

Math 308: Principal Component Analysis

Lily Samuel

2022-02-18

Task:

Perform a principal component analysis of this data using your preferred function. As part of this analysis, please be sure complete the following tasks:

- Report the eigenvalues for all 11 principal components.
- For the first two principal components, plot and interpret components in terms of the original variables. In particular, explain which variables are most highly correlated with each of these two components and how these components are different from each other.
- Choose the smallest number of principal components that you believe can be used to summarize the information from the data and justify your choice.

Solution:

```
library(tinytex)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.2      v readr      2.1.4
## v forcats    1.0.0      v stringr   1.5.0
## v ggplot2    3.4.2      v tibble    3.2.1
## v lubridate  1.9.2      v tidyr     1.3.0
## v purrr      1.0.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(FactoMineR)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg      ggplot2
```

```
## Rows: 1599 Columns: 12
## -- Column specification -----
## Delimiter: ","
## dbl (12): fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlo...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

## # A tibble: 6 x 12
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
##   <dbl>          <dbl>          <dbl>          <dbl>          <dbl>
## 1         7.4         0.7            0            1.9          0.076
## 2         7.8         0.88           0            2.6          0.098
## 3         7.8         0.76           0.04          2.3          0.092
## 4        11.2         0.28           0.56          1.9          0.075
## 5         7.4         0.7            0            1.9          0.076
## 6         7.4         0.66           0            1.8          0.075
## # i 7 more variables: free.sulfur.dioxide <dbl>, total.sulfur.dioxide <dbl>,
## #   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>

## Rows: 1,599
## Columns: 12
## $ fixed.acidity      <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5~
## $ volatile.acidity   <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ~
## $ citric.acid        <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0~
## $ residual.sugar     <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,~
## $ chlorides          <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ~
## $ free.sulfur.dioxide <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16~
## $ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,~
## $ density            <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0~
## $ pH                 <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3~
## $ sulphates          <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0~
## $ alcohol            <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10~
## $ quality            <dbl> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 7~
```

```
wine_chem<-wine%>%select(-quality)
head(wine_chem)
```

```
## # A tibble: 6 x 11
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
##   <dbl>          <dbl>          <dbl>          <dbl>          <dbl>
## 1         7.4         0.7            0            1.9          0.076
## 2         7.8         0.88           0            2.6          0.098
## 3         7.8         0.76           0.04          2.3          0.092
## 4        11.2         0.28           0.56          1.9          0.075
## 5         7.4         0.7            0            1.9          0.076
## 6         7.4         0.66           0            1.8          0.075
## # i 6 more variables: free.sulfur.dioxide <dbl>, total.sulfur.dioxide <dbl>,
## #   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>
```

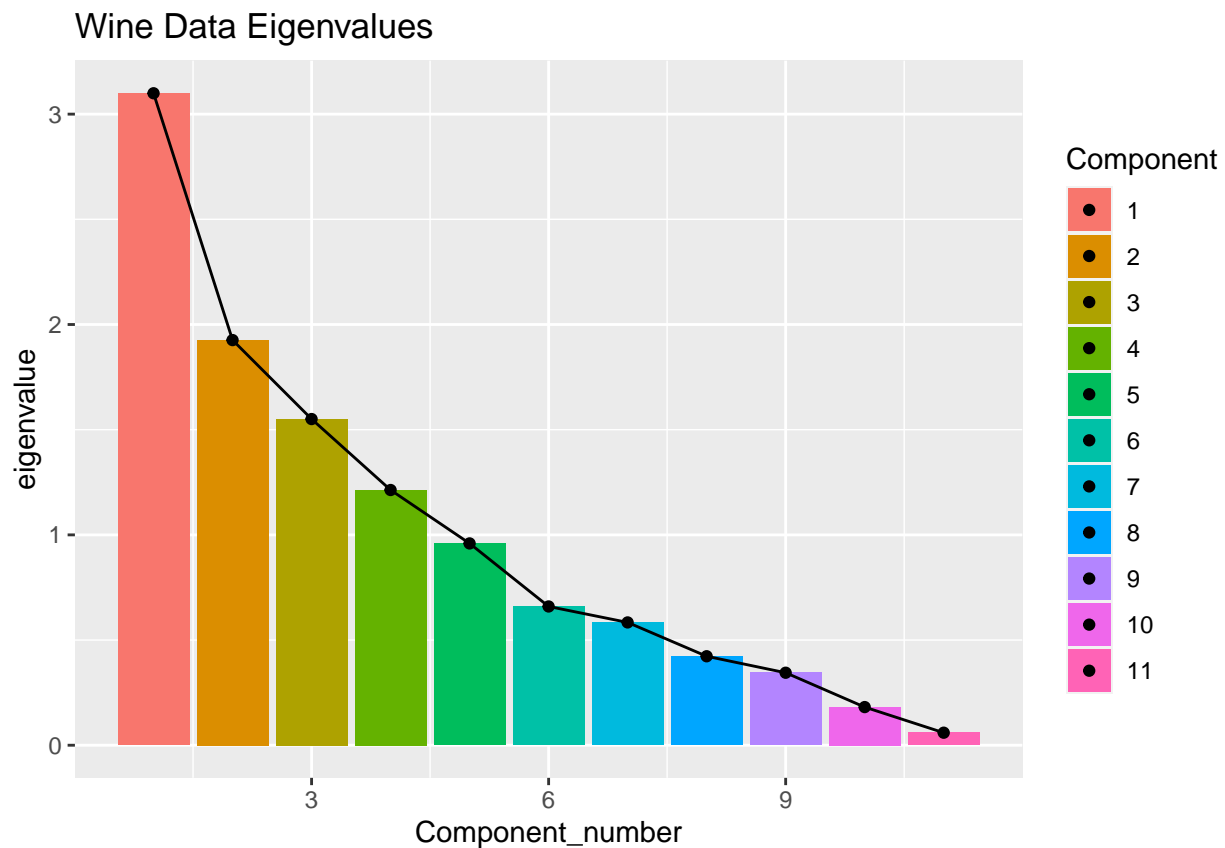
1) Eigenvalues:

```
wine_PCA<-PCA(wine_chem,graph=FALSE)
wine_PCA$eig[, "eigenvalue"]
```

```
##      comp 1      comp 2      comp 3      comp 4      comp 5      comp 6      comp 7
## 3.09913244 1.92590969 1.55054349 1.21323253 0.95929207 0.65960826 0.58379122
##      comp 8      comp 9      comp 10      comp 11
## 0.42295670 0.34464212 0.18133317 0.05955831
```

```
eigenvalues_wine<-as.data.frame(wine_PCA$eig) %>%
  rownames_to_column(var="Component")
eigenvalues_wine <- eigenvalues_wine %>%
  mutate(Component=map_chr(Component,~str_split(.x," ")[[1]][2]),
  Component_number=as.integer(Component),
  Component=factor(Component_number))

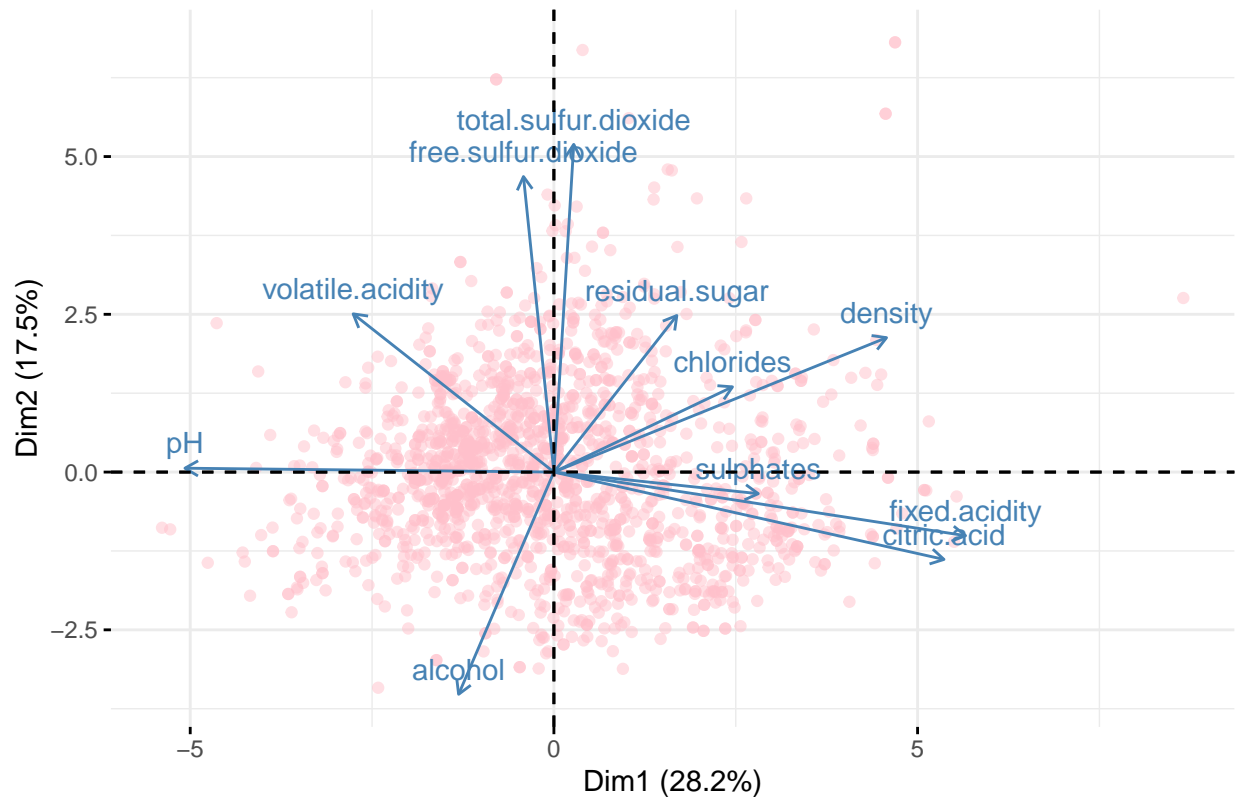
ggplot(eigenvalues_wine,aes(y=eigenvalue,x=Component_number,fill=Component)) +
  geom_bar(stat="identity") + geom_line(aes(fill=NULL)) + geom_point() +ggtitle("Wine Data Eigenvalues")
```



2) Component Analysis:

```
fviz_pca_biplot(wine_PCA,col.ind="pink",
  fill.ind="pink",label="var",
  alpha.ind=c(0.5))
```

PCA – Biplot



```
round(wine_PCA$var$coord[,c(1:2)],2)
```

##		Dim.1	Dim.2
##	fixed.acidity	0.86	-0.15
##	volatile.acidity	-0.42	0.38
##	citric.acid	0.82	-0.21
##	residual.sugar	0.26	0.38
##	chlorides	0.37	0.21
##	free.sulfur.dioxide	-0.06	0.71
##	total.sulfur.dioxide	0.04	0.79
##	density	0.70	0.32
##	pH	-0.77	0.01
##	sulphates	0.43	-0.05
##	alcohol	-0.20	-0.54

From the plot and the table above, we see that the first component mostly is a contrast between the pH level and the fixed acidity, citric acid and density values. These are the variables with the strongest correlations with the first component.

The second component mostly depends on the free and total sulfur dioxide measures, which are contrasted mostly with the alcohol content, as these three variables are most strongly associated with the second component.

3) Number of Components:

```
wine_PCA$eig
```

##	eigenvalue	percentage of variance	cumulative percentage of variance
## comp 1	3.09913244	28.1739313	28.17393
## comp 2	1.92590969	17.5082699	45.68220
## comp 3	1.55054349	14.0958499	59.77805
## comp 4	1.21323253	11.0293866	70.80744
## comp 5	0.95929207	8.7208370	79.52827
## comp 6	0.65960826	5.9964388	85.52471
## comp 7	0.58379122	5.3071929	90.83191
## comp 8	0.42295670	3.8450609	94.67697
## comp 9	0.34464212	3.1331102	97.81008
## comp 10	0.18133317	1.6484833	99.45856
## comp 11	0.05955831	0.5414392	100.00000

We see that using an 80% cumulative variance rule, we would decide to use either 5 or 6 components depending on how strictly we wanted to use the cutoff.