Unsupervised learning: PCA & CA

Contents

Introduction **1**

Principal components analysis **1**

Correspondence analysis **2**

Final assignment: High-dimensional PCA using SVD **4**

Introduction

In this practical, we will learn how to use principal components analysis and correspondence analysis.

We will use the package `ca`. For this, you will probably need to `install.packages("ca")` before running the `library()` functions.

```r
library(ISLR)
library(tidyverse)
library(ca)
```

Principal components analysis

---

1. **Load the questionnaire dataset and explore it.**

---

2. **Create a data frame with only the questionnaire columns, and standardise the dataset**

---

3. **Use the `prcomp()` function to create a principal components analysis for the scaled dataset. Save the result as `pca_mod`.**

4. **Are the first two principal components successful in explaining variance in the dataset? How many components do we need to explain 50% of the variation in the dataset?**

5. **Which original variable is most related to the first principal component? Which is the least relevant for the first principal component?**

6. **Create a scatter plot of the first two principal components. Map the sex of the respondents to the `colour` aesthetic. Is there a sex difference?**

Correspondence analysis

We've preprocessed a dataset from kaggle on song lyrics for the purpose of this practical. You can find the original dataset here or in the `data/` directory. If you want to know which preprocessing steps have been used and how it has been saved, you can take a look at the file `data/song_data_preproc.R`.

The `songs_ca` dataset is stored as a `.RData` file, a native file format from `R` which efficiently stores any `R` object. The `load()` function immediately loads the dataset `songs_ca` into your environment.

7. **Load the preprocessed `songs_ca` dataset into the environment from the `data/songs_ca.RData` file.**

8. **Use the `ca()` function from the `ca` package to create a correspondence analysis object.**

9. **Use the `summary()` function on this object. What can you conclude about the first two inertias? What can you say about the word “love” in this dataset?**

10. **Recreate using `ggplot` the biplot that results from the `plot()` method on this object. Hint: for this, you can use the `rowcoord` and `colcoord` elements of the object.**

11. **What can you conclude about Exo and Janis Joplin? Can you come up with reasonable explanations for this?**

12. **In which ways would the plot be different if we would use different artists?**

Final assignment: High-dimensional PCA using SVD

This is an advanced assignment. You will have to figure out how to create PC scores from the output of a singular value decomposition.

Principal components analysis can also be used to generate a low-dimensional number of features from a high-dimensional (p > n) dataset. One area where high-dimensional data frequently occurs is in chemometrics, assessing the properties of materials using spectroscopy (Wikipedia link).

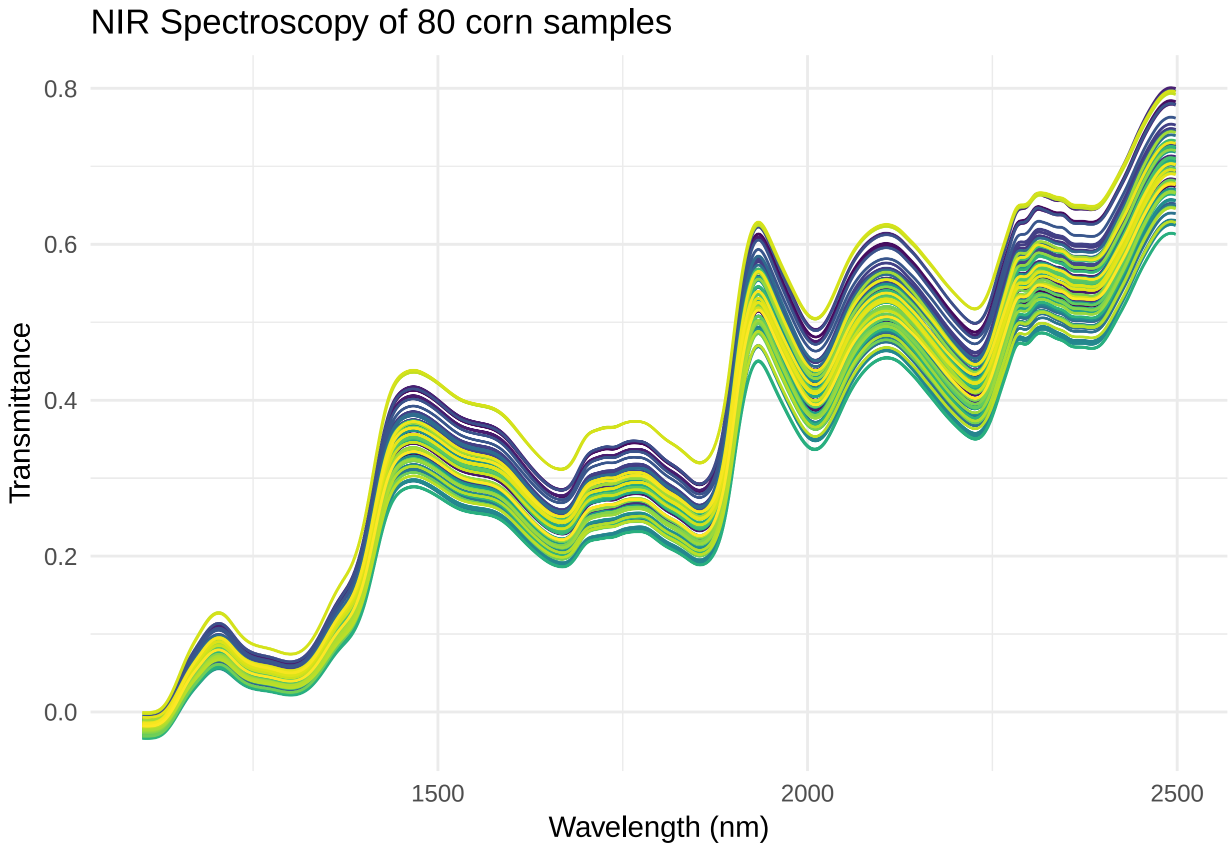
---

13. **Load the dataset "data/corn.RData" using the function `load()`.**

The first four columns contain properties of the corn samples (80 corn samples were analysed) and the remaining 700 columns indicate the measured transmittance at different near-infrared wavelengths. You can find more information about this dataset at the source website.

Here is a plot of wavelength versus transmittance, with one line for each of the 80 corn samples:

```r
t(corn[, -c(1:4)]) %>%
  as_tibble %>%
  gather(key = corn, value = signal) %>%
  mutate(wavelength = rep(seq(1100, 2498, 2), 80)) %>%
  ggplot(aes(x = wavelength, y = signal, colour = corn)) +
  geom_line() +
  theme_minimal() +
  scale_colour_viridis_d(guide = "none") +
  labs(x = "Wavelength (nm)",
       y = "Transmittance",
       title = "NIR Spectroscopy of 80 corn samples")
```



---

14. **Use the `svd()` function to run a principal components analysis on the spectroscopy part of the corn dataset. Save the PC scores and plot the first two principal components. Create four plots, each mapping one of the four properties to the `colour` aesthetic. Which of the properties relate most to the first two principal components? Base your answer on the plots only. Then, do the same thing for PCs 5 and 6.**
