# Factor analysis and multidimensional scaling

**Rui Zhu**

# Overview

1. Factor analysis
2. Multidimensional scaling

# Factor analysis

Multivariate data are often viewed as multiple indirect measurements arising from an underlying source, which typically cannot be directly measured.

- Educational and psychological tests use the answers to questionnaires to measure the underlying intelligence and other mental abilities of subjects.
- The trading prices of stocks change constantly over time, and reflect various unmeasured factors such as market confidence, external influences, and other driving forces that may be hard to identify or measure.
- EEG brain scans measure the neuronal activity in various parts of the brain indirectly via electromagnetic signals recorded at sensors placed at various positions on the head.

# Factor analysis

Factor analysis is a classical technique developed in the statistical literature to identify these latent sources.

We are going to identify the latent sources $S_1, S_2, \dots, S_q$ from the direct measurements $X_1, X_2, \dots, X_p$, where $p > q$.

$$
\begin{aligned}
X_1 &= a_{11}S_1 + \cdots + a_{1q}S_q + \epsilon_1 \\
X_2 &= a_{21}S_1 + \cdots + a_{2q}S_q + \epsilon_2 \\
&\cdots \\
X_p &= a_{p1}S_1 + \cdots + a_{pq}S_q + \epsilon_p
\end{aligned}
$$

- $S$: the vector of $q$ underlying latent variables or factors
- $\mathbf{A} \in \mathbb{R}^{p \times q}$: the matrix of factor loadings
- $\epsilon_j$: uncorrelated zero-mean disturbances

# Factor analysis

- $S_l$ are common sources of variation amongst $X_j$ and account for their correlation structure; each of them has zero mean and standard deviation of one
- $S_l$ are uncorrelated with each other
- $\epsilon_j$ are uncorrelated and unique to each $X_j$, and pick up the remaining unaccounted variation
- $S_l$ and $\epsilon_j$ are usually modelled as Gaussian random variables
- The variance of $X_j$ is composed of two parts: 1) $a_{j1}^2 + a_{j2}^2 + \ldots + a_{jq}^2$, the variance of $X_j$ explained by the common factors, and 2) $\sigma_j^2$, the specific variance

# Factor analysis

A linear two-factor model was fitted to some subject marks data.

The first factor measures overall ability in the six subjects, while the second contrasts humanities and mathematics subjects.

| Subject | $\hat{a}_{j1}$ | $\hat{a}_{j2}$ |
|---|---|---|
| Gaelic | 0.56 | 0.43 |
| English | 0.57 | 0.29 |
| History | 0.39 | 0.45 |
| Arithmetic | 0.74 | -0.28 |
| Algebra | 0.72 | -0.21 |
| Geometry | 0.60 | -0.13 |

# Factor analysis

A two-factor model was also fitted to some children's personality trait data.

The first factor represents some overall personality measure, while the second contrasts indicators, such as sociability, of how a child relates to other people with those which are internal to the individual, like guilt.

| Variable (personality trait) | $\hat{a}_{j1}$ | $\hat{a}_{j2}$ |
|---|---|---|
| Mannerliness | 0.65 | 0.57 |
| Approval seeking | 0.54 | 0.54 |
| Initiative | 0.61 | -0.45 |
| Guilt | 0.63 | -0.54 |
| Sociability | 0.56 | 0.54 |
| Creativity | 0.72 | -0.59 |
| Adult role | 0.67 | -0.45 |
| Cooperativeness | 0.64 | 0.60 |

# Factor analysis: communality

Communality: the proportion of the variance that is explained by the common factors.

The larger the communality, the better does the variable serve as an indicator of the associated factors.

| | Communalities |
|---|---|
| Gaelic | 0.49 |
| English | 0.41 |
| History | 0.36 |
| Arithmetic | 0.62 |
| Algebra | 0.56 |
| Geometry | 0.37 |

The sum of the communalities is the variance explained by the factor model. This is 2.81 or 47% of 6 which is the total variance for the subject marks data.

# Factor analysis: choice of the number of factors

PCA: the number of components needed is often a good guide to the number of factors.

# Factor analysis: rotation

When the first factor solution does not reveal a structure of the loadings, it is customary to apply rotation in an effort to find another set of loadings that fit the observations equally well but can be more easily interpreted. For example,

- Varimax procedure attempts to find an orthogonal rotation that is close to simple structure by finding factors with few large loadings and as many near-zero loadings as possible.
- Oblique rotation requires us to relax the original assumption of the linear factor model that the latent variables are uncorrelated. An oblique rotation leads to correlated factors. Although this complicates the interpretation of the factors, it is often reasonable to expect the latent variables to be correlated; e.g. one might expect a child's mathematical ability to be positively correlated with their verbal ability.

# Factor analysis: factor scores

- Factor score: an individual's score on the latent variable(s), perhaps for use in subsequent analysis.
- Not straightforward: the factors are random variables which have a probability distribution.

$$
\begin{aligned}
\hat{s}_{i1} &= c_{11}x_{i1} + c_{12}x_{i2} + \ldots + c_{1p}x_{ip} \\
\hat{s}_{i2} &= c_{21}x_{i1} + c_{22}x_{i2} + \ldots + c_{2p}x_{ip} \\
&\vdots \\
\hat{s}_{iq} &= c_{q1}x_{i1} + c_{q2}x_{i2} + \ldots + c_{qp}x_{ip}.
\end{aligned}
$$

$c$: estimated factor score coefficients

# FA vs. PCA

- PCA looks for linear combinations of the data that are uncorrelated and of high variance, whilst FA seeks unobserved linear combinations of the variables representing underlying fundamental quantities.
- PCA makes no assumptions, whilst FA assumes that the data comes from a well-defined model in which specific assumptions hold.
- FA models the correlation structure of $X_j$, while PCA models covariance structure. This is an important distinction between FA and PCA.

# FA vs. PCA

Generate 200 observations of three variates $X_1$, $X_2$ and $X_3$:

$$
\begin{aligned}
X_1 &\sim Z_1 \\
X_2 &= X_1 + 0.001 Z_2 \\
X_3 &= 10 Z_3
\end{aligned}
$$

where $Z_1$, $Z_2$ and $Z_3$ are independent standard normal variables.

| | PC1 | Factor 1 |
|---|---|---|
| $X_1$ | -0.007 | 0.999 |
| $X_2$ | -0.007 | 0.999 |
| $X_3$ | -1.000 | |

# Multidimensional scaling

**Multidimensional scaling aims to reveal the structure of a dataset by plotting points in one or two dimensions.**

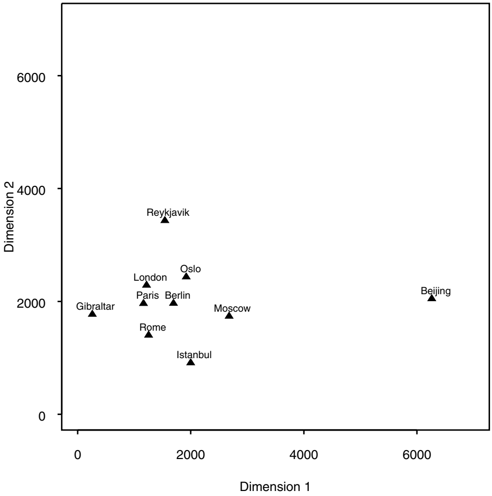
Multidimensional scaling explicitly tries to preserve all pairwise distances.

# Multidimensional scaling

Example: Reproducing a Two-Dimensional Map from Air Distances Between Pairs of Cities

| | London | Berlin | Oslo | Moscow | Paris | Rome | Beijing | Istanbul | Gibraltar | Reykjavik |
|---|---|---|---|---|---|---|---|---|---|---|
| London | - | | | | | | | | | |
| Berlin | 570 | - | | | | | | | | |
| Oslo | 710 | 520 | - | | | | | | | |
| Moscow | 1550 | 1000 | 1020 | - | | | | | | |
| Paris | 210 | 540 | 830 | 1540 | - | | | | | |
| Rome | 890 | 730 | 1240 | 1470 | 680 | - | | | | |
| Beijing | 5050 | 4570 | 4360 | 3600 | 5100 | 5050 | - | | | |
| Istanbul | 1550 | 1080 | 1520 | 1090 | 1040 | 850 | 4380 | - | | |
| Gibraltar | 1090 | 1450 | 1790 | 2410 | 960 | 1030 | 6010 | 1870 | - | |
| Reykjavik | 1170 | 1480 | 1080 | 2060 | 1380 | 2040 | 4900 | 2560 | 2050 | - |

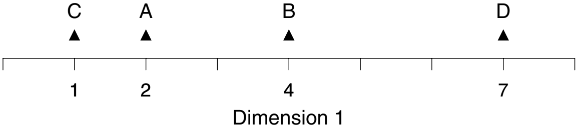
# Multidimensional scaling



# Multidimensional scaling

Suppose that we have four cities labelled A, B, C, and D and that the distances (in hundreds of miles) between the pairs of cities are as given by the following matrix:

$$
\begin{array}{c}
A \\ B \\ C \\ D
\end{array}
\begin{pmatrix}
- & & & \\
2 & - & & \\
1 & 3 & - & \\
5 & 3 & 6 & -
\end{pmatrix}
$$



# Multidimensional scaling

- $d_{ij}$: the inter-point distances in the configuration
- $\delta_{ij}$: the distances in the original distance matrix

# Multidimensional scaling

- Classical MDS: we seek a configuration such that the $d_{ij}$s will be approximately equal to the corresponding $\delta_{ij}$s.
- However, it is often the case, particularly in social science research, that the values of the $\delta_{ij}$ may be interpreted only in an ordinal sense as if, for example, the distances come from subjective similarity ratings.
- Ordinal MDS: the object is only to find a configuration such that the $d_{ij}$s are in the same rank order as the corresponding $\delta_{ij}$s.

# Classical MDS: distance metrics

Euclidean distance between two data points:

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}.$$

- Often the variables (columns in the data matrix) will be standardized prior to calculating distances.

# Classical MDS: distance metrics

Manhattan distance:

$$\delta_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|.$$

- This gives less relative weight to large differences.
- A large difference for a single variable can easily have a dominating effect on the Euclidean distance.

# Classical MDS: distance metrics

- If the features are binary, a measure of similarity could be based on the proportion of variables on which two objects (or individuals) match.
- For categorical data, where the features have more than two categories, a full set of dummy variables can be constructed for each feature.

# Classical MDS

We are looking for a new data matrix, with two or three columns, which is close to the original matrix in the sense that it gives rise to (nearly) the same distance matrix.

Minimise the stress function which is the normalised sum of squares $\sum_{i<j}(d_{ij} - \delta_{ij})^2$:

$$\sqrt{\frac{\sum_{i<j}(d_{ij} - \delta_{ij})^2}{\sum_{i<j} d_{ij}^2}}.$$

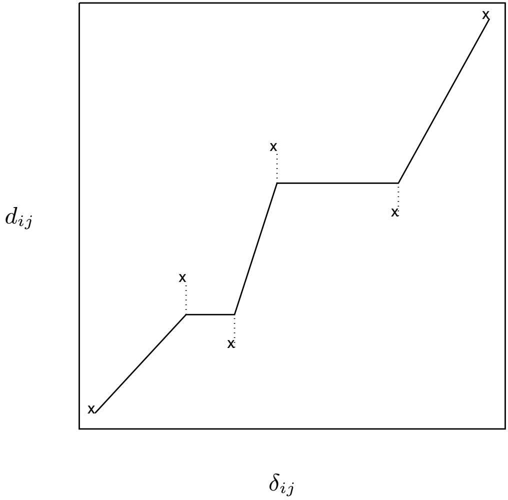
# Ordinal MDS

In some cases, $\delta_{ij}$s can be interpreted only in an ordinal sense, e.g. $\delta_{ij}$s are the result of an experiment where subjects are asked to give their subjective assessments of the distance between objects.

Disparities: $\hat{d}_{ij}$ in the same rank order as $\delta_{ij}$

- $\hat{d}_{ij}$ is the smoothed version of $d_{ij}$, which is achieved by monotonic regression.
- Monotonic regression is to fit a monotonic curve to the points $(d_{ij}, \delta_{ij})$, while making the sum of squared vertical deviations as small as possible.
- $\hat{d}_{ij}$ is the fitted or predicted value of $d_{ij}$ from the monotonic regression.

# Ordinal MDS



# Ordinal MDS

Minimise Kruskal's stress:

$$\sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}.$$

# Ordinal MDS

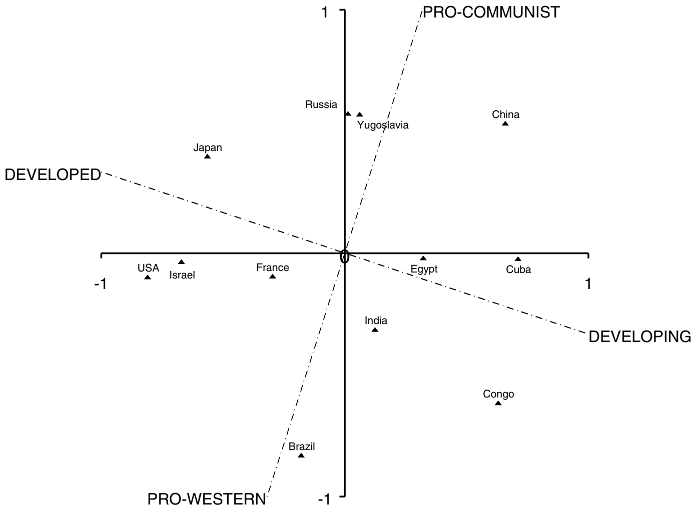
Example: An Attempt to Determine the Dimensions Underlying Similarity Judgements for Pairs of 12 Countries

In 1968, a group of 18 students was asked to rate the degree of similarity between each pair of 12 countries on a scale from 1 ("very different") to 9 ("very similar"). The mean similarity ratings were calculated across students to obtain the similarity matrix.

# Ordinal MDS

| | Brazil | Congo | Cuba | Egypt | France | India | Israel | Japan | China | Russia | USA | Yugoslavia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Brazil | - | | | | | | | | | | | |
| Congo | 4.83 | - | | | | | | | | | | |
| Cuba | 5.28 | 4.56 | - | | | | | | | | | |
| Egypt | 3.44 | 5.00 | 5.17 | - | | | | | | | | |
| France | 4.72 | 4.00 | 4.11 | 4.78 | - | | | | | | | |
| India | 4.50 | 4.83 | 4.00 | 5.83 | 3.44 | - | | | | | | |
| Israel | 3.83 | 3.33 | 3.61 | 4.67 | 4.00 | 4.11 | - | | | | | |
| Japan | 3.50 | 3.39 | 2.94 | 3.83 | 4.22 | 4.50 | 4.83 | - | | | | |
| China | 2.39 | 4.00 | 5.50 | 4.39 | 3.67 | 4.11 | 3.00 | 4.17 | - | | | |
| Russia | 3.06 | 3.39 | 5.44 | 4.39 | 5.06 | 4.50 | 4.17 | 4.61 | 5.72 | - | | |
| USA | 5.39 | 2.39 | 3.17 | 3.33 | 5.94 | 4.28 | 5.94 | 6.06 | 2.56 | 5.00 | - | |
| Yugoslavia | 3.17 | 3.50 | 5.11 | 4.28 | 4.72 | 4.00 | 4.44 | 4.28 | 5.06 | 6.67 | 3.56 | - |

# Ordinal MDS



1
-1
0
1
-1
PRO-COMMUNIST
PRO-WESTERN
DEVELOPED
DEVELOPING
Russia
Yugoslavia
China
Japan
USA
Israel
France
Egypt
Cuba
India
Congo
Brazil

# Ordinal MDS

- Assessing the fit of MDS: have a look at the stress, also consider the interpretability.
- Choose the dimension: use a scree plot of the stress.