Hauptkomponentenanalyse

Schmidt, Sophie C.

13 August, 2020

Die Hauptkomponentenanalyse ist in R base implementiert, es gibt unterschiedliche Möglichkeiten. Wir nutzen hier die Funktion prcomp.

Nehmen wir als Beispiel unsere Piraten und schauen, wie die metrischen Variablen Größe, Gewicht, Bartlänge und sword time mithilfe einer PCA auf zwei Faktoren reduziert werden können.

library(yarrr)  
data("pirates")

Da der Datensatz sehr groß ist (eigentlich ja gut), für die Visualisierung das aber schwierig wird, ziehen wir eine Stichprobe aus unserem Datensatz:

pir <- pirates[sample(nrow(pirates), 50), ] # aus der Anzahl der Zeilen im Datensatz (nrow) wird eine Zufallsstichprobe (sample) von 50 gezogen und dieses sample genutzt, um Zeilen auszuwählen -- in eckigen Klammern hinter dem eigentlichen Datensatz.

Hier haben wir schon das Problem, das angesprochen wurde: Die Bartlänge reicht nur bis 40, wohingegen Gewicht erst bei 33 beginnt und bis 105.6 geht. Also müssen wir bei der Berechnung der PCA die Werte normalisieren. Das machen wir, in dem wir die Argumente center und scale. auf TRUE setzen:

pirates.pca <- prcomp(pir[,c(4:5,14:15)], # Auswahl der metrischen stetigen Spalten  
 center = TRUE, # zentrieren  
 scale. = TRUE) # normalisieren

Jetzt haben wir die PCA einer Variablen zugewiesen und können uns Details mit dem üblichen summary anzeigen lassen:

summary(pirates.pca)

Wir haben “nur” 4 PCs, d.h. principal compontents (Hauptkomponenten), weil wir 4 Variablen in die Analyse einbezogen haben. Wir sehen jetzt die Standardabweichung, Anteil der erklärten Varianz und “cumulative proportion”, dh. den Wert, wenn wir die Varianz dieser und aller “vorhergehenden Hauptkomponenten” zusammenrechnen.

Wir sehen also, dass die erste Hauptkomponente 51% der Varianz erklärt, die zweite erklärt 27,5% und zusammen erklären sie 79% der Varianz. Ihr könnt ja einmal nachrechnen: 51,49% + 27,54% = 79,03 %. Stimmt. ;-)

Wenn ihr auf den kleinen Pfeil neben pirates.pca in der Environment klickt oder str(pirates.pca) ausführt, seht ihr, was noch alles in der Variablen “drinsteckt”

Das Zentrum (center), die (normalisierten) Mittelwerte (scale), Standardabweichung (sdev) jeder Hauptkomponente, die Faktorenladungen der ursprünglichen Ladungen im Bezug auf die Hauptkomponenten (rotation) sowie die Werte jedes Punktes beschrieben durch die neuen Achsen (x).

Wenn ihr euch head(pirates$x) einmal anschaut seht ihr die ersten 6 “neuen” Variablen. Es werden die gleichen Punkte beschrieben wie vorher durch height, weight, beard.length und sword.time, aber diesmal mit neuen Werten entlang neuer “hypothetischer” Achsen.

Das würden wir jetzt eigentlich gern visualisieren. Das “schönste” Paket dafür liegt unter Github und heißt ggbiplot.

library(devtools)  
install\_github("vqv/ggbiplot")

Jetzt plotten wir damit erst einmal ganz einfach unsere PCA:

library(ggbiplot)  
  
ggbiplot(pirates.pca)

Wie praktisch! Wir sehen: Die Achsen wurden für uns schon beschriftet, sogar mit der Beschreibung, wie viel Prozent der Varianz erklärt wird. Die roten Pfeile mit der roten Beschriftung zeigen in welche Richtungen die unterschiedlichen ursprünglichen Variablen die Daten “ziehen”. Wie man sieht, liegen height und weight fast direkt übereinander: Diese beiden korrelieren ja auch stark, wie wir schon wissen.

Jetzt hübschen wir das alles noch ein bisschen auf.