Multidimensional Scaling (MDS)

Objective: Produce a low dimensional display of high dimensional data that reflects similarities (or dissimilarities) among individuals (or items).

- Display similarities among pasta sauces using sensory judgments from human raters

- Display similarities among potential customers with respect to preferences for various products

- Display similarities among legislators with respect to voting records

- Display distances between storage facilities using lengths of transportation routes

Multidimensional Scaling

MDS is a collection of different algorithms, designed to arrive at optimal low-dimensional (usually 2 or 3 dimensions) representation of the data, whose inter-point distances (or *dissimilarities*) are as close as possible to that in the original space.

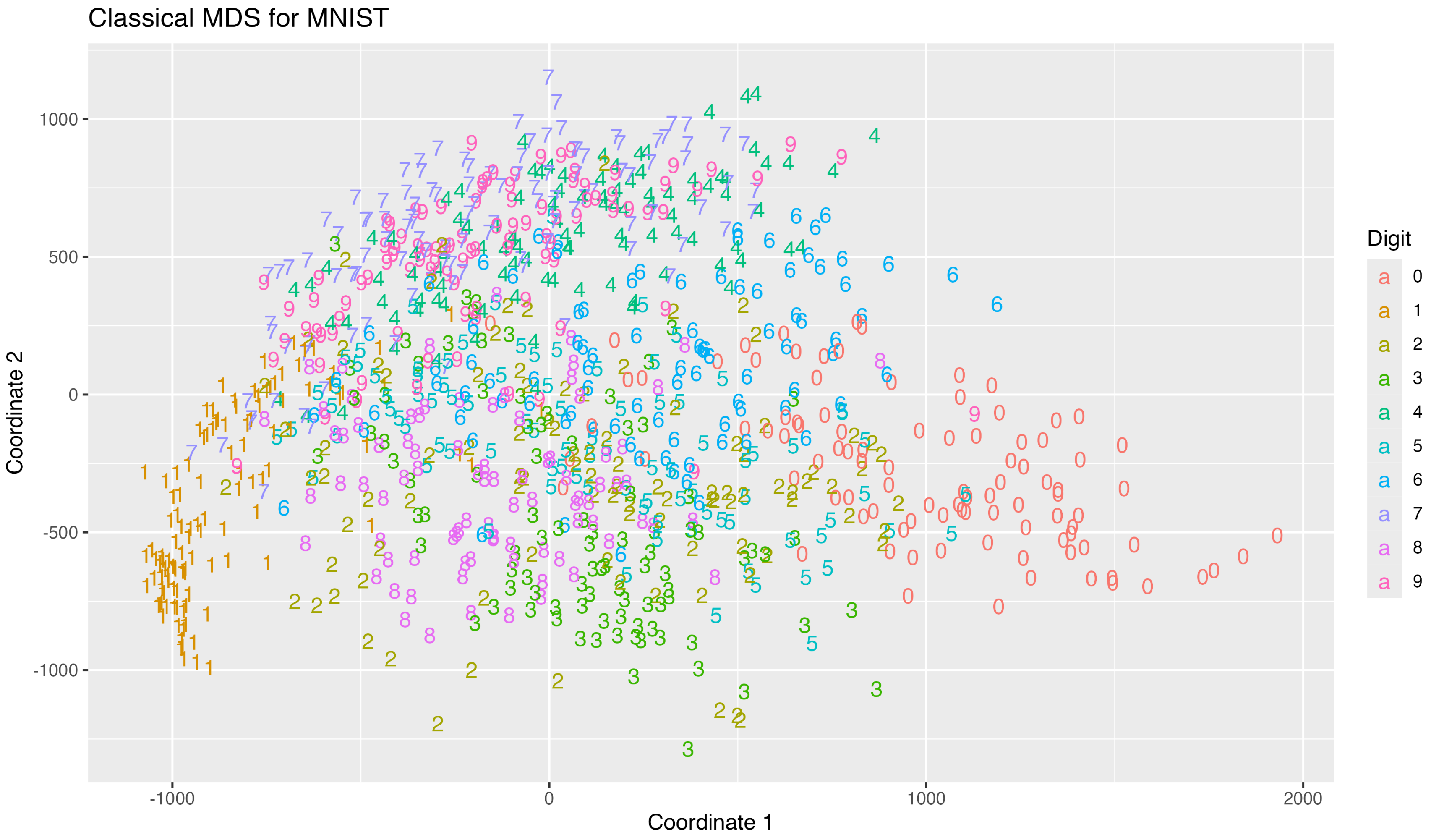
- Given a $n \times p$ data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]'$. (This is often not required for MDS!)
- Let $D = (d_{ij})_{i,j=1,\ldots,p}$ be a $p \times p$ *distance* matrix (for $\mathbf{X}$).
- MSD aims to find $m\,(< p)$ and a set of m-dimensional points $\mathbf{y}_1, \ldots, \mathbf{y}_n$ in the Euclidean space $\mathbb{R}^m$ such that

$$\text{distance}(\mathbf{y}_i, \mathbf{y}_j) \approx d_{ij} \text{ as close as possible.}$$

Motivating Example: MNIST Data

intensity
250
200
150
100
50
0

- MNIST consists of 60,000 handwritten images with $28\times28(=784)$ pixels.
- Every MNIST image can be thought of as an observation that is an array of numbers describing how dark each pixel is, with values between 0 and 255 representing grey scale.
- Can we project the images into 2D space such that we preserve their “similarity/dissimilarity”?



Classical MDS for MNIST

Multidimensional Scaling

- Multidimensional scaling can be applied to matrices of Euclidean distances between individuals, but it also be applied to more general matrices of similarity (or dissimilarity) measures.

- Similarity matrices are also called *proximity* matrices for pairs of individuals when more similar individuals are represented by a smaller proximity measure.

- Two main kinds of MDS methods/algorithms will be introduced:
  - Metric MDS (including classical MDS as a special case)
  - Non-metric MDS

Distance and Dissimilarity

- Distance, dissimilarity and proximity are defined for any pair of objects in any space. In mathematics, a distance function d (that gives a distance between two objects) satisfies

  1. $d(x,y) \geq 0$;
  2. $d(x,y) = 0$ if and only if $x = y$;
  3. $d(x,y) = d(y,x)$;

- A distance function d satisfying the triangle inequality

  $$d(x,z) \leq d(x,y) + d(y,z)$$

  is called a *metric* distance (or just metric).

- A distance function d does not satisfy the triangle inequality is called a *non-metric* distance.

Distance and Dissimilarity

- Two common choices of metric distances:
  - Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2$.
  - L_1 distance $\|\mathbf{x}_i - \mathbf{x}_j\|_1$, sometimes called the Manhanttan or taxicab metric.
- Common non-metric distances: rank order.

Classical MDS

- Classical MDS is based on Euclidean distances: $d_{ij} := d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}$.
- Classical MDS can be shown to be equivalent to PCA (See the posted notes)
- For this reason, classical MDS is also referred to as **principal coordinates analysis (PCoA)**.
- R code demonstration

MDS: Theory

- The objective of multidimensional scaling is to find a good value for m and the m-dimensional points

$$\mathbf{y}_1 = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{m1} \end{bmatrix} \quad \mathbf{y}_2 = \begin{bmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{m2} \end{bmatrix} \quad \cdots \quad \mathbf{y}_n = \begin{bmatrix} y_{1n} \\ y_{2n} \\ \vdots \\ y_{mn} \end{bmatrix}$$

- Let $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]'$ denote the $n \times m$ matrix.

- How do we find m and the m-dimensional points?

MDS: Theory

- First consider q-dimensional data vectors

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{q1} \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{q2} \end{bmatrix} \quad \cdots \quad \mathbf{x}_n = \begin{bmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{qn} \end{bmatrix}$$

  collected on the n individuals in the study.

- Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ denote the $n \times q$ data matrix.

- Let $D = (d_{ij})$ represent the $n \times n$ matrix of distances (e.g., Euclidean) between the n individuals in the study with

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ki} - x_{kj})^2}$$

MDS: Theory

- **Inner products matrix**: Define the $n \times n$ matrix of the inner products of the rows of X as $B = XX^{'}$ with each element $b_{ij} = \sum_{k=1}^{q} x_{ik}x_{jk}$.
- The goal then is to approximate YY' by using the observed data through XX'.
  - The approximation is $B = XX'$.
  - It can be shown that $d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$.
  - We want to get formulas for the b_{ij}'s in term of the d_{ij}'s, but no unique solution exists, unless a location constraint is introduced.
  - The most commonly used constraint centers the rows of X at the origin, i.e., $\sum_{i=1}^{n} x_{ij} = 0$ for all $j = 1, 2, ..., q$

MDS: Theory

- Those constraints imply that the sum of the elements in any row of B must be zero.

- This leads to the following set of equations:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} b_{ii} + nb_{jj}$$

$$\sum_{j=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} b_{ii} + nb_{ii}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2n \sum_{i=1}^{n} b_{ii}$$

Note that $\sum_{i=1}^{n} b_{ii} = trace(B)$

MDS: Theory

- The elements of B can now be found in terms of the squared distances between the rows of X:

$$b_{ij} = -0.5 \left(d_{ij}^2 - \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 \right)$$

- Having derived the elements of B in terms of distances between the rows of X, we must now factor B, our approximation for YY', to give the rows of X values.

- The spectral decomposition of B is $B = V \Lambda V^{'}$, where

  $\Lambda = Diag\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix containing the eigenvalues of B

  $V = [V_1, V_2, \ldots, V_n]$ is a matrix whose columns are the corresponding eigenvectors of B

MDS: Theory

- When D is computed from an $n \times q$ data matrix of full rank, then the rank of B is q, and the $n-q$ smallest eigenvalues of B will be zero.

- Then $B = \sum_{k=1}^{q} \lambda_k V_k V_k^{'}$.

- We want to approximate

$$B = XX^{'} = V \Lambda V^{'} = \sum_{k=1}^{m} \left(\sqrt{\lambda_k} V_k\right) \left(\sqrt{\lambda_k} V_k\right)^{'}$$

  which has rank $m < q$ and only has m positive eigenvalues.

- The approximation to Y_k is $\tilde{X}_k = \sqrt{\lambda_k} V_k$ for $k = 1, 2, ..., m$.

(Metric) MDS

- How do we pick the value of m?

- Compute the criterion

$$P_m = \frac{\sum_{k=1}^{m} \tilde{\lambda}_k}{\sum_{k=1}^{q} \tilde{\lambda}_k}$$

  for $m = 1, 2, 3..., q$. Values of P_m near 0.80 or above suggest a reasonable fit.

- This application of multidimensional scaling to a $n \times q$ data matrix is equivalent to computing the first m principal components from a covariance matrix (the scaling of the observations matters).

(Metric) MDS

- When the proximity matrix is not a matrix of Euclidean distances, the matrix B is not always positive definite.
  - Some of the eigenvalues of B may be negative.
  - If only a small fraction of the eigenvalues of B are negative, a useful representation of the proximity matrix may still be obtained using the m largest eigenvalues of B.
  - Consider the criterion

    $$P_m = \frac{\sum_{k=1}^{m} |\tilde{\lambda}_k|}{\sum_{k=1}^{n} |\tilde{\lambda}_k|} \quad \text{or} \quad P_m = \frac{\sum_{k=1}^{m} \tilde{\lambda}_k^2}{\sum_{k=1}^{n} \tilde{\lambda}_k^2}$$

    Values of P_m near 0.80 or above suggest a reasonable fit.
- If B has a considerable number of large negative eigenvalues, the MDS method described above should not be used, and non-metric scaling should be considered.

Non-metric MDS

- Used when subjects are only able to make ordinal comparisons, e.g.,
  - One color is brighter than another
  - One food is more salty than another
  - One speaker is more positive than another
- Non-metric scaling is based on rank ordering, not on quantitative values of responses.
- A "stress" criterion is minimized. Stress increase when the distances between individuals on the visual display disagrees more with the rank order of the individuals (or items). For example, the individuals who rank 1 and 2 should be closer together than the individuals who rank 1 and 5.

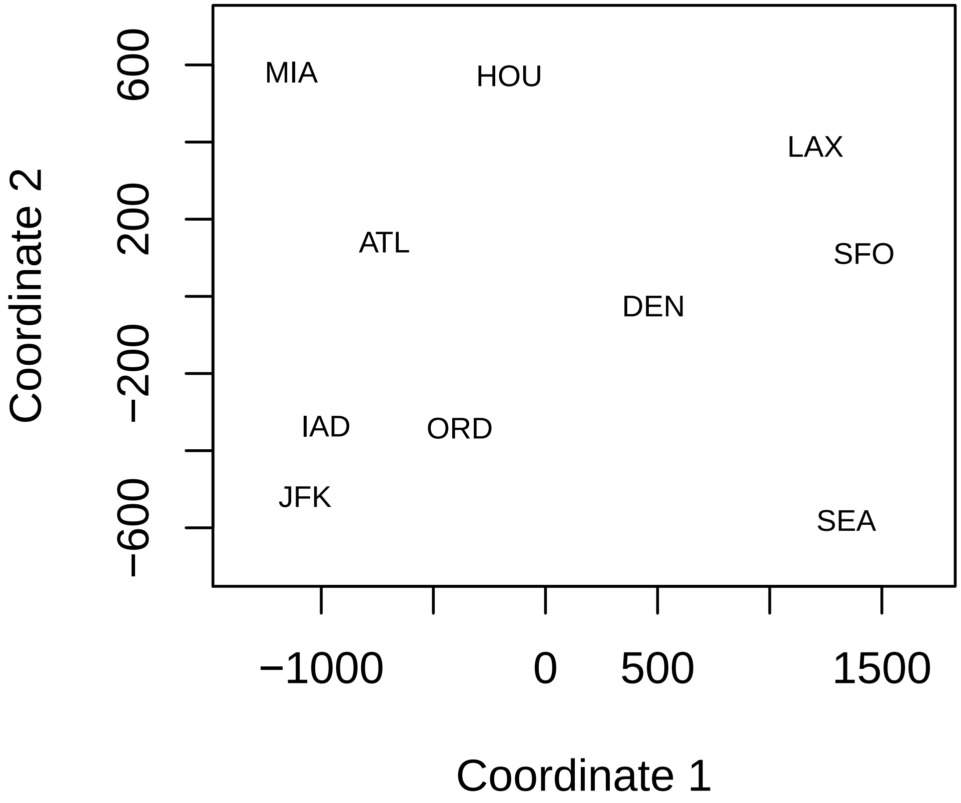
Example 1 in Section 4.4.2 of Everitt and Hothorn

Airline Distances between Cities

- We will start with a proximity matrix of airline distances between ten cities in the US
- These and essentially distances on a sphere
- They are not Euclidean distances
- Some eigenvalues of the distance matrix are negative.
- The distance matrix is presented on page 113 of Everitt and Hothorn (but one of the distances is incorrect).
- We will represent the distances between the ten cities on a two-dimensional graph

Airline Distances Example (Corrected)

- The correct distance between ATL and SEA is 2180 miles
- The multidimensional scaling display for the corrected proximity matrix is shown below. ATL is now in a more reasonable position relative to the other nine cities.



House of Representatives Voting

- Romesburg (1984) counted the number of times 15 U.S. congressmen from New Jersey voted differently on 19 environmental bills.

- Abstentions were not recorded, but two congressmen, Sandman (9 abstentions) and Thompson (6 abstentions), abstained more frequently than others

- Apply non-metric scaling to display dissimilarities in voting records for the individual congressmen and possibly reveal patterns.

- R code demonstration

Comparison to Other Methods for Dimension Reduction

- Multidimensional scaling
  - preserve the pairwise sample “dissimilarity”
  - no distribution assumption, could proceed with only dissimilarity metric
- Principal Component Analysis
  - maximize variance in a set of orthogonal projections
- Factor Analysis
  - explains the covariance matrix
  - assumes multivariate normality
  - identifying a small number of unobserved common factors