Principal Component Analysis

Introduction to Statistical Learning
Bachelor in *Global Governance*
University of Rome - Tor Vergata

Marco Stefanucci
Department of Economics and Finance
University of Rome - Tor Vergata
marco.stefanucci@uniroma2.it

Introduction

Suppose to have n observations of p variables, such that $\mathbf{X} = (X_1, \ldots, X_p)$.

- Graphical representations are useful with small number of variables
- Summarizing data in presence of too many variables can be difficult: Suppose we have data on athletes and their scores in 5 different specialties. How to summarize such information in order to rank them?
- Furthermore, the explanatory variables can be highly correlated: issues when applying multivariate techniques (e.g., regression)

Possible Solutions

- Simplest: take just one element and discard all others.
  $\Rightarrow$ loss of information and in interpretation

Possible Solutions

- Simplest: take just one element and discard all others.
  $\Rightarrow$ loss of information and in interpretation
- Consider a summary measure $\frac{1}{p}\sum_{j=1}^{p} X_j$ (simple average):
  $\Rightarrow$ same importance to all variables

Possible Solutions

- Simplest: take just one element and discard all others.
  $\Rightarrow$ loss of information and in interpretation
- Consider a summary measure $\frac{1}{p}\sum_{j=1}^{p} X_j$ (simple average):
  $\Rightarrow$ same importance to all variables
- Cosider a summary measure $\sum_{j=1}^{p} w_j X_j$ (weighted average) such that $\sum_{j=1}^{p} w_j^2 = 1$

Possible Solutions

- Simplest: take just one element and discard all others.
  $\Rightarrow$ loss of information and in interpretation
- Consider a summary measure $\frac{1}{p}\sum_{j=1}^{p} X_j$ (simple average):
  $\Rightarrow$ same importance to all variables
- Cosider a summary measure $\sum_{j=1}^{p} w_j X_j$ (weighted average) such that $\sum_{j=1}^{p} w_j^2 = 1$

  1. How do we choose the weights?
  2. Is a single summary measure enough?
  3. A more advanced solution: Principal Component Analysis (PCA)

Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised exploratory multivariate technique that aims at:

- Data reduction: reduce the dimension of the data matrix without (almost) losing information;
- Data visualization: identify a lower dimensional space where to project the points without (almost) deforming the distance between them;
- Data interpretation: by reducing the original data, we create a smaller number of new variables (artificial variables) that might have a more direct interpretation than the original ones.

The algebraic approach

Suppose to have a p-dimensional set of correlated variables $\mathbf{X} = (X_1, \ldots, X_p)$.

PCA is used to extract and express important information of the original variables with a set of few new variables $\mathbf{Z} = (Z_1, \ldots, Z_q)$, where $q \leq p$ called principal components (PCs), that:

1. correspond to linear combinations of $\mathbf{X}$:

$$Z_k = \sum_{j=1}^{p} a_{kj} X_j \quad k = 1, \ldots, q \quad (q \leq p)$$

2. are uncorrelated (orthogonal): $\mathbb{C}(Z_k, Z_l) = 0 \quad \forall k, l = 1, \ldots, q$
3. are build in a sequential way and ordered:
   - Z_1 is the variable with the highest variability;
   - Z_2 is orthogonal to Z_1 and $\mathbb{V}(Z_2) \leq \mathbb{V}(Z_1)$

The information of a data matrix $\mathbf{X}$ corresponds to its total variation:

$$I_{tot} = \sum_{j=1}^{p} \mathbb{V}(X_j) = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

and the variance-covariance matrix is denoted by $\mathbf{\Sigma}_x = \mathbf{X}^T\mathbf{X}$.

The information of a data matrix $\mathbf{X}$ corresponds to its total variation:

$$I_{tot} = \sum_{j=1}^{p} \mathbb{V}(X_j) = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

and the variance-covariance matrix is denoted by $\mathbf{\Sigma}_x = \mathbf{X}^T\mathbf{X}$.

PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension.

The information of a data matrix $\mathbf{X}$ corresponds to its total variation:

$$I_{tot} = \sum_{j=1}^{p} \mathbb{V}(X_j) = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

and the variance-covariance matrix is denoted by $\mathbf{\Sigma}_x = \mathbf{X}^T\mathbf{X}$.

PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension.

We identify directions (or principal components) along which the variation in the data is maximal.

- Z_1 is s.t. $\mathbb{V}(Z_1) = \mathbb{V}(\sum_{j=1}^{p} a_{1j} X_j)$ is maximal;
- Z_2 is s.t. $\mathbb{V}(Z_2) = \mathbb{V}(\sum_{j=1}^{p} a_{2j} X_j)$ is maximal s.t. $\mathbb{C}(Z_1, Z_2) = 0$

Stepback: eigenvalues and eigenvectors

Let $\mathbf{A}$ be a $p \times p$ square symmetric matrix. A real scalar λ is said to be an eigenvalue of $\mathbf{A}$ if there exist a non-zero vector $\mathbf{v}$ in $\mathbb{R}^p$ such that

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

The vector $\mathbf{v}$ is then referred to as an eigenvector associated with the eigenvalue λ. The eigenvector $\mathbf{v}$ is said to be normalized if $\mathbf{v}^T\mathbf{v} = 1$.

The interpretation of $\mathbf{v}$ is that it defines a direction along $\mathbf{A}$ behaves just like scalar multiplication. The amount of scaling is given by λ. (In German, the root "eigen", means "self" or "proper").

Stepback: eigenvalues and eigenvectors

The eigenvalues of the matrix $\mathbf{A}$ are characterized by the characteristic equation

$$det(\lambda \mathbf{I} - \mathbf{A}) = 0,$$

where the notation det refers to the determinant of its matrix argument and $\mathbf{I}$ is the $p \times p$ identity matrix.

The function with values $t \to p(t) := det(t\mathbf{I} - \mathbf{A})$ is a polynomial of degree p called the characteristic polynomial.

NB: eigenvectors are orthogonal one to each other, $\mathbf{v}_j^T \mathbf{v}_k = 0 \quad \forall j, k$

Stepback: the spectral theorem

Theorem

Any $p \times p$ symmetric matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j^T$$

where $\lambda_1, \ldots, \lambda_p$ are eigenvalues of the matrix $\mathbf{A}$ and $(\mathbf{v}_1, \ldots, \mathbf{v}_p)$ are the associated eigenvectors.

Finding PCs

"Interesting directions" are found through spectral decomposition of $\mathbf{\Sigma}_x$.

They are given by the eigenvectors $\boldsymbol{\phi}_k = (\phi_{k1}, \phi_{k2}, \ldots, \phi_{kp})$ corresponding to the largest eigenvalues λ of the covariance matrix $\mathbf{\Sigma}_x$.

1. Find the eigenvalues of $\mathbf{\Sigma}_x$ and order them:

$$\lambda_1 \geq \cdots \geq \lambda_q$$

2. Find the eigenvectors $\phi_1 \ldots \phi_q$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_q$ and define the first component as:

$$Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$$

3. Repeat the second step for each component.

Key properties

Principal components satisfy some interesting properties:

- The variance of each component is equal to the corresponding eigenvalue, $\mathbb{V}(Z_j) = \lambda_j$.
- The sum of all eigenvalues is equal to the total variability, $\sum_{j=1}^{p} \lambda_j = I_{tot}$.

How many PCs?

Usually, we consider $q \leq p$ PCs. In choosing q, we agree on a maximum number of information to lose. For instance, if we want to preserve 80% of the total information I_{tot} we select the first q components such that:

$$\frac{\mathbb{V}(Z_1) + \cdots + \mathbb{V}(Z_q)}{I_{tot}} \approx 0.8$$

How many PCs?

Usually, we consider $q \leq p$ PCs. In choosing q, we agree on a maximum number of information to lose. For instance, if we want to preserve 80% of the total information I_{tot} we select the first q components such that:

$$\frac{\mathbb{V}(Z_1) + \cdots + \mathbb{V}(Z_q)}{I_{tot}} \approx 0.8$$

Number of retained components depends on the threshold AND on data.

- If original variables $\mathbf{X}$ are strongly correlated, large dimensionality reduction is obtained with small q;
- If original variables $\mathbf{X}$ are almost uncorrelated, small dimensionality reduction is obtained, and $q \approx p$

Remarks

- In the end PCs are linear combinations of the original variables:

$$Z_{ik} = \phi_{k1} X_{i1} + \phi_{k2} X_{i2} + \cdots + \phi_{kp} X_{ip}$$

  where the generic coefficient ϕ_{kj} represents the weights that the variable X_j has in determining the k-th principal component;

- The component Z_k will be mostly represented by the variables with the larger coefficients.

Geometrical interpretation

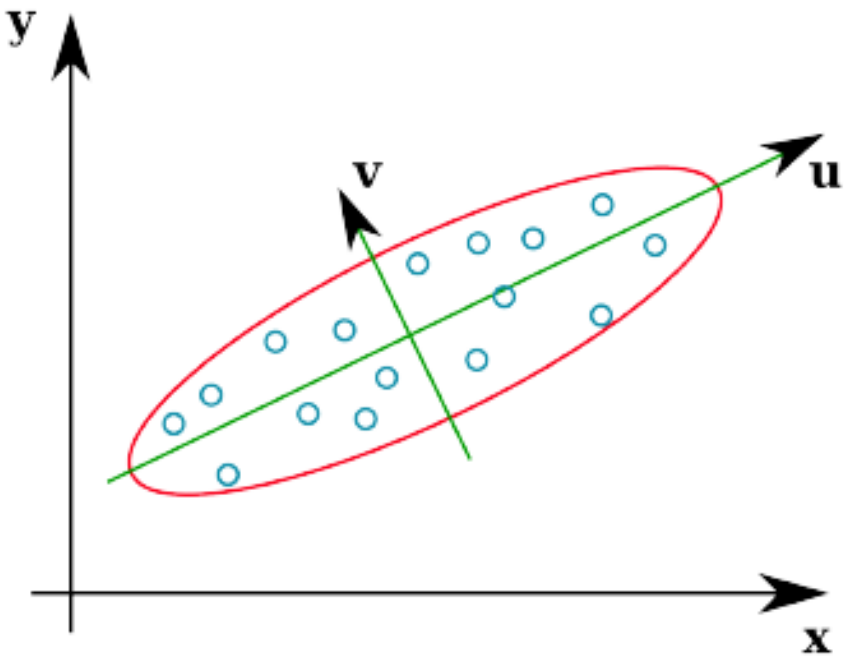
- PCs represent a selection of a new coordinate system obtained by rotating the original axes to a set of new axes (to provide a simpler structure).

  $\Rightarrow$ The first PC represents the direction of maximum variability, followed by the others, all orthogonal.

- "Best" fit hyper-plane: minimizes the sum of squared distances between points that represent cases and space defined by PCs

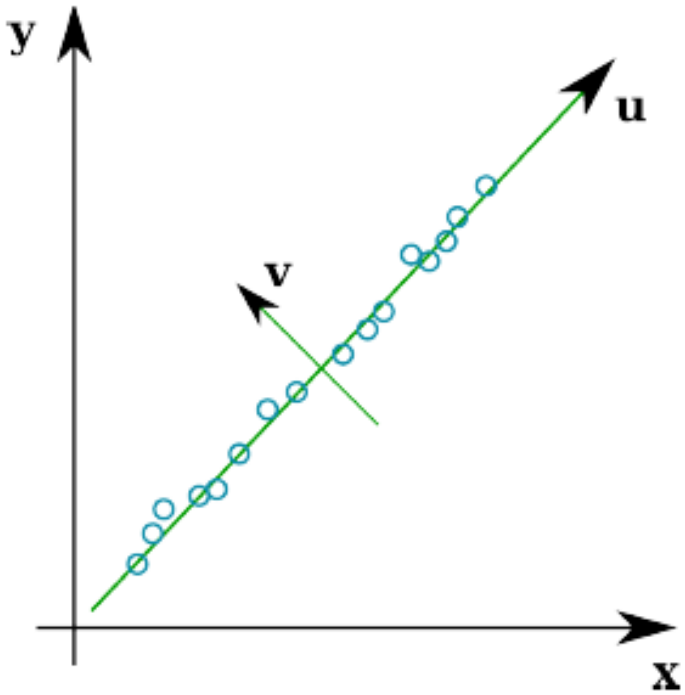
  $\Rightarrow$ The first PC defines a line; the first two PCs define a plane.

PCA for Data Representation



- The principal direction in which the data varies is shown by the U axis;
- The second most important direction is the V axis (orthogonal to U).
- If we place the $U - V$ axis system at the mean of the data it gives us a compact representation.

Data Representation and dimension Reduction



- If the variation in the data is caused by some other relationship then PCA gives us a way of reducing the dimensionality of a data set
- The principal direction in which the data varies is shown by the U axis
- In this case all the V coordinates are all very close to zero: we can represent the data set by one variable U and discard V

Summary

We start with the p quantitative original variables $(X_1, \ldots, X_p)$, and we obtain $q \leq p$ new variables, where q is a good compromise between

- the minimum number of variables (max dimensionality reduction)
- minimum lost of information (max variability)

$$(X_1, \ldots, X_p) \rightarrow (Z_1, \ldots, Z_q)$$

What Makes Party Systems Different? A Principal Component Analysis of 17 Advanced Democracies 1970–2013

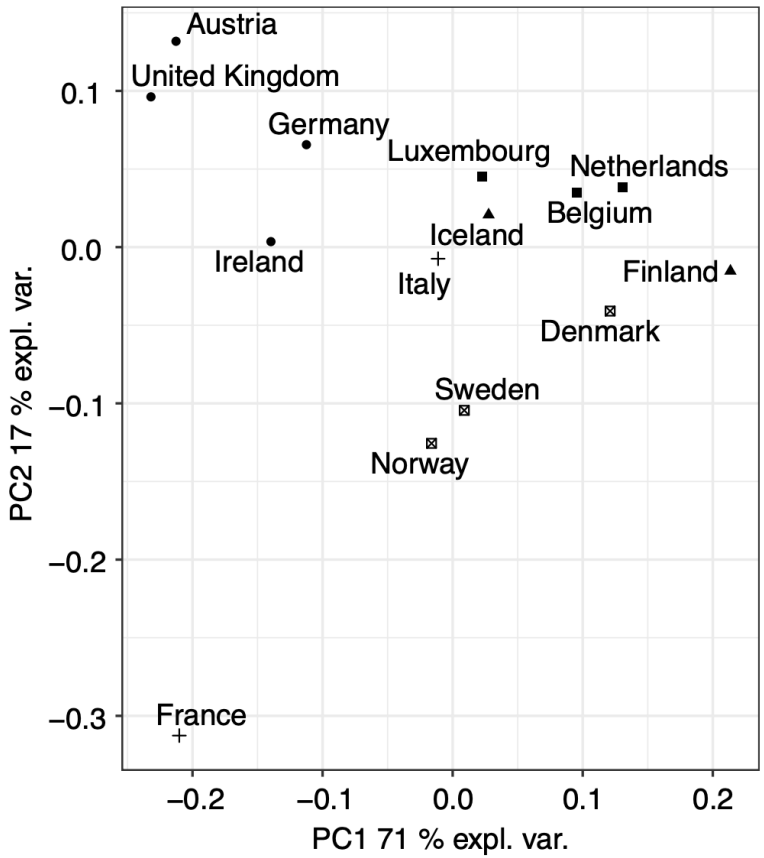
Zsuzsanna B. Magyar

University of Lucerne, Lucerne, Switzerland. Email: zmagyar@ucla.edu

Abstract

Party systems, that is, the number and the size of all the parties within a country, can vary greatly across countries. I conduct a principal component analysis on a party seat share dataset of 17 advanced democracies from 1970 to 2013 to reduce the dimensionality of the data. I find that the most important dimensions that differentiate party systems are: "the size of the biggest two parties" and the level of "competition between the two biggest parties." I use the results to compare the changes in electoral and legislative party systems. I also juxtapose the results to previous party system typologies and party system size measures. I find that typologies sort countries into categories based on variation along both dimensions. On the other hand, most of the current political science literature use measures (e.g., the effective number of parties) that are correlated with the first dimension. I suggest that instead of these, indices that measure the opposition structure and competition could be used to explore problems pertaining to the competitiveness of the party systems.

	Eigenval.	Expl. variance	Cum. variance
1	0.0194012464	71.04	71.04
2	0.0045310786	16.59	87.63
3	0.0020397785	7.47	95.10
4	0.0006762824	2.48	97.57
5	0.0003635705	1.33	98.91
6	0.0001867591	0.68	99.59
7	0.0000764051	0.28	99.87
8	0.0000170232	0.06	99.93
9	0.0000113822	0.04	99.97
10	0.0000036437	0.01	99.99
11	0.0000023248	0.01	100.00
12	0.0000009147	0.00	100.00
13	0.0000003156	0.00	100.00
14	0.0000000699	0.00	100.00
15	0.0000000216	0.00	100.00
16	0.0000000152	0.00	100.00
17	0.0000000039	0.00	100.00
18	0.0000000007	0.00	100.00
Sum	0.0260737		



Measuring real activity using a weekly economic index

Daniel J. Lewis[1] | **Karel Mertens**[2] | **James H. Stock**[3] | **Mihir Trivedi**[4]

[1]Federal Reserve Bank of New York, New York, New York, USA

[2]Federal Reserve Bank of Dallas, Dallas, Texas, USA

[3]Harvard University, Cambridge, Massachusetts, USA

[4]Amazon.com, Seattle, Washington, USA

Correspondence
Daniel J. Lewis, Federal Reserve Bank of New York, New York, NY, USA.
Email: daniel.lewis@ny.frb.org

Summary

This paper describes a weekly economic index (WEI) developed to track the rapid economic developments associated with the onset of and policy response to the novel coronavirus in the United States. The WEI is a weekly composite index of real economic activity, with eight of 10 series available the Thursday after the end of the reference week. In addition to being a weekly real activity index, the WEI has strong predictive power for output measures and provided an accurate nowcast of current-quarter GDP growth in the first half of 2020, with weaker performance in the second half. We document how the WEI responded to key events and data releases during the first 10 months of the pandemic.

KEYWORDS

Forecasting, High Frequency, Measurement of Economic Activity, Weekly Economic Index

Series	Weights baseline
Same-store retail sales	0.28
Consumer confidence	0.23
Initial claims	−0.37
Continued claims	−0.41
Staffing index	0.40
Tax withholding	0.30
Steel production	0.36
Fuel sales	0.22
Railroad traffic	0.34
Electricity output	0.12
Total variance explained	55.44

Example: Brand ratings data

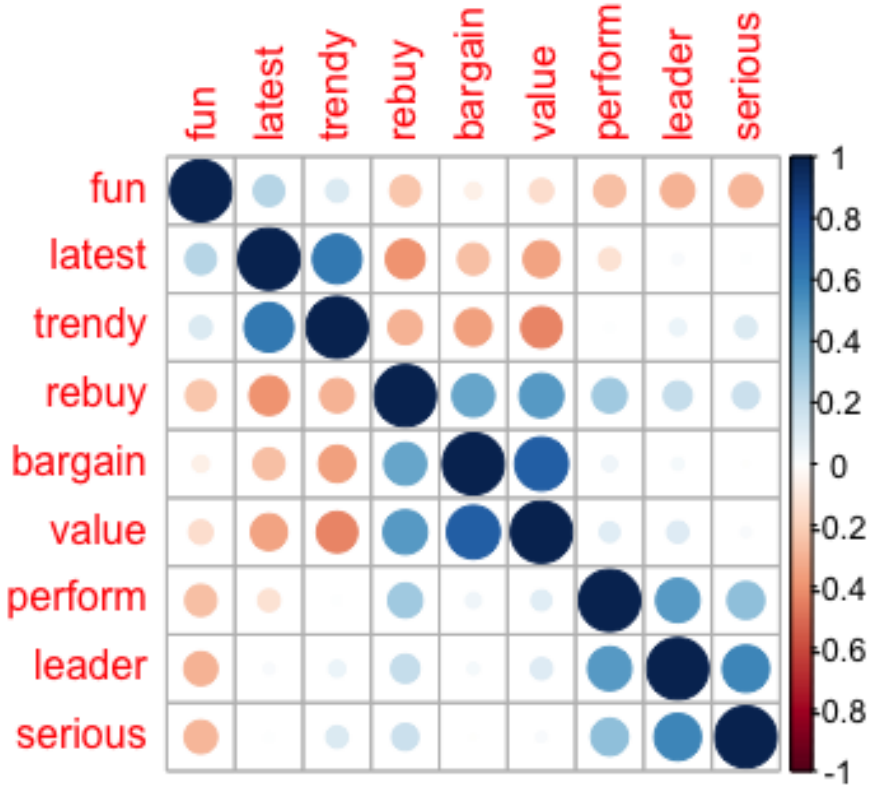
We investigate dimensionality using a simulated data set that is typical of consumer brand perception surveys.

The data comprise ratings of 10 brands (a to j) on 9 adjectives (`performance, leader, fun, etc`), for $n = 100$ respondents, as expressed on survey items with the following form: scale from 1 (least) to 10 (most).

Example:
How trendy is Intelligentsia Coffee?

```
> brand.ratings <- read.csv("http://goo.gl/IQl8nc")
> head(brand.ratings)
  perform leader latest fun serious bargain value trendy rebuy brand
1       2      4      8   8       2       9     7      4     6     a
2       1      1      4   7       1       1     1      2     2     a
3       2      3      5   9       2       9     5      1     6     a
4       1      6     10   8       3       4     5      2     1     a
5       1      1      5   8       1       9     9      1     1     a
6       2      8      9   5       3       8     7      1     2     a
```

```
> library(corrplot)
> corrplot(cor(brand.sc[, 1:9]), order = "hclust")
```

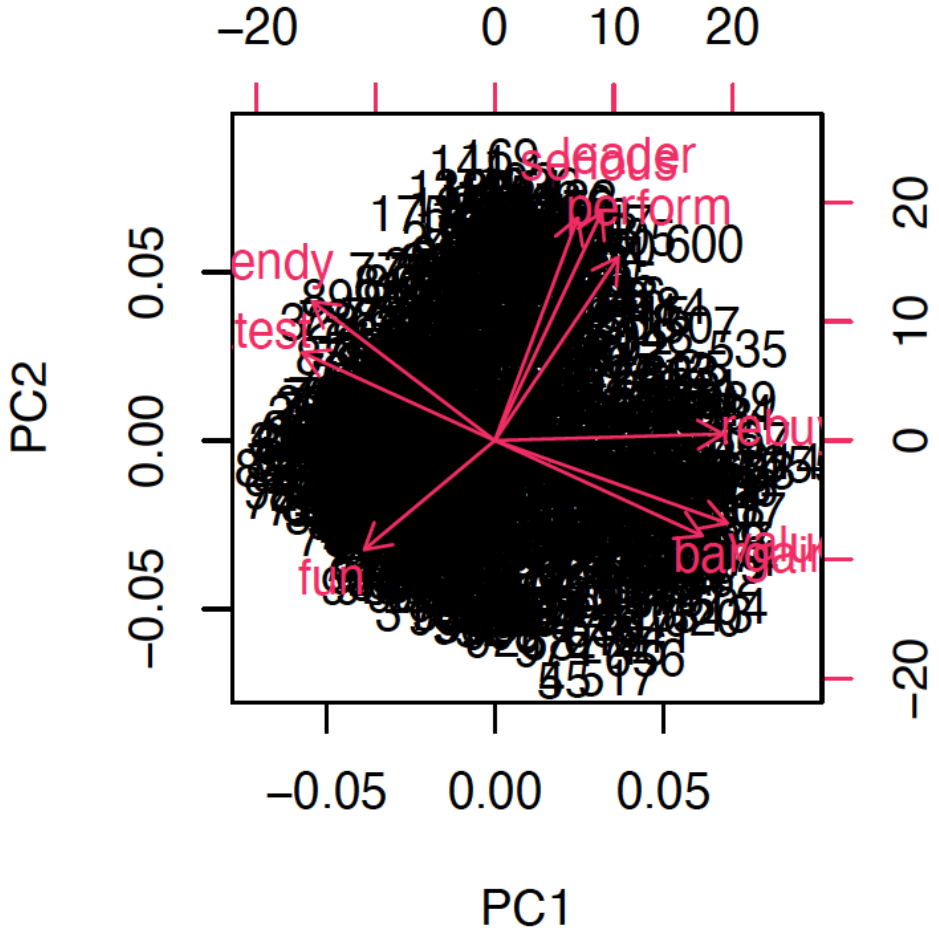


Perform PCA

```
> brand.pc <- prcomp(brand.sc[, 1:9])
> summary(brand.pc)

Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.726 1.4479 1.0389 0.8528 0.79846
Proportion of Variance 0.331 0.2329 0.1199 0.0808 0.07084
Cumulative Proportion  0.331 0.5640 0.6839 0.7647 0.83554
                          PC6     PC7     PC8     PC9
Standard deviation     0.73133 0.62458 0.55861 0.49310
Proportion of Variance 0.05943 0.04334 0.03467 0.02702
Cumulative Proportion  0.89497 0.93831 0.97298 1.00000
```

> biplot(brand.pc)



The plot of individual respondents' ratings is too dense and it does not tell us about the brand positions! Biplots are especially helpful when:

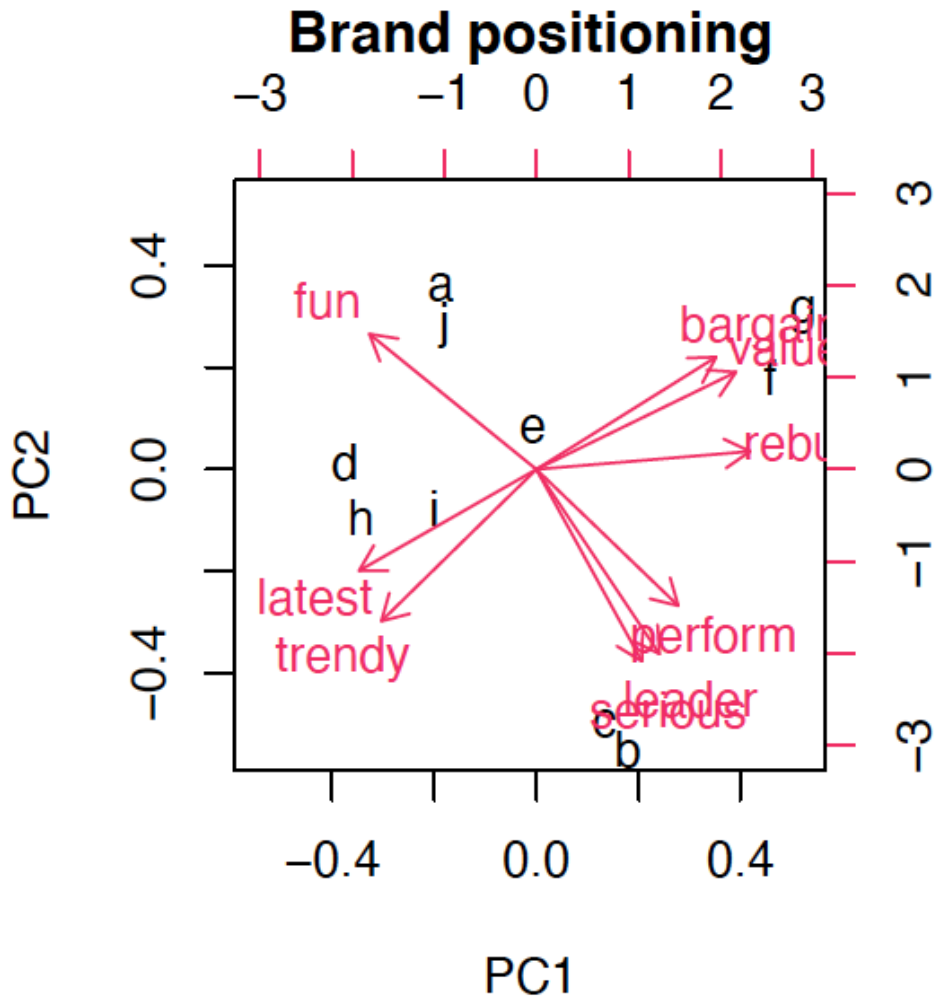
- there are a smaller number of points
- when there are clusters

Better solution: perform PCA using aggregated ratings by brand!

```
> brand.mu.pc <- prcomp(brand.mean[, 2:10], scale = TRUE)
> summary(brand.mu.pc)

Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     2.1345 1.7349 0.7690 0.61498 0.50983
Proportion of Variance 0.5062 0.3345 0.0657 0.04202 0.02888
Cumulative Proportion  0.5062 0.8407 0.9064 0.94842 0.97730
                          PC6     PC7     PC8     PC9
 Standard deviation    0.36662 0.21506 0.14588 0.04867
Proportion of Variance 0.01493 0.00514 0.00236 0.00026
Cumulative Proportion  0.99223 0.99737 0.99974 1.00000
```

> biplot(brand.mu.pc, main = "Brand positioning")



Interpretation

What does the map tell us? First we interpret the adjective clusters and relationships and see four areas with well differentiated sets of adjectives and brands that are positioned in proximity.

Brands f and g are high on value, for instance, while a and j are relatively high on fun, which is opposite in direction from leadership adjectives (leader and serious).

Let suppose that you are the brand manager for brand e. What does the map tell you?

Let suppose that you are the brand manager for brand e. What does the map tell you?

Your brand is in the center and thus appears not to be well-differentiated on any of the dimensions. That could be good or bad, depending on your strategic goals.

Let suppose that you are the brand manager for brand e. What does the map tell you?

Your brand is in the center and thus appears not to be well-differentiated on any of the dimensions. That could be good or bad, depending on your strategic goals.

If your goal is to be a safe brand that appeals to many consumers, then a relatively undifferentiated position like e could be desirable.

Let suppose that you are the brand manager for brand e. What does the map tell you?

Your brand is in the center and thus appears not to be well-differentiated on any of the dimensions. That could be good or bad, depending on your strategic goals.

If your goal is to be a safe brand that appeals to many consumers, then a relatively undifferentiated position like e could be desirable.

On the other hand, if you wish your brand to have a strong, differentiated perception, this finding would be unwanted.

Suppose you wanted to move in the direction of brand c. You could look at the specific differences from c in the data:

```
> brand.mean <- aggregate(. ~ brand, data = brand.sc, mean)
> brand.mean[3, -1] - brand.mean[5, -1]
   perform    leader     latest       fun  serious
3 1.214314 0.9699315 -0.5587936 -1.140567 1.180621
  bargain      value    trendy      rebuy
-1.158594 -0.8588416 -0.113052 -0.1689859
```

This shows you that e is relatively stronger than c on value and fun, which suggests dialing down messaging or other attributes that reinforce those. Similarly, c is stronger on perform and serious.