Multidimensional Scaling

Today . . .

- The basic goal of MDS
- Simple examples to introduce some concepts

The purpose of multidimensional scaling (MDS) is to map the distances between points in a high dimensional space into a lower dimensional space without too much loss of information.

Like principal components analysis and factor analysis, we use MDS to simplify a data set in order to understand its fundamental features.

The "points" that are represented in multidimensional space can be just about anything. These objects might be people, in which case MDS can identify clusters of people who are "close" versus "distant" in some real or psychological sense.

The objects might be physical objects (e.g., cities, cars, books) or psychological or cultural objects (e.g., personality traits, religions, political parties).

The only requirement is that some basis exists for rating or ranking the objects in terms of <u>similarity</u>.

As long as the "distance" between the objects can be assessed <u>in some fashion</u>, MDS can be used to find the lowest dimensional space that still adequately captures the distances between objects.

Once the number of dimensions is identified, a bigger challenge usually is identifying the meaning of those dimensions. That usually will rely on additional information about the objects.

The basic data representation in a standard MDS is a <u>dissimilarity matrix</u> that shows the distance between every possible pair of objects.

The goal of MDS is to faithfully represent these distances with the lowest possible dimensional space.

Like principal components, we seek a reference system that can capture the basic information in a data set—in this case a data set characterized by distances between objects.

MDS comes in several varieties. 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 or indirectly?
- What distance measure is used?
- Are the distances the average of several distance matrices? Is the variation in those matrices important?
- What information is available to assist interpretation of dimensions?

An intuitive understanding of MDS can most easily be gained by beginning with an application where the dimensionality is known, the objects easily understood, and the notion of distance easy to grasp: physical distances between cities.

This problem can be solved using ***metric multidimensional scaling*** because the data are at least interval level.

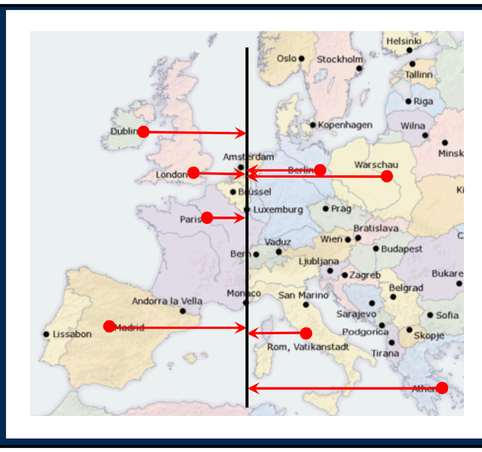
This table lists the distances between eight European cities. A multidimensional scaling of these data should be able to recover the two dimensions (North-South x East-West) that we know must underlie the spatial relations among the cities.

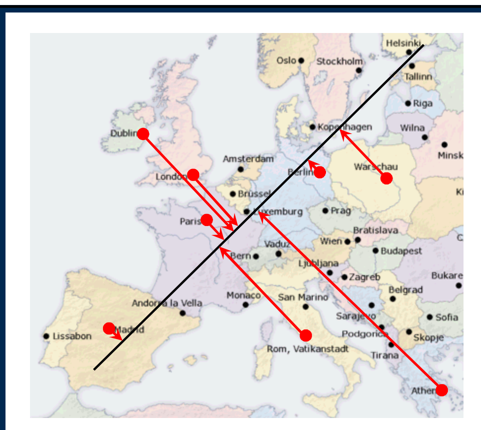
	Athens	Berlin	Dublin	London	Madrid	Paris	Rome	Warsaw
Athens	0	1119	1777	1486	1475	1303	646	1013
Berlin	1119	0	817	577	1159	545	736	327
Dublin	1777	817	0	291	906	489	1182	1135
London	1486	577	291	0	783	213	897	904
Madrid	1475	1159	906	783	0	652	856	1483
Paris	1303	545	489	213	652	0	694	859
Rome	646	736	1182	897	856	694	0	839
Warsaw	1013	327	1135	904	1483	859	839	0

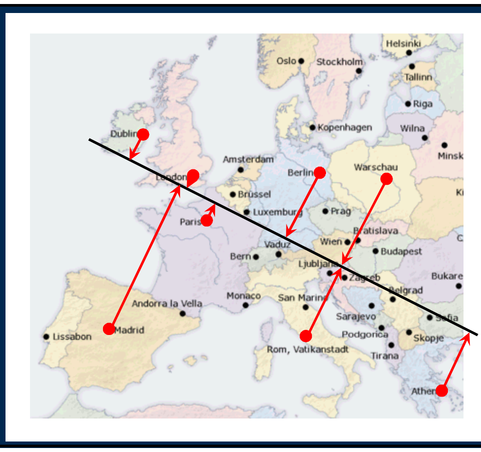
MDS begins by restricting the dimension of the space and then seeking an arrangement of the objects in that restricted space that minimizes the difference between the distances in that space compared to the actual distances. Sound familiar?

Imagine being restricted to a single dimension: a line with miles as the metric. The cities can be placed on that line in any order. Their positions on the line then dictate their distances from each other in one-dimensional space.

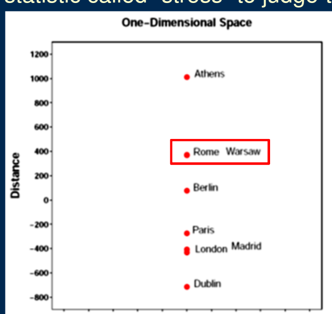
Or, imagine placing the line on a map and trying to find a location where projections of cities to the line best approximates their distances from each other.







This arrangement provides the best fit for a one-dimensional model. How good is the fit? We use a statistic called "stress" to judge the goodness-of-fit.



We don't really need it here. We know, for example, that Rome and Warsaw are not such near neighbors.

$$Stress = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(d_{ij} - \delta_{ij})^2}{\sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij}^2}}$$

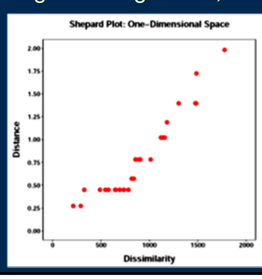
where d_{ij} is the distance in one-dimensional space between two cities and δ_{ij} is their actual distance.

In MDS, as in life, high stress is bad.

Smaller stress values indicate better fit. Some rules of thumb for degree of fit are:

Stress	***Fit***
.20	Poor
.10	Fair
.05	Good
.02	Excellent

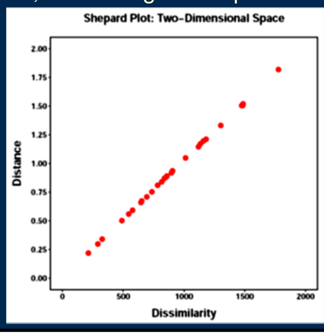
The stress for the one-dimensional model of the cities data is .31, clearly a poor fit. The poor fit can also be seen in a Shepard plot of the actual distances (called dissimilarity or disparity) versus the distances in the one-dimensional model. In a good fitting model, the points will lie along a line, sloping upward to the right, showing a one-to-one correspondence between distances in the model space and actual distances. Clearly not evident here.



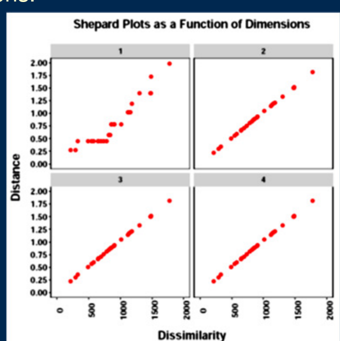
The analyses proceed by trying higher dimensional models, seeking the minimum number of dimensions necessary to get an acceptable fit.

Examination of stress for each model and the Shepard plots can sometimes identify a clear lower dimensional solution.

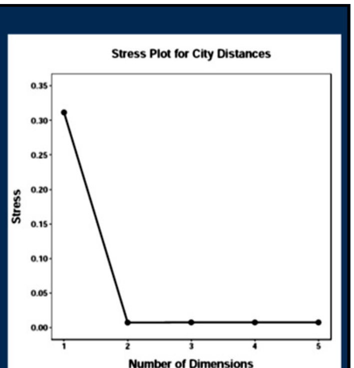
A two-dimensional model works very well. The stress value is quite small (.0072) indicating an exceptional fit. Of course, this is no great surprise for these data.



No room for improvement by including more than two dimensions.



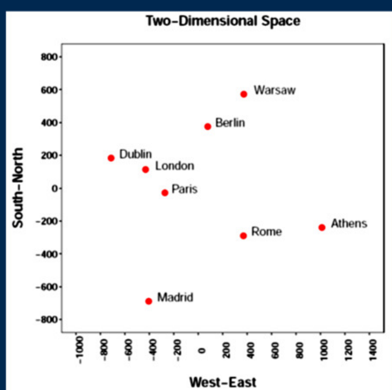
The stress values can be plotted and the “appropriate” number of dimensions determined in much the same way as in the scree plot, although here we choose the point at which stress no longer declines noticeably.

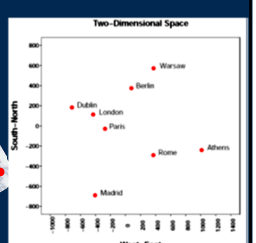


Once the appropriate number of dimensions are identified, the objects can be plotted in the multidimensional space to determine what objects cluster together. Why objects might cluster together often requires additional information.

In the cities data, the reason is quite clear. The dimensions should refer to the North-South x East-West surface area across which the cities are dispersed.

We would expect MDS to faithfully recreate the map relations among the cities.

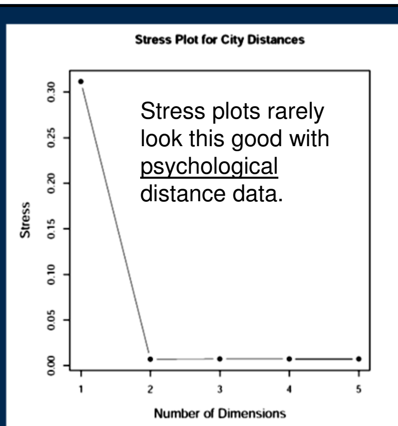




Multidimensional coordinates for a four-dimensional solution.

	D1	D2	D3	D4
Athens	1.03277	-0.24182	-0.017088	0.00135945
Berlin	0.08354	0.39023	-0.076468	-0.00061133
Dublin	-0.73174	0.18724	0.006095	-0.00937435
London	-0.44127	0.11654	0.006192	0.00614096
Madrid	-0.41294	-0.69342	-0.026475	0.00006271
Paris	-0.27868	-0.03252	0.032850	0.00967495
Rome	0.37723	-0.29783	0.045115	-0.00699100
Warsaw	0.37109	0.57158	0.029780	-0.00026139

Including third and fourth dimensions clearly does not add anything.



It is rare in the social sciences to have metric data on which to compare the objects. More typically, the “distances” are on psychological dimensions for which ordinal relations are the best scaling possible.

This creates the possibility of arriving at “distances” in quite a number of ways. Generally they can be classified into two major types:

- Direct Methods
- Indirect Methods

Direct methods require each pair of objects to be judged according to their similarity. The basis for the judgments can be either left up to the judge or provided as part of the instructions.

The MDS is more likely to uncover the underlying nature of the similarity judgments if the basis for them is <u>not constrained</u>. On the other hand, the structure may be clear but its origins uncertain.

Direct methods can be very time consuming because they require the rating of each and every pair of objects.

Indirect methods derive the distances between objects from other information provided by the judge. For example, the objects might each be rated on a small collection of rating scales. From these ratings, distances can be derived.

These methods are easier for the judge, but they are one step removed from a true similarity judgment. Objects are not being compared directly by the judge.

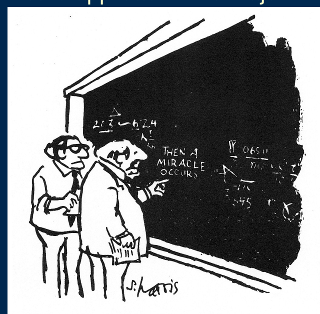
Also, the basis for the similarity calculations resides entirely in the ratings that were collected. There is no guarantee that a direct judgment would produce the same similarity configuration.

A single judge was asked to take each pair of 10 cars and rank order them according to their similarity. The rank of 1 was assigned to the most similar pair; the rank of 45 was assigned to the least similar pair.

	BMW	Ford	Infiniti	Jeep	Lexus	Chrysler	Mercedes	Saab	Porsche	Volvo
BMW	0									
Ford	34	0								
Infiniti	8	24	0							
Jeep	31	2	25	0						
Lexus	7	26	1	27	0					
Chrysler	43	14	35	15	37	0				
Mercedes	3	28	5	29	4	42	0			
Saab	10	18	20	17	13	36	19	0		
Porsche	6	39	41	38	40	45	32	21	0	
Volvo	33	11	22	12	23	9	30	16	44	0

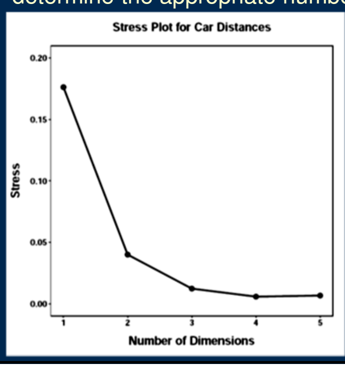
The general approach to non-interval data such as these is similar to the metric approach. The major difference is that calculating “distances” requires some additional assumptions and transformations.

Leaving aside those details for now . . .



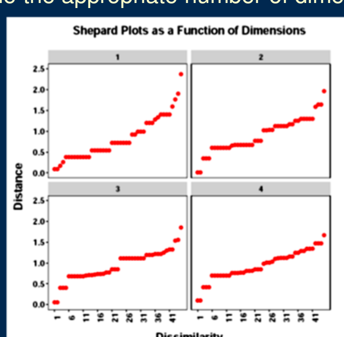
"I think you should be more explicit here in step two."

The Shepard plots and the plot of stress values determine the appropriate number of dimensions.

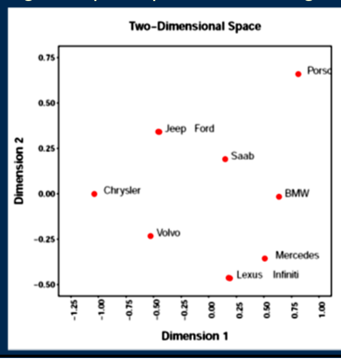


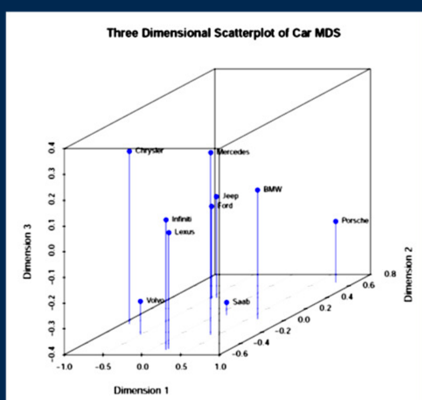
At least two, maybe three.

The Shepard plots and the plot of stress values determine the appropriate number of dimensions.

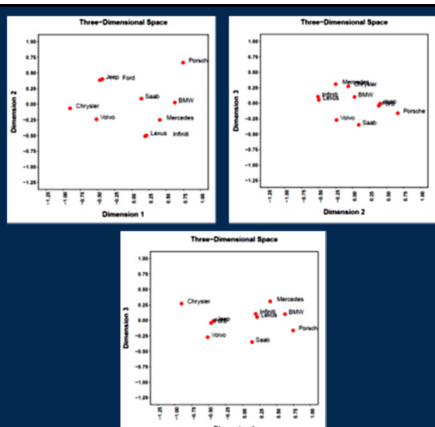


The objects can be plotted in the dimensional space to determine clustering and, perhaps, the meaning of the dimensions. What additional information would be helpful? What other approaches to similarity could be taken?





Do three dimensions help?



Suppose we wanted to know how people view the psychological distances between countries? Presidents? Ethnic groups?

What makes them similar or different?

How might we build a dissimilarity matrix?

What kind of direct and indirect approaches might be taken?

Next time . . .

- Alternative MDS approaches
- MDS details