Multidimensional Scaling

Today . . .

- Recap and more distinctions
- Expanded example

The nature of a particular MDS hinges on several features of the data:

- Are the distances at least interval in nature? Only ordinal?
- Are the distances measured directly (decompositional) or indirectly (compositional)?
- What distance measure is used (similarity, preference, attribute ratings, rankings, subjective, objective)?

The nature of a particular MDS hinges on several features of the data:

- Are the distances the average of several distance matrices (aggregate analysis)? Is the variation in those matrices important (disaggregate analysis)?
- What information is available to assist interpretation of dimensions?
- How are the objects selected?
- If indirect measurement is used, are the scales for different attributes similar?

The previous examples (city distances, car similarity) used direct assessment of similarity. The current example uses indirect assessment of similarity.

The data come from a 2017 C-SPAN survey (https://www.c-span.org/presidentsurvey2017/) of 91 historians who were asked to rank the presidents from George Washington through Barak Obama on 10 characteristics.

- Public persuasion (PP)
- Crisis leadership (CL)
- Economic management (EM)
- Moral authority (MA)
- International relations (IR)
- Administrative skills (AS)
- Relations with Congress (RC)
- Vision/Setting an agenda (VSA)
- Pursued equal justice for all (PEJ)
- Performance within context of times (PCT)

The composite ranks (across the 91 historians) are used in the analyses that follow.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Multidimensional_Scaling_II_R.Rnw* × | Presidents ×

Filter

	President	PP	CL	EM	MA	IR	AS	RC	VSA	PEJ	PCT
Washington	Washington	4	2	1	1	2	2	2	2	13	1
Adams	Adams	22	17	15	11	13	21	24	20	15	19
Jefferson	Jefferson	8	13	13	6	11	7	5	5	17	6
Madison	Madison	18	19	19	9	22	17	13	18	18	16
Monroe	Monroe	17	14	18	16	7	11	9	14	25	11
JQ_Adams	JQ_Adams	33	23	17	12	15	18	32	15	9	22
Jackson	Jackson	7	10	26	20	20	23	21	10	38	13
Van_Buren	Van_Buren	30	35	40	33	26	26	28	33	30	33
WH_Harrison	WH_Harrison	28	38	38	31	42	40	38	36	37	38
Tyler	Tyler	39	36	39	37	28	38	41	37	41	36
Polk	Polk	13	9	14	27	16	9	11	11	36	12
Taylor	Taylor	27	28	28	28	30	35	35	30	34	30
Fillmore	Fillmore	40	34	34	36	34	36	36	39	39	37
Pierce	Pierce	41	41	41	39	40	39	40	41	42	41
Buchanan	Buchanan	43	43	42	43	43	41	42	43	43	43
Lincoln	Lincoln	3	1	2	2	3	1	4	1	1	2
A_Johnson	A_Johnson	42	42	37	41	39	43	43	42	40	42
Grant	Grant	19	21	27	19	19	37	20	23	10	21
Hayes	Hayes	29	30	25	32	33	29	30	32	32	28
Garfield	Garfield	21	31	29	22	36	32	27	25	20	27

Showing 1 to 20 of 43 entries

Because the indirect method is being used, we must specify the distance measure to be calculated. Here we choose Euclidean distance, the most common choice and usually the default.

```
Presidents_Dist <- dist(Presidents[, 2:ncol(Presidents)], method = "euclidean",
    diag = TRUE)
Presidents_Dist

##              Washington   Adams Jefferson Madison  Monroe
## Washington        0.000
## Adams            49.629   0.000
## Jefferson        21.703  34.641     0.000
## Madison          46.098  16.733    27.129   0.000
## Monroe           37.014  25.080    20.616  21.048   0.000
## JQ_Adams         60.083  17.578    45.420  28.125  37.068
## Jackson          54.936  34.684    38.510  30.806  29.017
## Van_Buren        91.717  47.927    72.270  48.959  56.569
## WH_Harrison     107.620  62.490    88.312  63.474  74.740
## Tyler           108.849  62.921    90.405  66.821  73.851
## Polk             45.078  35.986    31.321  31.097  19.950
## Taylor           87.698  42.332    69.498  45.519  54.945
## Fillmore        106.353  60.581    87.966  63.403  71.868
## Pierce          118.966  73.212    99.910  75.260  84.350
## Buchanan        125.674  79.894   106.644  81.994  91.082
```

There are numerous distance measures that could be used. They depend on the kind of data that are collected (e.g., binary, ordinal, interval, count) and the features of the data that are allowed to have an influence (e.g., profile correlations, absolute distances). The boundaries between the types are often fuzzy (e.g., between ordinal and interval).

For ordinal (after normalization) and interval data, the most common choices are Euclidean, Euclidean^2, Manhattan, and Minkowski.

The general formula for Minkowski distance is:

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^{\lambda}}$$

When λ is 1, d_{ij} is Manhattan (or city block) distance. When λ is 2, d_{ij} is Euclidean distance.

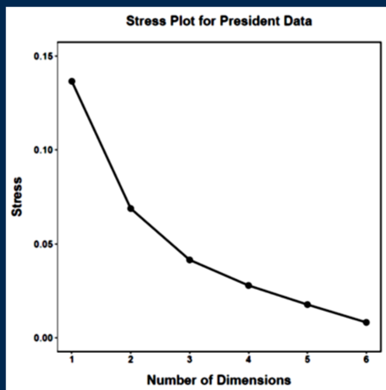
If the ratings have differences in scale (i.e., variances), then they should be standardized prior to calculating distance.

Metric MDS is similar to principal components analysis in that there is one best configuration in the chosen number of dimensions. This configuration will produce the lowest stress for that number of dimensions.

In non-metric MDS, the goal is to preserve the order of distances in the distance matrix but with a smaller number of dimensions. This approach is iterative, potentially dependent on the starting location.

Basic steps:

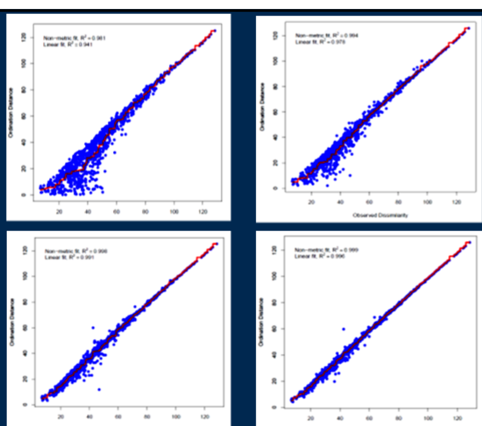
- Arrange the objects in a starting configuration.
- Calculate the distances among the objects (e.g., Euclidean metric).
- Regress the distances against the original distance matrix and get the predicted distances for each pair of objects.
- Goodness of fit is based on the sum of squared differences between ordination-based distances and the distances predicted by the regression (or stress or the rank-order correlation between ordination distances and original distances).
- Move the positions of objects in ordination space by a small amount and re-calculate goodness of fit.
- Stop when no further improvement (within tolerance) is found.



This stress plot is typical of rating-level data. A case could be made for 2 or 3 dimensions. Beyond that we don't get much simplification.

The isoMDS( ) function in the MASS package along with the stressplot( ) function in the vegan package can produce nice looking Shepard plots. These include linear and nonmetric fit indices.

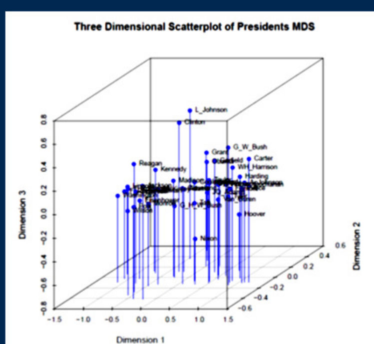
The linear fit index is the usual linear squared multiple correlation. The nonmetric fit index is 1-Stress^2.



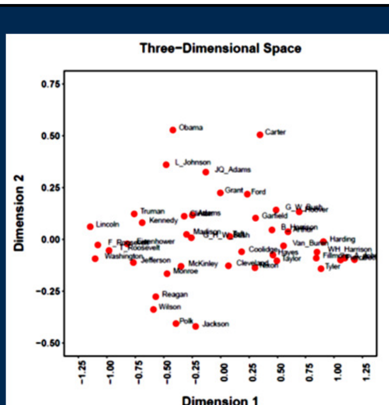
	PP	CL	EM	MA	IR	AS	RC	VSA
PP	1.0000	0.9186	0.8686	0.7448	0.7440	0.7159	0.8127	0.9281
CL	0.9186	1.0000	0.9002	0.8043	0.8706	0.7945	0.8408	0.9106
EM	0.8686	0.9002	1.0000	0.7545	0.7981	0.8163	0.7753	0.8807
MA	0.7448	0.8043	0.7545	1.0000	0.7419	0.7346	0.7085	0.8217
IR	0.7440	0.8706	0.7981	0.7419	1.0000	0.7596	0.7066	0.7735
AS	0.7159	0.7945	0.8163	0.7346	0.7596	1.0000	0.8025	0.7915
RC	0.8127	0.8408	0.7753	0.7085	0.7066	0.8025	1.0000	0.8043
VSA	0.9281	0.9106	0.8807	0.8217	0.7735	0.7915	0.8043	1.0000
PEJ	0.5512	0.5817	0.6616	0.6392	0.5159	0.5667	0.5435	0.6222
PCT	0.9230	0.9641	0.9166	0.8664	0.8443	0.8303	0.8774	0.9449

	PEJ	PCT
PP	0.5512	0.9230
CL	0.5817	0.9641
EM	0.6616	0.9166
MA	0.6392	0.8664
IR	0.5159	0.8443
AS	0.5667	0.8303
RC	0.5435	0.8774
VSA	0.6222	0.9449
PEJ	1.0000	0.6237
PCT	0.6237	1.0000

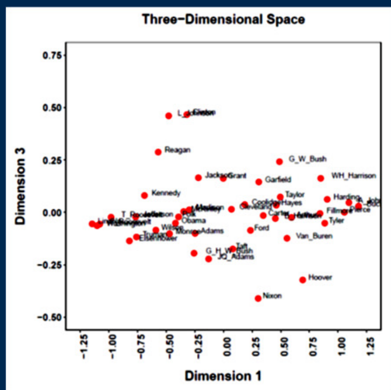
The rankings are very highly related, suggesting at least a general "greatness" dimension underlies the data.



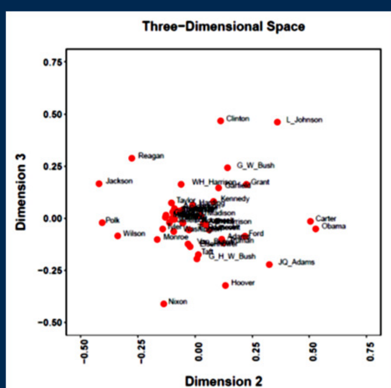
The three-dimensional plot identifies some of the more unusual presidents (Lyndon Johnson, Bill Clinton, Richard Nixon). Not particularly good for naming the dimensions.



The first dimension looks like "general greatness." The second dimension might be "justice."



The third dimension is harder to understand here. We'll gain some insights later.



The third dimension is harder to understand here. We'll gain some insights later.

The metaMDS( ) function in the vegan package can provide some other useful features. In particular it runs the analysis from multiple start points to find the best solution and insure that a local minimum has not been found.

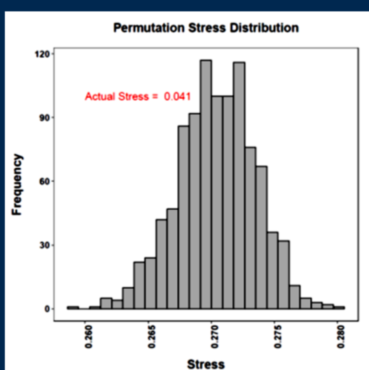
```
mds_4 <- metaMDS(Presidents_Dist, k = 3, distance = "euclidean", autotransform = FALSE,
    trymax = 100)

## Run 0 stress 0.04125
## Run 1 stress 0.04592
## Run 2 stress 0.0413
## ... Procrustes: rmse 0.01065  max resid 0.05273
## Run 3 stress 0.04245
## Run 4 stress 0.04484
## Run 5 stress 0.04782
## Run 6 stress 0.04245
## Run 7 stress 0.04645
## Run 8 stress 0.04615
## Run 9 stress 0.04436
```

The smacof package provides a permutation test that can be used to determine if the obtained stress value is different from what would be expected based on random data. It also provides a jackknife function to provide an indication of stability for the solution.

```
mds_5 <- smacofSym(Presidents_Dist, ndim = 3, verbose = FALSE, type = "ordinal",
    itmax = 1000)
```

```
perm_mds_5 <- permtest(mds_5, nrep = 1000, verbose = FALSE)
```

```
SMACOF Permutation Test
Number of objects: 43
Number of replications (permutations): 1000

Observed stress value: 0.041
p-value: <0.001
```



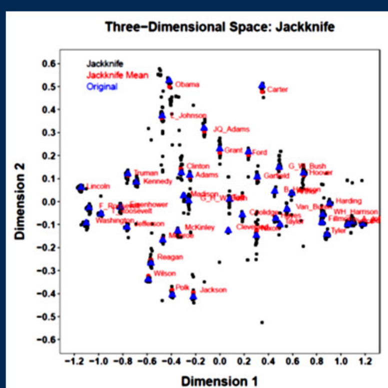
The obtained stress from a three-dimensional model is clearly unusual in the context of randomly re-arranged dissimilarities.

In the jackknife procedure (also called the leave-one-out procedure), each object is excluded from the analysis in turn and the remaining objects used to create the multidimensional space. The multiple solutions can then be plotted to determine how much variation exists, indicating how much the solution hinges on particular objects.

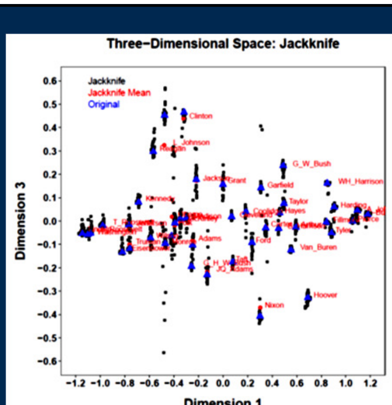
```
jackfit <- jackknife(mds_5, itmax = 1000)
jackfit

##
## Call: jackknife.smacofB(object = mds_5, itmax = 1000)
##
## SMACOF Jackknife
## Number of objects: 43
## Value loss function: 14.1
## Number of iterations: 3
##
## Stability measure: 0.9921
## Cross validity: 0.9997
## Dispersion: 0.0082
```

The stability measure indicates the between-object variability relative to total variability (akin to a variance accounted for estimate or intraclass correlation). The cross validity indicates how well the average location (centroid) from the jackknife samples matches the actual location. The dispersion estimates variability around the actual location (and = 2-[stability + cross validity]).

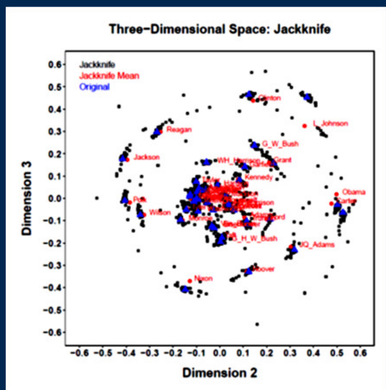
```
SMACOF Jackknife
Number of objects: 43
Value loss function: 14.1
Number of iterations: 3

Stability measure: 0.9921
Cross validity: 0.9997
Dispersion: 0.0082
```



Most of the instability occurs for the second dimension.



Here too there is more instability for the third dimension.

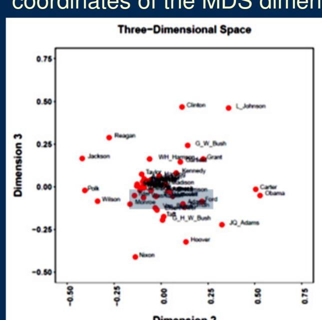


Uncovering the meaning of the dimensions can be assisted by correlating the original ratings with the coordinates of the MDS dimensions.

	D1	D2	D3
PP	0.9223566	0.141069	-0.2263773
CL	0.9633115	0.136533	-0.0106613
EM	0.9388875	-0.019069	-0.0532063
MA	0.8715320	-0.150442	0.1486345
IR	0.8657271	0.180475	0.3398971
AS	0.8708968	0.008660	0.2513751
RC	0.8803035	0.127574	-0.1577770
VSA	0.9508638	0.035587	-0.1019317
PEJ	0.6898128	-0.677203	-0.0473194
PCT	0.9860184	0.067788	-0.0545041
D1	1.0000000	0.001394	0.0009921
D2	0.0013940	1.000000	0.0430871
D3	0.0009921	0.043087	1.0000000

The stability of the first dimension is not surprising given that it contains information from all of the ranking scales.

Uncovering the meaning of the dimensions can be assisted by correlating the original ratings with the coordinates of the MDS dimensions.



	D2	D3
PP	0.141069	-0.2263773
CL	0.136533	-0.0106613
EM	-0.019069	-0.0532063
MA	-0.150442	0.1486345
IR	0.180475	0.3398971
AS	0.008660	0.2513751
RC	0.127574	-0.1577770
VSA	0.035587	-0.1019317
PEJ	-0.677203	-0.0473194
PCT	0.067788	-0.0545041
D1	0.001394	0.0009921
D2	1.000000	0.0430871
D3	0.043087	1.0000000

Next time . . .

Individual differences