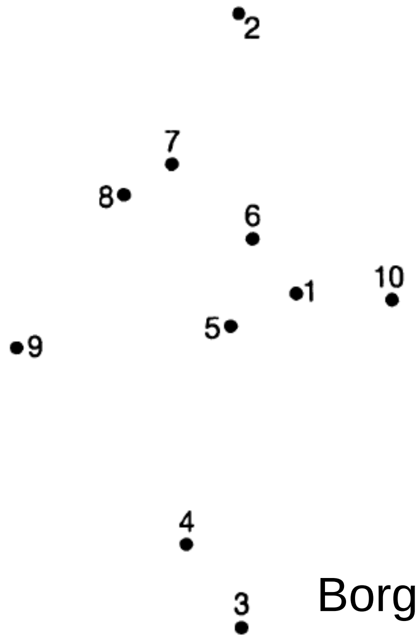
Borg and Groenen (1997)

Multidimensional scaling (MDS)

TABLE 2.1. Distances between ten cities.

	1	2	3	4	5	6	7	8	9	10
1	0	569	667	530	141	140	357	396	570	190
2	569	0	1212	1043	617	446	325	423	787	648
3	667	1212	0	201	596	768	923	882	714	714
4	530	1043	201	0	431	608	740	690	516	622
5	141	617	596	431	0	177	340	337	436	320
6	140	446	768	608	177	0	218	272	519	302
7	357	325	923	740	340	218	0	114	472	514
8	396	423	882	690	337	272	114	0	364	573
9	569	787	714	516	436	519	472	364	0	755
10	190	648	714	622	320	302	514	573	755	0

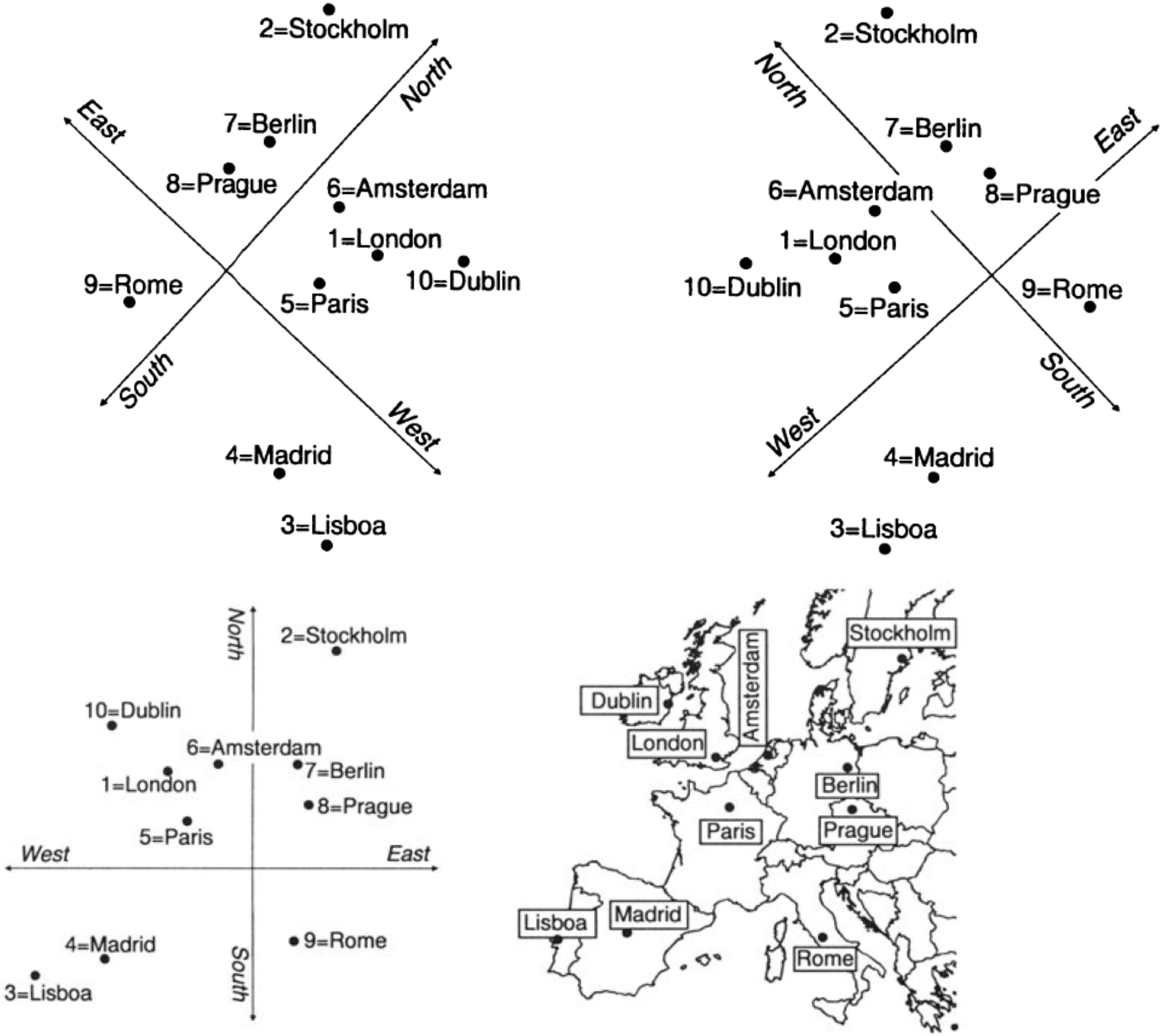
- Final map
- Constrained by multiple pairs (multiple distances)
  - e.g., location of 9 constrained by 9 pairs
  - etc



Borg and Groenen (1997)

Multidimensional scaling (MDS)

- Rotate, Reflect, Scale
- Its all good!



Borg and Groenen (1997)

Nonmetric MDS

- Assumes that the degree of separation is not as important as the relative ranking of the samples
- Works on the same basis as metric
  - But narrows down *zones* of occupation

TABLE 2.3. Ranks for data in Table 2.1; the smallest distance has rank 1.

	1	2	3	4	5	6	7	8	9	10
1	–	26	34	25	3	2	14	16	27	5
2	26	–	45	44	31	20	11	17	41	33
3	34	45	–	6	29	40	43	42	36	36
4	25	44	6	–	18	30	38	35	23	32
5	3	31	29	18	–	4	13	12	19	10
6	2	20	40	30	4	–	7	8	24	9
7	14	11	43	38	13	7	–	1	21	22
8	16	17	42	35	12	8	1	–	15	28
9	27	41	36	23	19	24	21	15	–	39
10	5	33	36	32	10	9	22	28	39	–

Comparison of metric and nonmetric MDS

- Usually very similar

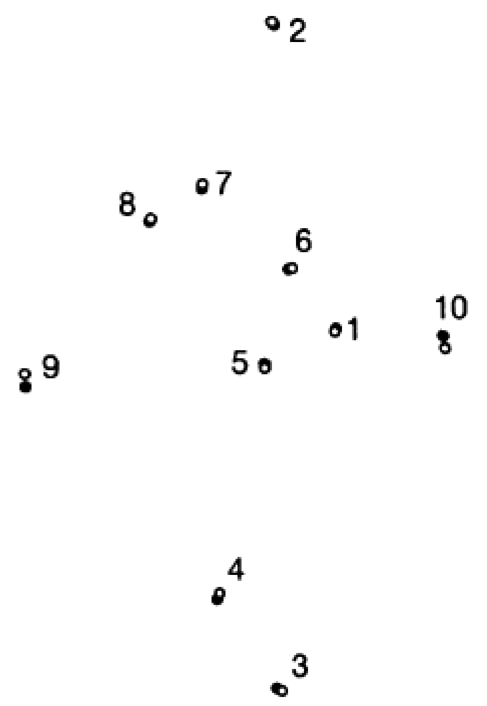
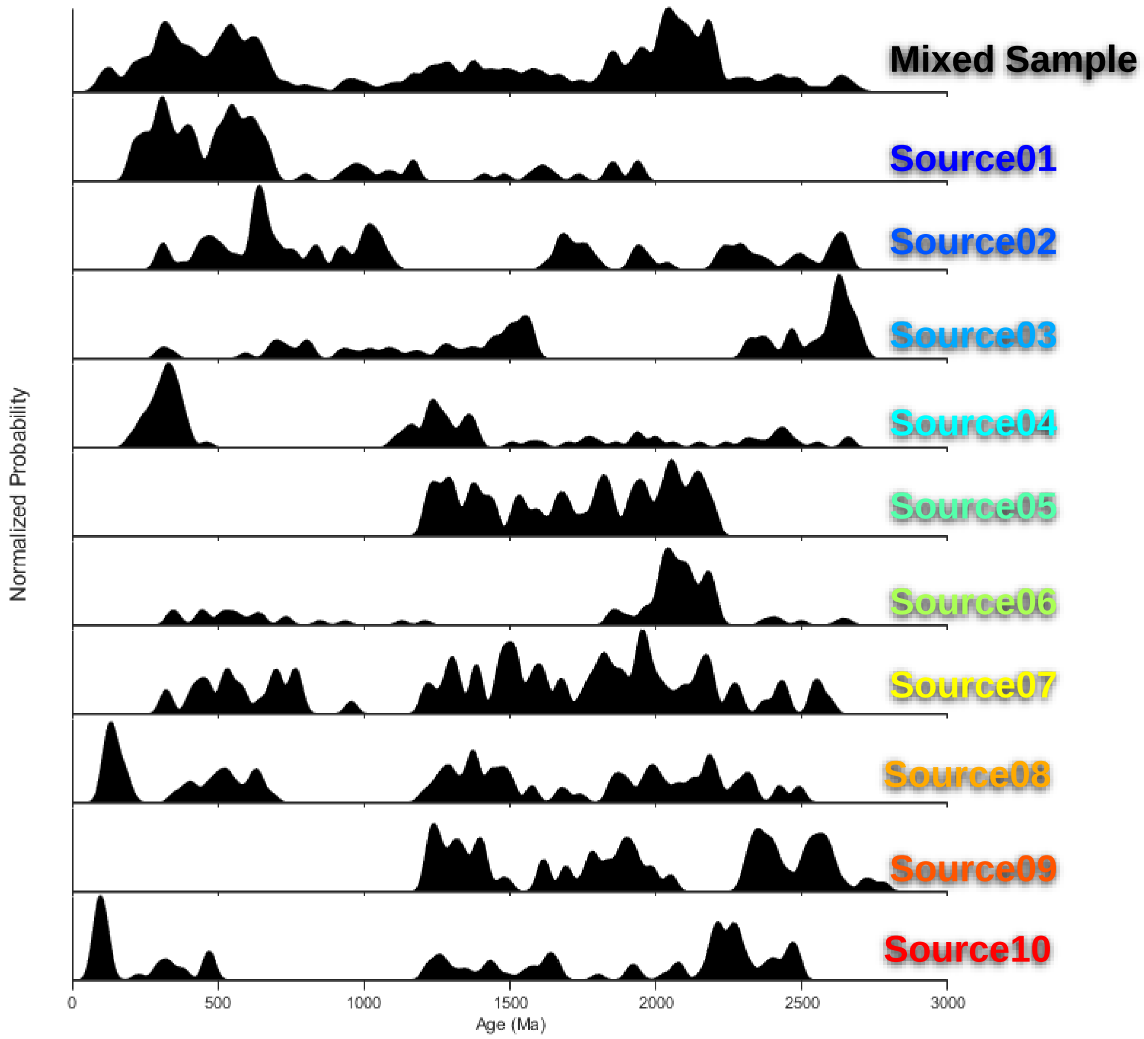


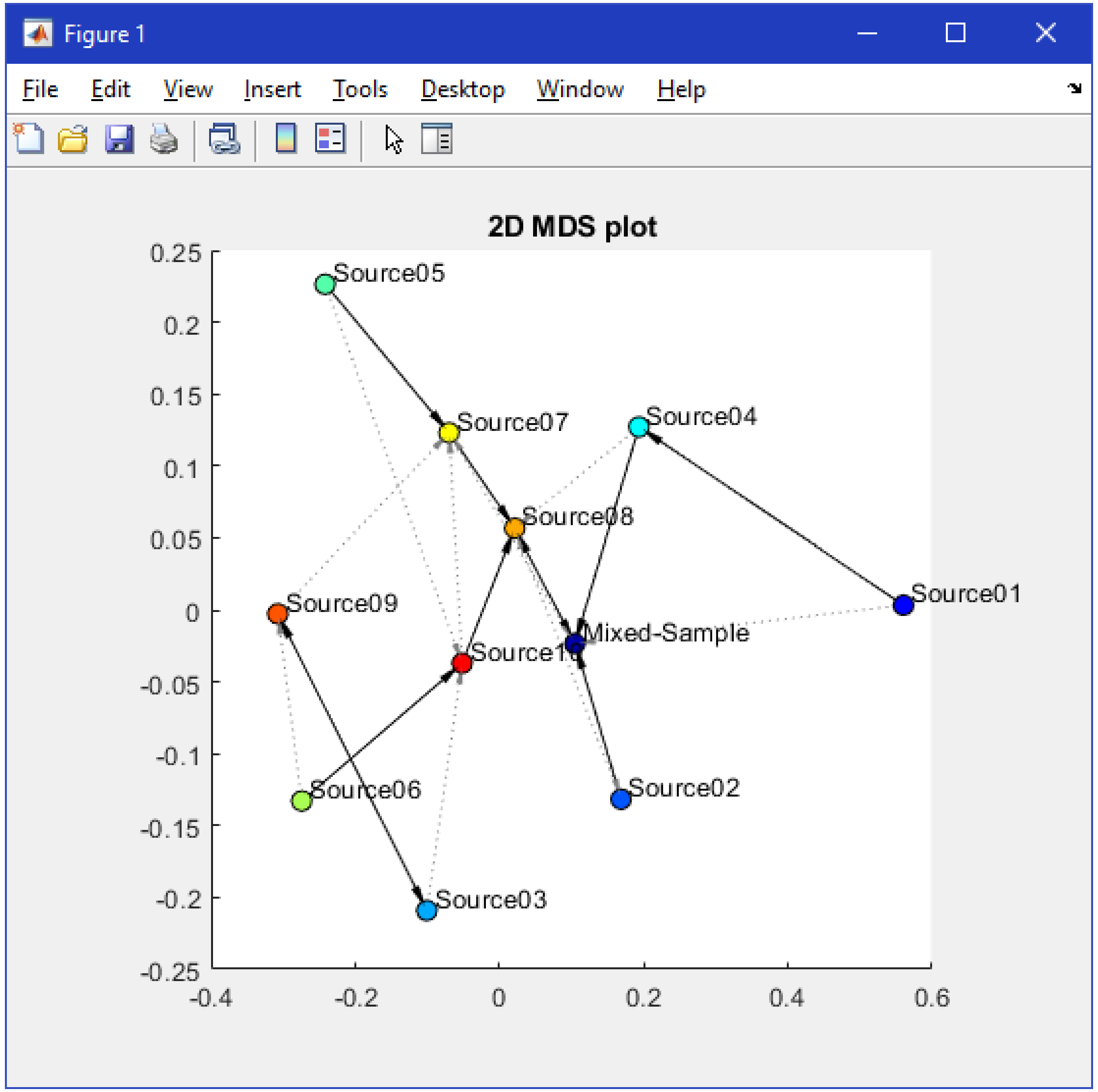
FIGURE 2.14. Comparing ratio MDS (solid points) and ordinal MDS (open circles) after fitting the latter to the former.

Borg and Groenen (1997)

Assessing the quality of the MDS

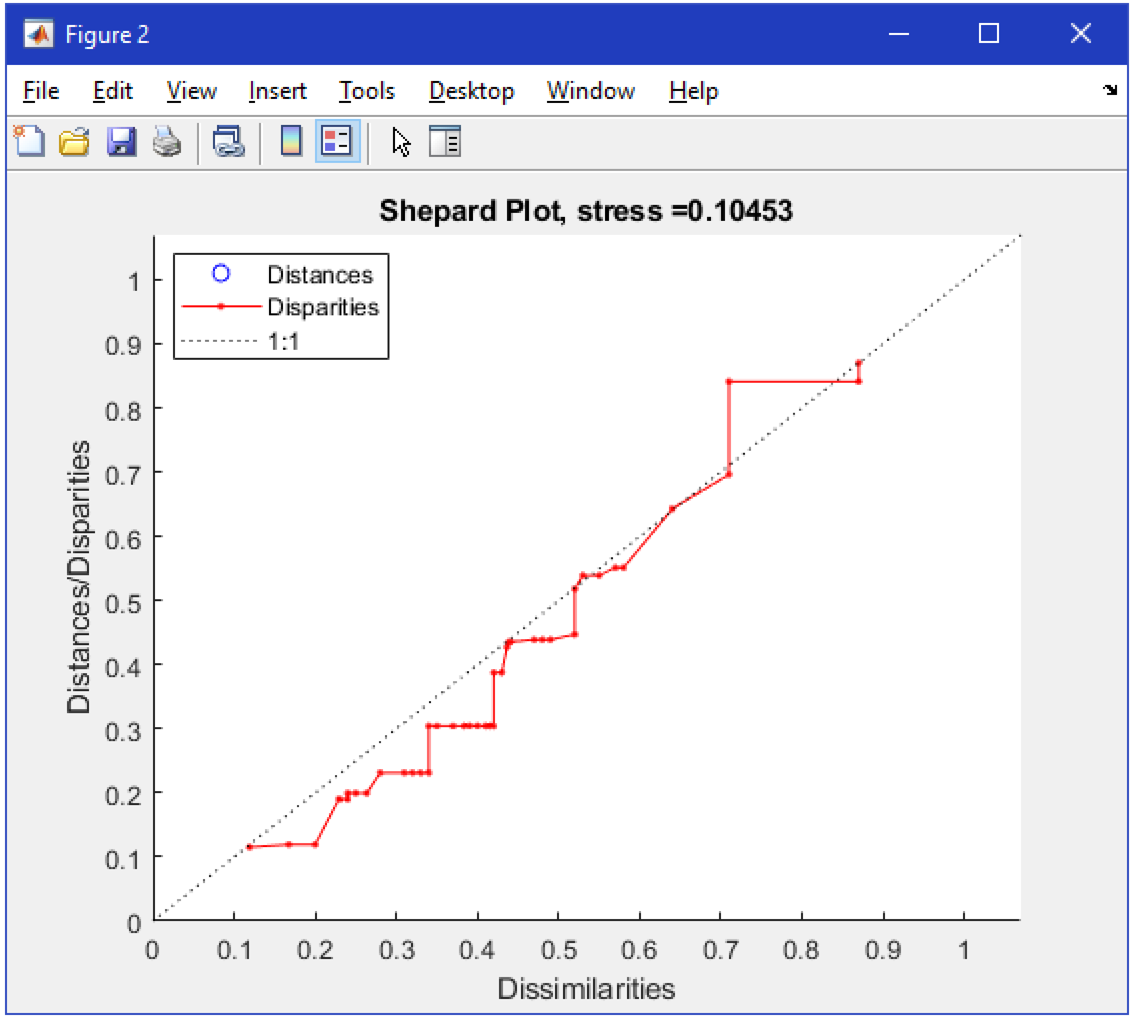
- Nonmetric MDS based on K-S D value





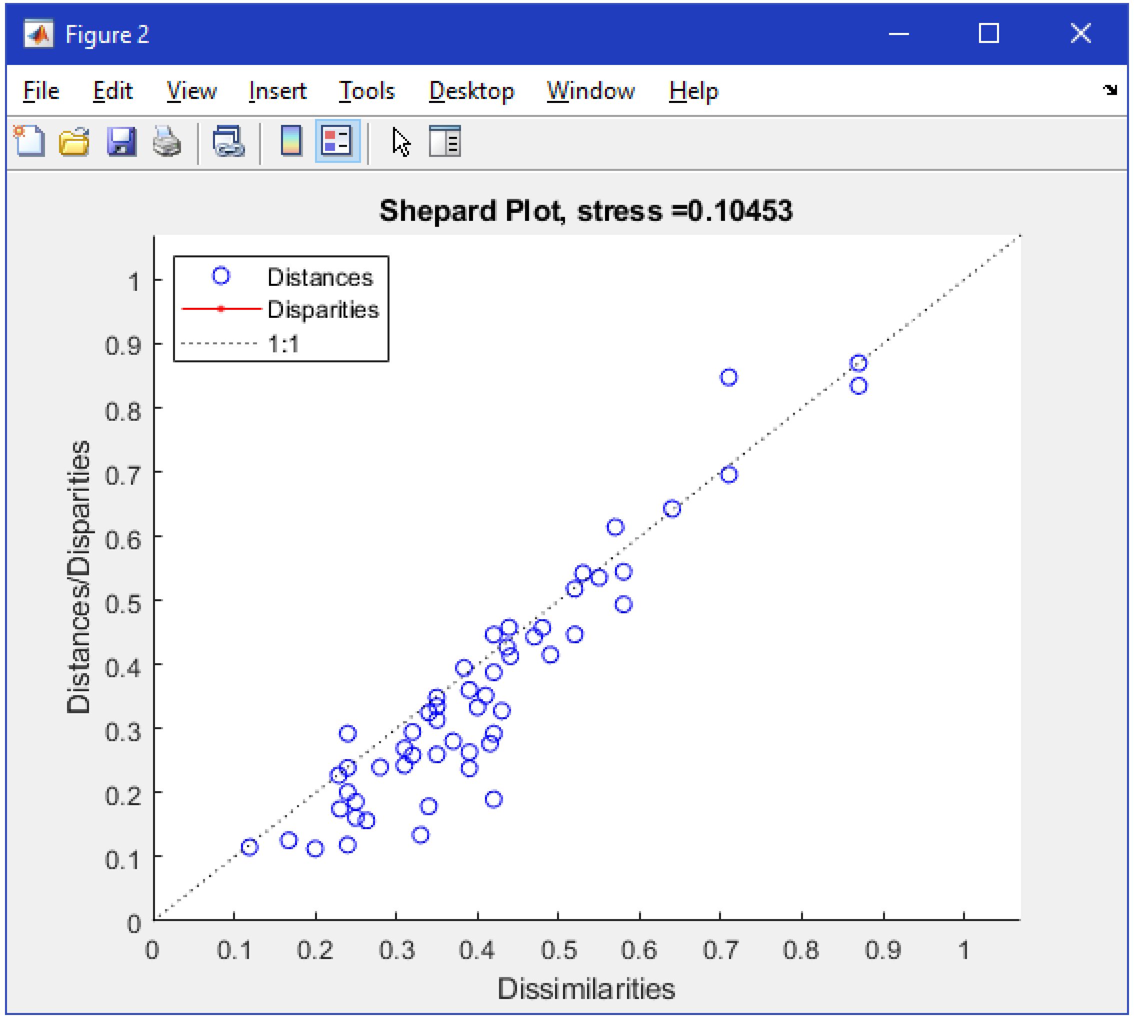
Assessing the quality of the MDS

- Nonmetric MDS
- Based on K-S D value
- x : $p(i,j)$
  - dissimilarity, rank in this case
- y : $\hat{d}(i,j)$
  - disparity



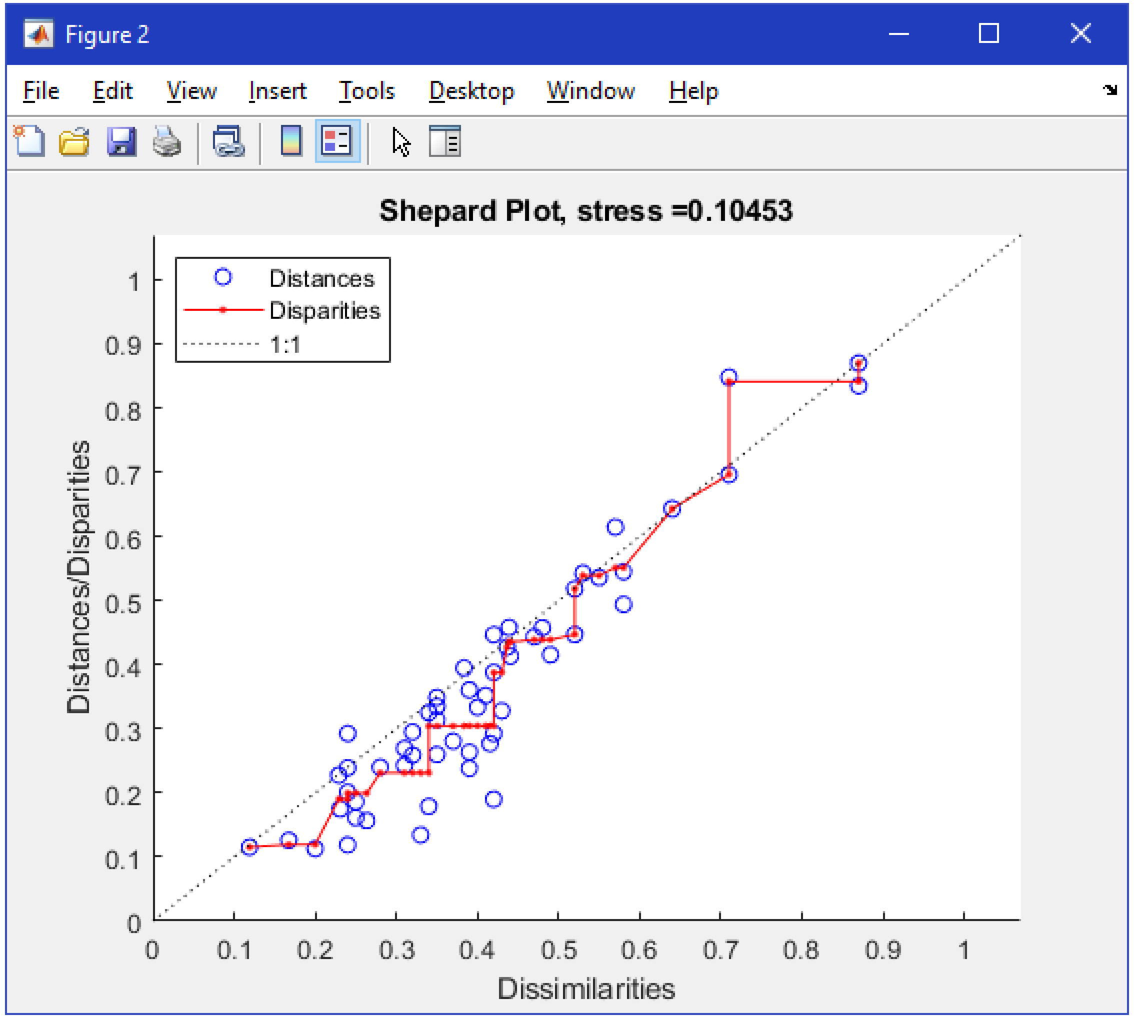
Assessing the quality of the MDS

- Nonmetric MDS
- Based on K-S D value
- x : $p(i, j)$
  - dissimilarity
- y : $d(i, j)$
  - distance



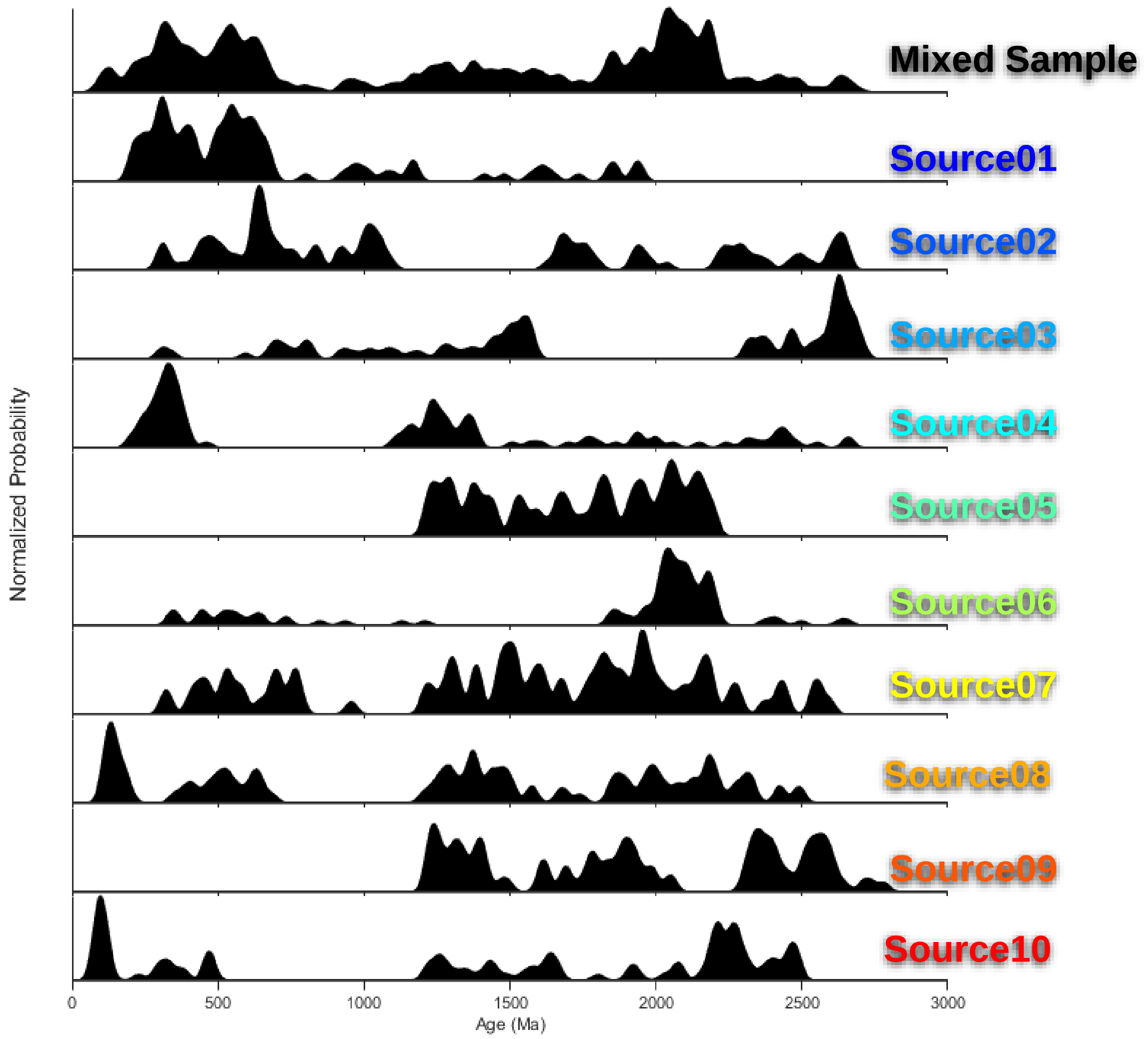
Assessing the quality of the MDS

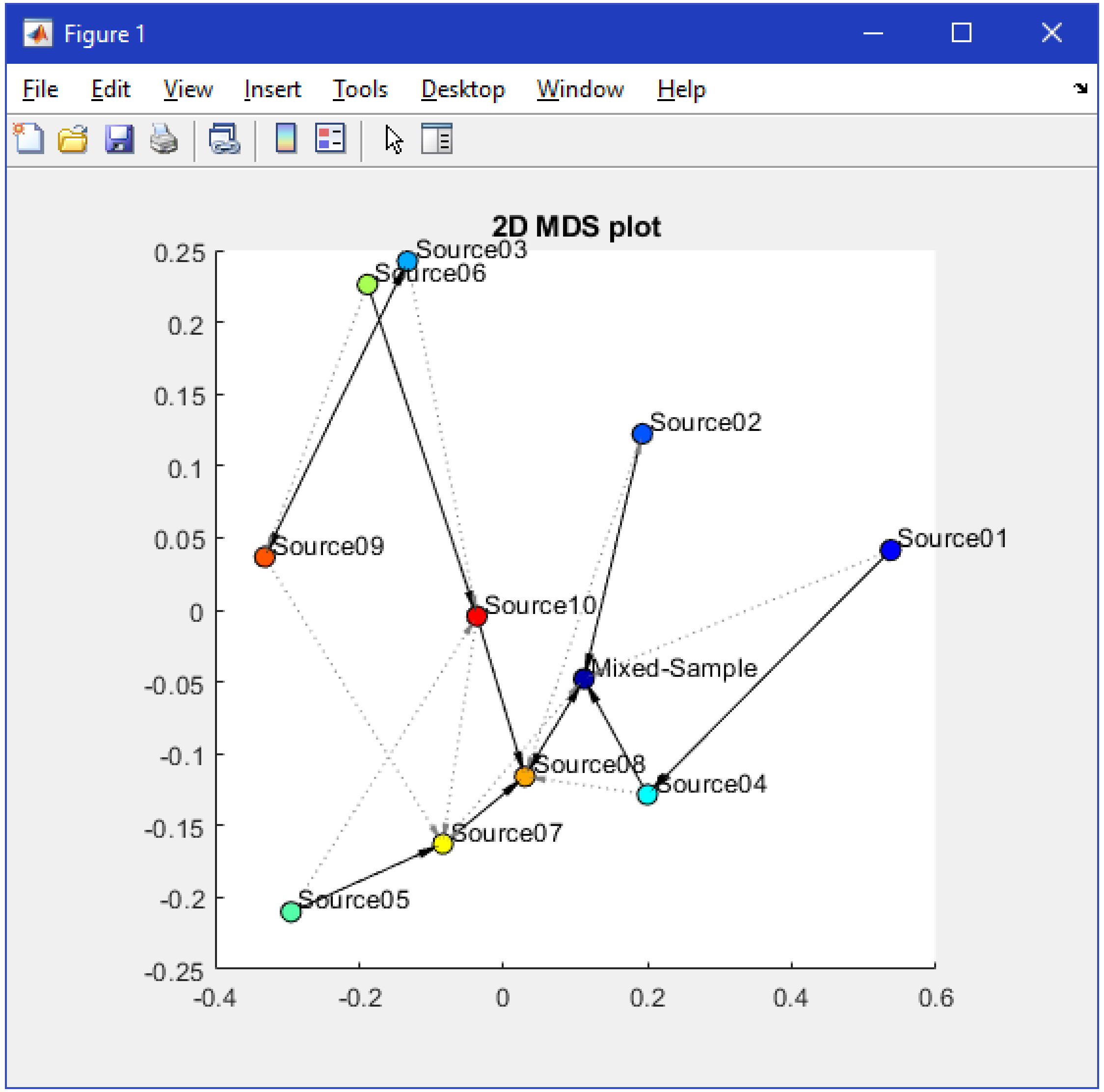
- Nonmetric MDS
- Based on K-S D value
- x : $p(i,j)$
  - dissimilarity
- y : $\hat{d}(i,j)$
  - disparity
- y : $d(i,j)$
  - distance



Assessing the quality of the MDS

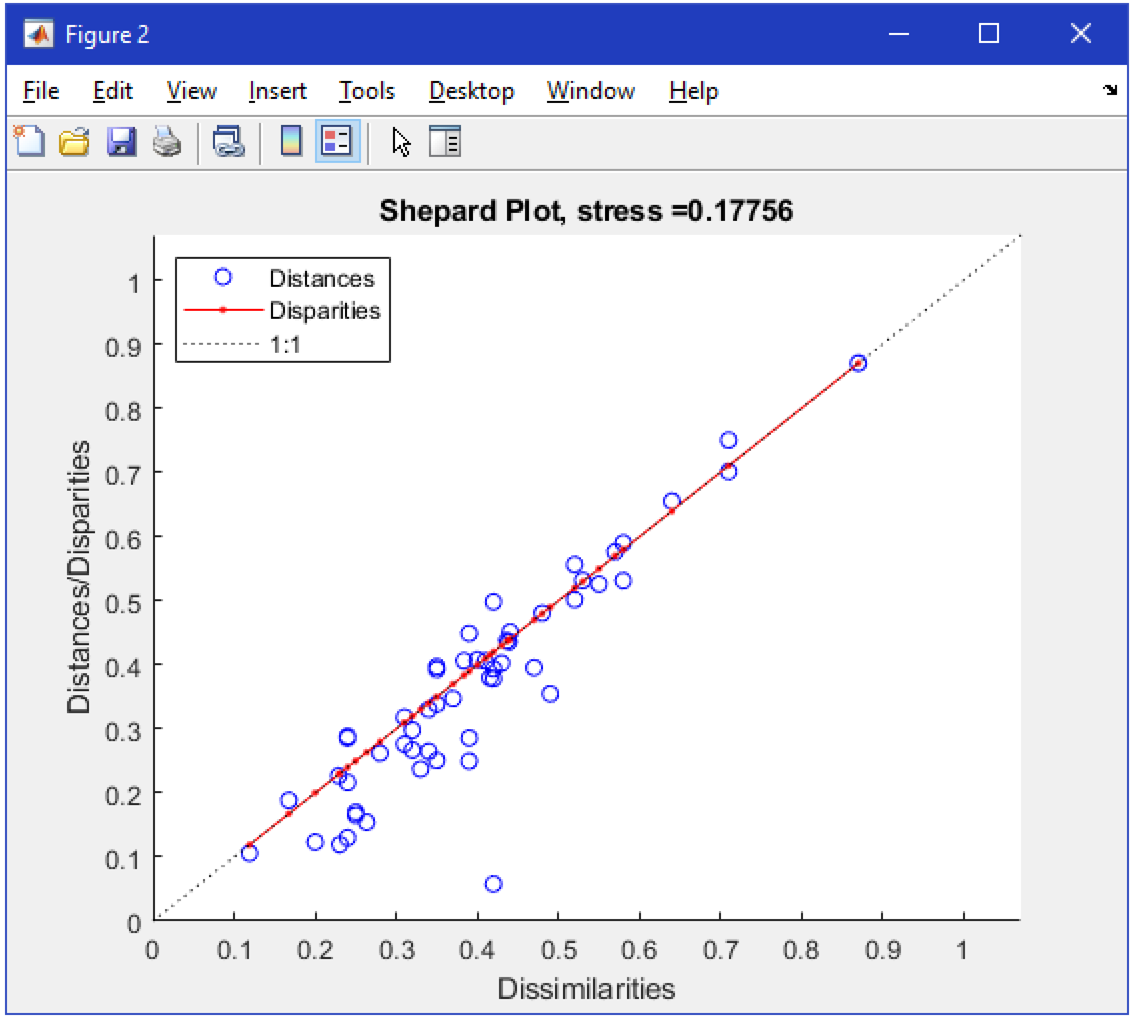
- Metric MDS based on K-S D value



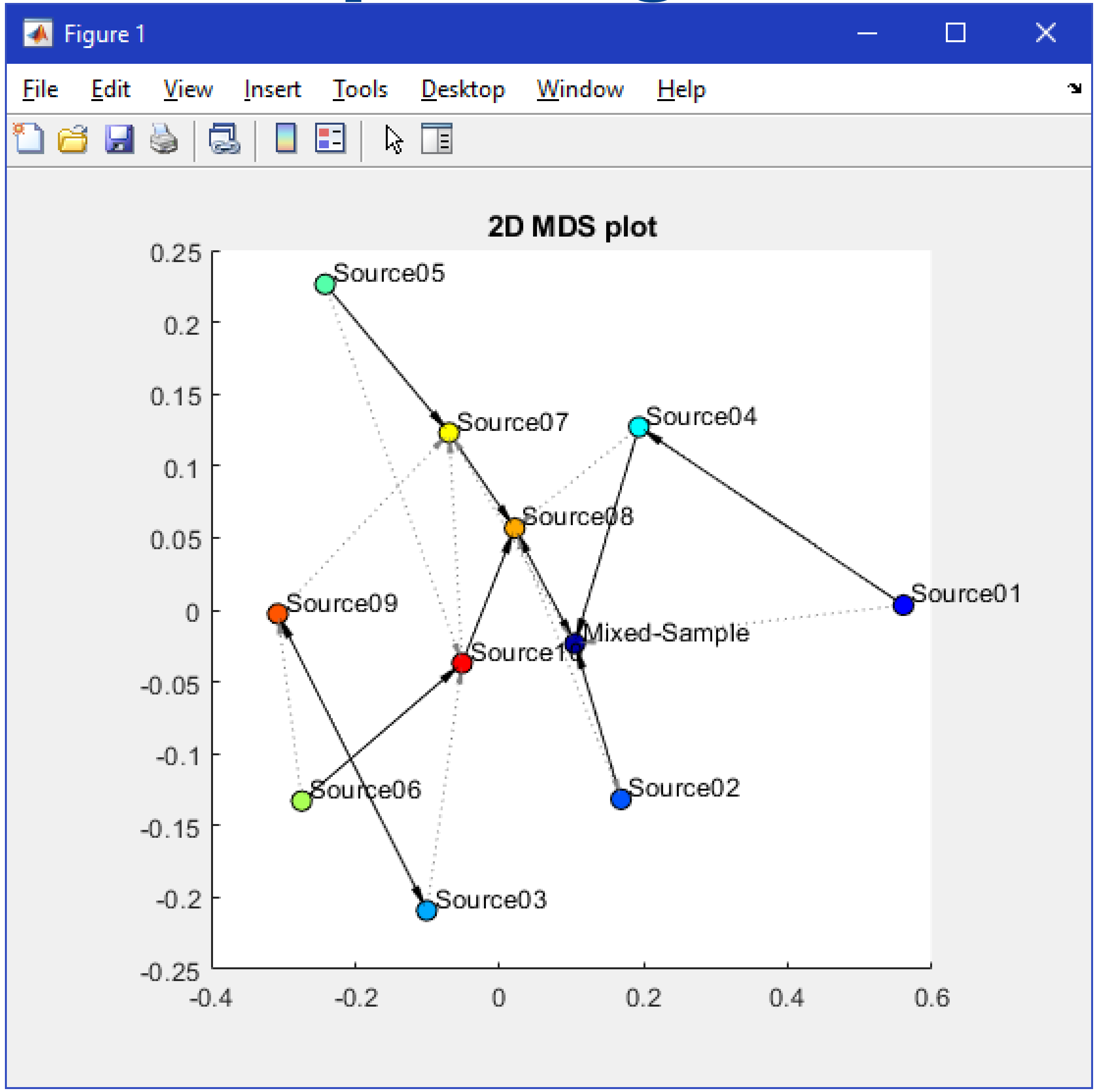


Assessing the quality of the MDS

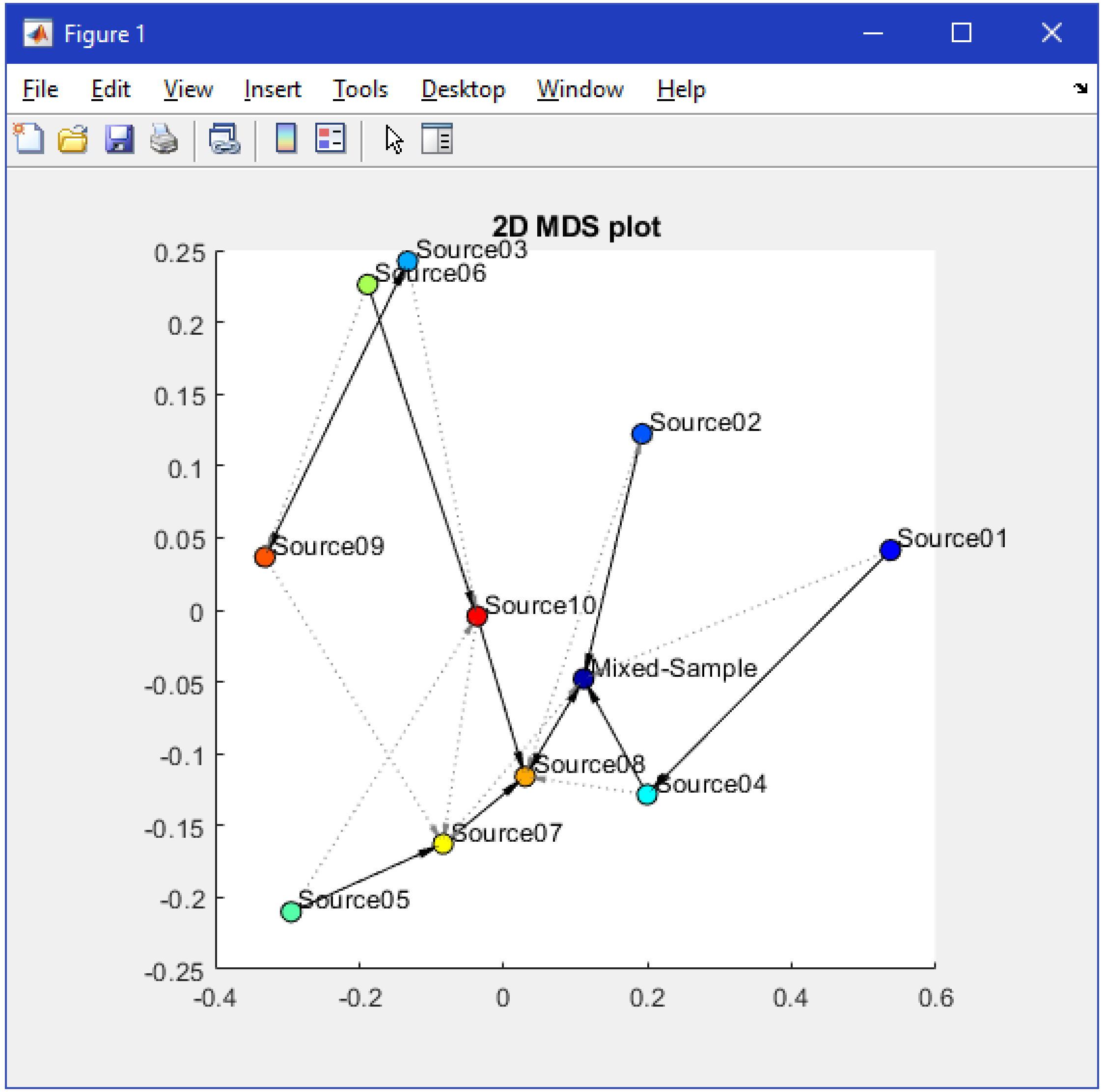
- Metric MDS
- Based on K-S D value
- Stress squared
- x : $p(i,j)$
  - dissimilarity
- y : $\hat{d}(i,j)$
  - Disparity
  - Lie on 1:1 line because it is a linear transformation of *p(i,j)*
- y : $d(i,j)$
  - distance



Comparing Nonmetric and Metric

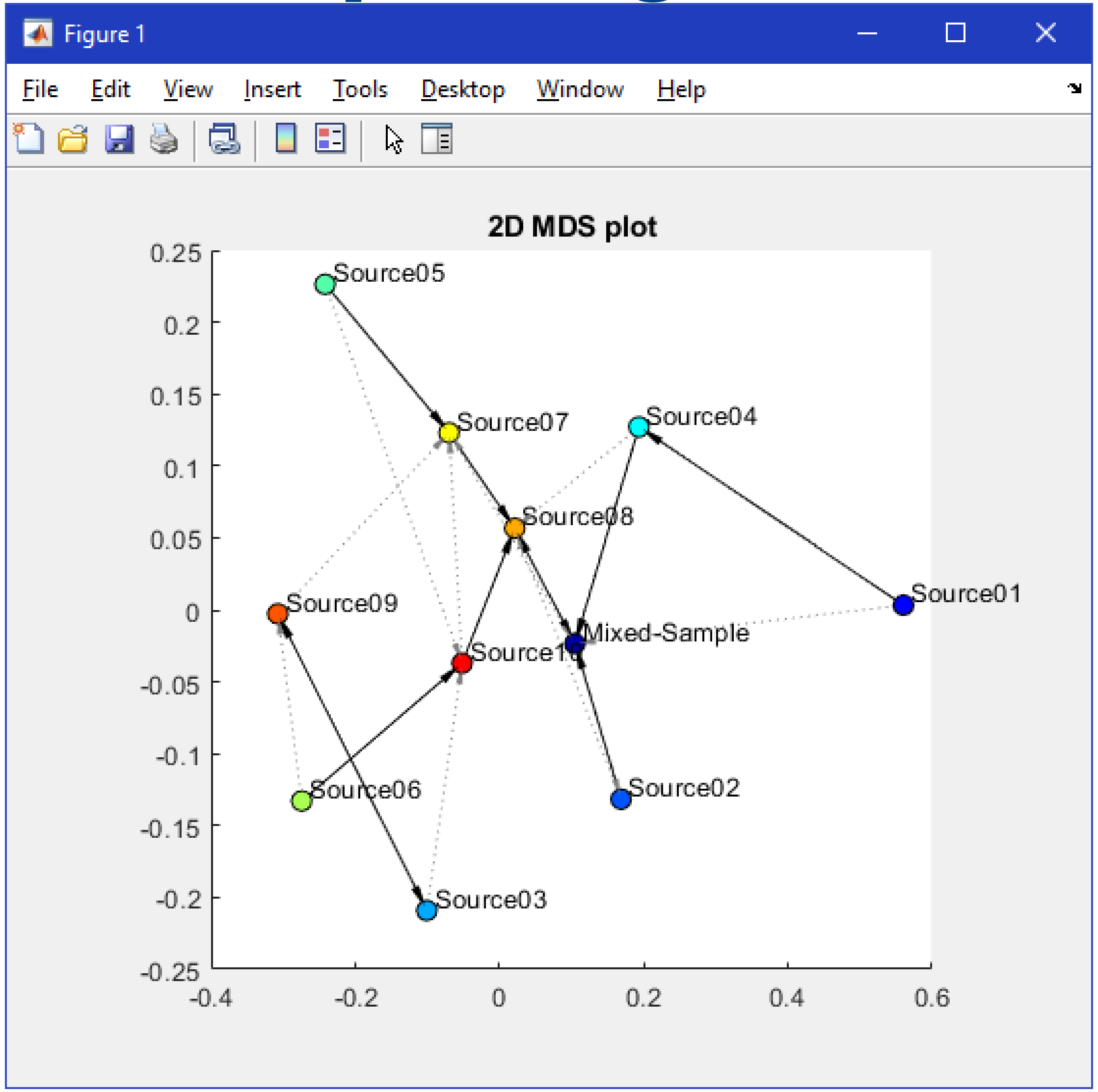


Nonmetric MDS



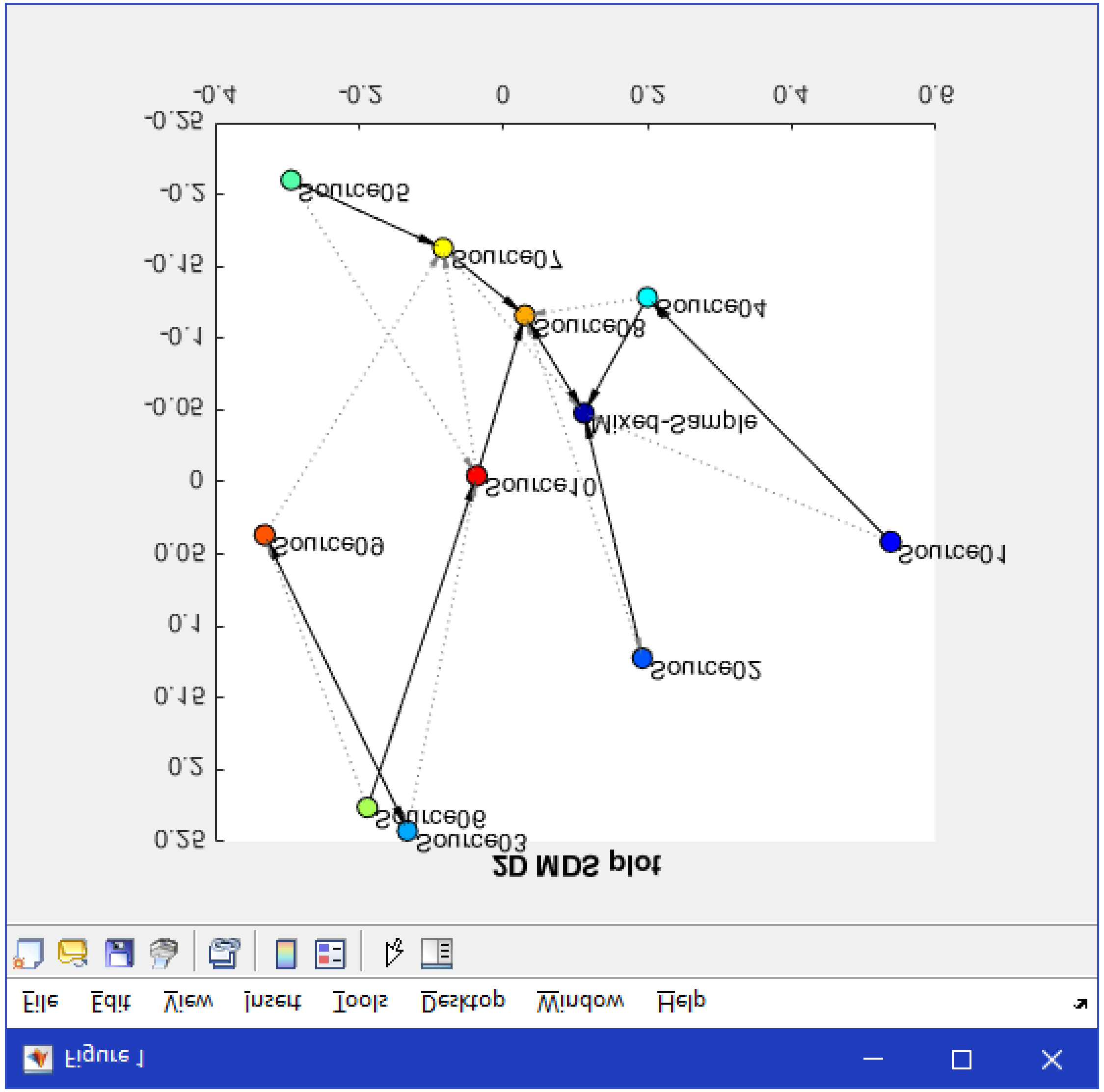
Metric MDS

Comparing Nonmetric and Metric



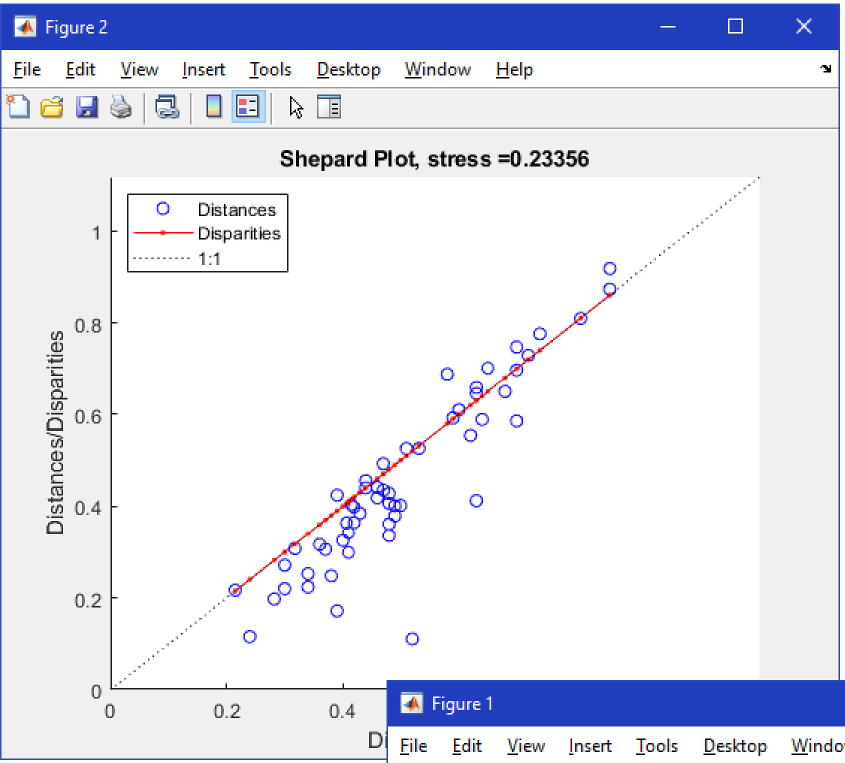
Nonmetric MDS

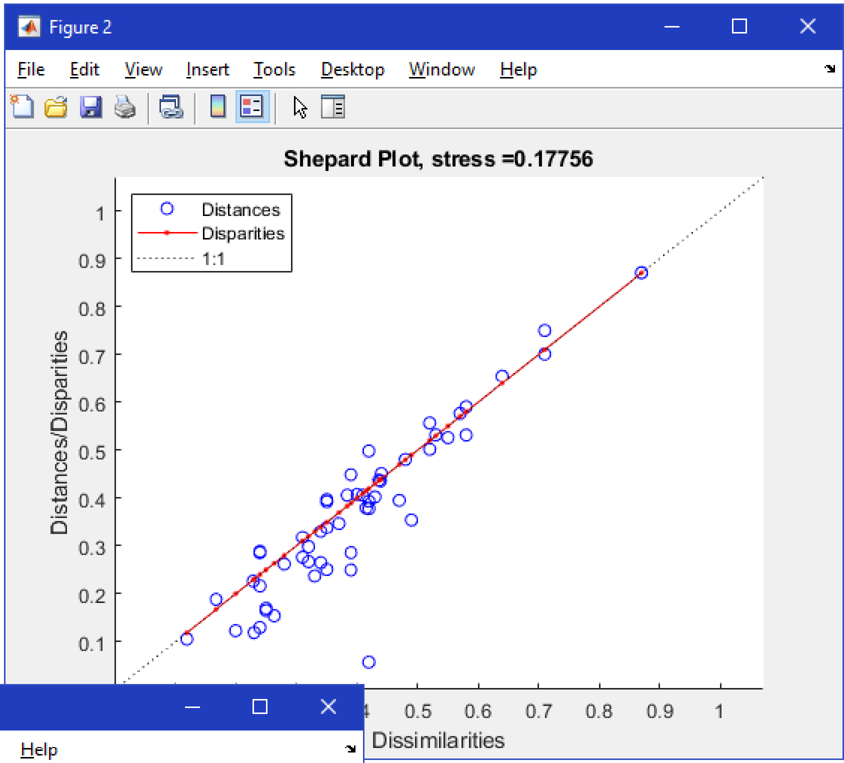
Metric MDS



Comparing Metrics

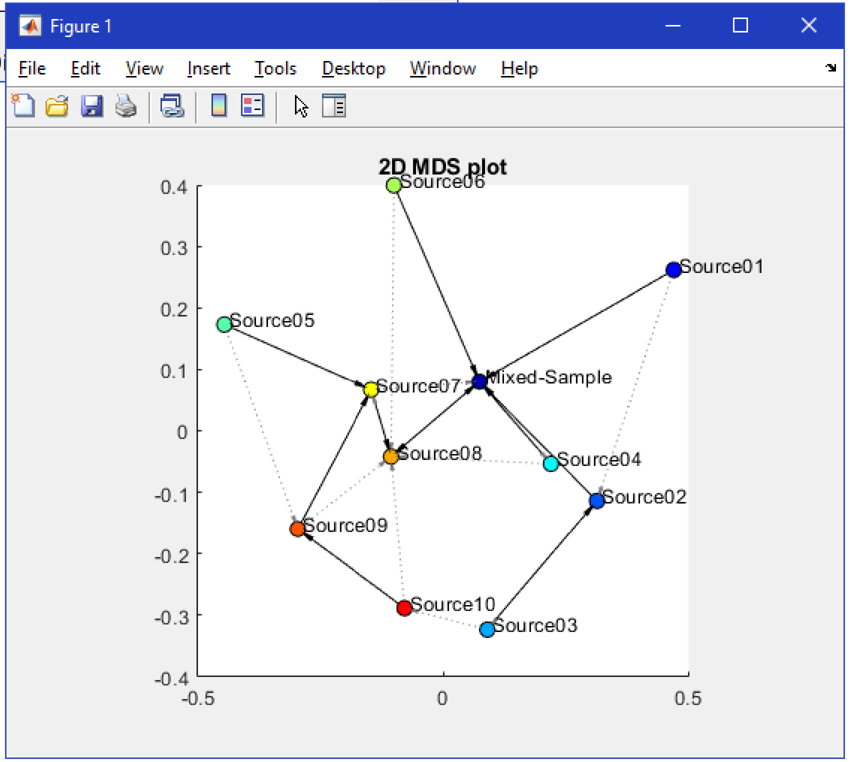
- Metric MDS

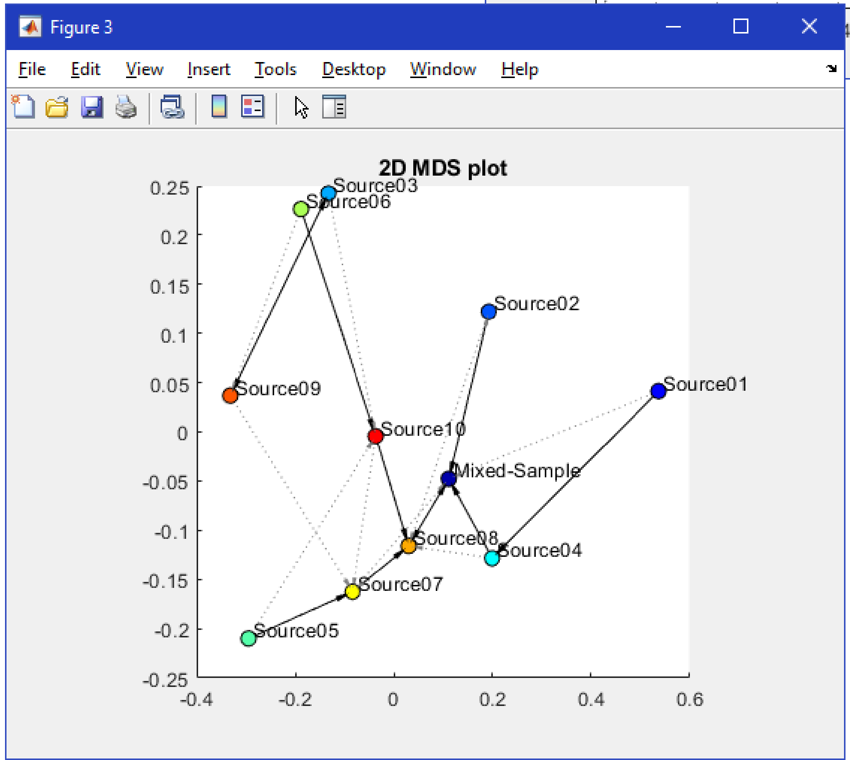




Kuiper V Value
Stress squared

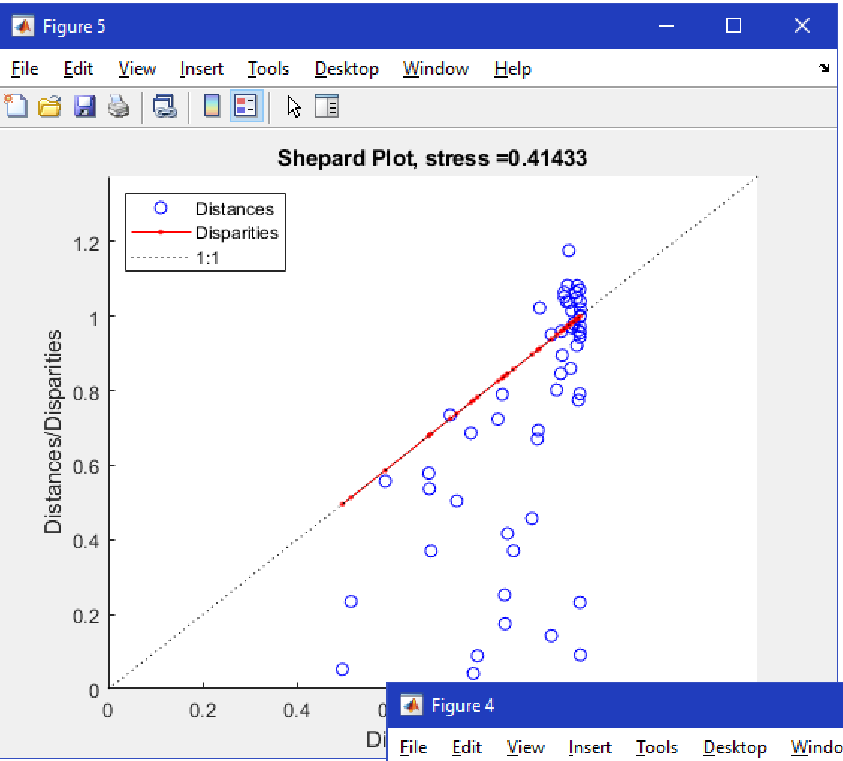
K-S D Value
Stress squared





Comparing Metrics

- Metric MDS



Cross-correlation
Stress squared

K-S D Value
Stress squared