Data Analytics for Social Science
Principal components and multidimensional scaling

Johan A. Elkink

School of Politics & International Relations
University College Dublin

14 November 2017

Outline

Unsupervised learning

So far we have looked at models that in the machine learning literature are called **supervised learning**.

The idea here is that the algorithm learns to find some pattern in the data, whereby—at least for the training data—the correct answer is known.

E.g. we know which countries are democratic and which are not, but try to predict based on a set of variables.

With **unsupervised learning**, there is no *a priori* labelling of the data.

Dimension reduction

We can use **dimension reduction** techniques to get low dimensional representations of high dimensional data, for visualisation or analysis.

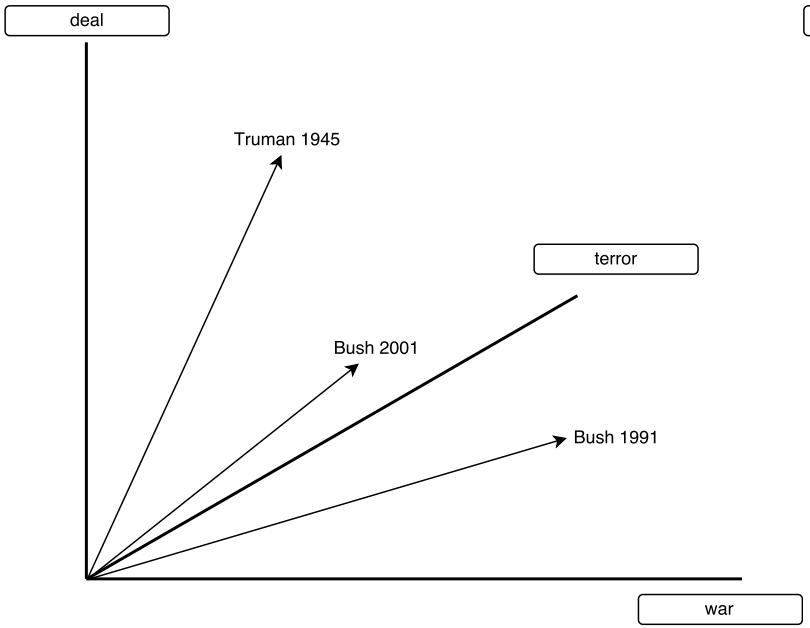
Multidimensional scaling: find a low-dimensional coordinate system based on a distance matrix, so that in the new space, distances between observations remain as similar as possible.

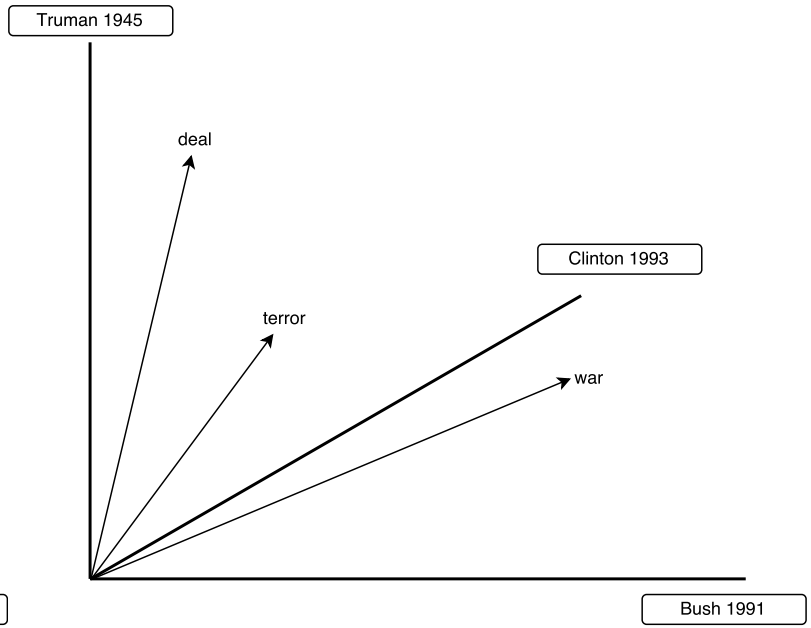
Principal component analysis: find a low-dimensional coordinate system (dimensions) such that as much as possible of the variance in the data is captured by the location on these coordinates.

Factor analysis: find a latent coordinate system (factors) such that the position on these factors can be seen as underlying dimensions that explain the data.

Text in high-dimensional space

Our speech data is an example of high-dimensional data that we want to visualise using low-dimensional representations of the data.



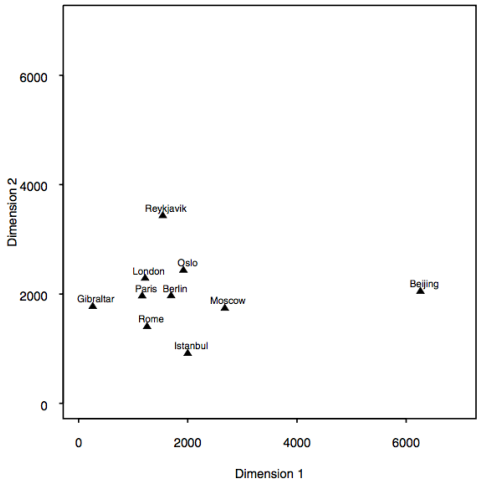


Outline

Multidimensional scaling

Imagine we have a table with distances between cities and we want to reconstruct the original map. We can use **multidimensional scaling (MDS)** to do so.

	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	–



Distances can be any kind of dissimilarity matrix, for example Euclidean distances in the space defined by terms of speeches.

(Bartholomew et al., 2008, 55–58)

MDS: implementation

To calculate distances between observations, we can use the `dist()` function:

```
D <- dist(X)
```

We can then use the output in a **classical MDS** analysis as follows:

```
cmdscale(D)
```

Alternative, **non-metric MDS** procedures exist for variables that are not at a scale level of measurement.

MDS: development example

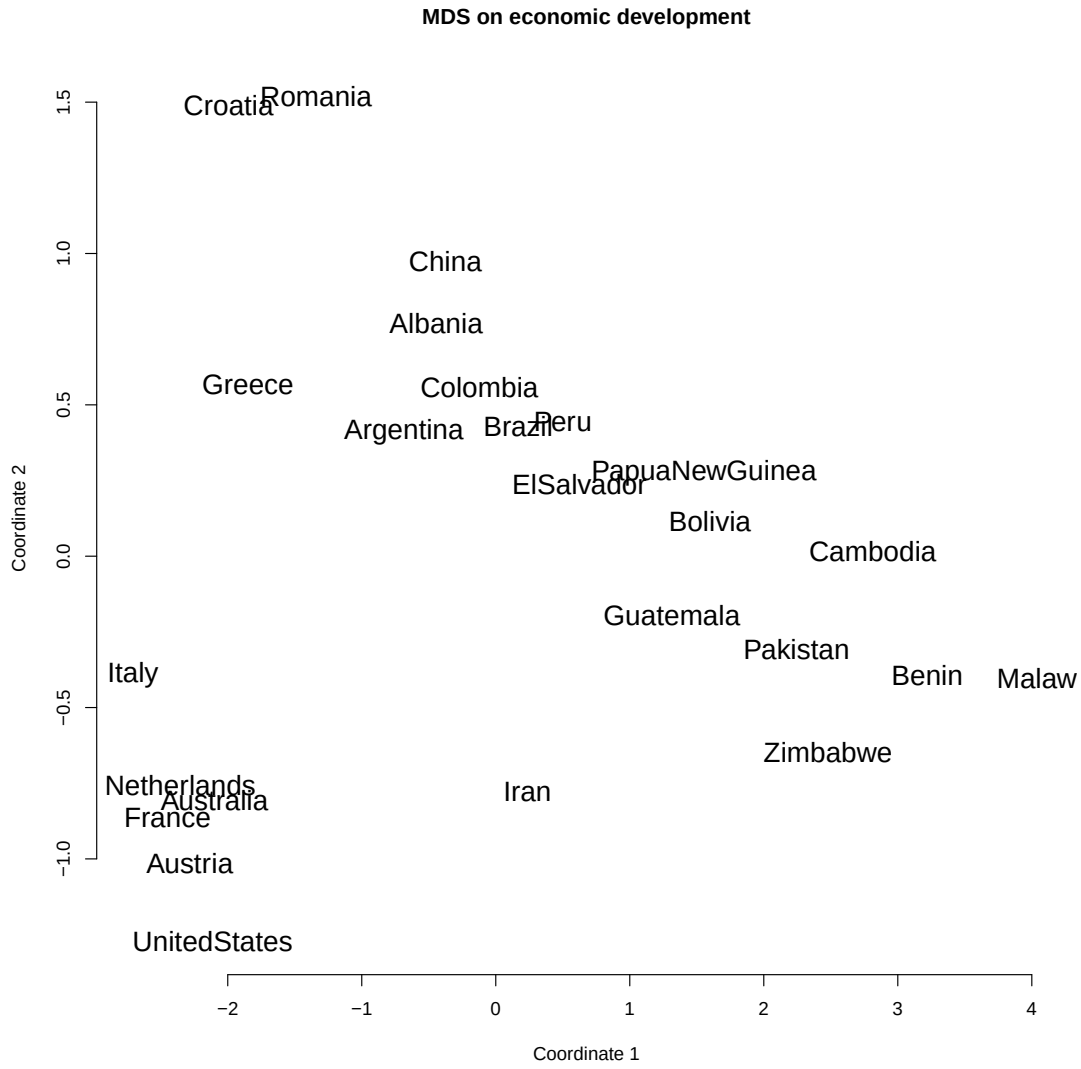
	increase	life	imr	tfr	gdp
Albania	1.20	69.20	30.00	2.90	659.91
Argentina	1.20	68.60	24.00	2.80	4343.04
Australia	1.10	74.70	7.00	1.90	17529.98
Austria	1.00	73.00	7.00	1.50	20561.88
Benin	3.20	45.90	86.00	7.10	398.21
Bolivia	2.40	57.70	75.00	4.80	812.19
Brazil	1.50	64.00	58.00	2.90	3219.22
Cambodia	2.80	50.10	116.00	5.30	97.39
China	1.10	66.70	44.00	2.00	341.31
Colombia	1.70	66.40	37.00	2.70	1246.87
Croatia	-1.50	67.10	9.00	1.70	5400.66

Data set on 25 countries in 1997 containing measures of growth rate, life expectancy, infant mortality, fertility, and GDP.

(Bartholomew et al., 2008, 71)

MDS: development example

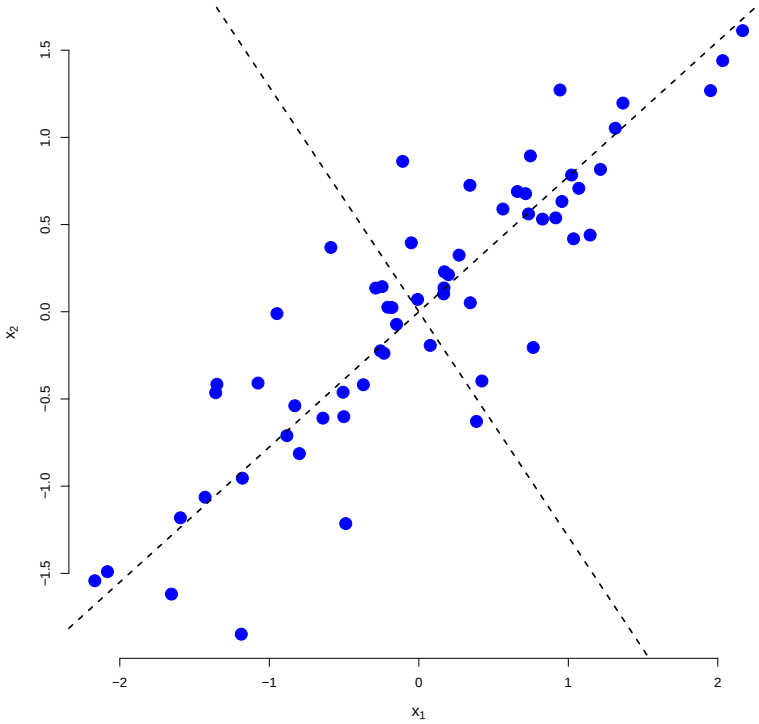
```
plot(cmdscale(dist(scale(econ25))))
```

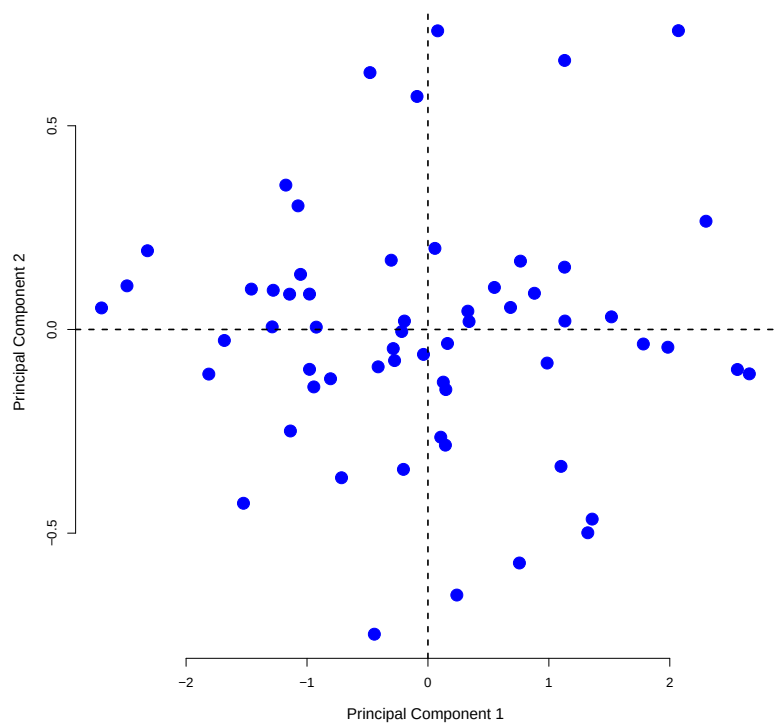


Outline

Principal component analysis

"The main aim of **principal components analysis (PCA)** is to replace p metrical correlated variables by a much smaller number of uncorrelated variables which contain most of the information in the original set."

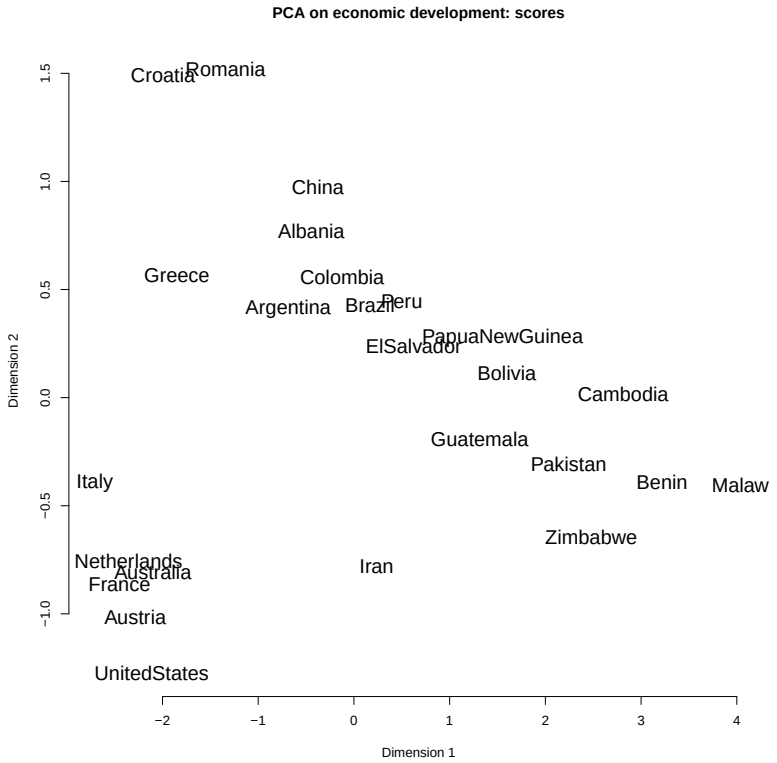


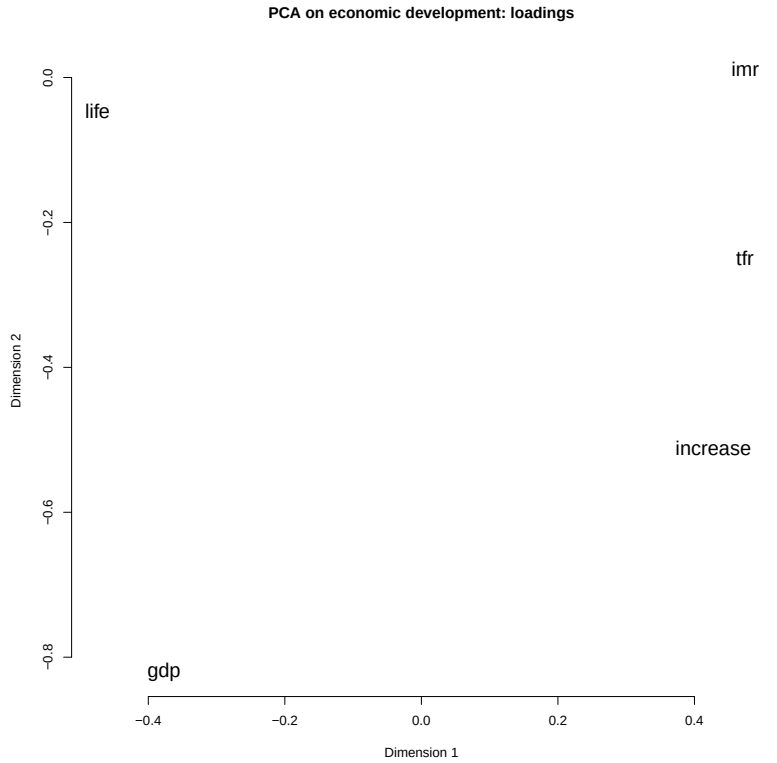


(Bartholomew et al., 2008, 117)

PCA: development example

```
plot(princomp(scale(econ25))$scores)
```





PCA uses the full matrix and MDS only the distances, but when using Euclidean distances, MDS and PCA give the same answer.

(Bartholomew et al., 2008, 135)

PCA: terminology

The underlying dimensions are called the **principal components**.

The projection of each data point on the principal components are called the **scores**.

The projection of each variable on the principal components are called the **loadings**.

The loadings help in substantively, *post hoc*, interpreting the meaning of the dimensions—but this is not really what PCA is designed for.

Outline

Factor analysis

In **factor analysis** we are looking for **latent variables**, unobserved variables that are underlying the structure we observe.

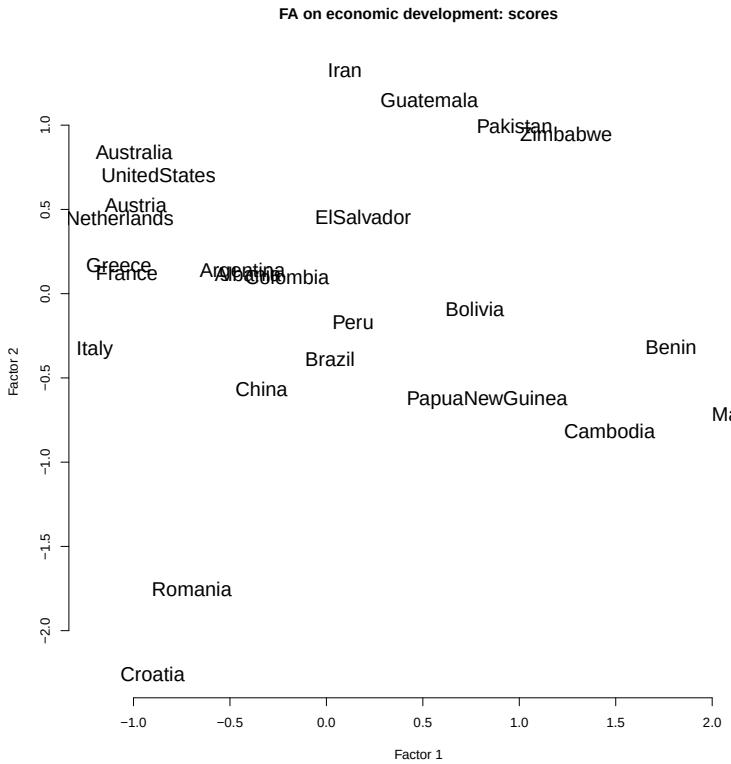
For example, we might have data on the answers of a set of students on a range of different test questions, but in the end most of the test scores are driven by two latent traits, mathematical and verbal skills.

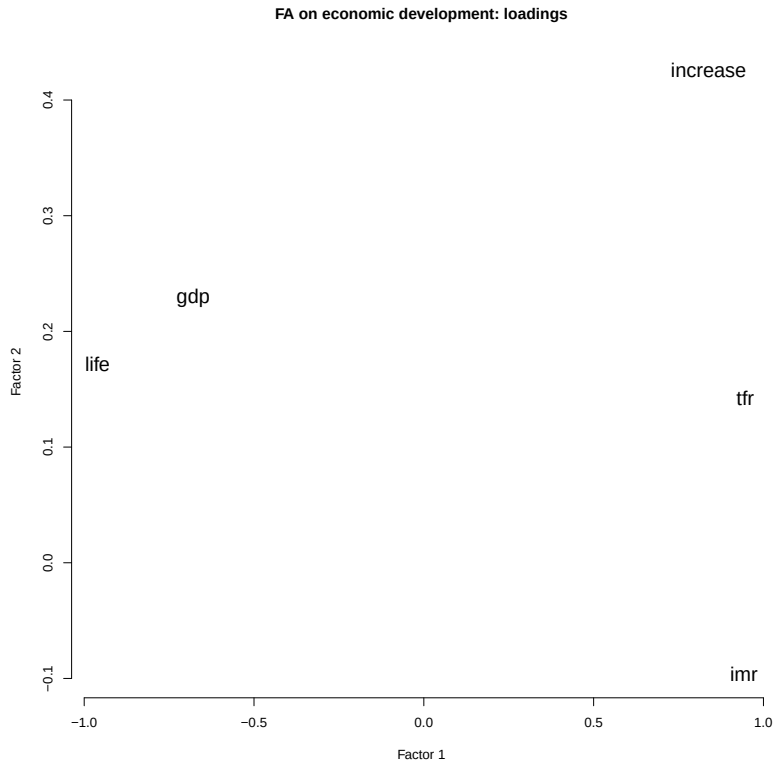
In factor analysis, the variation in the data (e.g. test scores) is seen as a combination of variation in underlying traits (e.g. foundational skills), variation in specific features of the particular test, and error variation.

(van de Geer, 1967, 126–129)

Factor analysis: development example

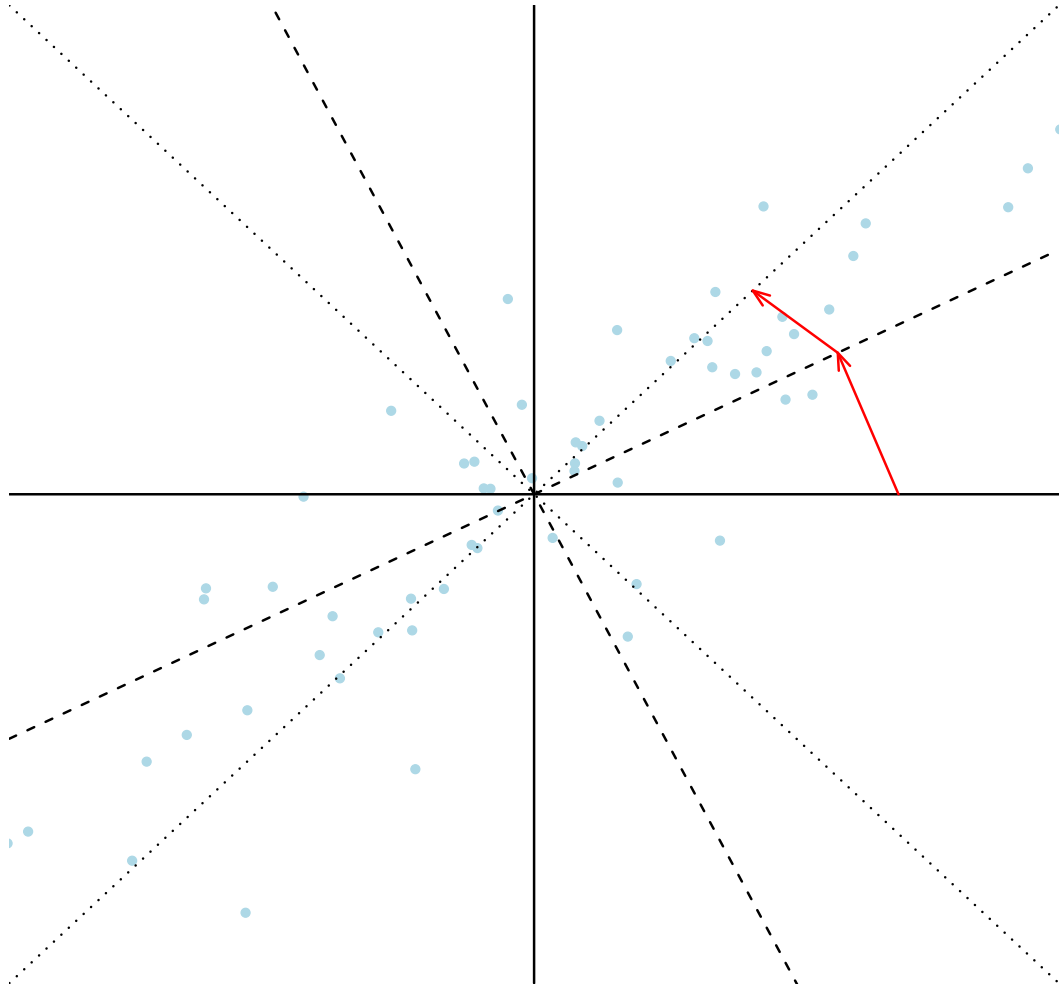
```
library(psych)
plot(fa(scale(econ25), nfactors = 2,
        rotate = "none"))
```





Factor analysis: orthogonal rotation

The factors define a space in which we can locate the observations, but the axes can be freely rotated without altering this space.



Factor analysis: rotation

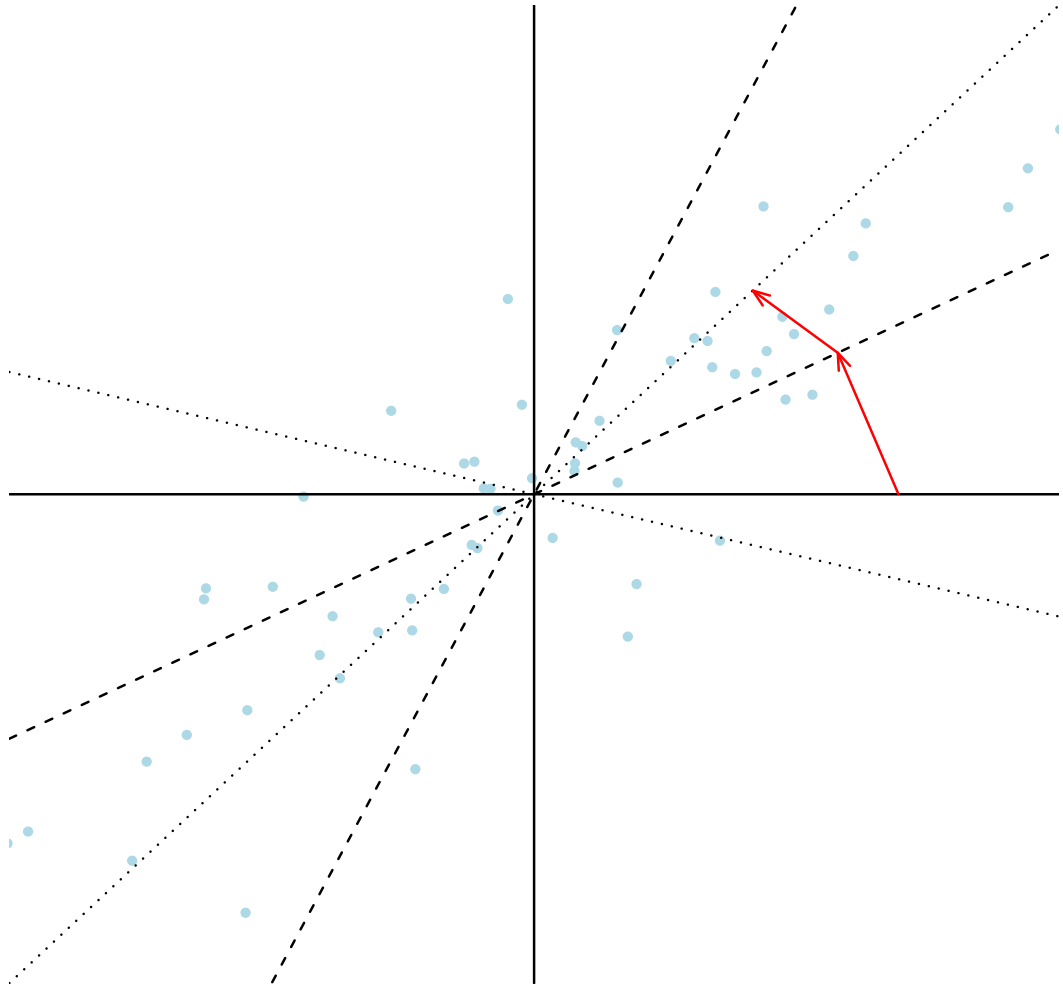
To find potential underlying traits, we tend to search for a rotation such that some variables correlate strongly with a particular factor (high loadings) and other variables correlate weakly (low loadings), avoiding factors where a lot of variables correlate somewhat.

Varimax is the most commonly used rotation algorithm if we insist on orthogonal axes, i.e. uncorrelated factors, called **orthogonal rotation**.

Oblimin is one of a list of options when we allow factors that are correlated, called **oblique rotation**.

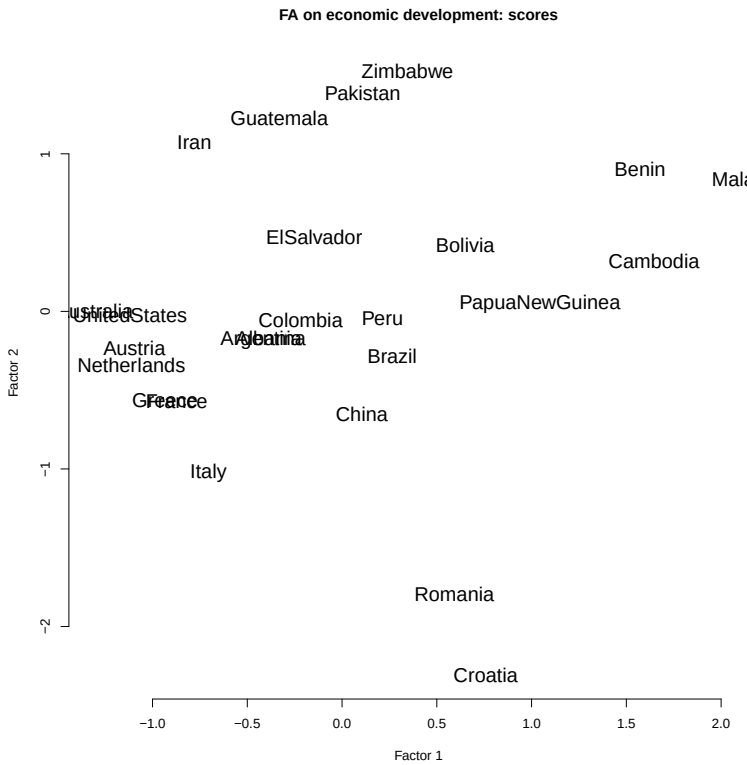
Factor analysis: oblique rotation

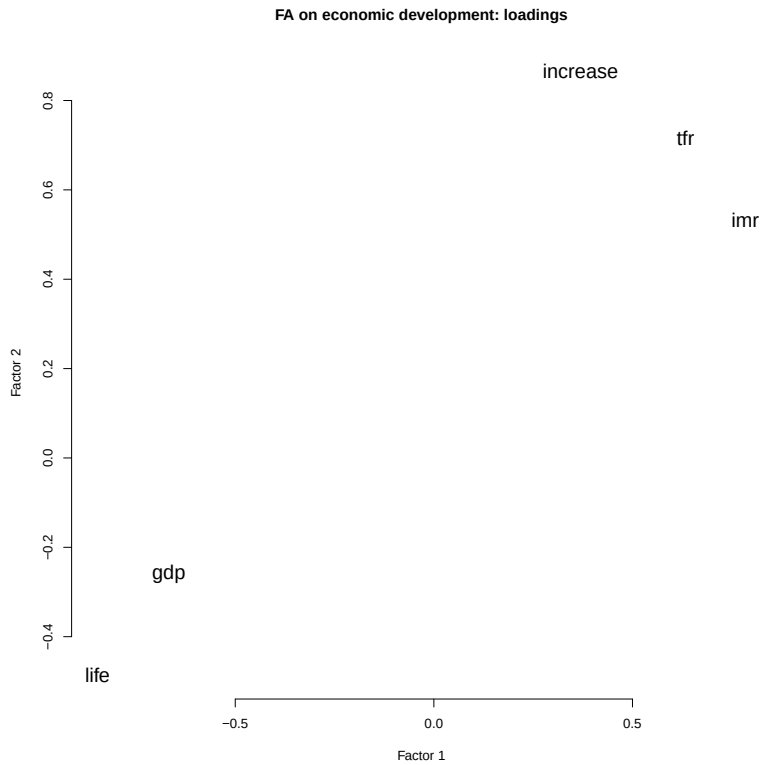
The factors define a space in which we can locate the observations, but the axes can be freely rotated without altering this space, even when we allow for correlate axes.



Factor analysis: development example

```
plot(fa(scale(econ25), nfactors = 2,
        rotate = "varimax"))
```





This assumes orthogonal, uncorrelated factors (using varimax).

Factor analysis: development example

```
fa(scale(econ25), nfactors = 2, rotate = "none")
fa(scale(econ25), nfactors = 2, rotate = "varimax")
```

Loadings without rotation:

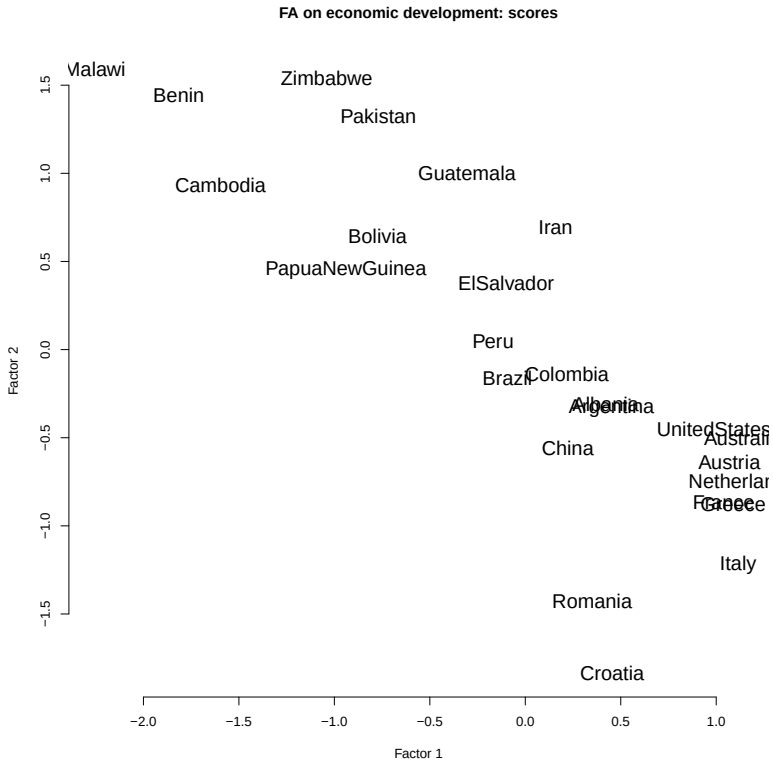
	F1	F2
increase	0.84	0.43
life	-0.96	0.17
imr	0.94	-0.10
tfr	0.95	0.14
gdp	-0.68	0.23

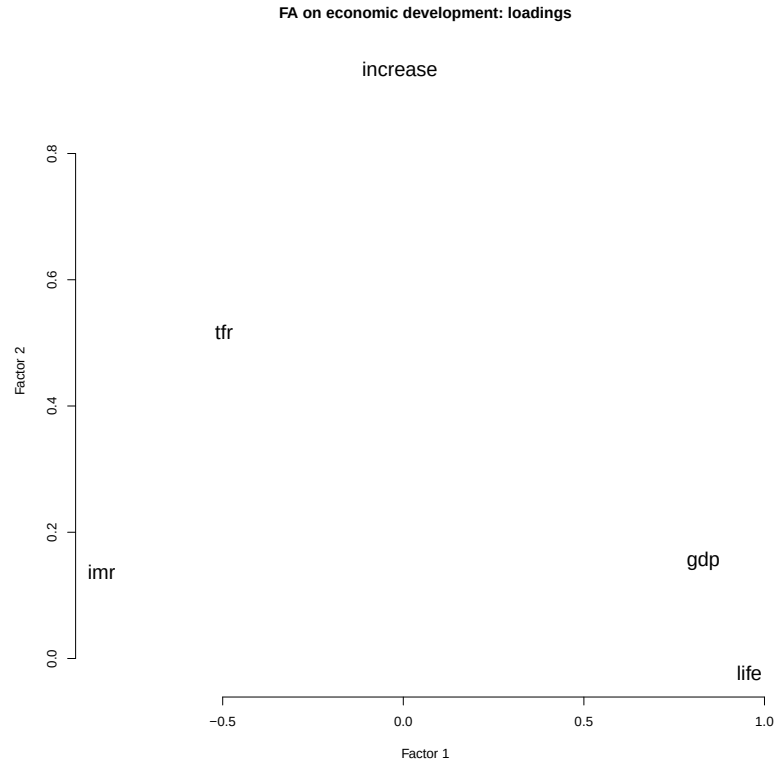
After varimax rotation:

	F1	F2
increase	0.37	0.87
life	-0.85	-0.49
imr	0.78	0.53
tfr	0.63	0.72
gdp	-0.67	-0.26

Factor analysis: development example

```
plot(fa(scale(econ25), nfactors = 2,
        rotate = "oblimin"))
```





This allows for correlated factors (using oblimin).

Factor analysis: terminology

The underlying dimensions are called the **factors**.

The projection of each data point on the factors are called the **scores**.

The projection of each variable on the factors are called the **loadings**.

The loadings help in substantively, *post hoc*, interpreting the meaning of the factors.

Factor analysis: Lisbon example

In a survey after the 2008 Lisbon Referendum, respondents were asked for a number of items whether they thought this was or was not part of the referendum proposal.

		Campaign emphasis		Perceptions of treaty contents	
	Correct	NO campaign	YES campaign	Factor 1	Factor 2
Loss of Irish Commissioner	Yes	Yes		.09	.23
Ending right to decide its own corporate tax	No	Yes		**.56**	.05
Conscription to a European army	No	Yes		**.63**	.04
Reduction of voting strength in Council	Yes	Yes		**.42**	.15
End of control over policy on abortion	No	Yes		**.80**	-.02
Erosion of Irish neutrality	Ambiguous	Yes		**.76**	.03
Improved efficiency of EU decision-making	Yes		Yes	.03	**.60**
Strengthening Europe's role in world	Yes		Yes	-.09	**.63**
Improved protection of workers' rights	Ambiguous		Yes	.02	**.55**
Strengthening role of National Parliaments	Yes		Yes	.13	**.56**
The Charter of Fundamental Rights	Yes		Yes	.07	**.69**

(Elkink and Sinnott, 2009)

PCA vs factor analysis

PCA	Factor analysis
Models all variation	Separates common and unique variation
Summarizing method	Modelling method
Not affected by number of dimensions extracted	Affected by number of factors extracted
Preserves distances	Preserves correlations
Exact analytical solution	Approximate solution (many methods available)
Within space of variables	Transcends space of variables
No underlying assumptions	Assumes multivariate normally distributed data

http://stats.stackexchange.com/questions/95038/how-does-factor-analysis-explain-the-covariance-while-pca-explains-the-variance/

Confirmatory vs exploratory factor analysis

Here we use PCA and factor analysis as dimension reduction techniques to summarize data and find structure.

Factor analysis is also often used to either:

1. find latent variables that are expected to be there *a priori*,
2. or to see if there are underlying latent variables.

The latter is heavily criticised as "Tom Swift and His Electric Factor Analysis Machine", because without substantive understanding of the variables and their measurement, it is risky to put too much weight on findings in factor analysis.

(Armstrong, 1967)

Dimension reduction choices

Find a low-dimensional representation that attempts to preserve correlations (and model latent traits): **factor analysis**.

Find a low-dimensional representation that attempts to preserve distances: **multidimensional scaling** or PCA.

Find a low-dimensional representation that attempts to preserve variation: **principal component analysis**.

Note that the simplest form of MDS happens to be PCA (but less simple forms are not) and FA often uses PCA as part of its iterative algorithm.

`http://stats.stackexchange.com/questions/94048/pca-and-exploratory-factor-analysis-on-the-same-dataset-differences-and-similar/`

Armstrong, J. Scott. 1967. "Derivation of Theory by Means of Factor Analysis or Tom Swift and His Electric Factor Analysis Machine." *The American Statistician* 21(5):17–21.

Bartholomew, David J., Fiona Steele, Irini Moustaki and Jane I. Galbraith. 2008. *Analysis of multivariate social science data*. Boca Raton: CRC Press.

Elkink, Johan A. and Richard Sinnott. 2009. "Political knowledge and campaign effects in the 2008 Irish referendum on the Lisbon Treaty." *Electoral Studies* 38:217–225.

van de Geer, J.P. 1967. *Inleiding in the multivariate analyse*. Arnhem: Van Loghum Slaterus.