Introduction to Principal Component Analysis (PCA) in R

Understanding Dimensionality Reduction

Mahesh Divakaran

Amity University, Lucknow

2024-10-07

Section 1

Introduction

Why Do We Need PCA?

- In industries like retail, we often work with high-dimensional data.
- Imagine analyzing customer data with 5 characteristics:
  - *Monthly expense, Age, Gender, Purchase frequency, Product rating*
- **Challenges with High-Dimensional Data:**
  - Visualizing 5 dimensions is not intuitive for human understanding.
  - It becomes difficult to identify patterns and relationships among variables.
  - Increased complexity can lead to overfitting in models, reducing generalization.
- **Principal Component Analysis (PCA) simplifies this problem:**
  - Reduces the dimensionality of the data.
  - Captures the most important information by identifying patterns.
  - Helps visualize complex data structures in 2D or 3D.
  - **Example:** By reducing 5 dimensions into 2 or 3 principal components, we can create scatter plots that highlight key groupings or trends among customers.
- **Benefits of PCA:**
  - Enhances interpretability while preserving the variability in the data.
  - Facilitates data exploration and analysis, aiding in decision-making processes.
  - Useful for subsequent machine learning algorithms by reducing noise and improving performance.

Challenges with High-Dimensional Data

- **Curse of Dimensionality:**
  - As dimensions increase, it becomes harder to analyze and interpret data.
  - Models become more complex and may suffer from overfitting.
  - **Impact:** Increased computational cost and time for training models.
- **Redundant Information:**
  - High-dimensional data often contains correlated variables, adding noise.
  - **Example:** Monthly expense and purchase frequency might be correlated, leading to redundant insights.
- **Visual Limitations:**
  - Human beings are limited to visualizing up to 3 dimensions effectively.
  - Higher dimensions can lead to misinterpretations.
- **PCA helps address these challenges:**
  - By condensing the data into fewer, uncorrelated dimensions.
  - **Result:** Easier to visualize and interpret relationships among variables.

What is PCA?

- PCA is a statistical technique for **dimensionality reduction**.
- It transforms the original data into a new set of variables (principal components), which are:
  - **Uncorrelated** with each other.
  - **Ordered** by the amount of variance they explain in the data.
  - **Example:** The first principal component accounts for the largest variance, while the second captures the next largest variance.
- **Key idea:** Keep the components that explain the most variance and discard the rest.
- **Visualization Aid:** PCA can generate 2D or 3D plots that illustrate how data points cluster, allowing for easier interpretation of customer segments.
  - **Practical Use:** Marketers can identify target segments based on purchasing behaviors.

Real-life Example: Retail Data

- Consider a dataset with the following customer information:
  - Monthly expense: $300
  - Age: 27 years
  - Gender: Female
  - Purchase frequency: 12 times/month
  - Product rating: 4.5/5
- PCA would identify which of these characteristics contribute the most to customer behavior:
  - **For instance:** Monthly expense and purchase frequency might be the most significant.
  - **Customer Segmentation:** PCA can help categorize customers into segments based on spending behavior.
- **PCA Visualization:**
  - After PCA transformation, we might find that these two dimensions allow us to plot customers in a way that highlights spending patterns and purchasing behaviors.

How Does PCA Work? (5 Steps)

1. **Data Normalization**
   - Ensure all variables contribute equally by standardizing the data (mean 0, standard deviation 1).
2. **Covariance Matrix**
   - Compute the covariance matrix to capture relationships between variables.
3. **Eigenvectors and Eigenvalues**
   - Eigenvectors: Directions of maximum variance.
   - Eigenvalues: Amount of variance explained by each eigenvector.
4. **Select Principal Components**
   - Choose the eigenvectors with the highest eigenvalues (most variance).
5. **Transform Data**
   - Project the original data onto the new space formed by the principal components.

Applications of PCA

- **Finance:**
  - Analyze stock prices by reducing hundreds of variables (features) to a handful of principal components.
  - Forecast future prices by focusing on the most important factors.
- **Image Processing:**
  - Use PCA for image compression—reduce the size of an image while retaining important features.
  - In **face recognition**, PCA can identify key distinguishing features.
- **Healthcare:**
  - Dimensionality reduction in MRI scans to visualize large, complex datasets.
  - Use in **disease detection** from medical images.
- **Security:**
  - **Fingerprint recognition** systems apply PCA to extract the most relevant features from the fingerprint data (e.g., texture, ridge patterns).

Applications of PCA

- **Marketing:**
  - Segment customers based on purchase behaviors, identifying the principal factors that influence buying decisions.
- **Genomics:**
  - Analyze gene expression data to identify the most important genes driving a particular condition.
- **Sports Analytics:**
  - Evaluate player performance by reducing multiple performance metrics into principal components, focusing on key indicators.
- **E-commerce:**
  - Optimize product recommendations by identifying underlying customer preferences using PCA.

General Methods for Principal Component Analysis

- **Spectral Decomposition**: Examines the covariances/correlations between variables.
- **Singular Value Decomposition (SVD)**: Examines the covariances/correlations between individuals.

In R:

- `princomp()` uses spectral decomposition.
- `prcomp()` and `PCA()` [FactoMineR] use singular value decomposition (SVD).

Note: SVD has better numerical accuracy, hence `prcomp()` is generally preferred over `princomp()`.

`prcomp()` vs `princomp()` in R

prcomp()	princomp()	Description
sdev	sdev	Standard deviations of the principal components
rotation	loadings	Matrix of variable loadings (columns are eigenvectors)
center	center	Variable means (subtracted before PCA)
scale	scale	Variable standard deviations (scaling applied)
x	scores	Coordinates of observations on principal components

Arguments for `prcomp()` and `princomp()`

Arguments for `prcomp()`:

- `x`: Numeric matrix or data frame.
- `scale`: Logical, scales variables to unit variance before analysis.

Arguments for `princomp()`:

- `x`: Numeric matrix or data frame.
- `cor`: Logical, centers and scales data if `TRUE`.
- `scores`: Logical, calculates coordinates on principal components if `TRUE`.

Output of `prcomp()` and `princomp()`

Output Elements:

- **`sdev`**: Standard deviations of principal components.
- **`rotation`** (`loadings` for `princomp()`): Matrix of variable loadings.
- **`center`**: Variable means (centered values).
- **`scale`**: Variable standard deviations.
- **`x`** (`scores` for `princomp()`): Coordinates of observations on the principal components.

Section 2

Exploratory Data Analysis (EDA)

Getting the Data

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

# Structure of the 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
```

Data Summary

```
# Summary statistics of the dataset
summary(iris) %>% kable()
```

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

The iris dataset contains 150 observations and 5 variables:

- Sepal Length
- Sepal Width
- Petal Length
- Petal Width
- Species (setosa, versicolor, virginica)
- Summary statistics show the range, quartiles, and mean values for each numerical variable.

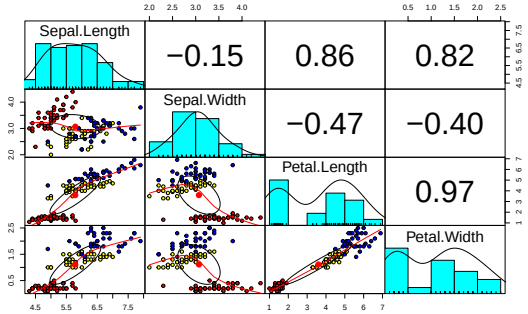
Partition the Data

```r
# Set seed for reproducibility
set.seed(111)

# Split the data into training (80%) and testing (20%) sets
ind <- sample(2, nrow(iris), replace = TRUE, prob = c(0.8, 0.2))

# Partition the data
training <- iris[ind == 1,]
testing <- iris[ind == 2,]
```

- The dataset is split into training and testing sets to build and evaluate models.
- Training Set: 80% of the data.
- Testing Set: 20% of the data.

Exploratory Data Analysis (EDA) - Correlation

```r
# Load the psych package for scatter plots and correlation analysis
library(psych)

# Create pair panels excluding the species factor
pairs.panels(training[,-5],
             gap = 0,
             bg = c("red", "yellow", "blue")[training$Species],
             pch = 21)
```



Exploratory Data Analysis (EDA) - Correlation

The scatter plots (lower triangle) and correlations (upper triangle) between numerical variables show relationships:

- Highly Correlated Pairs:
- Petal Length and Petal Width
- Sepal Length and Petal Length
- Sepal Length and Petal Width
- These correlations indicate multicollinearity, which can negatively affect model performance.

Handling Multicollinearity with PCA

Problem

- Multicollinearity can lead to misleading results in predictive models.

Solution

- Apply Principal Component Analysis (PCA) to remove redundancy and extract uncorrelated components.
- PCA transforms the original data into a set of principal components that explain the majority of the variance while addressing multicollinearity.
- The first few components can capture most of the variability in the data, reducing the need for all original variables.

Check for Null Values

The presence of missing values can bias the result of PCA.

- Before performing PCA, it is crucial to ensure there are no missing values in the dataset.
- In the iris dataset, no missing values are present as shown by the output of `colSums(is.na(iris))`.

```
# Check for missing values
colSums(is.na(iris))
#> Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
#>            0            0            0            0            0
```

Section 3

Principal Component Analysis (PCA)

Applying PCA

```
# Apply PCA on training data
pca_iris <- prcomp(training[,-5], scale = TRUE)

# Summary of PCA
summary(pca_iris)
#> Importance of components:
#>                           PC1    PC2     PC3    PC4
#> Standard deviation     1.7173 0.9404 0.38432 0.1371
#> Proportion of Variance 0.7373 0.2211 0.03693 0.0047
#> Cumulative Proportion  0.7373 0.9584 0.99530 1.0000
```

Interpretation:

- PC1 explains about 73.7% of the total variance in the data.
- PC2 adds another 22.1%, leading to a cumulative proportion of 95.8% of the variance explained by the first two principal components.
- PC3 and PC4 together contribute less than 5% of the remaining variance, meaning most of the information is captured by PC1 and PC2.

Thus, for visualization purposes, PC1 and PC2 are sufficient to represent the dataset without significant loss of information.

Access to the PCA results

```
# Load the factoextra package
library(factoextra)
# Results for Variables
res.var <- get_pca_var(pca_iris)
res.var$coord  %>% kable()        # Coordinates
```

	Dim.1	Dim.2	Dim.3	Dim.4
Sepal.Length	0.8839386	-0.3744264	0.2783529	0.0312587
Sepal.Width	-0.5024996	-0.8588093	-0.0982892	-0.0167318
Petal.Length	0.9913349	-0.0275715	-0.0674651	-0.1092861
Petal.Width	0.9657279	-0.0758483	-0.2366677	0.0748665

```
res.var$contrib %>% kable()            # Contributions to the PCs
```

	Dim.1	Dim.2	Dim.3	Dim.4
Sepal.Length	26.493283	15.8544805	52.456399	5.195838
Sepal.Width	8.561758	83.4089554	6.540618	1.488669
Petal.Length	33.322098	0.0859684	3.081525	63.510408
Petal.Width	31.622861	0.6505957	37.921458	29.805086

Access to the PCA results

```
res.var$cos2  %>% kable()              # Quality of representation
```

	Dim.1	Dim.2	Dim.3	Dim.4
Sepal.Length	0.7813474	0.1401951	0.0774804	0.0009771
Sepal.Width	0.2525058	0.7375535	0.0096608	0.0002800
Petal.Length	0.9827448	0.0007602	0.0045515	0.0119435
Petal.Width	0.9326304	0.0057530	0.0560116	0.0056050

Access to the PCA results

```
# Results for individuals
res.ind <- get_pca_ind(pca_iris)
res.ind$coord  %>% kable()            # Coordinates
```

	Dim.1	Dim.2	Dim.3	Dim.4
1	-2.0942621	-0.3904799	0.1117896	0.0155965
2	-1.9004407	0.7013950	0.2145513	0.0931822
3	-2.1852900	0.4016807	-0.0630100	0.0292277
4	-2.1194349	0.6457879	-0.1148752	-0.0610428
5	-2.2205378	-0.5411573	-0.0319019	-0.0386754
6	-1.9160708	-1.3579482	-0.0397399	0.0030169
7	-2.2686138	0.0396553	-0.3537552	-0.0239723
8	-2.0607048	-0.1447322	0.0697462	-0.0300848
9	-2.1492140	1.1422048	-0.1694966	-0.0186376
10	-2.0063320	0.5113963	0.2304915	-0.0502480
11	-2.0019748	-0.9352522	0.2543675	0.0008732
12	-2.1534233	-0.0496619	-0.1159887	-0.1300380
13	-2.0373705	0.7604250	0.2080917	-0.0067508
14	-2.4468051	1.0071236	-0.2022257	-0.0113928
15	-2.0401384	-1.7208508	0.4684543	0.1655990
16	-2.1119798	-2.4948975	-0.0349015	0.0422916
17	-2.0452623	-1.3513862	-0.0004528	0.1813739
19	-1.7392003	-1.2933069	0.3613617	0.0404016
20	-2.1788452	-0.9998935	-0.1467341	-0.0365116
21	-1.7460763	-0.3414348	0.4019288	-0.0085355
22	-2.0406559	-0.8115354	-0.1724961	0.0623294
23	-2.5997621	-0.3411737	-0.3444410	0.0289538
24	-1.6465682	-0.0293598	-0.0506968	0.1517617
26	-1.7733367	0.6497586	0.2828642	0.0316857
27	-1.8795636	-0.1677220	-0.1030696	0.0698280
28	-1.9994560	-0.4404758	0.1899243	-0.0013109

Access to the PCA results

```
res.ind$contrib %>% kable()          # Contributions to the PCs
```

	Dim.1	Dim.2	Dim.3	Dim.4
1	1.2392884	0.1436929	0.0705064	0.0107792
2	1.0205140	0.4636211	0.2597095	0.3847688
3	1.3493621	0.1520547	0.0223998	0.0378550
4	1.2692597	0.3930227	0.0744523	0.1651213
5	1.3932424	0.2759847	0.0057420	0.0662833
6	1.0373694	1.7378182	0.0089100	0.0004033
7	1.4542245	0.0014820	0.7060434	0.0254656
8	1.1998913	0.0197410	0.0274452	0.0401078
9	1.3051778	1.2294925	0.1620867	0.0153927
10	1.1374071	0.2464638	0.2997336	0.1118849
11	1.1324722	0.8243192	0.3650472	0.0000338
12	1.3102952	0.0023243	0.0759027	0.7493320
13	1.1728713	0.5449426	0.2443065	0.0020195
14	1.6916445	0.9558802	0.2307271	0.0057517
15	1.1760603	2.7907719	1.2381123	1.2152034
16	1.2603462	5.8660172	0.0068725	0.0792577
17	1.1819752	1.7210634	0.0000012	1.4577512
19	0.8546918	1.5763082	0.7367326	0.0723322
20	1.3414148	0.9422050	0.1214752	0.0590738
21	0.8614633	0.1098635	0.9114314	0.0032285
22	1.1766571	0.6206587	0.1678743	0.1721551
23	1.9097545	0.1096955	0.6693533	0.0371487
24	0.7660724	0.0008124	0.0145006	1.0206058
26	0.8885722	0.3978707	0.4514210	0.0444897
27	0.9982156	0.0265105	0.0599358	0.2160693
28	1.1296244	0.1828446	0.2035105	0.0000761
29	1.0943456	0.0541933	0.3682504	0.2163190
30	1.2323561	0.1483515	0.0482478	0.4842832

Access to the PCA results

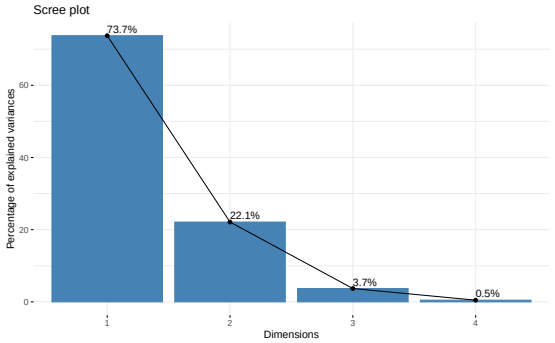
```
res.ind$cos2   %>% kable()            # Quality of representation
```

	Dim.1	Dim.2	Dim.3	Dim.4
1	0.9636982	0.0335024	0.0027459	0.0000534
2	0.8685366	0.1183055	0.0110698	0.0020881
3	0.9663733	0.0326504	0.0008034	0.0001729
4	0.9119027	0.0846619	0.0026789	0.0007564
5	0.9434833	0.0560357	0.0001947	0.0002862
6	0.6654645	0.3342476	0.0002863	0.0000016
7	0.9758642	0.0002982	0.0237287	0.0001090
8	0.9937478	0.0049020	0.0011384	0.0002118
9	0.7759539	0.2191616	0.0048261	0.0000584
10	0.9269605	0.0602242	0.0122339	0.0005814
11	0.8101187	0.1768027	0.0130784	0.0000002
12	0.9929702	0.0005281	0.0028808	0.0036209
13	0.8697543	0.1211628	0.0090733	0.0000095
14	0.8501425	0.1440318	0.0058072	0.0000184
15	0.5647163	0.4017885	0.0297746	0.0037207
16	0.4173342	0.5823845	0.0001140	0.0001673
17	0.6923092	0.3022463	0.0000000	0.0054444
19	0.6262984	0.3463261	0.0270375	0.0003380
20	0.8227647	0.1732728	0.0037315	0.0002310
21	0.9163814	0.0350401	0.0485566	0.0000219
22	0.8574635	0.1356098	0.0061268	0.0007999
23	0.9662774	0.0166412	0.0169615	0.0001199
24	0.9903334	0.0003149	0.0009388	0.0084129
26	0.8620579	0.1157332	0.0219336	0.0002752
27	0.9878006	0.0078657	0.0029704	0.0013634
28	0.9455779	0.0458900	0.0085317	0.0000004
29	0.9680907	0.0143741	0.0163151	0.0012202
30	0.9610210	0.0346865	0.0018843	0.0024081

Visualizing PCA Results

- **Scree Plot:** Shows variance explained by each component.

```
# Visualize the percentage of variance explained by each component
fviz_eig(pca_iris, addlabels = TRUE)
```



Section 4

Visualizing PCA

Visualizing PCA Results

Scree Plot

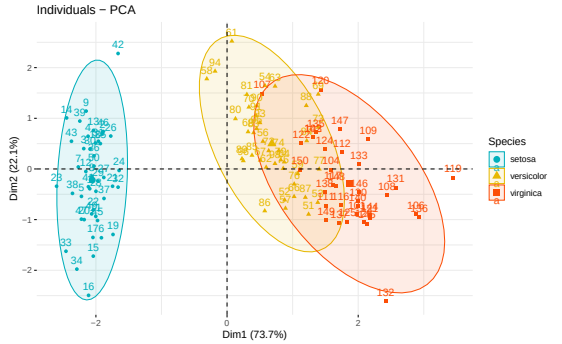
Interpretation

- The scree plot shows the percentage of variance explained by each principal component (PC).
- PC1 explains around 73.7% of the variance, which means it captures most of the variability in the data.
- PC2 explains an additional 22.1% of the variance.
- Together, PC1 and PC2 explain about 95.8% of the total variance. This suggests that the first two principal components are sufficient for representing the dataset, reducing the dimensionality from 4 to 2 with minimal loss of information.

Visualizing PCA Results

- **PCA Plot:** Visualizes species in the reduced PCA space, aiding interpretation.

```
# Plot individuals based on principal components
fviz_pca_ind(pca_iris,
             col.ind = training$Species,
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE,
             legend.title = "Species")
```



Visualizing PCA Results

PCA Individual Plot

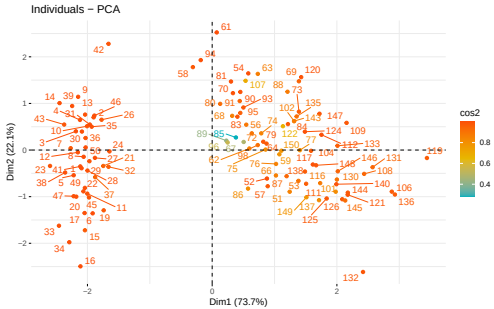
Interpretation

- The PCA plot visualizes the relationship between data points (individuals) in the reduced principal component space (PC1 and PC2).
- The three species of the iris dataset (`setosa`, `versicolor`, and `virginica`) are represented by different colors and symbols.
- The points are clustered according to species, with clear separations between `setosa` (on the left), `versicolor` (center), and `virginica` (right).
- Ellipses represent the confidence regions for each species, showing that `setosa` is distinctly separated from the other two species, while there is slight overlap between `versicolor` and `virginica`.
- This plot shows how PCA has reduced the data to two dimensions while retaining the ability to distinguish between species.
- This analysis indicates that PCA is effective in simplifying the dataset while maintaining most of the variance and keeping the species well-separated in the reduced feature space.

Graph of Individuals

The plot below shows the individuals in the PCA space, where individuals with similar profiles are grouped together. The coloring is based on the cos2 value, representing the quality of representation for each individual.

```
# Visualize individuals' PCA
fviz_pca_ind(pca_iris,
             col.ind = "cos2", # Color by quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```

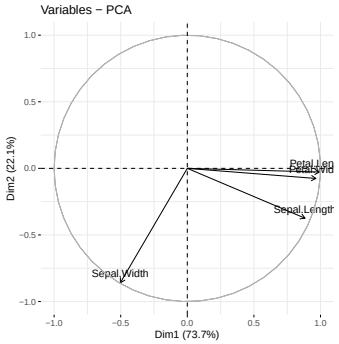


Biplot of the Attributes

With the biplot, we can visualize the similarities and dissimilarities between the samples, and it further shows the impact of each attribute on each of the principal components.

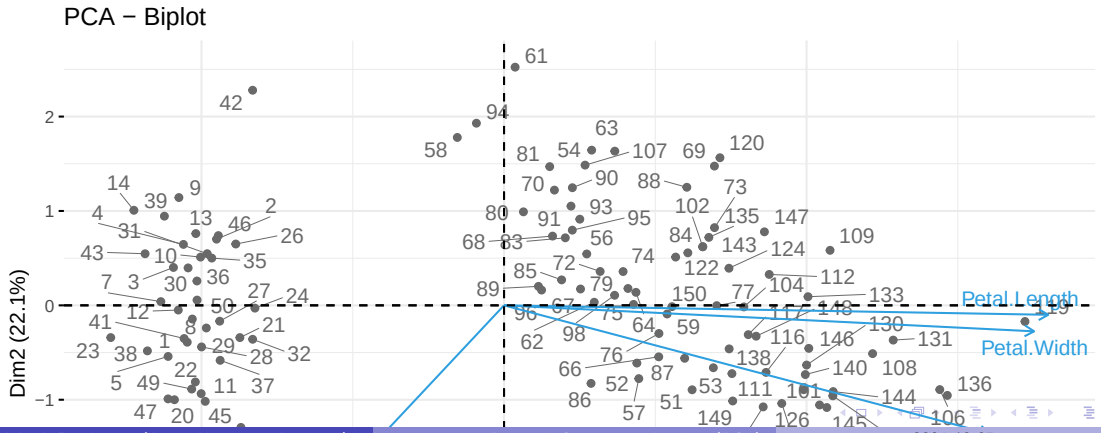
Graph of the Variables

```
# Biplot of the variables
fviz_pca_var(pca_iris, col.var = "black")
```



Biplot of individuals and variables

```
fviz_pca_biplot(pca_iris, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
                )
```



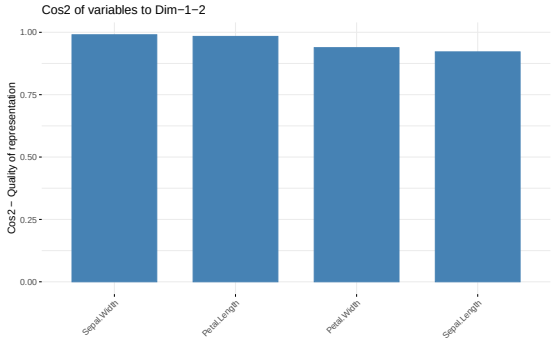
Biplot of the Attributes

Interpretation

- The biplot provides a way to visualize the relationship between the variables and the principal components.
- Here, Sepal Length and Petal Length are positively correlated, as indicated by their proximity on the plot, whereas Sepal Width shows a negative correlation with Sepal Length.
- Petal Width has a smaller contribution compared to the other variables along the first principal component (Dim 1), which explains 73.7% of the variation. Dim 2 explains an additional 22.1% of the variation.

Variables Contribution to Principal Components

```
# Visualize variables' contribution to the components
fviz_cos2(pca_iris, choice = "var", axes = 1:2)
```



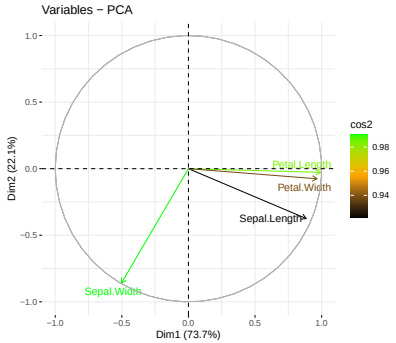
Variables Contribution to Principal Components

Interpretation

- The plot above shows the Cos2 values for each variable with respect to the first two principal components.
- Variables such as Sepal Width and Petal Length have higher Cos2 values, meaning they contribute significantly to the first two principal components.
- A high Cos2 value indicates good representation of the variable on the corresponding principal component.

Biplot combined with Cos2 score for better representation

```
fviz_pca_var(pca_iris, col.var = "cos2", gradient.cols = c("black", "orange", "green"), repel = TRUE)
```



Biplot combined with Cos2 score for better representation

Interpretation

- This plot combines the biplot with Cos2 information. The colors represent the quality of representation (Cos2) for each variable.
- Variables with higher Cos2 values are shown in green, indicating better representation in the biplot, such as `Sepal Width` and `Petal Length`.
- The variables with lower Cos2 values, such as `Sepal Length`, are shown in black, indicating a relatively lower contribution to the principal components.

Access to the PCA results

```
library(factoextra)
# Eigenvalues
eig.val <- get_eigenvalue(pca_iris)
eig.val%>% kable()
```

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.9492285	73.7307121	73.73071
Dim.2	0.8842617	22.1065429	95.83725
Dim.3	0.1477043	3.6926074	99.52986
Dim.4	0.0188055	0.4701377	100.00000

PCA Conclusion

The PCA conducted on the Iris dataset has yielded insightful results. Below is a comprehensive summary, including the visual interpretations from earlier images and the principal component loadings:

Key Findings:

Explained Variance:

The first two principal components (PC1 and PC2) capture a substantial amount of variance in the data, specifically 73.7% and 22.1% respectively, which together account for approximately 95.8% of the total variance. This suggests that these two components provide a very good approximation of the original dataset in lower dimensions.

PCA Conclusion

Key Findings:

Variable Contributions (Cos2 and Loadings):

- The Cos2 values and variable contribution plots indicate that Sepal Width and Petal Length contribute significantly to the first two principal components, while Petal Width and Sepal Length also play a role but to a slightly lesser extent.
- The variable loadings (rotation) reveal that Sepal Length and Petal Length are positively correlated with PC1, while Sepal Width has a negative correlation with PC1 and a strong negative correlation with PC2. This is reflected in the PCA biplot where the vectors corresponding to these variables point in different directions, showing that these features contribute differently to the variance captured by each component.

PCA Conclusion

Key Findings:

PCA Biplot (Variables):

- The PCA biplot of variables shows that Sepal Length, Petal Length, and Petal Width are strongly aligned with PC1, indicating they have a higher influence on this principal component. On the other hand, Sepal Width is primarily aligned with PC2, demonstrating its influence is stronger on this component.
- This separation of variables across components allows for better interpretation of how the original features are spread across the lower-dimensional space.

Cos2 Quality of Representation:

- The Cos2 plot confirms that all variables are well represented by the first two principal components, with Sepal Width and Petal Length showing particularly high values, indicating a strong contribution to these components.

Section 5

Conclusion

PCA Conclusion

Key Findings:

Graph of Individuals:

- The graph of individuals illustrates that samples (individuals) with similar profiles are grouped together in the PCA plot. The gradient colors (ranging from blue to orange) indicate the quality of representation based on Cos2 values. Blue-colored individuals are well represented, while orange-shaded individuals are not as well represented by the first two components.
- The separation of species is also visible in this plot, with clear groupings of different species reflecting their inherent differences in terms of the measured features (sepal length, sepal width, petal length, petal width).

PCA Conclusion

Conclusion

- The PCA analysis of the `Iris dataset` shows that the majority of the variance can be captured by the first two components, with `PC1` being dominated by `Petal Length` and `Sepal Length`, and `PC2` being largely influenced by `Sepal Width`.
- This dimensionality reduction allows for a simplified yet effective visualization of the dataset, where species are naturally grouped according to their feature similarities. The results suggest that `Sepal Width` is a key feature for distinguishing between species in the context of the second component.
- The visualizations such as the `Cos2 plot`, `PCA biplot of variables`, and the `PCA individual plot` further aid in interpreting the relationships between features and individuals, providing a strong foundation for deeper analysis or classification tasks using the reduced dataset.