Principal Component Analysis (PCA) on the Iris Data Set

Author: Keaton LaBorde

Date: 12/10/2024

Introduction to PCA

- What is PCA?
  - A technique for dimensionality reduction
  - Transforms data into principal components
  - Retains as much variability as possible
- Why Use PCA?
  - Reduce dimensionality in correlated datasets
  - Improve efficiency, reduce multicollinearity, enhance interpretability

Load the iris dataset and review it

```
# Load in the dataset
data(iris)

# View Iris dataset
head(knitr::kable(iris))
```

```
## [1] "| Sepal.Length| Sepal.Width| Petal.Length| Petal.Width|Species    |"
## [2] "|------------:|-----------:|------------:|-----------:|:----------|"
## [3] "|          5.1|         3.5|          1.4|         0.2|setosa     |"
## [4] "|          4.9|         3.0|          1.4|         0.2|setosa     |"
## [5] "|          4.7|         3.2|          1.3|         0.2|setosa     |"
## [6] "|          4.6|         3.1|          1.5|         0.2|setosa     |"
```

```
# Structure of the iris dataset
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Summary statistics for each variable
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
# Check for missing values
any(is.na(iris))
```

```
## [1] FALSE
```

Iris Dataset Overview

- Dataset Summary:
  - Observations: 150
  - Features: 4
  - Species: 3
- Basic Data Check:
  - No missing values
  - Summary statistics and structure

Exploratory Data Analysis

- Correlation Check:
  - Scatter plots to visualize pairwise relationships
  - Correlation matrix to quantify relationships
- Objective: Identify correlated features to guide PCA

Pairwise Scatter Plots

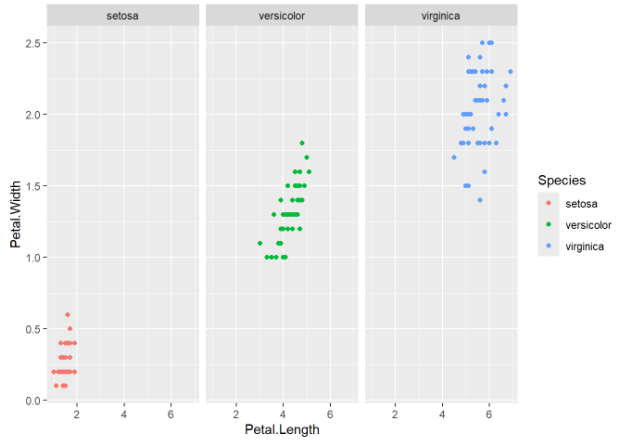
- Scatter Plots:
  - Visualize the relationship between pairs of features
  - Explore how features correlate for different species
- Observation: Petal Length and Petal Width are strongly correlated

```r
# Load necessary libraries
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tid
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to
```

```r
library(ggbiplot)
```

```r
ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point() +
  facet_wrap(~ Species)
```



Correlation Matrix

Calculate the correlation matrix

```
cor_matrix <- cor(iris[, 1:4])
cor_matrix
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

- Matrix of correlations between features:
  - High correlation between Petal Length and Petal Width
  - Lower correlations between Sepal Length and Sepal Width
- PCA Objective: Reduce redundant features

Remove any categorical data (Species)

```
flowerData <- subset(iris, select = -Species)

# View the first few rows of the flowerData
head(flowerData)
```

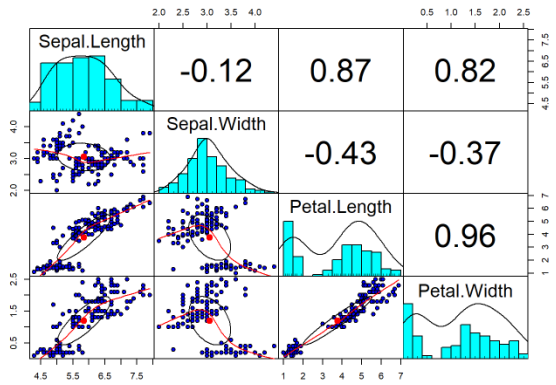
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.4.2
```

Check association between independent variables

```
pairs.panels(flowerData,
             gap = 0,
             bg = "blue",
             pch=21)
```




```
iris.pca <- prcomp(flowerData, scale = TRUE,
                   center = TRUE, retx = T)
iris.pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.7083611 0.9560494 0.3830886 0.1439265
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2         PC3         PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Principal Component Analysis (PCA)

- PCA Basics:
  - Eigenvalues and eigenvectors represent directions of maximum variance
  - Principal Components (PCs) are uncorrelated
- First Two Components:
  - PC1 explains 72.96% of the variance
  - PC2 explains 22.85% of the variance
  - Together: 95.81% of variance explained

PCA Results

- PCA Object Summary:
  - Standard deviations (eigenvalues)
  - Center (means subtracted)
  - Rotation (eigenvectors)
- First Two PCs:
  - PC1: Key driver of variability in petal-related features
  - PC2: Captures additional variance, especially between species

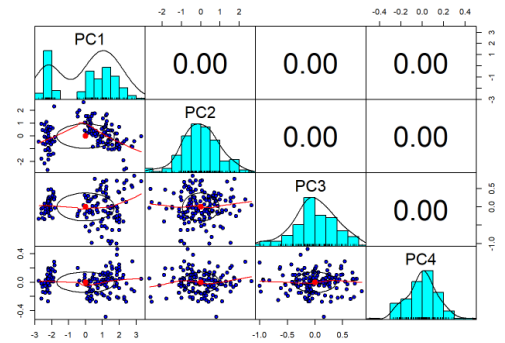
```
# preview our object with summary
summary((iris.pca))
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

The **first PC**, `PC~1~`, accounts for **72.96%** of the variability in the original data. The **second PC**, `PC~2~`, accounts for **22.85%**. Together, they account for **95.81%** of the variability in the original `m` = 5 variables. This shows that the first two components capture most of the important information from the original data, reducing the need for all original features.

- Principal components (PCs) are constrained to be ***uncorrelated/orthogonal/independent*** with each other.

```
library(psych)
pairs.panels(iris.pca$x,
             gap=0,
             bg = "blue",
             pch=21)
```



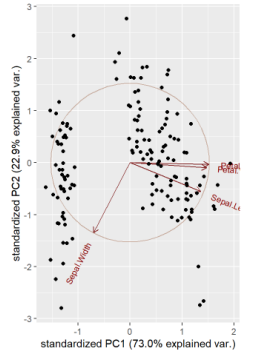
Now the correlation coefficients are zero, so we can get rid of multi-collinearity issues.

PCA Visualization

- Biplot:
  - Principal components plotted with variables and observations
  - Clear separation of species
  - Setosa is distinct, Versicolor and Virginica are closer
- Interpretation:
  - Vector angles indicate correlation between variables (e.g., Petal Length & Petal Width)

Biplot using `ggbiplot` function:

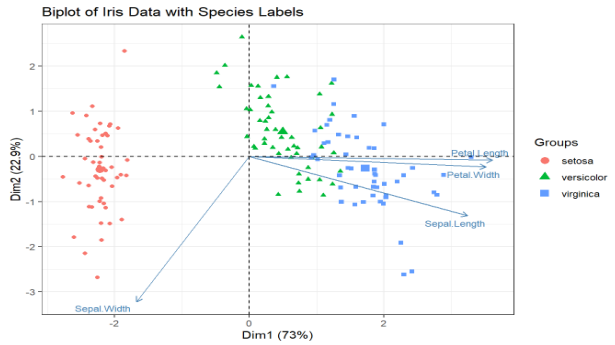
```
library(devtools)

install_github("vqv/ggbiplot")

library(ggbiplot)

# circle: The correlation circle is a visualisation displaying how much the original variables are correlated with the first two principal components.
ggbiplot(iris.pca, groups = iris$species, ellipse = TRUE, ellipse.prob = 0.8, circle = TRUE)
```




```
# Load necessary library
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.4.2
```

```
# Plot the results with species labels
fviz_pca(iris.pca,
         repel = TRUE,
         label = "var",          # Remove default numeric labels
         habillage = iris$Species, # Color points by species
         labelsize = 3) +
  theme_bw() +
  labs(title = "Biplot of Iris Data with Species Labels")
```



Orthogonality of Principal Components

- Property of PCA: PCs are uncorrelated (orthogonal)
- Dot Product Check:
  - Confirming orthogonality by computing the dot product between principal components
  - PCs are independent

Check orthogonality of the principal components

```
dot_product <- cor(iris.pca$x)
dot_product
```

```
##              PC1          PC2           PC3          PC4
## PC1  1.000000e+00 6.417823e-16 -1.680803e-15  9.358505e-16
## PC2  6.417823e-16 1.000000e+00  4.117641e-16  2.143073e-15
## PC3 -1.680803e-15 4.117641e-16  1.000000e+00 -1.757211e-15
## PC4  9.358505e-16 2.143073e-15 -1.757211e-15  1.000000e+00
```

PCA Biplot cont.

- Interpretation:
  - Vectors represent original features
  - Small angle = strong correlation between features
  - Points represent observations, colors represent species
  - The results show close to zero off the diagonal
  - Iris.pca is truly orthogonal
  - Captures independent and non-redundant information

Advantages of PCA

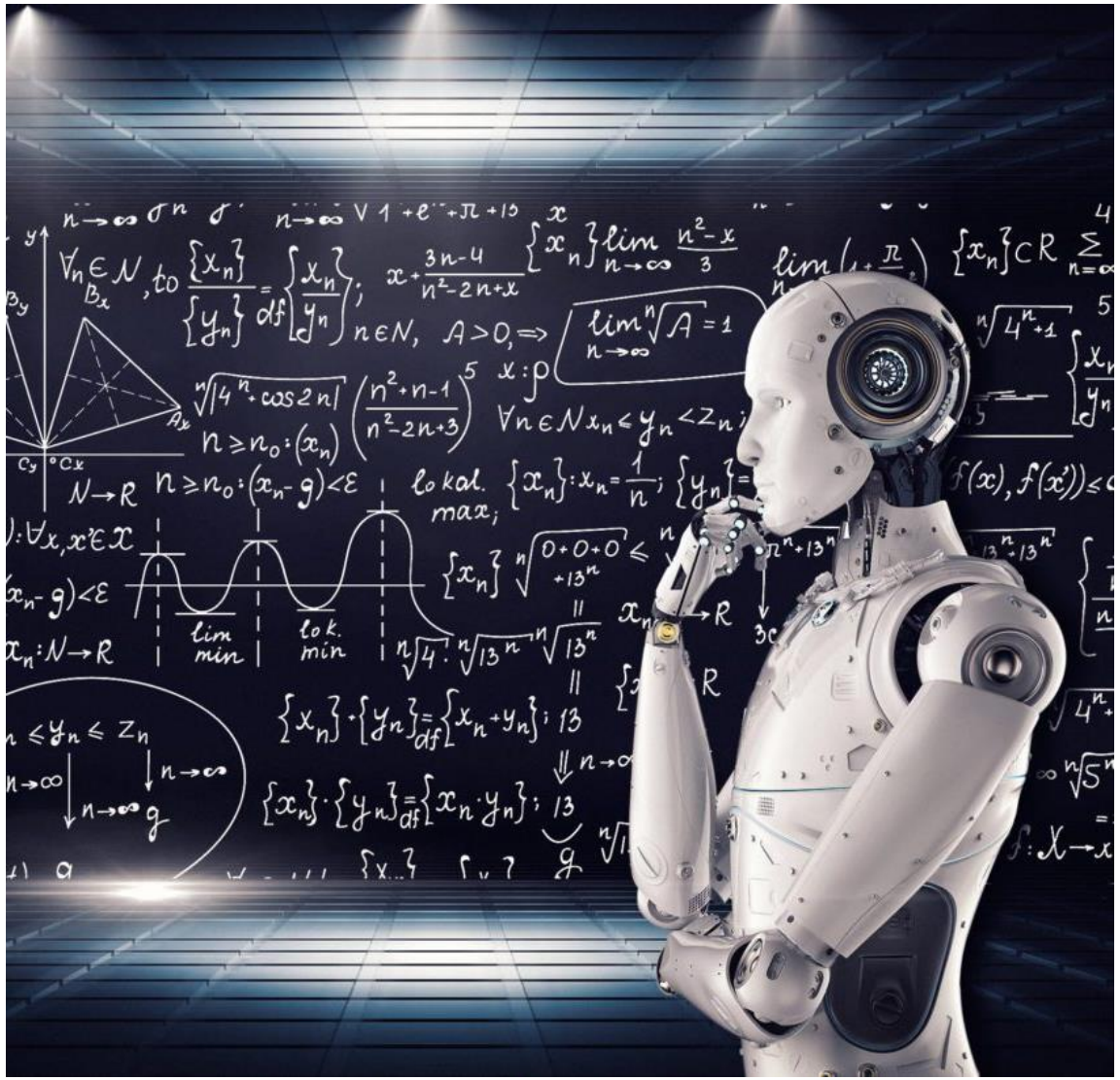
- **Key Benefits**:
  - Reduces dimensionality
  - Resolves multicollinearity
  - Facilitates data visualization
  - Enhances computational efficiency

Disadvantages of PCA

- **Key Limitations**:
  - Loss of interpretability (PCs are linear combinations)
  - Assumes linearity, which may not apply to all data
  - Sensitive to scaling and outliers

Applications of PCA

- Machine Learning: Reduce features for better model performance and prevent overfitting
- Image Processing: Reduce image size while retaining important features
- Finance: Dimensionality reduction for portfolio optimization and risk management



Conclusion

- Summary:
  - PCA effectively reduced dimensionality of the Iris dataset
  - Retained 95.81% of the variability in 2 principal components
  - Clear separation of species, aiding in classification
- Future Considerations:
  - Use PCA for feature reduction in machine learning models

References

- Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics.
- Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers. Sage.