PCA Class Activity

Vedanivas

18/03/2024

Contents

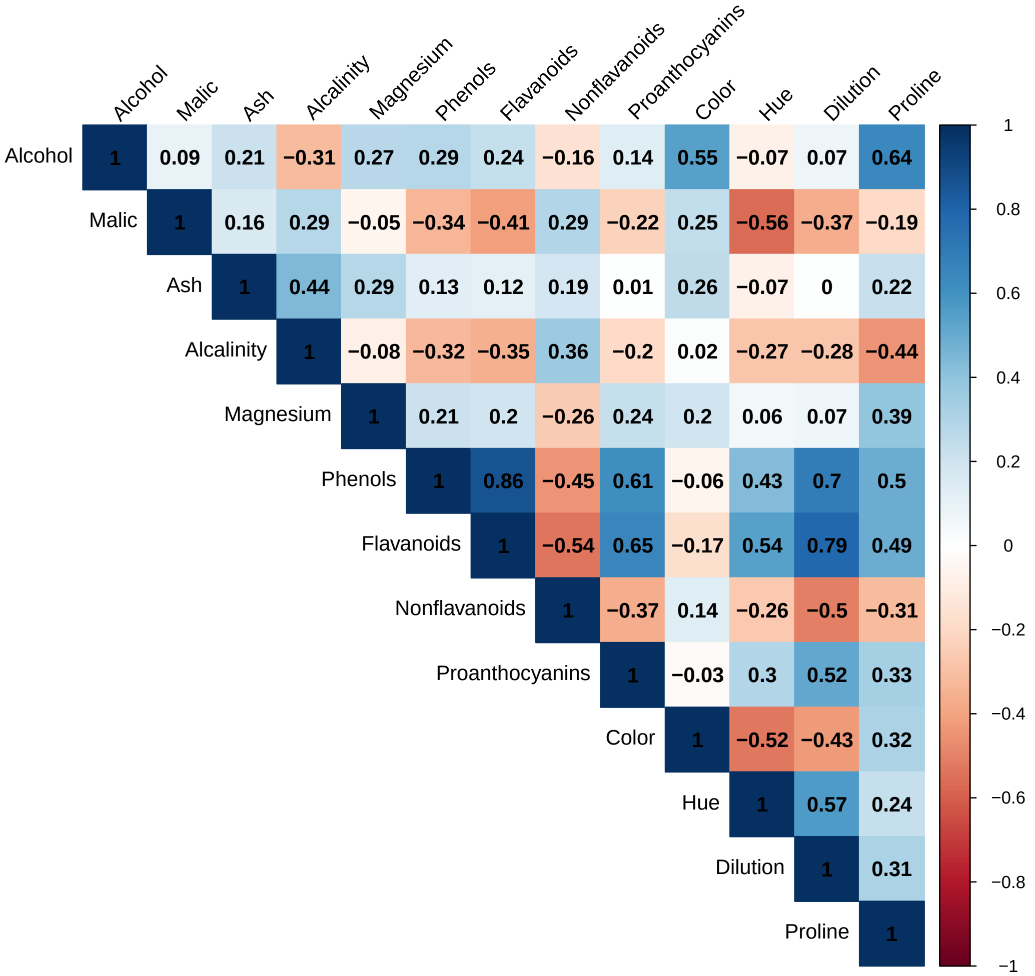
1 Wine Dataset

```r
library(openxlsx)

# load wine dataset
file_path <- "wine.xlsx"
data <- read.xlsx(file_path)
```

2 Principal Component Analysis

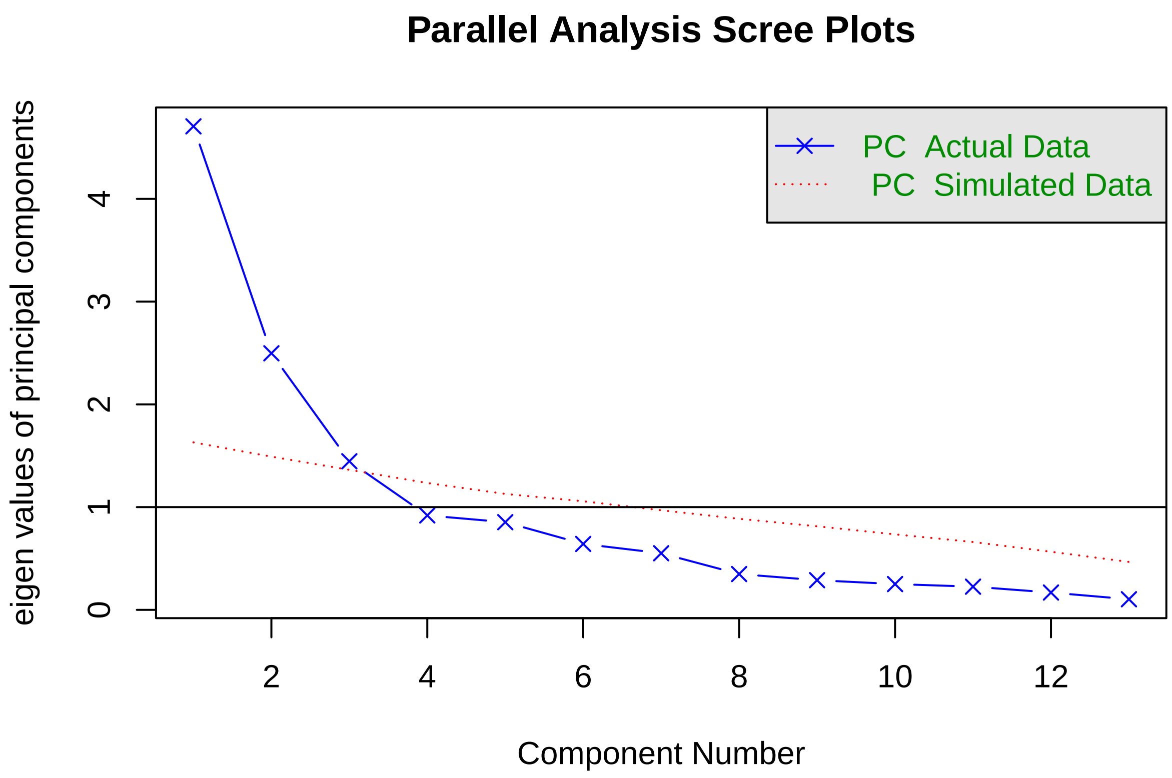
1. Correlation Heat Map

```r
library(corrplot)
library(ggplot2)

# calculate the correlation matrix
X_wine <- data[-1]
y_wine <- data$Type
corr_matrix <- cor(X_wine)

# plotting the correlation heat map
corrplot(corr_matrix, method = "color", type= "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```



2. Parallel Analysis

```
library(psych)

# Perform parallel analysis for PCA (principal components analysis)
set.seed(123) # For reproducibility
fa.parallel(corr_matrix, fa="pc")
```



```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  2
```

3. Proportion of Variance

```
# Standardize the features
X_wine_scaled <- scale(X_wine)

# Perform PCA
pca_result <- prcomp(X_wine_scaled, scale = TRUE)

# Proportion of variance explained by each principal component
prop_var <- (pca_result$sdev^2) / sum(pca_result$sdev^2)

# Proportion of variance explained by the first 3 components
percent <- sum(prop_var[1:3] * 100)
print(prop_var[1:3] * 100)
```

```
## [1] 36.19885 19.20749 11.12363
```

```
cat("The first 3 principal components cumulatively explains ",
    percent, "% of the \ninformation in the dataset", sep = "")
```

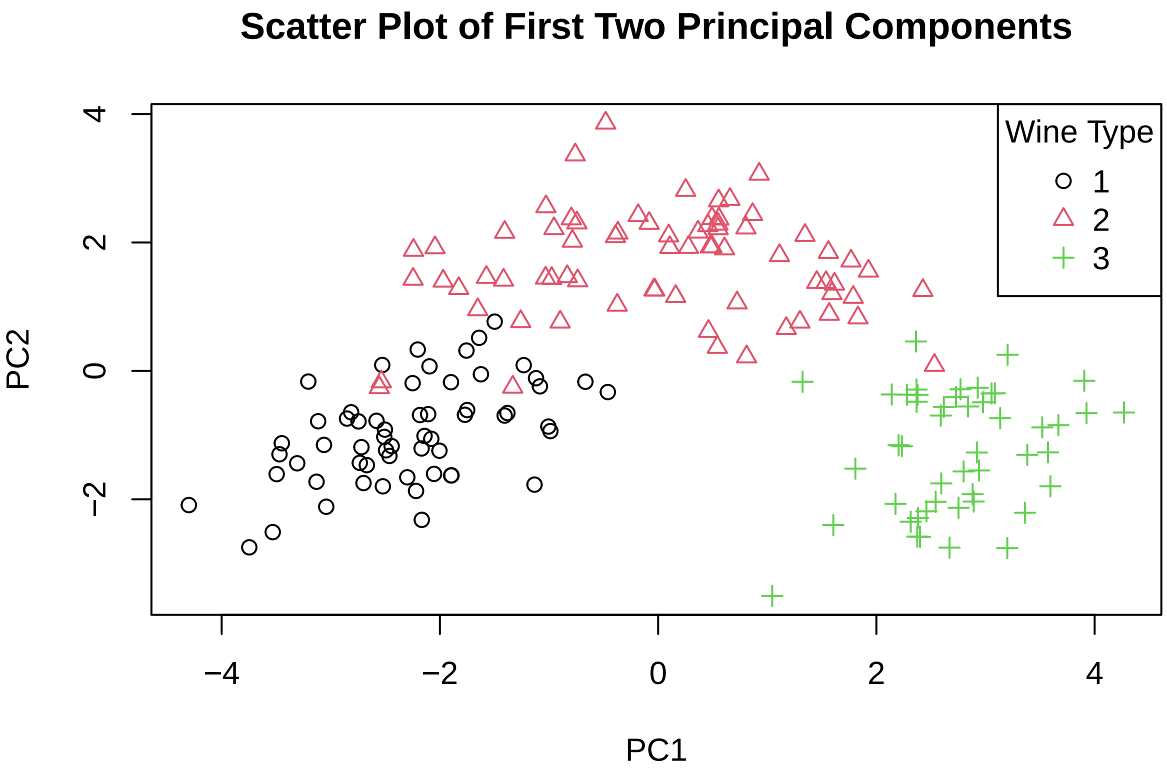
```
## The first 3 principal components cumulatively explains 66.52997% of the
## information in the dataset
```

- **Parallel Analysis**: Recommends retaining 3 principal components.
- **Proportion of Variance Explained Criterion**: The first 3 principal components cumulatively explain 66.52997% of the information in the dataset, which is sufficient.

Comparison: Both methods suggest retaining 3 principal components.

4. Scatter Plots for the first two Principal Components

```r
# Extract scores for the first two principal components
scores <- predict(pca_result)[, 1:2]

# Plot first two principal components
plot(scores[, 1], scores[, 2], pch = as.numeric(y_wine), col = as.numeric(y_wine),
     xlab = "PC1", ylab = "PC2",
     main = "Scatter Plot of First Two Principal Components")

# Add legend
legend("topright", legend = levels(factor(y_wine)),
       col = 1:3, pch = 1:3, title = "Wine Type")
```



Interpretation: There is a significant overlap between wine types 1 and 2, and minor overlap between wine types 2 and 3. The observed overlaps indicate that the first two principal components may not adequately separate the wine types. This suggests that other factors beyond the first two principal components may contribute to the variability in the dataset and the distinction between wine types.

To improve separation between the wine types, we might explore including additional principal components.

3 Chi-Square Test - Goodness of Fit

1. Hypotheses

- **Null Hypothesis (H0)**: The observed distribution of wine types does not significantly differ from an equal distribution (each type representing 1/3 of the total wines).

- **Alternative Hypothesis (Ha)**: The observed distribution of wine types significantly deviates from an equal distribution.

2. Performing the Chi-Square Goodness of Fit Test

```r
# Observed frequencies of wine types
observed <- table(y_wine)

# Expected frequencies assuming equal distribution
expected <- rep(1/3, 3)

# Perform Chi-Square Goodness of Fit Test
chisq_test <- chisq.test(observed, p = expected)

# Extract chi-square statistic and p-value
chi_square_statistic <- chisq_test$statistic
p_value <- chisq_test$p.value

# Print results
cat("Chi-Square Statistic:", chi_square_statistic, "\np-value:", p_value, "\n")
```

```
## Chi-Square Statistic: 4.460674
## p-value: 0.1074922
```

3. Interpretation

- If the p-value is less than the significance level, we reject the null hypothesis. This suggests there's significant evidence that the wine type distribution deviates from an equal one.
- If the p-value is greater than significance level, we fail to reject the null hypothesis. This indicates a lack of strong enough evidence to suggest significant deviation from equal distribution.

```r
# Set significance level
alpha <- 0.05

# Interpret the results
if (p_value < alpha) {
  cat("Since the p-value ", p_value, " is less than the\nsignificance level ",
      alpha, ", we reject the null hypothesis", sep = "")
} else {
  cat("Since the p-value ", p_value,
      " is greater than or equal to the\nsignificance level ",
      alpha, ", we fail to reject the null hypothesis", sep = "")
}
```

```
## Since the p-value 0.1074922 is greater than or equal to the
## significance level 0.05, we fail to reject the null hypothesis
```

4 Chi-Square Test - Independence

1. Categorise Alcohol Content

```
# Categorize 'Alcohol' content into three levels: 'Low', 'Medium', and 'High'
alcohol_categories <- cut(X_wine$Alcohol,
                          breaks = quantile(X_wine$Alcohol, probs = c(0, 1/3, 2/3, 1)),
                          labels = c("Low", "Medium", "High"))
```

2. Contingency Table

```
contingency_table <- table(y_wine, alcohol_categories)
contingency_table
```

```
##       alcohol_categories
## y_wine Low Medium High
##      1   0     17   42
##      2  53     14    3
##      3   6     28   14
```

3. Hypotheses

- **Null hypothesis (H0)**: The choice of wine type is independent of the alcohol content
- **Alternative hypothesis (Ha)**: The choice of wine type is dependent on the alcohol content

4. Performing Chi-Squared Test

```
chisq_result <- chisq.test(contingency_table)

# Extract values
chisq_statistic <- chisq_result$statistic
p_value <- chisq_result$p.value
df <- chisq_result$parameter

cat("Chi-Square Statistic:", chisq_statistic, "\n",
    "p-value:", p_value, "\n",
    "Degrees of Freedom:", df, "\n")
```

```
## Chi-Square Statistic: 120.0613
##  p-value: 5.182946e-25
##  Degrees of Freedom: 4
```

5. Interpretation

- If the p-value is less than the significance level, we reject the null hypothesis. This suggests there is significant evidence of an association between wine type preference and alcohol content level (Low, Medium, High).
- If the p-value is greater than the significance level, we fail to reject the null hypothesis. This means we cannot conclude there is a significant association between wine type preference and alcohol content level (Low, Medium, High) based on our data.

```r
# Set significance level
alpha <- 0.05

# Interpret the results
if (p_value < alpha) {
  cat("Since the p-value ", p_value, " is less than the\nsignificance level ",
      alpha, ", we reject the null hypothesis", sep = "")
} else {
  cat("Since the p-value ", p_value,
      " is greater than or equal to the\nsignificance level ",
      alpha, ", we fail to reject the null hypothesis", sep = "")
}
```

```
## Since the p-value 5.182946e-25 is less than the
## significance level 0.05, we reject the null hypothesis
```