Chapter 5: Multidimensional Scaling and Correspondence Analysis

- Recall that we used *distances* to measure how different multivariate observations were from each other.
- In Chapter 1, we took a multivariate data set (a set of q-dimensional vectors) and calculated distances between pairs of vectors.
- Both *multidimensional scaling* and *correspondence analysis* are techniques related to distances.
- *Multidimensional Scaling* can be viewed as a way of generating a geometric representation of some observed *proximity matrix*.
- The proximity matrix could contain *similarity* values for pairs of observations or *dissimilarity* values, but we will typically work with dissimilarities (i.e., distances).
- With *multidimensional scaling*, we begin with a distance matrix and produce a "possible data set" that could have yielded such a distance matrix.

Classical Multidimensional Scaling (MDS)

- Given a $n \times n$ distance matrix, the goal is to construct a "map" (geometrical model) containing multivariate points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.
- Each point represents one of the individuals in the original data set.
- Two goals in determining this map:
  1. What is an appropriate dimension q for the points on the map?
  2. Where should the points be placed on the map in order to "fit" the observed distances well?

Multidimensional Scaling Example

- A subject was asked to taste 10 colas, and, for each pair of colas, to rate how different the two colas were, on a scale of 0 to 100.
- A "dissimilarity" of 0 would mean the two colas tasted exactly the same, and a dissimilarity of 100 would mean the two colas tasted completely different.
- A 10×10 distance matrix can be constructed based on the subject's judgments.
- Multidimensional scaling allows us to represent the 10 colas as points in a q-dimensional space and visually examine the similarities and differences among the colas.

Nonuniqueness of the Coordinates

- Note that there is no uniquely best solution for where to place the points on the map.
- If we have a set of coordinates that "best" fits the distances, we can get an equally good set of coordinates by shifting, rotating, or reflecting the points in the q-dimensional space.
- Partial Solution: We can constrain the solution so that the mean vector of the points lies at the origin $(0, 0, \ldots, 0)'$.
- We can then choose the rotation (via some orthogonal transformation) of the points so that the solution is most easily interpreted.

Mathematics behind Classical MDS

- First assume our distance matrix $\mathbf{D}$ contains Euclidean distances derived from an (unknown) data matrix $\mathbf{X}$.
- Define the $n \times n$ matrix $\mathbf{B} = \mathbf{X}\mathbf{X}'$.
- The squared Euclidean distances d_{ij}^2 between the rows of $\mathbf{X}$ can be written in terms of the elements of $\mathbf{B}$:

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}.$$

- Constrain $\bar{\boldsymbol{x}}$ to be the zero vector; then summing over i, over j, and over i and j gives a series of equations with which we can solve for the b_{ij} values in terms of the d_{ij}^2 values:

$$b_{ij} = -0.5\Big[d_{ij}^2 - n^{-1}\sum_j d_{ij}^2 - n^{-1}\sum_i d_{ij}^2 + n^{-2}\sum_i\sum_j d_{ij}^2\Big].$$

- Since we know the distances d_{ij}, we can write the whole matrix $\mathbf{B}$.

Mathematics behind Classical MDS (Continued)

- Now we must factor $\mathbf{B}$ to obtain the matrix of coordinate values $\mathbf{X}$.
- We can use the singular value decomposition $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{B}$, and the columns of $\mathbf{V}$ are the orthonormal eigenvectors of $\mathbf{B}$.
- If the original $\mathbf{X}$ is $n \times q$ and of full rank, then the last $n - q$ of the eigenvalues of $\mathbf{B}$ are zero.
- Hence $\mathbf{B} = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1'$, where $\mathbf{\Lambda}_1$ and $\mathbf{V}_1$ contain the *nonzero* eigenvalues and the corresponding eigenvectors.
- Then let $\mathbf{X} = \mathbf{V}_1\mathbf{\Lambda}_1^{1/2}$, so that $\mathbf{B} = \mathbf{X}\mathbf{X}'$ as needed.

Some Practical Considerations with Classical MDS

- We can reduce the dimension of the solution by restricting attention to the k largest eigenvalues.
- If the distances are not Euclidean, $\mathbf{B}$ is not positive definite and some eigenvalues of $\mathbf{B}$ will be negative.
- In this case, we can still choose the dimension corresponding to the k largest *positive* eigenvalues.
- See the *trace criterion* or *magnitude criterion* suggested by Sibson (1979) on page 109-110 of the book.

MDS on Euclidean Distances between Multivariate Observations

- In some situations, the distance matrix may be obtained by calculating Euclidean distances between observed q-variate observations.
- Question: If we already had the data set, why use MDS to create an "artificial data set" that reflects the distance structure?
- Perhaps the original number of variables is large, and we want our "map" to be a lower-dimensional representation of the data.
- The map would consist of k-dimensional points, with $k < q$ — the goal is dimension reduction, similar to PCA.
- In fact, when the distances in MDS are Euclidean distances derived from a data matrix, the coordinates of the MDS solution equal the PC scores from using PCA on $\mathbf{S}$.
- This usage of MDS is sometimes called *principal coordinates analysis*.

Determining the Amount of Data Reduction

- When using MDS to "reduce the dimensionality" from q to k, what is a proper choice of k?
- If there are k "relatively large" eigenvalues of $\mathbf{B}$, this is evidence that a k-dimensional solution is appropriate.
- We could base the choice of k on the sizes of the first few eigenvalues $\lambda_1, \lambda_2 \ldots$ (listed in decreasing order).
- We could calculate (for each possible k):

$$P_k = \frac{\sum_{i=1}^{k} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|}$$

  (or a similar measure with the absolute values replaced by squares).
- Values of k that yield a P_k near 1 (say, at least 0.8) would give a good representation.
- The `cmdscale` function in `R` prints this criterion when the `eig=T` option is specified.

Other Methods of Determining k

- Another option: For each possible value of k, try to minimize

$$\phi = \sum_{r,s}(d_{rs}^2 - \hat{d}_{rs}^2),$$

  where d_{rs}^2 is the Euclidean distance between the r-th and s-th observations in the (full) q-dimensional space, and $\hat{d}_{rs}^2$ is the Euclidean distance between the r-th and s-th observations in the (reduced) k-dimensional space.

- As k increases, the minimum value of ϕ will decrease monotonically.

- We can plot this minimum against the various values of k and pick the k value at the "elbow" of the plot.

- Takane et al. (1977) suggested a scaled version of ϕ called *SStress* that always lies between 0 and 1:

$$SStress = \left[\frac{\sum_{r<s}(d_{rs}^2 - \hat{d}_{rs}^2)^2}{\sum_{r<s} d_{rs}^4}\right]^{1/2}$$

- Values of *SStress* below 0.1 represent a good fit.

Nonmetric Multidimensional Scaling

- Sometimes we may not be able to assign precise numerical dissimilarities to pairs of observations, but we could *rank* the pairs of observations in terms of how dissimilar they are.
- Or we may not trust the exact numerical dissimilarities, but we believe basically in their ordering.
- *Nonmetric multidimensional scaling* (or isometric multidimensional scaling) uses only the rank orders of the distances to arrive at an MDS solution.
- The `R` function `isoMDS` in the `MASS` package performs nonmetric scaling.

Correspondence Analysis

- A two-way *contingency table* presents sample values for two *categorical variables*.
- Commonly, we test whether the two classifications are *independent* or *dependent* using a chi-squared test.
- *Correspondence Analysis* (CA) can be used to supplement such a chi-squared test.
- It presents the categorical data graphically, based on a decomposition of the chi-squared test statistic using *chi-squared distances*.
- The two categorical variables are often called the *row variable* and the *column variable*, based on their placement in the contingency table.
- Correspondence analysis finds and plots coordinates that represent the categories of both the row and column variables.
- We can then interpret the pattern of association based on the plot.

Contingency Table Notation

- Suppose the counts in a $r \times c$ contingency table are represented as follows:

	1	2	$\cdots$	c	Row Totals
1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
Column Totals	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	N

Chi-Squared Distances

- We can then define a $r \times c$ table of column proportions, with entries $p_{ij}^{(col)}, i = 1, \ldots, r, j = 1, \ldots, c$, where

$$p_{ij}^{(col)} = n_{ij}/n_{i\cdot}$$

- We can also define a $r \times c$ table of row proportions, with entries $p_{ij}^{(row)}, i = 1, \ldots, r, j = 1, \ldots, c$, where

$$p_{ij}^{(row)} = n_{ij}/n_{\cdot j}$$

- The chi-squared distance between two columns is a weighted Euclidean distance (with the rarer column categories weighted more heavily).
- The chi-squared distance between two rows is a weighted Euclidean distance (with the rarer row categories weighted more heavily).
- Page 105 gives formulas for the *squared* chi-squared distances between rows and between columns.

Plotting Coordinates

- We perform a classical MDS on the distance matrix for columns and a classical MDS on the distance matrix for rows.
- We plot the first two coordinates for column categories and the first two coordinates for row categories on the same axis.
- Each point should be labeled according to its category, for ease of interpretation.
- Note: In correspondence analysis, an exact representation of the chi-squared distances in K-dimensional space is possible, where $K = \min(r-1, c-1)$.
- Thus if *both* the number of rows and the number of columns are *greater than* 3, an exact 2-D representation is not possible.
- The 2-dimensional representation in that case is only an approximation.
- We could check the fit of the 2-dimensional representation using measures such as P_k or $SStress$.

Interpreting the Correspondence Analysis

- In a 2-dimensional plot, all row categories and all column categories are labeled on the plot.
- Two row categories that are near each other on the plot would have similar conditional distributions across the columns.
- Two column categories that are close together on the plot would have similar profiles down the rows.
- A row category and a column category that are close together on the plot would tend to appear together more often than would be expected under independence.

Interpreting the Correspondence Analysis (Continued)

- A one-dimensional solution (in which each category has a *single* coordinate value) can often yield useful interpretations as well.
- When a row category and a column category have coordinates that are large in magnitude and have the same sign, this row-column combination tends to appear *more often* than would be expected under independence.
- When a row category and a column category have coordinates that are large in magnitude and have different signs, this row-column combination tends to appear *less often* than would be expected under independence.
- When a row category and a column category have coordinates whose product is near zero, this row-column combination tends to appear about as often as would be expected under independence.