Module 9 Lecture 1 Transcript

Slide 1: [Title Slide]

Slide 2: Ordination techniques are a broad group of analysis tools that are used to represent high-dimensional multivariate data in lower dimensions. Most commonly, this is done for the purpose of visualizing high-dimensional data.

The ordination technique that we will focus on is called ‘multidimensional scaling’. There are other techniques (such as principal components analysis, or PCA) that can also be used for the same purpose of representing high-dimensional data with just a few dimensions.

Slide 3: Multidimensional scaling (or MDS) is also sometimes referred to as ‘principal coordinate analysis’. It is an ordination method that produces a low-dimensional representation (typically two or three dimensional) of data such that the pairwise distances in the representation are close to the pairwise distances in the original data.

Sometimes MDS is performed using just an ordering of the pairwise distances rather than the distances themselves. That is, we might only know that the distance between A and B is larger than the distance between B and C, which is larger than the distance between C and A. In this setting where only the ordering of distances is known, the method is referred to as ‘Non-metric multidimensional scaling’ (NMS or NMDS).

If the actual distances are used, the method is called metric multidimensional scaling.

Slide 4: The setting for MDS is the same as the setting for many of the methods we have covered in this class. We have a sample of n observation units, where for each unit we have measured p different variables. We let bold $\mathbf{X}_i$ denote the vector of values for the ith experimental unit. $_i$

Truly, to perform MDS all we need is the (n x n) pairwise distance matrix that gives us the pairwise distances d_{ij} between observations $\mathbf{X}_i$ and $\mathbf{X}_j$.

Our goal is to find a set of n points $\mathbf{Z}_1$, $\mathbf{Z}_2$, …, $\mathbf{Z}_n$ in q dimensions (where q is typically very small, like 2 or 3) such that $\mathbf{Z}_i$ represents $\mathbf{X}_i$, and the pairwise distances between the **Z**s are as close as possible to the pairwise distances between the **X**s.

Slide 5: Consider the toy example presented in the plot below: we have $n = 4$ observations in $p = 2$ dimensions.

Slide 6: We can compute the Euclidean distance matrix for the four points shown in the plot on the previous slide. The resulting distance matrix is shown here.

Note that points 2 and 4 are the furthest apart, with a distance of 2.14, and points 2 and 3 are closest together, with a distance of 1.12.

Slide 7: Using multidimensional scaling, we can find a representation of these points in ***one*** dimension such that the pairwise distance between each pair of points is preserved as much as possible. The resulting solution is shown on the number-line below.

Note that in this one-dimensional solution, the representations of points 2 and 3 are quite close to each other, and the representations of points 2 and 4 are the furthest from each other.

Slide 8: We can calculate the Euclidean distance matrix based on the 1-dimensional representation of the points by using the coordinates for each point in the 1-dimensional representation.

For instance, the representation of point 2 was located at roughly the value -0.95, and the representation of point 3 was located at roughly the value -0.7. The distance between these two representations is therefore about 0.25.

Slide 9: We can then compare the original distance matrix to the distance matrix we get from the MDS solution. We notice that some of the distances are preserved quite well—for instance, the distance between points 1 and 2 is 1.8 in the original matrix, and is 1.79 in the MDS solution.

However, other distances are more distorted: the distance between points 1 and 4 was 1.25, but the distance between the MDS representations of points 1 and 4 is just 0.06.

Such discrepancies are generally unavoidable when we try to represent a higher-dimensional space in just a few dimensions.

Slide 10: Here we illustrate the point pairs whose distances are least-well-represented in the MDS solution.

Again, note that the distances for the MDS solution come from the 1-dimensional representations of the points.

Slide 11: Another, more realistic (and I think pretty cool!) example of the application of multidimensional scaling takes the straight-line distances between a collection of 11 cities in the US, and uses these distances to choose a two-dimensional representation of how these cities are arranged in space.

Below is the pairwise distance matrix for these 11 cities.

Slide 12: The map below shows the true locations of these 11 cities, superimposed on a map of the United States.

Slide 13: The coordinates that we get for the representation of these cities using multidimensional scaling are shown as green dots on the map of the US.

I did have to rotate the resulting coordinates a bit to find the best orientation to match the true US map, but otherwise MDS did all the work in reconstructing the map based solely on the pairwise distances.

Slide 14: There are several different solutions to the MDS problem, but here we will consider just the classical solution as one example of a typically MDS algorithm.

This solution is found by first constructing centered, negative squared distances (the b_{ij} values shown in step 2).

Slides 15 & 16: Then we find the eigendecomposition of the matrix **B** consisting of these centered negative squared pairwise distances.

The first *q* eigenvectors of **B**, each multiplied by the square-root of its corresponding eigenvalue, provide the q-dimensional coordinates for the MDS solution—that is, the coordinates in q-dimensions for points that best reconstruct the distances in the given distance matrix.

Slide 17: Note that if we are using Euclidean distance, the classical MDS solution is exactly the scores for the first q principal components.

Slide 18: Whereas metric MDS tries to preserve the values of the pairwise distances, non-metric MDS just tries to produce a solution where the pairwise distances are in roughly the same order for the low-dimensional solution as they are in the original data.

Slide 19: The non-metric MDS solution is obtained using an iterative algorithm, where we start by randomly placing *n* points in q-dimensional space. Then we calculate the distances between those *n* points in the q-dimensional representation.

Slide 20: Next we assess how far the ordering of the pairwise distances in the q-dimensional space is from the ordering in the original space. This is called the 'stress' of the current solution; higher stress means that the current solution does *not* do a good job of reproducing the ordering of distances.

After assessing the stress, we move the points in the q-dimensional space to try to improve the fit.

These steps are repeated until we cannot improve the solution any further.

Slide 21: Non-metric MDS is used frequently in data where the observations might be qualitative rather than quantitative, or where we have a dissimilarity matrix rather than a true distance matrix. Such occasions arise, for instance, from ordered categorical data such as might be obtained from Likert scales or user evaluations.