MONASH University

MONASH BUSINESS SCHOOL

Principal Components Analysis

High Dimensional Data Analysis

Anastasios Panagiotelis

Lecture 3

Motivation

Motivation

- In **marketing** surveys we may ask a large number of questions about customer experience.
- In **finance** there may be several ways to assess the credit worthiness of firms.
- In **economics** the development of a country or state can be measured in different ways.

A real example

- Consider a dataset with the following variables for the 50 States of the USA
  - Income
  - Illiteracy
  - Life Expectancy
  - Murder Rate
  - High School Graduation Rate
- You can access this via moodle from the file *StateSE.RData*
- Let's do some exploratory data analysis

Life Exp v High School Grad.

Three dimensions

All five variables

- Can not do a scatter plot in five dimensions.
- Can we find a single variable to summarise the information in this data set?
- Going beyond that can we *rotate* the data to find a good 2-dimensional representation?
- Both of these ideas can be achieved using **Principal Components**

Summarising many variables

- Often we aim to combine many variables into a single index?
  - In finance a credit score summarises all the information about the likelihood of bankruptcy for a company.
  - In marketing we require a single overall measure of customer experience.
  - In economics the Human Development Index is a single measure that takes income, education and health into account.

Weighted linear combination

- A convenient way to combine variables is through a *linear combination* (LC)
  - For example, your grade for this unit:

$$w_1 \text{Assign. Marks} + w_2 \text{Exam Mark}$$

  - Here w_1 and w_2 are called *weights*
  - In this unit, the weight for the Assignments is *50%* and for the Examination is *50%*
- What is a good way to choose weights?

Maximise variance

- The purpose of grading students is to differentiate the best perfoming students from the weakest performing students
- The index should have *large variance*.
- The LC with the highest variance is the **first Principal Component** of the data.
- The first principal component is a new variable that *explains* as much variance as possible in the original variables.

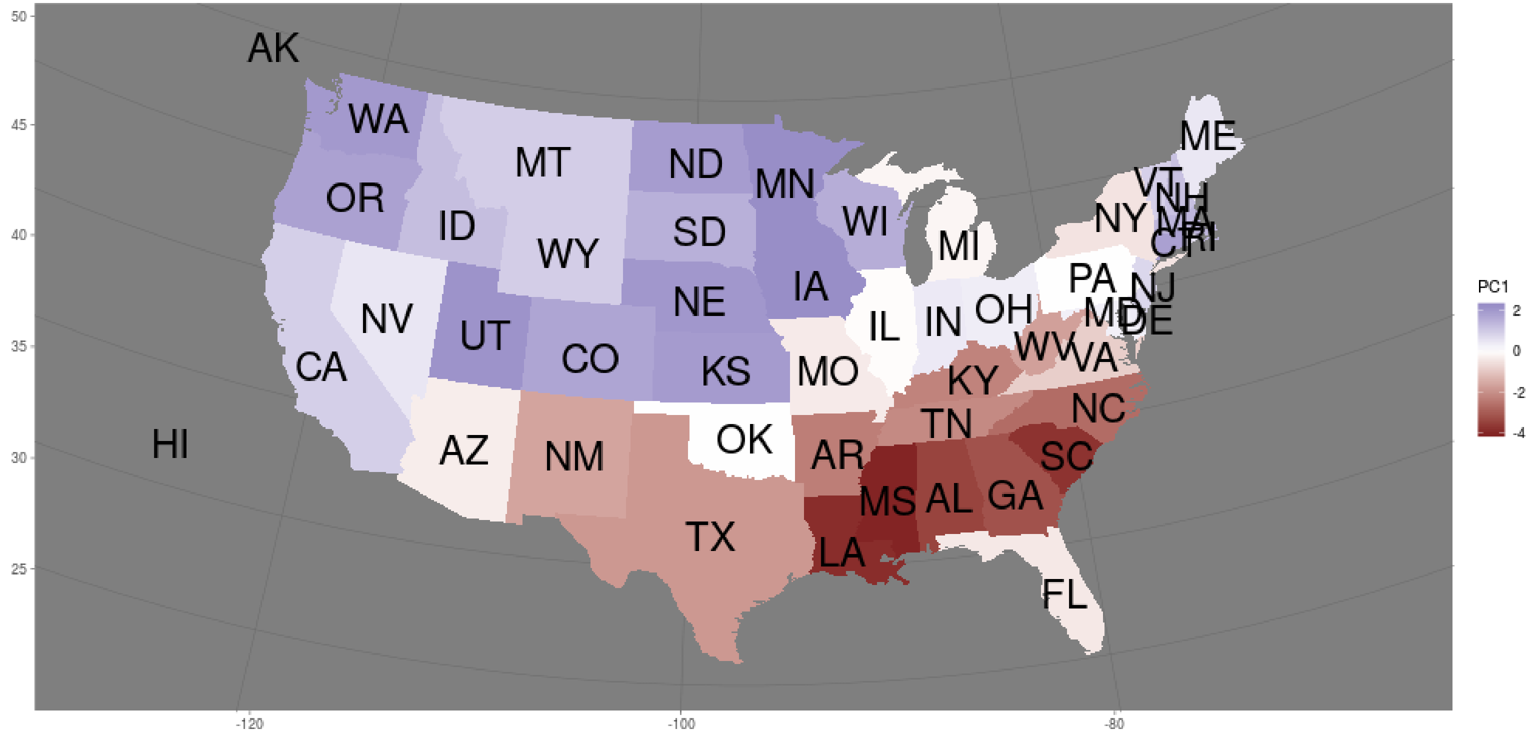
Original Data

State	Income	Illiteracy	LifeExp	M
alabama	3624	2.1	69.05	
alaska	6315	1.5	69.31	
arizona	4530	1.8	70.55	
arkansas	3378	1.9	70.66	
california	5114	1.1	71.71	
colorado	4884	0.7	72.06	
connecticut	5348	1.1	72.48	

First PC

State	PC1
alabama	-3.4736429
alaska	0.5523458
arizona	-0.3218179
arkansas	-2.3518240
california	0.9138319
colorado	1.7319349
connecticut	1.8293070
[illegible]	[illegible]

First PC on Map



Second Principal Component

- Sometimes a single index still oversimplifies the data.
- The second principal component is an LC that
  - Is uncorrelated with the first PC.
  - Has the highest variance out of all LCs that satisfy condition 1.
- Since there is no need for PC2 to *explain* any variance already explained by PC1, PC2 and PC1 are uncorrelated.
- We can plot the first two principal components on a scatter plot.

Scatter-plot of PCs

The weights

	PC1	PC2
Income	0.3473146	0.7315324
Illiteracy	-0.4803318	0.0693093
LifeExp	0.4685523	-0.3243911
Murder	-0.4594049	0.4916219
HSGrad	0.4669687	0.3363552

- A high (low) weight indicates a strong positive (negative) association between a variable and the corresponding PC.

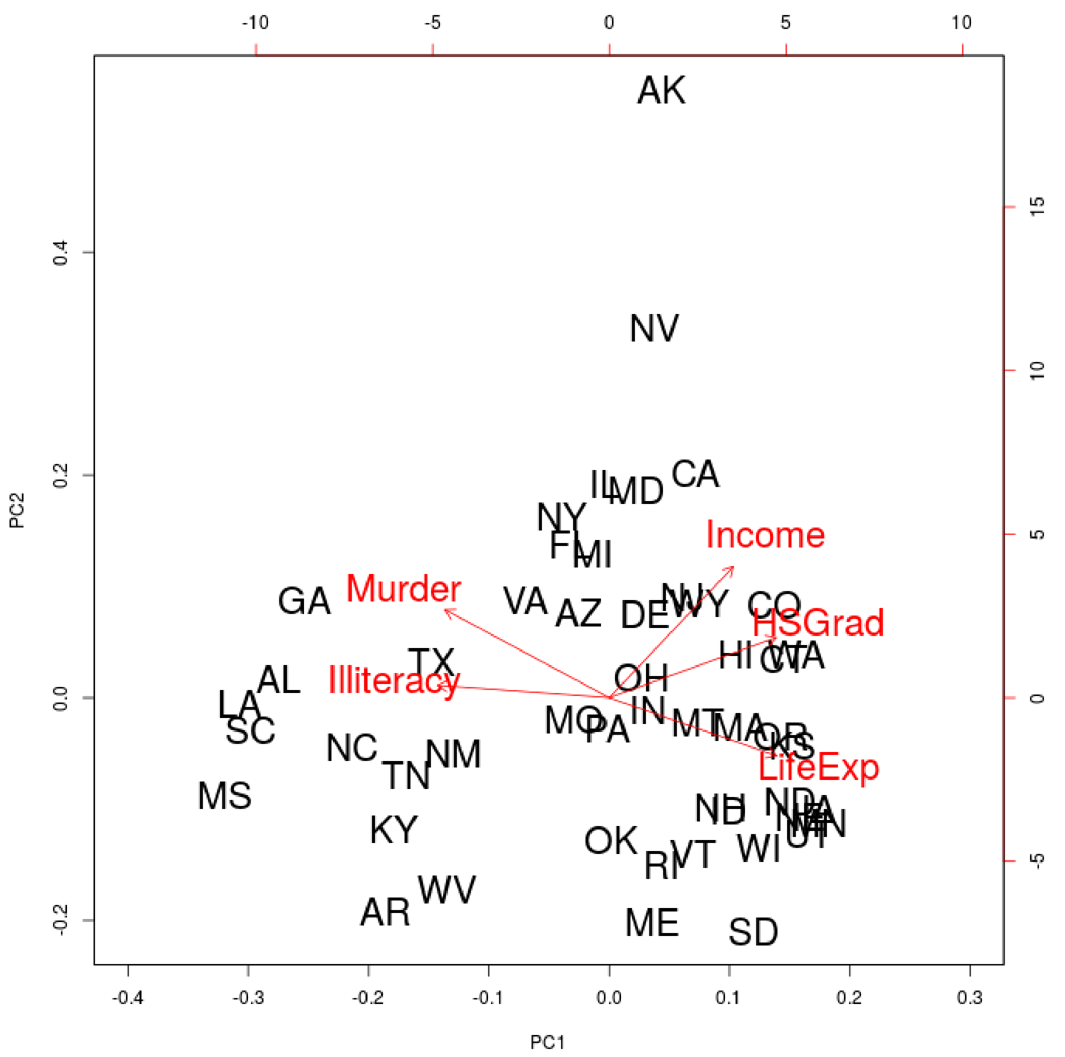
Biplot

- The weight vectors can be plotted on the same scatterplot as the data.
- This is called a biplot.
- We can do several useful things with a biplot
  - See how the observations relate to one another
  - See how the variables relate to one another
  - See how the observations relate to the variables

Types of biplot

- There are multiple ways to draw a biplot.
- We will look at two versions
  - Distance Biplot
  - Correlation Biplot

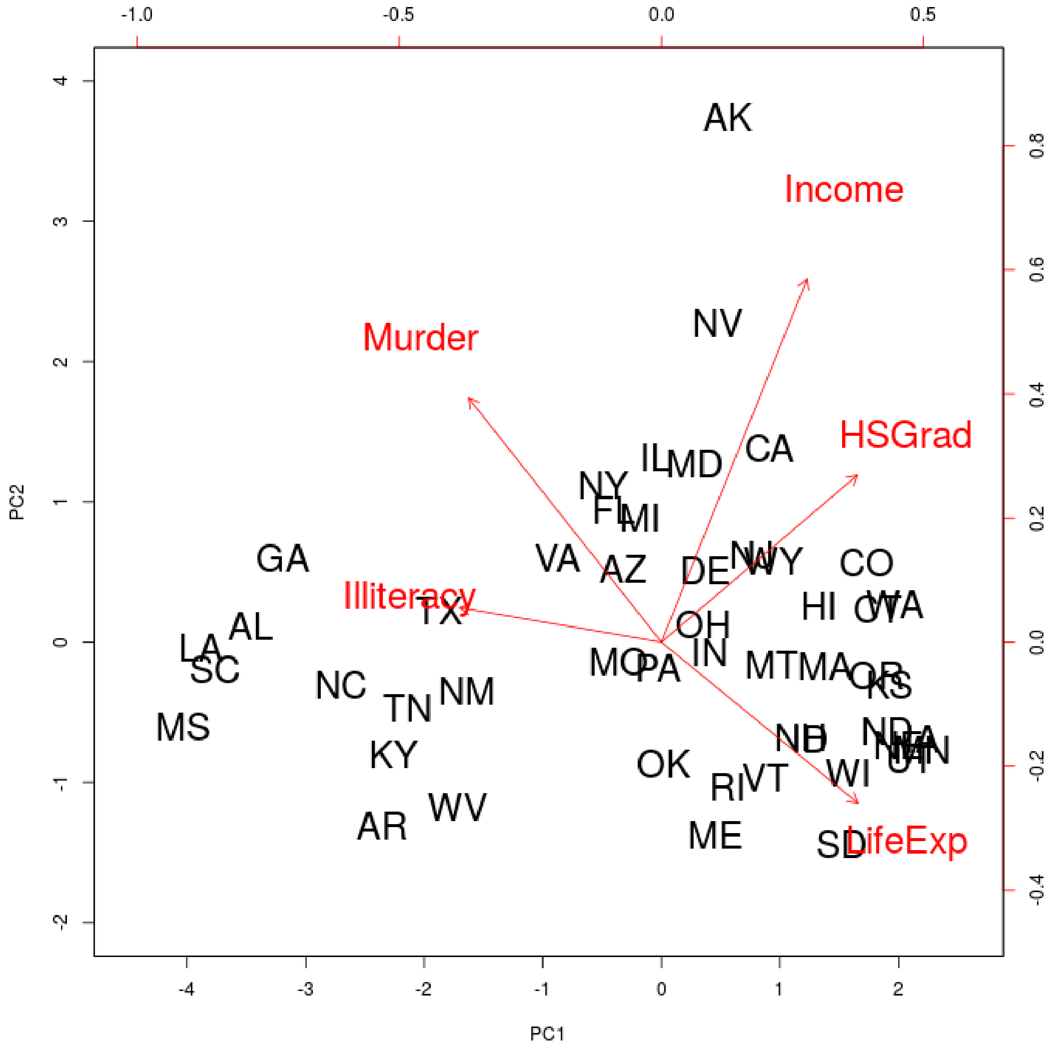
Distance Biplot



Distance Biplot

- The distance between observations implies similarity between observations
  - Louisiana (LA) and South Carolina (SC) are close therefore are similar.
  - Arkansas (AR) and California (CA) are far apart and therefore different.
- If the variables are ignored this is identical to a scatter plot of principal components.

Correlation Biplot



Correlations

	Income	Illiteracy	LifeExp	Murder	HSGrad
Income	1.000	-0.437	0.340	-0.230	0.620
Illiteracy	-0.437	1.000	-0.588	0.703	-0.657
LifeExp	0.340	-0.588	1.000	-0.781	0.582
Murder	-0.230	0.703	-0.781	1.000	-0.488
HSGrad	0.620	-0.657	0.582	-0.488	1.000

Correlation Biplot

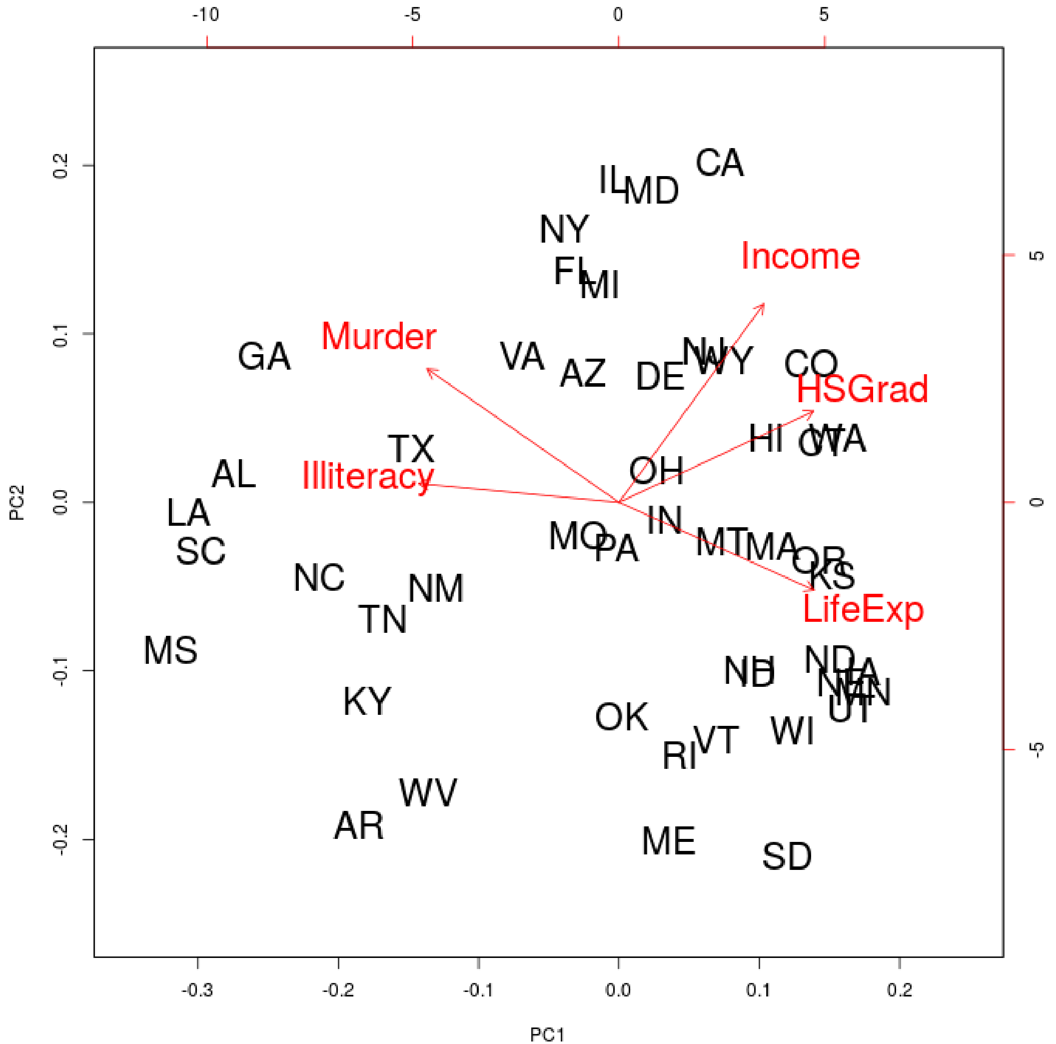
- The angles between variables tell us something about correlation (approximately)
  - Income and HSGrad are highly positively correlated. The angle between them is close to zero.
  - LifeExp and Income are close to uncorrelated. The angle between them is close 90 degrees.
  - Murder and LifeExp are highly negatively correlated. The angle between them is close 180 degrees.

More comparison

- The biplot also allows us to compare observations to variables.
- Think of the variables as axes.
- Draw the shortest line from each point to the axis.
- The position along that axis gives an approximation to the actual value of the variable for that observation.

Usual scatter plot

Biplot



-10
-5
0
5
0.2
0.1
0.0
-0.1
-0.2
-0.3
PC2
PC1
CA
IL
MD
NY
Income
FL
MI
Murder
GA
VA
AZ
DE
NJ
WY
CO
HSGrad
HI
UT
WA
TX
AL
Illiteracy
OH
LA
SC
IN
MO
PA
MT
MA
OR
KS
NC
NM
TN
LifeExp
MS
NH
ND
IA
NE
MN
KY
OK
WI
RI
VT
WV
AR
ME
SD

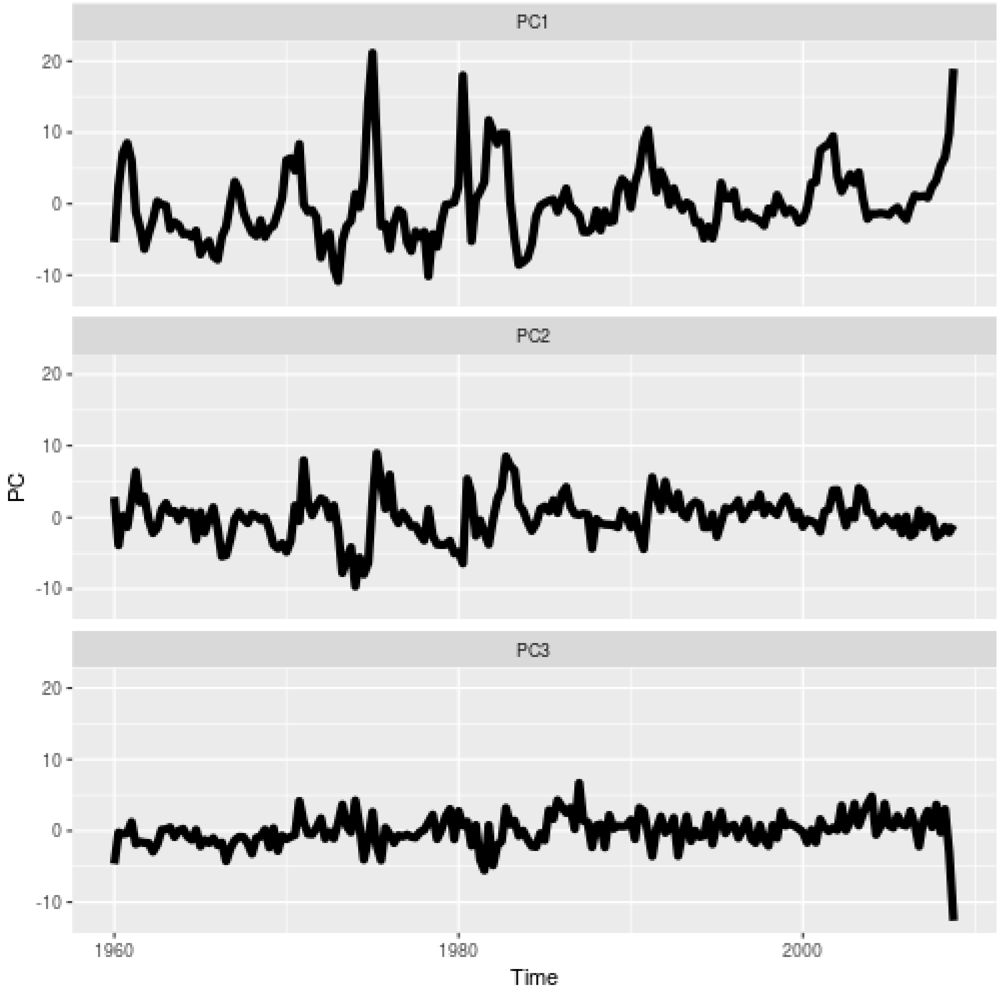
More PCs

- We can find a third PC, which has the highest variance, while being uncorrelated with PC1 and PC2.
- We cannot visualise this with a biplot, but there are alternatives depending on the structure of the data.
- Now a time series example where we consider 3 principal components.

A Time Series Example

- The Stock and Watson dataset contains data on 109 macroeconomic variables in the following categories
  - Output
  - Prices
  - Labour
  - Finance
- One cannot look at 109 time series plots to visualise general macroeconomic conditions.
- However, one can look at time series plots of the principal components of these variables.

Plots of PCs



All PCs

- There are as many principal components as there are variables.
- Together all p principal components explain all of the variation in all p original variables.

$$\sum_{j=1}^{p} \mathrm{Var}(C_j) = \sum_{j=1}^{p} \mathrm{Var}(Y_j)$$

- Where C_j is principal component j and Y_j is variable j

So why PCs

- However a small number of principal components can often explain a large proportion of the variance
  - In the first example, 2 PCs explain 84% of the total variation of 5 variables.
  - In our second example, 3 PCs explain 35% of the total variation of 109 variables.

Summary

- Principal components analysis is useful for
  - Creating a single index
  - Seeing how variables are associated with observations on a single biplot.
  - Visualising high-dimensional time series.
- How do we do it?

Implementation of PCA

Restriction

- Recall that the objective is to find an LC with a large variance. How could we ‘cheat’ ?
  - For a single variable
    $\mathrm{Var}(wY) = w^2\mathrm{Var}(Y)$
  - The variance can be made large by choosing a huge value of w.
- For this reason the following restriction (normalization) is used

$$w_1^2 + w_2^2 \ldots + w_p^2 = 1.$$

Standardisation

- A similar logic applies to the units that the variables are measured in.
- In the states dataset, income varies from \$3000 to \$6000, life expectancy varies from 67 years to 73 years.
  - Which variable will probably have the larger variance?
- Income likely to have a larger variance.

Different units

- If income is measured in \$ ’000s then it will vary from about 3 to 6, If Life Expectancy in measured in days rather than years it will vary from about 24800 days to 26900 days
  - Which variable will have the larger variance now?
- The weights can be influenced by the units of measurement.

Effect of standardisation

	Std	Unstd	DifUnits
Income	0.3473	1.0000	0.0004
Illiteracy	-0.4803	-0.0004	-0.0007
LifeExp	0.4686	0.0007	0.9999
Murder	-0.4594	-0.0014	-0.0059
HSGrad	0.4670	0.0081	0.0096

Standardise or not?

- While the normalisation $w_1^2 + w_2^2 + \ldots + w_p^2 = 1$ is always implemented in any software that does PCA, the decision to standardise is up to you.
- If the variables are measured in the *same* units then
  - *No* need to standardise.
- If the variables are measured in the *different* units then
  - *Standardise* the data.

Principal Components in R

- There are several functions for doing Principal Components Analysis in R. We will use `prcomp`
- We can scale in two ways
  - Scale the data using the function scale
  - Include the option `scale.=TRUE` when calling the function `prcomp`
- Now we will do PCA on the states dataset using R

Principal Components in R

```
StateSE%>%
  select_if(is.numeric)%>% #Only use nume
  prcomp(scale. = TRUE)->pcaout #Do pca
summary(pcaout) #summary of information
```

```
## Importance of components:
##                           PC1    PC2    PC
## Standard deviation     1.7892 0.9686 0.631
## Proportion of Variance 0.6403 0.1876 0.079
## Cumulative Proportion  0.6403 0.8279 0.907
```

Principal Components in R

- The output of the `prcomp` function is a prcomp object.
- It is a list that contains a lot of information. Of most interest are
  - The principal components which are stored in `x`
  - The weights which are stored in `rotation`

Biplot

- The biplot can be produced by:

```
biplot(pca)
```

- To have the state abbreviations on the plot they need to be attached to the matrix `pca$x`

```
rownames(pca$x)<-use_series(StateSE,State
biplot(pca)
```

- Try it!

Correlation biplot

- By default `biplot` produces the distance biplot.
- To produce the correlation biplot try

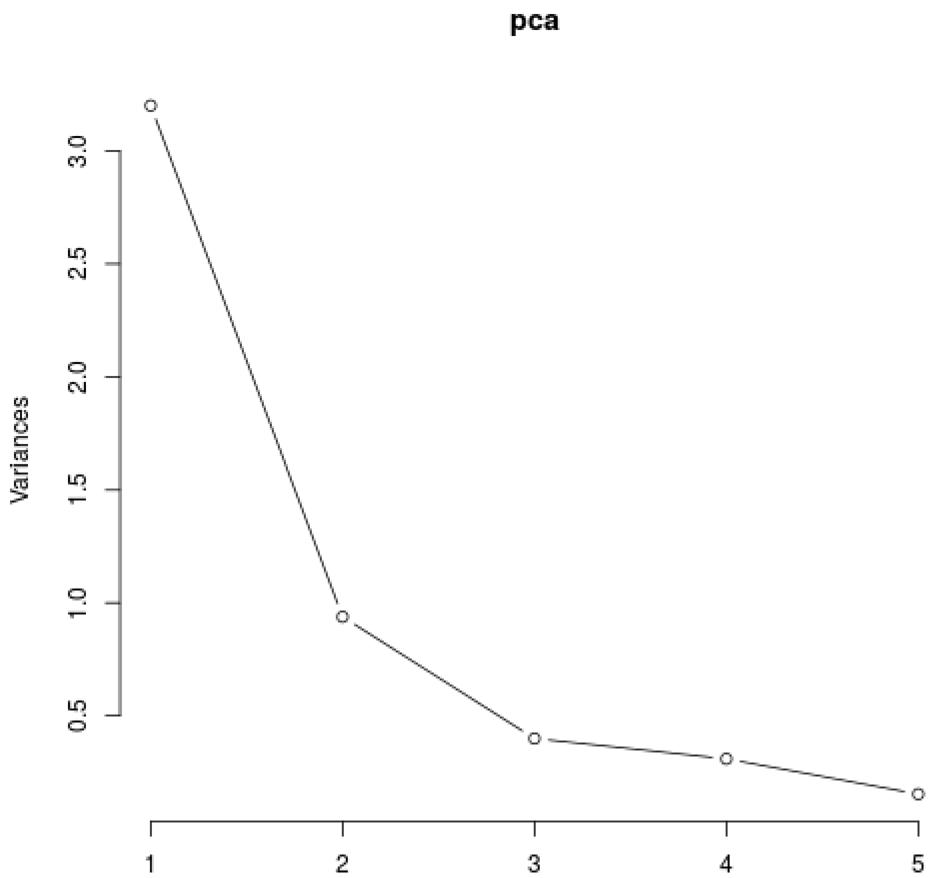
```
biplot(pca,scale = 0)
```

Scree Plot

- Another plot that is easy to create is the Scree plot.
- Along the horizontal axis is the Principal Component.
- Along the vertical axis is the variance corresponding to each Principal Component.
- The Scree plot indicates how much each PC explains the total variance of the data.

```
screeplot(pca,type="lines")
```

Scree Plot



Selecting the number of PCs

- The Scree plot can be used to select the number of Principal Components.
- Look for a part where the plot flattens out also called the elbow of the Scree Plot.
- Another criterion used for standardised data is Kaiser’s Rule. The rule is to select all PCs with a variance greater than 1.

Number of PCs

- The way PCs are selected depend on the nature of the analysis.
- For a visualisation via the biplot, two PCs must be selected.
- In this case check the proportion of variance explained by those PCs
- The higher this number the more accurate the biplot

Towards Factor Analysis

- For survey data it is often the case that multiple survey questions are measures of the same underlying factor.
- For example, at the end of semester you evaluate this unit.
- Typically you will be asked many questions.
- This is no different from any other customer satisfaction survey

Underlying factors

- Although you are asked many questions perhaps there are two underlying factors that drive
  - The quality of the course materials
  - The quality of the teaching staff
- Perhaps the quality of assessment is a third factor.
- For survey data, Scree plots and Kaiser's rule can be used to select the number of underlying factors.