STAT 730 Chapter 14: Multidimensional scaling

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Data Analysis

Basic idea

We have n objects and a matrix of distances between each object; d_{rs} is the distance between objects r and s and $\mathbf{D} = [d_{ij}] \in \mathbb{R}^{n \times n}$ is the distance matrix. Multidimensional scaling (MDS) seeks to create points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^k$ s.t. $d_{rs} \approx ||\mathbf{x}_r - \mathbf{x}_s||$. The points are then plotted to gauge how "similar" objects or variables are, with "like" objects/variables near each other in the plot. Often $d_{rs} = ||\mathbf{z}_r - \mathbf{z}_s||$ where $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathbb{R}^p$ where $k << p$. Distances can be any measure of dissimilarity, e.g. actual Euclidean distance, Mahalanobis distance, genetic distance.

Example from Marden (2013): Louis Roussos recorded data on 130 sets of individual ranks on seven sports. Each subject allocated the numbers $1, 2, 3, 4, 5, 6, 7$ to the seven sports from wanting to participate in the most to the least. The sports are baseball, football, basketball, tennis, cycling, swimming and jogging.

Sports rankings

Here $d_{rs} = ||\mathbf{z}_{(r)} - \mathbf{z}_{(s)}||$ for $1 \leq r, s \leq 7$ where $\mathbf{z}_{(1)}, \ldots, \mathbf{z}_{(7)} \in \mathbb{R}^{130}$. We want to find $\mathbf{x}_1, \ldots, \mathbf{x}_7 \in \mathbb{R}^2$, i.e. $k = 2$, such that $d_{rs} = ||\mathbf{z}_{(r)} - \mathbf{z}_{(s)}|| \approx ||\mathbf{x}_r - \mathbf{x}_s||$. Here, the traditional idea of objects and variables have been reversed. The $n = 7$ objects are different sporting activities, and the $p = 130$ measurements taken on each sport are simply 130 individual rankings.

```
source("http://www.stat.sc.edu/~hansont/stat730/Marden_Rcode.txt")
sportsranks
cor(sportsranks)
D=dist(t(sportsranks)) # 7 by 7 matrix Euclidean dist. between columns
```

The `daisy` function in the `cluster` package creates dissimilarity matrices for mixed data: continuous and categorical (both ordinal and nominal).

Objects vs. variables

We have been considering data matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ where typically $n >> p$. The matrix $\mathbf{X}$ has n objects, often subjects in a study, and p variables. Think of $p = 4$ N/W/S/E cork weight measurements on $n = 28$ trees; or $p = 5$ exam scores and $n = 88$ students. In the first scenario we may want to have a two-dimensional map showing relative distances, with closer meaning more "similar," between the four compass directions, or alternatively among the 88 trees. In the second scenario we may want to have a map among the students, or a map showing relative distances among the exam types.

MDS can work on either objects or variables. For example, among students we may want to cluster the students to see if meaningful groups arise; we would need to know the students names and other covariate information about them for this purpose. Alternatively we may want to know if two exam types are similar, i.e. do they carry the same information in some sense. If we are administering a battery of tests, and some tests carry duplicate information, we can get rid of some of them.

MDS

Start with a "distance" or dissimilary matrix $\mathbf{D} = [d_{rs}] \in \mathbb{R}^{n \times n}$. The goal is find $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^k$ such that $\hat{\mathbf{D}} = [\hat{d}_{rs}] \in \mathbb{R}^{n \times n}$ is close to $\mathbf{D}$, where $\hat{d}_{rs} = ||\mathbf{x}_r - \mathbf{x}_s||$. If $k = 2$ we can plot the $\mathbf{x}_1, \dots, \mathbf{x}_n$ to get a picture of relative "location" amongst the n objects.

Note that $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$ where $\mathbf{A}\mathbf{A}' = \mathcal{I}_k$ also satisfies $\hat{d}_{rs} = ||\mathbf{y}_r - \mathbf{y}_s|| = ||\mathbf{x}_r - \mathbf{x}_s||$. So any solution will be translation/rotation invariant.

Classical (metric) MDS constructs an $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]' \in \mathbb{R}^{n \times p}$ that leads to $\mathbf{D}$ exactly; this is part (b) of the theorem on the next slide. If we want to use $k < p$, we simply use n points obtained from the first k columns of $\mathbf{X}$ instead. This is justified via PCA.

Def'n Euclidean for $\mathbf{D}$ and theorem

def'n: $\mathbf{D}$ is Euclidean if for some p there exists $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ s.t. $d_{rs}^2 = ||\mathbf{x}_r - \mathbf{x}_s||^2$.

thm: Let $\mathbf{A} = [-\frac{1}{2}d_{rs}^2]$, $\mathbf{H} = \mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$, and $\mathbf{B} = \mathbf{HAH}$. $\mathbf{D}$ Euclidean $\Leftrightarrow$ $\mathbf{B} \geq 0$.

(a) If $\mathbf{D}$ is Euclidean, i.e. $d_{rs} = ||\mathbf{z}_r - \mathbf{z}_s||$ where $1 \leq r, s \leq n$, then $b_{rs} = (\mathbf{z}_r - \bar{\mathbf{z}})'(\mathbf{z}_s - \bar{\mathbf{z}})$. In matrix terms $\mathbf{B} = (\mathbf{HZ})(\mathbf{HZ})'$, hence $\mathbf{B} \geq 0$.

(b) If $\mathbf{B} \geq 0$ and $\text{rank}(\mathbf{B}) = p$ then write $\mathbf{B} = \mathbf{\Gamma}_p\mathbf{\Lambda}_p\mathbf{\Gamma}_p'$ where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p \times p}$ and $\mathbf{\Gamma}_p = [\gamma_{(1)} \cdots \gamma_{(p)}] \in \mathbb{R}^{n \times p}$. Create $\mathbf{X} = [\gamma_{(1)}\sqrt{\lambda_1} \cdots \gamma_{(p)}\sqrt{\lambda_p}] \in \mathbb{R}^{n \times p}$. Then $\mathbf{D}$ formed from $d_{rs} = \mathbf{x}_r'\mathbf{x}_s$ ($\bar{\mathbf{x}} = \mathbf{0}$) has corresponding $\mathbf{B}$.

Proof: $\mathbf{D}$ Euclidean $\Rightarrow \mathbf{B} \geq 0$

$\mathbf{D} = [d_{rs}]$ where $d_{rs} = ||\mathbf{z}_r - \mathbf{z}_s||$. $\mathbf{B} = \mathbf{HAH}$ where $a_{rs} = -\frac{1}{2} d_{rs}^2$.

$$[\mathbf{AH}]_{ij} = \begin{bmatrix} -\frac{1}{2} d_{i1}^2 & \cdots & -\frac{1}{2} d_{in}^2 \end{bmatrix} \begin{bmatrix} -\frac{1}{n} \\ -\frac{1}{n} \\ \vdots \\ 1 - \frac{1}{n} \leftarrow j \\ \vdots \\ -\frac{1}{n} \end{bmatrix} = \frac{1}{2n} \sum_{s=1}^{n} d_{is}^2 - \frac{1}{2} d_{ij}^2.$$

$$\begin{aligned} [\mathbf{HAH}]_{ij} &= \begin{bmatrix} -\frac{1}{n} & -\frac{1}{n} & \cdots & \overbrace{1 - \frac{1}{n}}^{i} & \cdots & -\frac{1}{n} \end{bmatrix} \begin{bmatrix} \frac{1}{2n} \sum_{s=1}^{n} d_{1s}^2 - \frac{1}{2} d_{1j}^2 \\ \vdots \\ \frac{1}{2n} \sum_{s=1}^{n} d_{ns}^2 - \frac{1}{2} d_{nj}^2 \end{bmatrix} \\ &= -\frac{1}{2n^2} \sum_{s=1}^{r} \sum_{t=1}^{n} d_{rs}^2 + \frac{1}{2n} \sum_{r=1}^{n} d_{rj}^2 + \frac{1}{2n} \sum_{s=1}^{n} d_{is}^2 - \frac{1}{2} d_{ij}^2 \\ &= \bar{a}_{\cdot\cdot} - \bar{a}_{\cdot j} - \bar{a}_{i\cdot} + a_{ij}. \end{aligned}$$

Now MKB 14.2.1 implies $b_{rs} = [\mathbf{HAH}]_{rs} = (\mathbf{z}_r - \bar{\mathbf{z}})'(\mathbf{z}_s - \bar{\mathbf{z}})$. Thus $\mathbf{B} = (\mathbf{HZ})(\mathbf{HZ})'$ so $\mathbf{B} \geq 0$.

Proof: $\mathbf{D}$ Euclidean $\Leftarrow \mathbf{B} \geq 0$

We'll build an $\mathbf{X}$. Let $p = \text{rank } \mathbf{B}$, $\mathbf{B} \geq 0$. Recall $\mathbf{B} \in \mathbb{R}^{n \times n}$. Let $\lambda_1 \geq \cdots \geq \lambda_p > 0$ be the non-zero e-values, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p \times p}$, and $\boldsymbol{\Gamma} = [\gamma_{(1)} \cdots \gamma_{(p)}] \in \mathbb{R}^{n \times p}$.

Take $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2} = [\sqrt{\lambda_1}\gamma_{(1)} \cdots \sqrt{\lambda_p}\gamma_{(p)}]$. Then $\mathbf{B} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}' = \mathbf{X}\mathbf{X}'$. So then $b_{rs} = \mathbf{x}_r'\mathbf{x}_s$ where $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]' \in \mathbb{R}^{n \times p}$.
Now

$$
\begin{aligned}
||\mathbf{x}_r - \mathbf{x}_s|| &= \mathbf{x}_r'\mathbf{x}_r - 2\mathbf{x}_r'\mathbf{x}_s + \mathbf{x}_s'\mathbf{x}_s \\
&= b_{rr} - 2b_{rs} + b_{ss} \\
&= a_{rr} - 2a_{rs} + a_{ss} \\
&= -2a_{rs} = d_{rs}^2,
\end{aligned}
$$

Because $b_{rs} = \bar{a}_{\cdot\cdot} - \bar{a}_{\cdot s} - \bar{a}_{r\cdot} + a_{rs}$ and $a_{ii} = 0$ for $i = 1, \ldots, n$.

How this works in practice

Since $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_{1\times p}$, the columns have mean zero. This holds because $\mathbf{1}_n$ is an e-vector of $\mathbf{B}$ with e-value zero, orthogonal to the columns of $\mathbf{\Gamma}$. Thus a plot of the $\mathbf{x}_1, \dots, \mathbf{x}_n$ has $\mathbf{0}$ at its center.

To find k-dimensional $\mathbf{x}_1, \dots, \mathbf{x}_n$ that represent objects from $\mathbf{D}$, create $\mathbf{A} = [-\frac{1}{2} d_{rs}^2]$ and form $\mathbf{B} = \mathbf{HAH}$. Take $\lambda_1 \geq \cdots \geq \lambda_k$ and $\mathbf{\Gamma} = [\gamma_{(1)} \cdots \gamma_{(k)}]$. Then $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]' = [\sqrt{\lambda_1}\gamma_{(1)} \cdots \sqrt{\lambda_k}\gamma_{(k)}]$. If $p = \text{rank}(\mathbf{B}) = k$ then $||\mathbf{x}_r - \mathbf{x}_s|| = d_{rs}$, otherwise $||\mathbf{x}_r - \mathbf{x}_s|| \approx d_{rs}$.

It is common to take $k = 2$ and plot the $\mathbf{x}_1, \dots, \mathbf{x}_n$ in the plane labeled by the object names. If the first two e-values eat up most of the variability (as in PCA) the Euclidean distances $||\mathbf{x}_r - \mathbf{x}_s||$ in the plane should closely coincide with the actual d_{rs}.

Driving distance between cities in S.C.

```
# driving distance
# charleston
# columbia    112
# florence    110  80
# hilton head 108 164 181
# myrtle beach 94 143  70 203
# spartanburg 201  93 153 253 223

D=matrix(c(  0,112,110,108, 94,201,
           112,  0, 80,164,143, 93,
           110, 80,  0,181, 70,153,
           108,164,181,  0,203,253,
            94,143, 70,203,  0,223,
           201, 93,153,253,223,  0),6,6)
f=cmdscale(D,k=2)
par(mfrow=c(1,2))
plot(f,xlab="Coord. 1",ylab="Coord 2",main="Classical MDS",type="n")
text(f,labels=c("Charleston","Columbia","Florence","Hilton Head",
 "Myrtle Beach","Spartanburg"),cex=.7)
t=-pi/3; R=matrix(c(cos(t),sin(t),-sin(t),cos(t)),2,2)
plot(f%*%t(R),xlab="Coord. 1",ylab="Coord 2",main="Classical MDS",type="n")
text(f%*%t(R),labels=c("Charleston","Columbia","Florence","Hilton Head",
 "Myrtle Beach","Spartanburg"),cex=.7)
```

Similarity measures instead of distance

If $\mathbf{C}$ is a similarity matrix (larger values mean objects are more similar or "closer") then $\mathbf{C} = \mathbf{C}'$ and $c_{rs} \leq c_{rr}$ & $c_{rs} \leq c_{ss}$.

def'n: The standard transformation from $\mathbf{C}$ to $\mathbf{D}$ is $d_{rs} = \sqrt{c_{rr} - 2c_{rs} + c_{rr}}$.

thm: If $\mathbf{C} \geq 0$ then $\mathbf{D}$ as defined above is Euclidean with $\mathbf{B} = \mathbf{HCH}$. See MKB 402–403 for a proof.

In practice, one transforms $\mathbf{C}$ to $\mathbf{D}$ and then performs the classical multidimensional scaling described thus far.

A nice similarity matrix for continuous outcomes is the sample correlation $\mathbf{D} = \mathbf{R}$.

Duality between PCA and classical MDS

Take $\mathbf{Z} \in \mathbb{R}^{n \times p}$ where rows are p measurement values on an object. If we wish to MDS on the n objects note the following relationship between MDS and PCA:

thm: If $\mathbf{D} = [d_{rs}]$ where $d_{rs} = ||\mathbf{z}_r - \mathbf{z}_s||$, the Euclidean distance between rows r and s of $\mathbf{Z}$, then the principal coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ are equal to the first k principal components: $\mathbf{x}_i' = (y_{i1}, \ldots, y_{ik})$. Proof is on pp. 405–406 on MKB.

Heuristically, recall the approximation $\mathbf{z}_i \approx \boldsymbol{\mu} + \boldsymbol{\Gamma}_k \boldsymbol{\Gamma}_k' (\mathbf{z}_i - \boldsymbol{\mu})$. Then

$$||\mathbf{z}_r - \mathbf{z}_s||^2 = (\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s) \approx ||\boldsymbol{\Gamma}_k'(\mathbf{z}_r - \mathbf{z}_s)||^2 = \sum_{j=1}^{k} (y_{rj} - y_{sj})^2.$$

The points $\mathbf{x}_i = (y_{i1}, \ldots, y_{ik}) \in \mathbb{R}^k$ approximately preserve the distances among the $\mathbf{z}_i \in \mathbb{R}^p$.

Doesn't matter if we consider n objects on p variables (using $\mathbf{Z}$) or p variables on n objects (using $\mathbf{Z}'$).

Duality between PCA and classical MDS

MKB starts with $\mathbf{D}$ and finds a constructive approach to finding $\mathbf{x}_1, \ldots \mathbf{x}_n \in \mathbb{R}^q$ s.t. $d_{rs} \approx ||\mathbf{x}_r - \mathbf{x}_s||$. Marden (2013) starts with a data matrix $\mathbf{Z} \in \mathbb{R}^{n \times p}$ and seeks the solution $\mathbf{x}_i = \mathbf{A}\mathbf{z}_i$ for $\mathbf{A} \in \mathbb{R}^{q \times p}$, i.e. $\mathbf{X} = \mathbf{Z}\mathbf{A}'$, satisfying

$$\mathbf{A} = \text{argmin}_{\mathbf{A} \in \mathbb{R}^{q \times p}} \sum_{r=1}^{n} \sum_{s=1}^{n} \left| ||\mathbf{z}_r - \mathbf{z}_s||^2 - ||\mathbf{A}\mathbf{z}_r - \mathbf{A}\mathbf{z}_s||^2 \right|.$$

Marden shows $\mathbf{A}' = [\gamma_{(1)} \cdots \gamma_{(q)}]$ where $\mathbf{Z}'\mathbf{H}\mathbf{Z} = n\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$. MKB (p. 405) show $\mathbf{B} = \mathbf{H}\mathbf{Z}\mathbf{Z}'\mathbf{H}$.

Marden then goes on to consider $\mathbf{D}$ arriving without a data matrix, where MKB start.

Sports rankings

```
source("http://www.stat.sc.edu/~hansont/stat730/Marden_Rcode.txt")
sportsranks
cor(sportsranks)
D=dist(t(sportsranks)) # 7 by 7 matrix Euclidean dist. between columns
par(mfrow=c(1,2))
f=cmdscale(D,k=2)
par(mfrow=c(1,2))
plot(f,xlab="Coord. 1",ylab="Coord. 2",main="Classical MDS",type="n")
text(f,labels=colnames(sportsranks),cex=.7)

library(MASS)
f=isoMDS(D) # Kruskal's non-metric multidimensional scaling, pp. 413-415
x=f$points[,1]; y=f$points[,2]
plot(x,y,xlab="Coord. 1", ylab="Coord. 2",main="Nonmetric MDS", type="n")
text(x,y,labels=colnames(sportsranks),cex=.7)
```

Non-metric approach

The classical approach thus far uses a metric, i.e. the d_{rs} are distances between points. Kruskal (1964) considers construction of $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ from $\mathbf{D}$ based on ranks; any monotone transformation of the distances in $\mathbf{D}$ gives the same answer. A data matrix $\hat{\mathbf{X}} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ with corresponding interpoint distances $\hat{d}_{rs}$ is constructed s.t. $d_{rs} \approx ||\mathbf{x}_r - \mathbf{x}_s||$.

Start with $\mathbf{D}$; there are $m = \frac{n(n-1)}{2}$ distances ordered $d_{r_1 s_1} \leq \cdots \leq d_{r_m s_m}$ where $r_i < s_i$. Define $\mathcal{D} = \{\tilde{d}_{rs} : d_{rs} < d_{uv} \Rightarrow \tilde{d}_{rs} \leq \tilde{d}_{uv}, r < s, u < v\}$. For a fixed q define the stress

$$S^2(\hat{\mathbf{X}}) = \min_{\tilde{\mathbf{d}} \in \mathcal{D}} \frac{\sum_{r<s} (\tilde{d}_{rs} - \hat{d}_{rs})^2}{\sum_{r<s} \hat{d}_{rs}^2}.$$

This minimization is carried out via isotonic regression, typically using the *pool adjacent violators algorithm*.

Non-metric approach

An $\hat{\mathbf{X}}$ is found to minimize $S^2(\mathbf{X})$, e.g. using the method of steepest descent.

The dimension k can be picked to give reasonable stress; Kruskal says, roughly, $S \approx 20\%$ poor agreement, $S \approx 10\%$ fair, $S \approx 5\%$ good, and $S \approx 0$ perfect.

The non-metric approach can handle missing distances in $\mathbf{D}$, ties, and ordinal data. See MKB 413–415 for a partial description. The method can be carried out via the `isoMDS` function in the `MASS` package.