PCA III: a tool for multidimensional data exploration and a cautionary note

Data Science 101 Team
When the data has many variables

- We have seen that R automatically looks at all eigenvectors and eigenvalues and outputs many principal components.
- PC are ordered by the size of the relative eigenvalues.
- Each of them defines a linear combination of $X$ that is orthogonal to the others, and explain a portion of the total variance

$$\sum_i \lambda_i = \text{trace}(X'X) = \sum_i \text{Var}(X_i)$$

- We have seen in the lab, that if a dataset has a true dimension that is lower than the number of variables, PCA is able to recover this.
- PCA analysis can be used to study the “dimension” of a dataset.
A “fake” 3D dataset

```r
x <- rnorm(100)
y <- rnorm(100)
z <- x * 3 + y * 3
dataset <- cbind(x, y, z)
dataset.pc <- prcomp(dataset, center = TRUE)
summary(dataset.pc)
```

```r
## Importance of components:
##                  PC1     PC2     PC3
## Standard deviation 4.2894 0.92398 5.367e-16
## Proportion of Variance 0.9557 0.04434 0.000e+00
## Cumulative Proportion 0.9557 1.00000 1.000e+00
```
library(scatterplot3d)
par(mfrow=c(1,2))
sd3<-scatterplot3d(dataset,angle=25,box=FALSE,
                  pch=20,cex.lab=1.3)
sd3<-scatterplot3d(dataset,angle=120,box=FALSE,
                  pch=20,cex.lab=1.3)
How our data looks in 3D

It sits on a plane!
How the data looks in the PCA coordinate system
Another example

```r
x <- rnorm(100)
y <- rnorm(100)
z <- x^2 + y^2
dataset <- cbind(x, y, z)
```

dataset.pc

**Variances**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

**Graphs**

- Scatter plot of `x` vs `y` with `z` as an extra dimension.
- Bar chart of `z` values.
http://mypersonality.org/

Thanks to M. Kosinski we have data on 10000 individuals who took the *IPIP proxy for Costa and McCrae’s NEO-PI-R domains (Five Factor Model)*

Personality theory talks about 5 traits (OCEAN)

**Openness** to experience describes a dimension of personality that distinguishes imaginative, creative people from down-to-earth, conventional people.

**Conscientiousness** concerns the way in which we control, regulate, and direct our impulses.

**Extraversion** is marked by pronounced engagement with the external world, versus being comfortable with your own company.

**Agreeableness** reflects individual differences in concern with cooperation and social harmony.

**Neuroticism** refers to the tendency to experience negative emotions.
The questions

The data contains answers to the first 20 questions

The questions presumably probe the different 5 dimensions as follows:

You can take the test
A quick look at the demographics

![Histogram of Age](image1)

- **Age**
  - Frequency: 0 - 5000
  - Years: 20 - 80

![Histogram of Age of Males](image2)

- **Age of males**
  - Frequency: 0 - 2000
  - Years: 20 - 80

![Histogram of Gender](image3)

- **Gender**
  - Frequency: 0 - 5000
  - Male: 0 - 5000
  - Female: 0 - 5000

![Histogram of Age of Females](image4)

- **Age of females**
  - Frequency: 0 - 2500
  - Years: 30 - 70
Principal components of the Personality questions

Variance of the principal components

Variance
Principal components

1 2 3 4 5

1 2 3 4 5

1 2 3 4 5

1 2 3 4 5
Coefficients of questions and their association to OCEAN
Projection of subjects onto the first 6 PCA
Looking at more than one PCA

- Because they allow to summarize data in a few dimensions, PCA are a very handy tool to visualize high dimensional data.
- How much time can you spend looking at the scatterplots of 10,000 variables? How meaningful would that be?
- It is often handy to look at the scatterplots of the first few principal components: this gives us an idea of the spread of the data, of the possible existence of clusters, or relationships
Novembre et al. (2008) Genes mirror geography within Europe, Nature

- 1387 European individuals (number of observations)
- Genotypes at 197,146 Single Nucleotide Polymorphism (number of variables)
- A genotype is here coded as a quantitative variable with values 0, 1, or 2 – corresponding to the number of copies of the more rare allele that an individual has the SNP in question.
- Each subject in the study had all 4 grandparents coming from the same country within Europe
- Plotting each subject on the coordinates of the first 2 principal components, coloring each point by the country of origin.
PCA for genetics data

- The first two principal components pretty much capture the spread of individuals across latitude and longitude.
- Routinely used in genetics to account for global genetic variation that is correlated with the population of origin.
- It is very useful to identify subjects with “outlier” genetic background (for example, askenazi jewish).
- It is tempting to use these genetic principal components to describe the population history, but one has to be somewhat careful.
Reification fallacy

- For any collection of variables, one can find the linear combination that has the maximal variance, aka the first principal component.
- Recall that in the artificial model we created for father and sons heights, the principal component did a good job at reconstructing the common element $z$. If multivariate data is generated in a similar fashion, the first principal component will reconstruct the underlying unmeasured factor.
- However, this is true only if there is an underlying factor.
- It is often tempting to believe that because we can calculate a first principal component, this must capture an underlying factor. But this is not the case.
- This is called reification fallacy. (from res latin for “thing”, making something real)
A book on this topics

*S.J. Gould “The mismeasure of man.”*