Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Master thesis

Multidimensional Scaling for Big Data

Cristian Pachón García

Advisor: Pedro Delicado Useros

Dept. d'Estadística i Investigació Operativa

Contents

Chapter 1

Classical Multidimensional Scaling

1.1 Introduction to Multidimensional Scaling

Multidimensional Scaling (MDS) is a method that represents measurements of similarity (or dissimilarity) among pairs of objects as distances between points of a low-dimensional space. The data, for example, may be correlations among intelligence tests and the MDS representation is a plane that shows the tests as points. The graphical display of the correlations provided by MDS enables the data analyst to literally ”look” at the data and to explore the structure visually. This often shows regularities that remain hidden when studying arrays of numbers. Another application of MDS is to use some of its mathematical as models for dissimilarities judgements. For example, given two objects of interest, one may explain their perceived dissimilarity as the result of a mental arithmetic that mimics the distance formula. According to this model, the mind generates impression of dissimilarity by adding up the perceived differences of the two objects over this properties.

Given a square matrix $\mathbf{D}$ $n \times n$, the goal of MDS is to obtain a set of orthogonal variables $y_1, ..., y_p$ called *principal coordinates*, where $p < n$, such that the euclidean distances of the elements with respect of these variables are equal to the matrix $\mathbf{D}$. Therefore, the aim is to obtain a matrix $\mathbf{X}$ $n \times p$ that could be interpreted as the matrix of p variables for the n observations, where the euclidean distance between the elements could be approximated by $\mathbf{D}$.

This approach arises two questions: is it (always) possible to find these variables? How are they obtained? In general, it is not possible to find a set of p variables that reproduces *exactly* the initial distance. However, it is possible to find a set of variables which distance is approximately the initial distance matrix $\mathbf{D}$.

A classical example, consider the distances between European cities as in the table 1.1. One would like to get a representation in a 2-dimensional space such that the distances would be almost the same as in the table 1.1. The representation of these corrdinates are displayed in figure 1.1.

MDS methods can be divided into two groups: *Metric MDS* and *Non-metric MDS*. Metric MDS, also known as principal coordinates, use the differences between similarities. However, Non-metric MDS states that if a is more similar to b than c, then a is closer to b than c, but the differences between the similarities ab and ac do not have any interpretation. This thesis is focused on the Metric MDS.

	Athens	Barcelona	Brussels	Calais	Cherbourg
Athens	0.00	3313.00	2963.00	3175.00	3339.00
Barcelona	3313.00	0.00	1318.00	1326.00	1294.00
Brussels	2963.00	1318.00	0.00	204.00	583.00
Calais	3175.00	1326.00	204.00	0.00	460.00
Cherbourg	3339.00	1294.00	583.00	460.00	0.00

Table 1.1: Distances between European cities

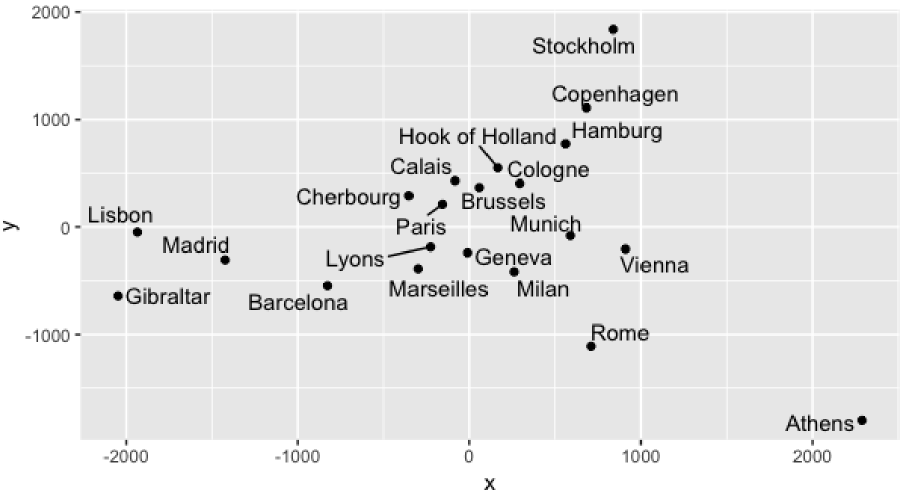


Figure 1.1: MDS on the Eurepean cities.

1.2 Principal coordinates

Given a Matrix $\mathbf{X}$ $n \times p$, the matrix of n individuals over p variables, it is possible to obtain a new one with mean 0 from the previous one:

$$\widetilde{\mathbf{X}} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X} = \mathbf{P}\mathbf{X}$$

where

$$\mathbf{P} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)$$

This new matrix, $\widetilde{\mathbf{X}}$, has the same dimensions as the orginial one but it is centered in $\mathbf{0}$. From this matrix, it is possible to build two square semi-positive definite matrices: the covariance matrix $\mathbf{S}$, defined as $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/n$ and the cross-prodructs matrix $Q = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$. The last matrix can be interpreted as a similarity matrix between the n elements. The term ij is obtained as follows:

$$q_{ij} = \sum_{s=1}^{p} x_{is}x_{js} = \mathbf{x_i'x_j} \tag{1.1}$$

where $\mathbf{x_i'}$ is the i-th row from $\widetilde{\mathbf{X}}$. Given the scalar product formula, $\mathbf{x_i'x_j} =| \mathbf{x_i} || \mathbf{x_i} | cos\theta_{ij}$, if the elements i and j have similar coordinates, then $cos\theta_{ij} \simeq 1$ and q_{ij} will be large. On the contrary, if the elements are very different, then $cos\theta_{ij} \simeq 0$ and q_{ij} will be small. So, $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ can be interpreted as the similarity matrix between the elements.

The distances between elements can be deduced from the similarity matrix. The euclidean distance between two elements is calculated in the following way:

$$d_{ij}^2 = \sum_{s=1}^{p} (x_{is} - x_{js})^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js} \tag{1.2}$$

This expression can be obtained directly from the matrix $\mathbf{Q}$:

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} \tag{1.3}$$

We have just seen that, given the matrix $\widetilde{\mathbf{X}}$, it is possible to get the similarity matrix $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ and from it, to get the distance matrix $\mathbf{D}$. Let $diag(\mathbf{Q})$ be the vector that contains the diagonal terms of $\mathbf{Q}$ and $\mathbf{1}$ be the vector of ones, the matrix $\mathbf{D}$ is given by:

$$\mathbf{D} = diag(\mathbf{Q})\mathbf{1}' + \mathbf{1}diag(\mathbf{Q})' - 2\mathbf{Q}$$

The problem we are dealing with goes in the opposite direction. We want to rebuild $\widetilde{\mathbf{X}}$ from a square distance matrix $\mathbf{D}$, with elements d_{ij}^2. The first step is to obtain $\mathbf{Q}$ and afterwards, to get $\widetilde{\mathbf{X}}$. *Daniel Peña* develops in his book[1] the theory needed to get the solution. Here, we summarise it.

The first step is to find out a way to obtain the matrix $\mathbf{Q}$ given $\mathbf{D}$. We can assume without loss of generality that the mean of the variables is equal to 0. This is a consequence of the fact that the distance between two points remains the same if the variables are expressed in terms of the mean:

$$d_{ij}^2 = \sum_{s=1}^{p} (x_{is} - x_{js})^2 = \sum_{s=1}^{p} [(x_{is} - \overline{x_s}) - (x_{js} - \overline{x_s})]^2 \tag{1.4}$$

The previous condition means that we are lookig for a matrix $\widetilde{\mathbf{X}}$ such that $\widetilde{\mathbf{X}}'\mathbf{1} = 0$. It also means that $\mathbf{Q1} = 0$, i.e, the sum of all the elements of a column of $\mathbf{Q}$ is 0. Since the matrix is symmetric, the previous condition should state for the rows as well.

To establish these constrains, we sum up 1.3 at row level:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + nq_{jj} = t + nq_{jj} \tag{1.5}$$

where $t = \sum_{i=1}^{n} q_{ii} = trace(\mathbf{Q})$, and we have used that the condition $\mathbf{Q1} = 0$ implies $\sum_{i=1}^{n} q_{ij} = 0$. Summing up 1.3 at column level:

$$\sum_{i=1}^{n} d_{ij}^2 = t + nq_{ii} \tag{1.6}$$

Summing up 1.5 we obtain:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nt \tag{1.7}$$

Replacing in 1.3 q_{jj} obtained in 1.5 and q_{ii} obtained in 1.6, we have the following expression:

$$d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} - 2q_{ij} \tag{1.8}$$

Let $d_{i.}^2 = \frac{1}{n}\sum_{j=1}^n d_{ij}^2$ and $d_{.j}^2 = \frac{1}{n}\sum_{i=1}^n d_{ij}^2$ be the row-mean and column-mean. Using 1.7, we have that:

$$d_{ij}^2 = d_{i.}^2 + d_{.j}^2 - d_{..}^2 - 2q_{ij} \tag{1.9}$$

where $d_{..}$ is the mean of all the elements of $\mathbf{D}$, given by:

$$d_{..}^2 = \frac{1}{n^2}\sum\sum d_{ij}^2$$

Finally, from 1.9 we get the following expression:

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2) \tag{1.10}$$

The previous expression shows how to build the matrix of similarities $\mathbf{Q}$ from the distance matrix $\mathbf{D}$.

The next step is to obtain the matrix $\mathbf{X}$ given the matrix $\mathbf{Q}$. Let's suppose that the similarity matrix is positive definite of range p, it can be represented by:

$$\mathbf{Q} = \mathbf{V\Lambda V'}$$

where $\mathbf{V}$ is a $n \times p$ matrix that contains the eigenvectors with eigenvalues not nulls of $\mathbf{Q}$, $\mathbf{\Lambda}$ is a diagonal matrix $p \times p$ that contains the eigenvalues.

Re-writing the previous expression, we obtain:

$$\mathbf{Q} = (\mathbf{V\Lambda}^{1/2})(\mathbf{\Lambda}^{1/2}\mathbf{V'}) \tag{1.11}$$

Getting:

$$\mathbf{Y} = \mathbf{V\Lambda}^{1/2}$$

we have obtained a matrix with dimension $n \times p$ with p uncorrelated variables that reproduce the initial metric. It is important to notice that if one starts from $\mathbf{X}$ (i.e $\mathbf{X}$ is known) and calculates from these variables the distance matrix in 1.2 and after that it is applied the method explained, the matrix obtained is not the same as $\mathbf{X}$, but its principal components. This happens since the distance between elements does not change if:

- The mean values are modified.
- Poits are rotated, i.e, multiplications by orthogonal matrices.

By 1.3, the distance is a function of the terms of the similarity matrix $\mathbf{Q}$ and this matrix is invariant given any rotation, reflexion or translation of the variables:

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

for any orthogonal $\mathbf{A}$ matrix. The matrix $\mathbf{Q}$ only contains information about the space generated by the variables $\mathbf{X}$. Any rotation, reflexion or translation keeps the distance unchaged. Therefore, the solution is not unique

1.3 Building principal coordinates

In general, the distance matrix is not compatible with an euclidean metric but usually the similarity matrix obtained from it has p positive eigenvalues and greater than the other ones. If the rest $n-p$ not null eigenvalues are much less than the other ones, it is possible to obtain an (approximated) representation using the p eigenvectors associated with the first p eigenvalues of the similarity matrix.

Let's suppose that we have a square distance matrix $\mathbf{D}$. The process to obtain the *principal coordinates* is:

1. Build the matrix $\mathbf{Q} = -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{PDP}$ of cross-products.

2. Obtain the eigenvalues of $\mathbf{Q}$. Take the r greatest eigenvalues. Since $\mathbf{P1} = 0$, where $\mathbf{1}$ is a vector of ones, $range(\mathbf{Q}) = n-1$, being the vector $\mathbf{1}$ an eigenvector with eigenvalue 0.

3. Obtain the coordinates of the individuals in the variables $\mathbf{v_i}\sqrt{\lambda_i}$, where λ_i is an eigenvalue of $\mathbf{Q}$ and $\mathbf{v_i}$ is the associated unitary eigenvector. This implies that $\mathbf{Q}$ is apporximated by:

$$\mathbf{Q} \approx (\mathbf{V_r}\mathbf{\Lambda}^{1/2})(\mathbf{\Lambda_r}^{1/2}\mathbf{V_r'})$$

4. Take as coordinates of the points the following variables:

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

The method can also be applied if the initial information is not a distance matrix but a similarity matrix. A *similarity function* between two element i and j s_{ij} is defined as:

- $s_{ii} = 1$
- $0 \leq s_{ij} \leq 1$
- $s_{ij} = s_{ji}$

If the initial information is $\mathbf{Q}$, a similariy matrix, then $q_{ii} = 1$, $q_{ij} = q_{ji}$ and $0 \leq q_{ij} \leq 1$. The associated distance matrix (by 1.3):

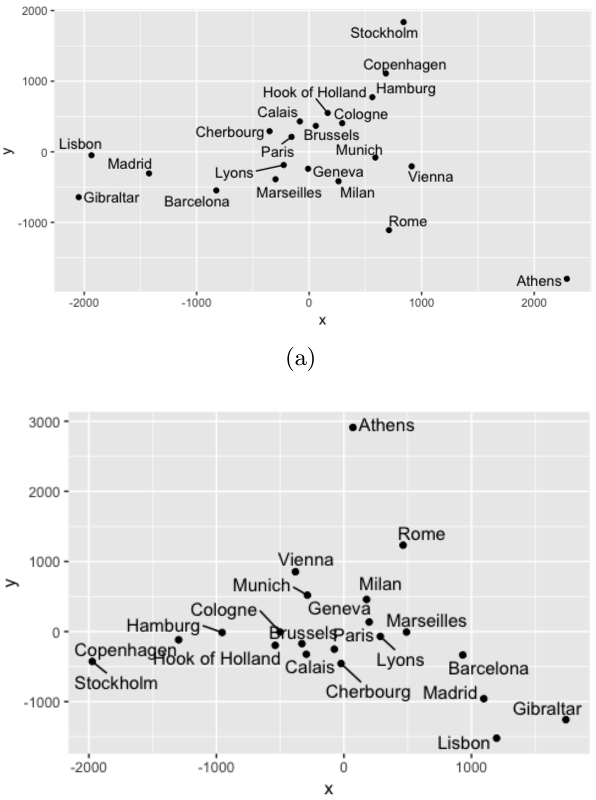
$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{qij} = 2(1 - q_{ij})$$

and it is easy to see that $\sqrt{2(1-q_{ij})}$ is a distance and it verifies the triangle inequality.

1.4 Procrustes transformation

As we have mentioned before, the MDS solution is not unique. One example of it is shown in figure 1.2.

Since rotations, translations, reflections and dilations are distance-preserving functions, one can find two different MDS configurations for the same set of data. How



(a)

(b)

Figure 1.2: Two different solutions of MDS

is it possible to align both solutions? This problem is solved by means of *Procrustes transformations.*

The Procrustes problem is concern with fitting a configuration (testee) to another (target) as closely as possible. In the simple case, both configurations have the same dimensionality and the same number of points, which can be brought into 1-1 correspondence by substantive considerations. Under orthogonal transformations, the testee can be fitted it to the target. In addition to such rigid motions, one may also allow for dilations and for shifts.

Ingwer Borg and *Patrik Groenen* detail all the steps needed to obtain the solution[2]. This is out of the scope of this thesis. However, since it has been a repeatdly used tool, we briefly summarise it.

Let **A** and **B** be two different MDS configurations for the same set of data. Without loss of generality, let's suppose that the target is **A** and the testee is **B**. One wants to obtain s, **T** and t such that:

$$\mathbf{A} = s\mathbf{B}\mathbf{T} + \mathbf{1}\mathbf{t}'$$

where **T** is an orthogonal matrix. As mentioned before, *Ingwer Borg* and *Patrik Groenen* develop all the theory[2] that allows to get these parameters.

1.5 Multidimensional Scaling with R

All the algorithms have been coded in R, since it has a widely statistics packages already implemented. We have used two packages for developing our MDS approaches:

- Package: stats. From this one we have used the function cmdscale to do the MDS. The output of this function is:
  - The new coordinates for the individuals.
  - All the eigenvalues found.
- Package: MCMCpack. From this one we have used the function procrustes to do the Procrustes transformation. The output of this function is:
  - The dilation coeficient s.
  - The orthogonal matrix $\mathbf{T}$.
  - The translation vector $\mathbf{t}$.

Chapter 2

Algorithms for Multidimensional Scaling with Big Data

2.1 Divde and Conquer Multidimensional Scaling

2.2 Fast Multidimensional Scaling

2.3 Multidimensional Scaling based on Gower interpolation

Chapter 3

Simulation study

Chapter 4

Conclusions

Bibliography

[1] Daniel Peña. *Análisis de datos multivariantes.* McGraw Hill, Madrid, Spain, 2002.

[2] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer, 2005.

Appendix A

Code