Multivariate Statistics

Chapter 5: Multidimensional scaling

Pedro Galeano
Departamento de Estadística
Universidad Carlos III de Madrid
pedro.galeano@uc3m.es

Course 2017/2018

Master in Mathematical Engineering

Introduction

- As we have seen in previous chapters, principal components and factor analysis are important dimension reduction tools.
- However, in many applied sciences, data is recorded as ranked information.
- For example, in marketing, one may record "product A is better than product B".
- Multivariate observations therefore often have mixed data characteristics and contain information that would enable us to employ one of the multivariate techniques presented so far.
- Multidimensional scaling (MDS) is a method based on proximities between objects, subjects, or stimuli used to produce a spatial representation of these items.
- MDS is a dimension reduction technique since the aim is to find a set of points in low dimension (typically two dimensions) that reflect the relative configuration of the high-dimensional data objects.

Introduction

- The proximities between objects are defined as any set of numbers that express the amount of similarity or dissimilarity between pairs of objects.
- In contrast to the techniques considered so far, MDS does not start from a $n \times p$ dimensional data matrix, but from a $n \times n$ dimensional dissimilarity or distance matrix, D, with elements $\delta_{ii'}$ or $d_{ii'}$, respectively, for $i, i' = 1, \ldots, n$.
- Hence, the underlying dimensionality of the data under investigation is in general unknown.

Introduction

- MDS techniques are often used to understand how people perceive and evaluate certain signals and information.
- For instance, political scientists use MDS techniques to understand why political candidates are perceived by voters as being similar or dissimilar.
- Psychologists use MDS techniques to understand the perceptions and evaluations of speech, colors and personality traits, among other things.
- Marketing researchers use MDS techniques to shed light on the way consumers evaluate brands and to assess the relationship between product attributes.

Introduction

- What lies behind MDS is the concept of distance.
- Therefore, we first review briefly the most important statistical distances.
- After this, we are going to present two different MDS solutions:
  - The Metric MDS solution is concerned with a representation of the distance matrix in Euclidean coordinates where the projections are obtained via a spectral decomposition of a distance matrix.
  - The Non-metric MDS is a more sophisticated solution particularly useful when the proximities are measured in an ordinal scale.

Statistical distances

- As mentioned before, MDS depends on the concept of statistical distance.
- Distances also play an important role in other multivariate techniques such as cluster analysis that will be presented in Chapter 6 and some of the methods for classification in Chapter 7.
- We already know some distances between multivariate observations: the Euclidean distance and the Mahalanobis distance.
- Next, we present alternative distances.

Statistical distances

- We begin with the definition of distance.
- For simplicity we focus on distances between random variables defined in the same probability space.
- Definition: A distance between two independent random variables $x_{i\cdot}$ and $x_{i'\cdot}$ is a positive random variable which satisfies:
  1. $d\left(x_{i\cdot}, x_{i'\cdot}\right) \geq 0$;
  2. $d\left(x_{i\cdot}, x_{i'\cdot}\right) = 0$, if and only if $x_{i\cdot} = x_{i'\cdot}$;
  3. $d\left(x_{i\cdot}, x_{i'\cdot}\right) = d\left(x_{i'\cdot}, x_{i\cdot}\right)$; and
  4. $d\left(x_{i\cdot}, x_{i'\cdot}\right) \leq d\left(x_{i\cdot}, x_{i''\cdot}\right) + d\left(x_{i''\cdot}, x_{i'\cdot}\right)$, for any other independent random variables $x_{i''\cdot}$.

Statistical distances

- The two most common distances in Statistics for quantitative multivariate random variables are the Euclidean distance and the Mahalanobis distance.
- The Euclidean distance, d_E, between $x_{i\cdot}$ and $x_{i'\cdot}$, is given by:

$$d_E\left(x_{i\cdot}, x_{i'\cdot}\right) = \left[\left(x_{i\cdot} - x_{i'\cdot}\right)' \left(x_{i\cdot} - x_{i'\cdot}\right)\right]^{1/2}$$

- The Mahalanobis distance, d_M, between $x_{i\cdot}$ and $x_{i'\cdot}$, is given by:

$$d_M\left(x_{i\cdot}, x_{i'\cdot}\right) = \left[\left(x_{i\cdot} - x_{i'\cdot}\right)' \Sigma_x^{-1} \left(x_{i\cdot} - x_{i'\cdot}\right)\right]^{1/2}$$

  where Σ_x is the common covariance matrix of $x_{i\cdot}$ and $x_{i'\cdot}$.
- Note that the Euclidean distance coincides with the Mahalanobis distance if $\Sigma_x = I_p$.

Statistical distances

- The Minkowski distance, d_p, between $x_{i\cdot}$ and $x_{i'\cdot}$, is given by:

$$d_p\left(x_{i\cdot}, x_{i'\cdot}\right) = \left[\sum_{j=1}^{p} \left|x_{ij} - x_{i'j}\right|^p\right]^{1/p}$$

- If $p = 1$, d_p is called the Manhattan distance.
- If $p = 2$, d_p is the Euclidean distance.
- If $p = \infty$, d_p is the maximum distance or the Chebychev distance, $d_{\max}$, that can be written as:

$$d_{\max}\left(x_{i\cdot}, x_{i'\cdot}\right) = \max_{j=1,\ldots,p} \left|x_{ij} - x_{i'j}\right|$$

Statistical distances

- The Canberra distance, d_{Canb}, between $x_{i\cdot}$ and $x_{i'\cdot}$, is given by:

$$d_{Canb}\left(x_{i\cdot}, x_{i'\cdot}\right) = \sum_{j=1}^{p} \frac{|x_{ij} - x_{i'j}|}{|x_{ij}| + |x_{i'j}|}$$

- The Bhattacharyya distance, d_{Bhat}, between $x_{i\cdot}$ and $x_{i'\cdot}$, is given by:

$$d_{Bhat}\left(x_{i\cdot}, x_{i'\cdot}\right) = \sum_{j=1}^{p} \left(x_{ij}^{1/2} - x_{i'j}^{1/2}\right)^2$$

Statistical distances

- The cosine distance (or dissimilarity), $d_{\cos}$, between $x_{i\cdot}$ and $x_{i'\cdot}$, is given by:

$$d_{\cos}\left(x_{i\cdot}, x_{i'\cdot}\right) = 1 - \cos\left(x_{i\cdot}, x_{i'\cdot}\right)$$

  where $\cos\left(x_{i\cdot}, x_{i'\cdot}\right)$ is the cosine of the included angle of the two random vectors, given by:

$$\cos\left(x_{i\cdot}, x_{i'\cdot}\right) = \frac{x_{i\cdot}' x_{i'\cdot}}{\|x_{i\cdot}\| \, \|x_{i'\cdot}\|}$$

  and $\|\cdot\|$ denotes the Euclidean norm of a vector.

- The correlation distance (or dissimilarity), d_{cor}, between $x_{i\cdot}$ and $x_{i'\cdot}$, for $i, i' = 1, \ldots, n$, is given by:

$$d_{cor}\left(x_{i\cdot}, x_{i'\cdot}\right) = 1 - \rho_{ii'}$$

  where $\rho_{ii'}$ is the correlation coeficient betwen $x_{i\cdot}$ and $x_{i'\cdot}$.

Statistical distances

- The Hamming distance, d_{Hamm}, can be used for binary random variables with entries 0 and 1.
- The Hamming distance between $x_{i\cdot}$ and $x_{i'\cdot}$, for $i, i' = 1, \ldots, n$, is given by:

$$d_{Hamm}\left(x_{i\cdot}, x_{i'\cdot}\right) = \frac{\#\left\{x_{ij} \neq x_{i'j} : 1 \leq j \leq p\right\}}{p}$$

Statistical distances

- The Gower distance, d_{Gow}, can be used for random variables with quantitative and qualitative entries.
- The Gower distance can be computed as follows:
  1. Express the qualitative variables as indicator variables (as seen in Chapter 1).
  2. Standardize all variables individually such that the sample mean of each variable is 0 and the sample variance is 1.
  3. Compute the distance between observations using the Manhattan (or the Euclidean) distance.

Metric MDS

- Metric MDS begins with a $n \times n$ distance matrix D with elements $d_{ii'}$, where $i, i' = 1, \ldots, n$.
- The goal of Metric MDS is to find a configuration of points such that the coordinates of the n points along p dimensions yields a Euclidean distance matrix whose elements are as close as possible to the elements of the given distance matrix D.
- Before introducing metric MDS, we show how to obtain, from a data matrix X, the matrix of squared Euclidean distances between the observations of X.
- Then, we will be ready to present metric MDS.

Metric MDS

- Let X be a $n \times p$ data matrix and let $\widetilde{X}$ be the $n \times p$ centered data matrix.
- Let $D_E^{(2)}$ be the matrix of squared Euclidean distances between the observations of X, with elements $d^2_{E,ii'} = (x_{i\cdot} - x_{i'\cdot})' (x_{i\cdot} - x_{i'\cdot})$.
- Then, $D_E^{(2)}$ can be written as follows:

$$D_E^{(2)} = diag\,(Q)\, 1_n' + 1_n diag\,(Q)' - 2Q$$

where:

$$Q = \widetilde{X}\widetilde{X}'$$

and $diag\,(Q)$ is a column vector with the diagonal elements of Q.

- In particular, the squared Euclidean distance between $x_{i\cdot}$ and $x_{i'\cdot}$ can be written in the terms of the elements of Q as follows:

$$d^2_{E,ii'} = Q_{ii} + Q_{i'i'} - 2Q_{ii'}$$

Metric MDS

- Therefore, we can write the matrix $D_E^{(2)}$ in terms of the matrix Q, or, in other words, in terms of $\widetilde{X}$ and then in terms of X.
- The question is: is it possible to reconstruct the data matrix X from the matrix $D_E^{(2)}$ of squared Euclidean distances?
- In order to do that, the goal is, first, to obtain Q, then, the matrix $\widetilde{X}$, and then, the matrix X, if possible.

Metric MDS

- We begin by studying how to obtain the matrix Q given the matrix $D_E^{(2)}$.
- First, noting that $Q1_n = 0_p$ as $\widetilde{X}'1_n = 0_p$, we get:

$$\sum_{i=1}^{n} Q_{ii'} = \sum_{i'=1}^{n} Q_{ii'} = 0$$

- Then,

$$\sum_{i=1}^{n} d_{E,ii'}^2 = \sum_{i=1}^{n} (Q_{ii} + Q_{i'i'} - 2Q_{ii'}) = Tr(Q) + nQ_{i'i'} \Longrightarrow$$

$$\Longrightarrow Q_{i'i'} = \frac{1}{n}\sum_{i=1}^{n} d_{E,ii'}^2 - \frac{Tr(Q)}{n} = d_{i'} - \frac{Tr(Q)}{n}$$

where:

$$d_{i'} = \frac{1}{n}\sum_{i=1}^{n} d_{E,ii'}^2$$

Metric MDS

- Similarly,

$$\sum_{i'=1}^{n} d_{E,ii'}^2 = nQ_{ii} + Tr\left(Q\right) \Longrightarrow$$

$$\Longrightarrow Q_{ii} = \frac{1}{n}\sum_{i'=1}^{n} d_{E,ii'}^2 - \frac{Tr\left(Q\right)}{n} = d_i - \frac{Tr\left(Q\right)}{n}$$

where:

$$d_i = \frac{1}{n}\sum_{i'=1}^{n} d_{E,ii'}^2$$

Metric MDS

- On the other hand,

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{E,ii'}^2 = \sum_{i=1}^{n} (nQ_{ii} + Tr(Q)) = 2nTr(Q) \Longrightarrow$$

$$\Longrightarrow Tr(Q) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{E,ii'}^2 = \frac{n}{2}d$$

where:

$$d = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{E,ii'}^2$$

Metric MDS

- Therefore,

$$d^2_{E,ii'} = Q_{ii} + Q_{i'i'} - 2Q_{ii'} = d_i - \frac{Tr\,(Q)}{n} + d_{i'} - \frac{Tr\,(Q)}{n} - 2Q_{ii'} =$$

$$= d_i + d_{i'} - d - 2Q_{ii'} \Longrightarrow Q_{ii'} = -\frac{1}{2}\left(d^2_{E,ii'} - d_i - d_{i'} + d\right)$$

  which allows to construct the matrix Q by means of the matrix $D^{(2)}_E$.

- Indeed, from the last expression, it can be checked that:

$$Q = -\frac{1}{2}\left(I_n - \frac{1}{n}1_n 1_n'\right) D^{(2)}_E \left(I_n - \frac{1}{n}1_n 1_n'\right) = -\frac{1}{2} P_n D^{(2)}_E P_n$$

  where P_n is the projection matrix.

- Consequently, we can recover the matrix Q from the matrix $D^{(2)}_E$.

Metric MDS

- Now we turn to the problem of how to obtain the matrix X when the matrix Q is given.
- Assuming that Q is positive definite of rank p (remember that $Q = \widetilde{X}\widetilde{X}'$ and $\widetilde{X}$ has dimension $n \times p$ with $p < n$), it can be represented by:

$$Q = V_p \Lambda_p V_p'$$

where V_p is a $n \times p$ matrix containing the eigenvectors corresponding to nonzero eigenvalues of Q and Λ_p is a $p \times p$ diagonal matrix containing the eigenvalues.

- We write:

$$Q = \left(V_p \Lambda_p^{1/2}\right)\left(V_p \Lambda_p^{1/2}\right)' = Y_p Y_p'$$

where $Y_p = V_p \Lambda_p^{1/2}$ is a $n \times p$ matrix with p uncorrelated variables that reproduce the initial metric.

Metric MDS

- Is Y_p the matrix $\widetilde{X}$ that has lead to the matrix Q that leads to the distance matrix $D_E^{(2)}$?
- The answer is no almost surely, because there is an indeterminacy in the problem when the only information available are the distances.
- In fact, the distances between variables do not vary if:
  1. We modify the means of the variables.
  2. We multiply $\widetilde{X}$ by an orthogonal matrix A as follows:

$$Q = \widetilde{X}\widetilde{X}' = \widetilde{X}AA'\widetilde{X}'$$

- Therefore, from $D_E^{(2)}$, it is only possible to obtain a rotation from the data matrix $\widetilde{X}$ given by the matrix Y_p, which is called the matrix of principal coordinates.

Metric MDS

- In practice, the distance matrix D may not be the matrix of Euclidean distances.
- However, the metric MDS solution considers the way of obtaining the principal coordinates as if D is the matrix of Euclidean distances.
- Therefore, in practice we assume that we have a $n \times n$ distance matrix D.
- The procedure to obtain the principal coordinates is:
  1. Obtain the matrix of squared distances or dissimilarities $D^{(2)}$ just computing the square of each element of D.
  2. Construct the matrix $Q = -\frac{1}{2} P_n D^{(2)} P_n$.
  3. Obtain the eigenvalues of Q. Take the r largest eigenvalues, where r is chosen so that the remaining $n - r$ eigenvalues are much smaller than the first ones.
  4. Define the matrix of r principal coordinates:

$$Y_r = V_r \Lambda_r^{1/2}$$

Metric MDS

- Note that given the matrix D there is no reason to think that all the eigenvalues of $Q = -\frac{1}{2} P_n D^{(2)} P_n$ should be positive.
- Then, in the algorithm, we consider the r largest eigenvalues of the matrix Q and discard the others, including the negative ones.
- This is necessary because, in order to compute the principal coordinates, we need to compute the square root of the Λ_r, and thus, the eigenvalues of Q considered should be be nonnegative.
- A precision measure of the principal components obtained from the r positive eigenvalues of the squared distance matrix $D^{(2)}$ by means of the following coefficient:

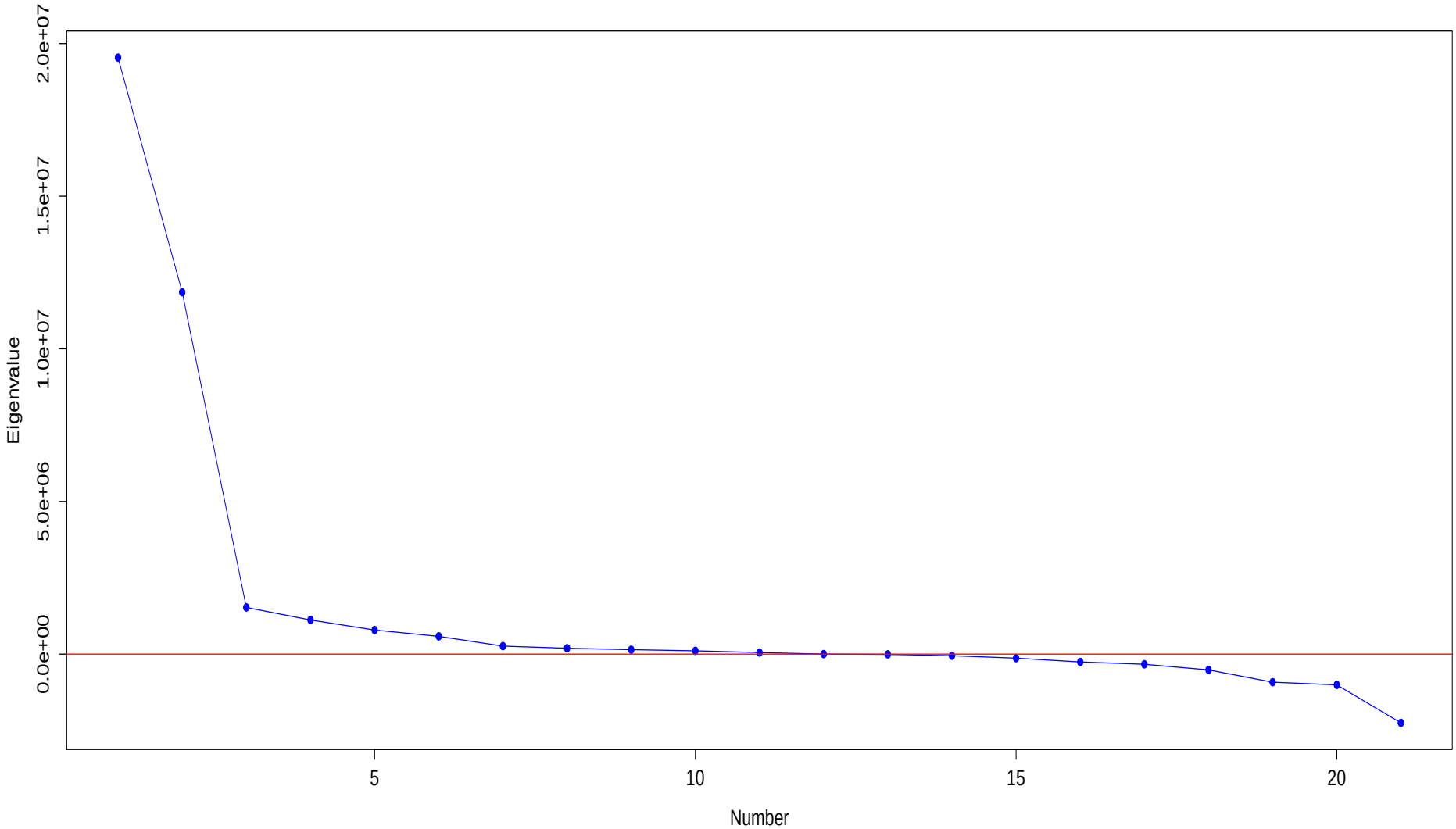
$$m_r = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|}$$

Illustrative example (I)

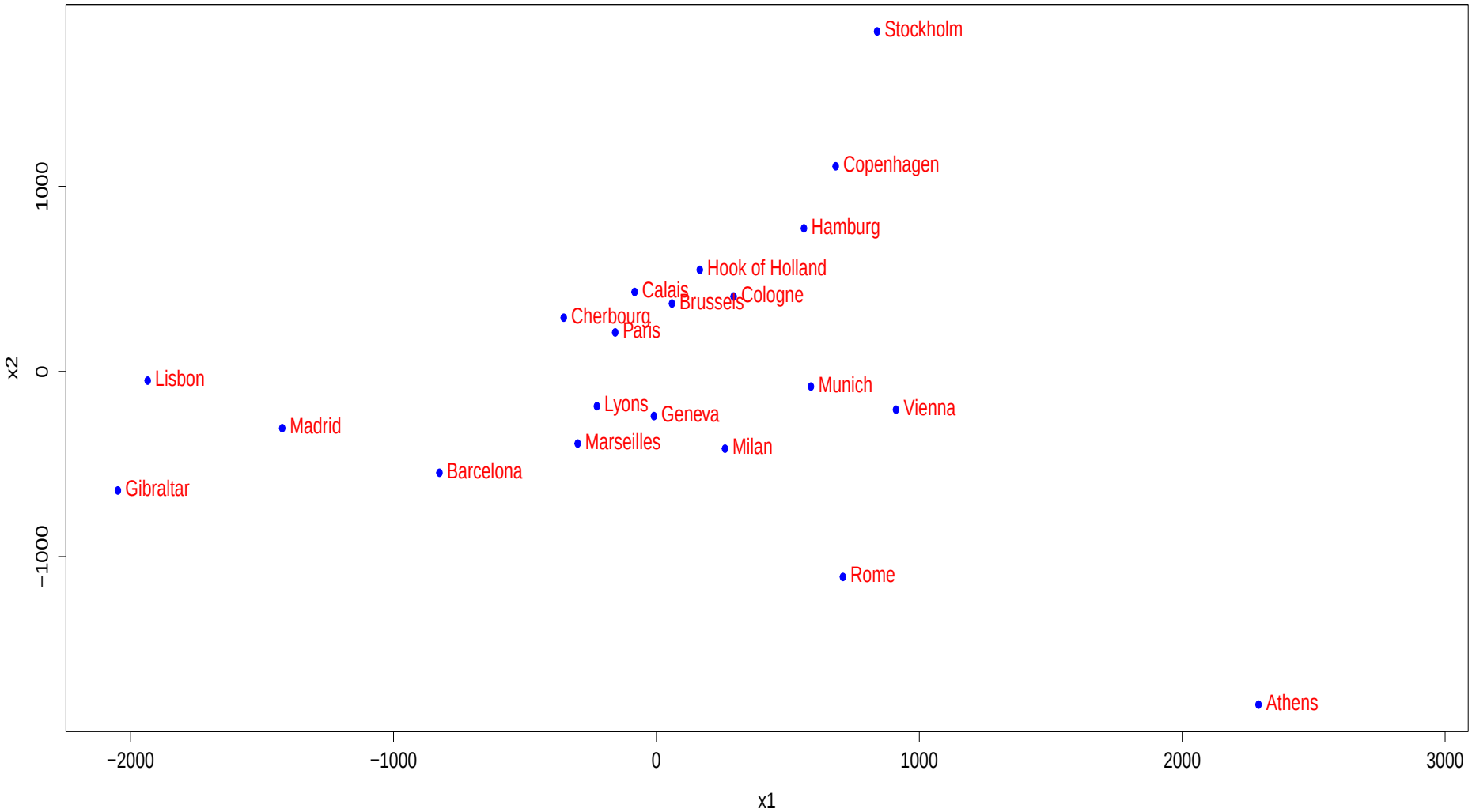
- A good example of how metric MDS works is the following.
- We have a distance matrix with the road distances (in km) between 21 cities in Europe.
- The problem is to recreate the map that has generated the road distances.
- Metric MDS is a method for solving this problem in arbitrary dimensions.
- The next two slides shows the eigenvalues of the matrix Q obtained using metric MDS and the corresponding solution taking $r = 2$.
- The precision of this solution is:

$$m_2 = \frac{31394932}{41651413} = 0.7537$$

Illustrative example (I)



Illustrative example (I)



Metric MDS

- When the original data are in the matrix $\widetilde{X}$ and we construct the matrix $D^{(2)}$ using the Euclidean distances between the points with the original variables, then the principal coordinates obtained from matrix $D^{(2)}$ are equivalent to the principal components of the variables.
- Indeed, if $Q = \widetilde{X}\widetilde{X}'$, it is not difficult to show that the r-th principal component of X, Z_r, is proportional to the r-th principal coordinate, i.e., $Y_r = aZ_r$, for certain value of a.

Non-metric MDS

- Problems of non-metric MDS start from a matrix of differences or dissimilarities between objects that have been obtained via enquiries or from procedures of ordering the elements.
- For example, non-metric MDS is applied to the study of dissimilarities between people's attitudes, preferences or perceptions about political or social affairs or in order to evaluate preferences for products or services in marketing and quality.
- It is thought that the dissimilarity matrix is related to the (real) distance matrix, but in a complex way.
- For instance, it is accepted that the judges, in their assessment, use certain variables or dimensions, however, this also means that the data include elements of error and personal variability.

Non-metric MDS

- Therefore, the variables that explain the dissimilarities between the elements being compared will determine the (true) distances between them, $d_{ii'}$, which are related to the dissimilarities given, $\delta_{ii'}$, by means of an unknown function:

$$\delta_{ii'} = f\left(d_{ii'}\right)$$

  where the only constraint imposed is that f is a monotonous function, meaning that:

$$\delta_{ii'} > \delta_{ii''} \iff d_{ii'} > d_{ii''}$$

- The objective is to try to find the principal coordinates corresponding to the unknown distances $d_{ii'}$, for $i, i^{prime} = 1, \ldots, n$, using only the constraint of monotonicity and using the dissimilarities given.

Non-metric MDS

- For that the usual approach is to use the Shepard-Kruskal algorithm:
  1. Use the metric MDS in the dissimilarities to obtain an initial set of principal coordinates.
  2. Compute the Euclidean distances between the obtained principal coordinates.
  3. Regress these Euclidean distances on the dissimilarities taking into account the monotonicity constraint (not entering into details here).
  4. Compare the Euclidean distances with the predicted values given by the regression using the STRESS:
  $$S^2 = \frac{\sum_{i<i'} \left(\delta_{ii'} - \widehat{d}_{E,ii'}\right)^2}{\sum_{i<i'} \delta_{ii'}^2}$$
  where $\widehat{d}_{E,ii'}$ are the predicted Euclidean distances from the regression.
  5. Replace the original Euclidean distances with the predicted distances and repeat the process until the STRESS is very small.

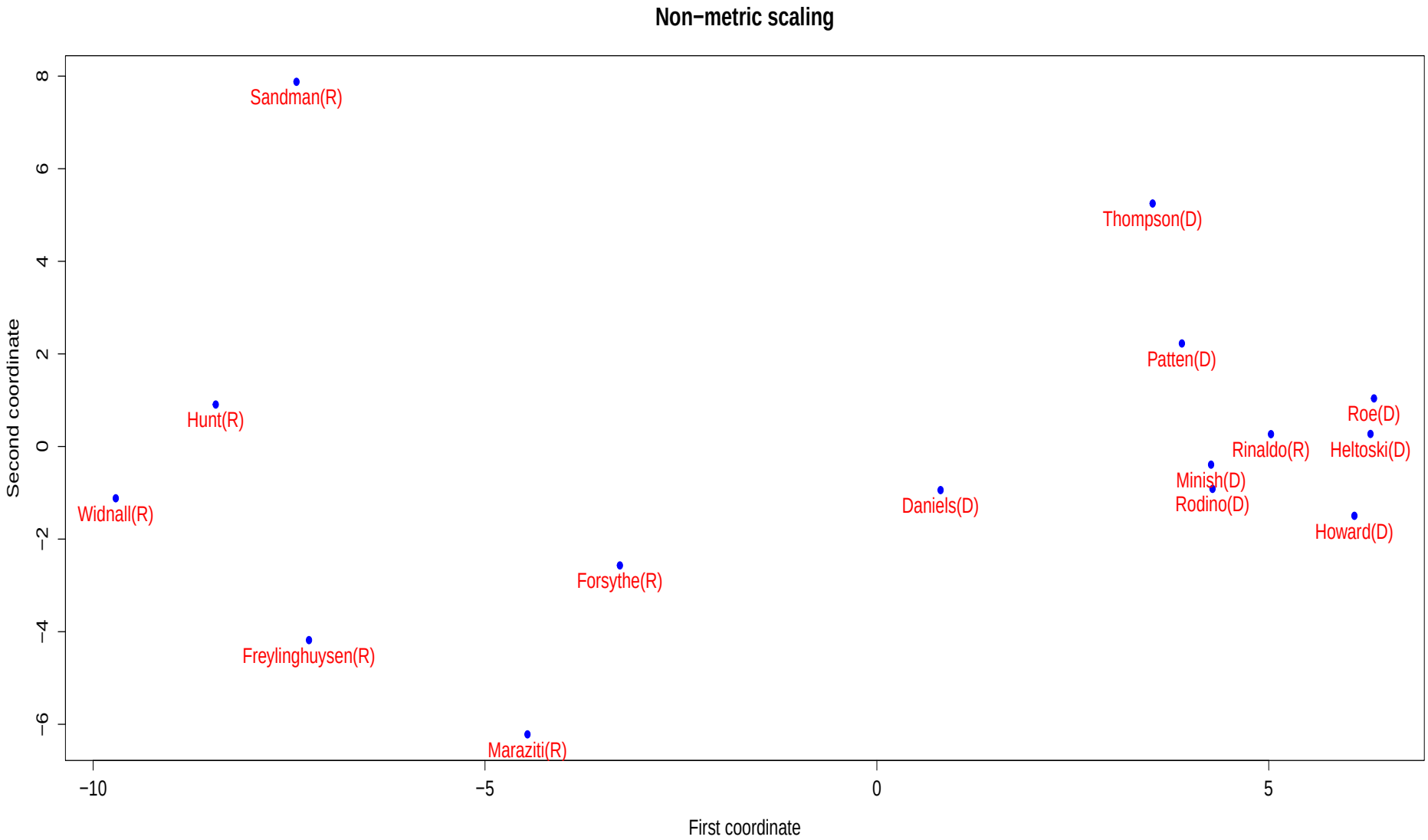
Non-metric MDS

- The dimension taken is usually $r = 2$, in order to ease the graphical representation of the data.
- Anyway, the number of dimensions needed for a good representation of the data can be estimated by testing different values of r and studying the evolution of the criterion in a similar way to that of determining the number of principal components.

Illustrative example (II)

- We consider a data set that shows the number of times 15 congressmen from New Jersey voted differently in the House of Representatives on 19 environmental bills.
- Abstentions are not recorded.
- The question is whether party affiliations can be detected in the data.
- We apply non-metric scaling to the voting behavior shown in the data set. We plot in the next slide the two-dimensional solution.
- The figure suggests that voting behavior is essentially along party lines, although there is more variation among Republicans.
- The voting behavior of one of the Republicans (Rinaldo) seems to be closer to his Democratic colleagues rather than to the voting behavior of other Republicans.

Illustrative example (II)



Chapter outline

- We are ready now for:

Chapter 6: Cluster analysis