Arbeitsblatt 1

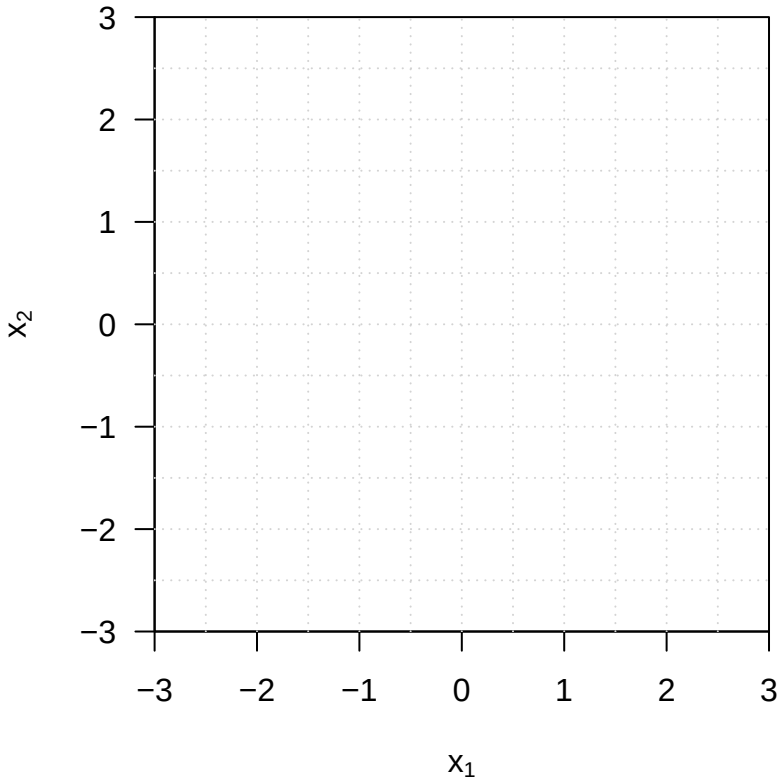
Aufgabe 1: PCA von Hand

Wir wollen eine PCA mit Lineal und Papier durchführen. Wir haben die Koordinaten von 5 Beobachtungen:

Tabelle 1: Datensatz

	X_1	X_2
1	-1.5	-0.5
2	-2.0	-1.5
3	-1.0	0.0
4	2.5	0.5
5	2.0	1.5

a) Zeichnen Sie die 5 Beobachtungen in das vorgedruckte Koordinatensystem (benutzen Sie als Punktbezeichnung die Nummer der Beobachtung). Wenn Ihnen das manuelle Zeichnen zu mühsam ist, können Sie alternativ die Aufgabe auch direkt mit dem shinyApp_PCA_Rotation.R bearbeiten.



b) Zeichnen Sie in das Diagramm ein gedrehtes Koordinatensystem ein, so dass die erste Achse (PC1) die grösste Varianz hat. Die zweite Achse (PC2) steht senkrecht dazu.

c) Welche Koordinaten (Scores) haben die Punkte im neuen Koordinatensystem (grobe Abschätzung genügt).

d) Vergleichen Sie Ihre Ergebnisse mit dem R output.

```
        PC1        PC2
1 -1.5673743  0.2081770
2 -2.4551500 -0.4714216
3 -0.8987892  0.4383811
4  2.4661635 -0.6465581
5  2.4551500  0.4714216
```

Aufgabe 2: Körpergrössen

Laden Sie den R-Datensatz *body*. Er enthält die Körpergrösse, das Gewicht und die Schuhgrösse von 600 Personen (simulierte Daten).

a) Visualisieren Sie die Daten. Was sieht man?

b) Berechnen Sie die Kovarianz-Matrix der Daten (R-Befehl `cov()`). Berechnen Sie die Summe der Varianzen der 3 Variablen (d.h. Diagonale der Kovarianz-Matrix).

c) Führen Sie mit der Funktion `prcomp` die Hauptkomponentenanalyse durch. Bei der Hauptkomponentenanalyse arbeiten wir mit zentrierten Daten. Sie können das manuell machen oder bei der Funktion `prcomp` das Argument `center` auf `TRUE` setzen (Defaulteinstellung).

d) Visualisieren Sie die Hauptkomponenten. Auf die Hauptkomponenten können Sie mit dem PCA-Objekt `pca$x` zugreifen.

e) Sie erfahren, dass die letzten 100 Beobachtungen von Frauen stammen, während der Rest der Daten zu jungen Männern gehört. Visualisieren Sie diese Beobachtungen in den Hauptkomponenten mit einer anderen Farbe.

f) Berechnen Sie die Kovarianzmatrix der Hauptkomponenten. Was sehen Sie? Berechnen Sie zusätzlich die Summe der Varianzen und vergleiche Sie das Ergebnis mit b). Wie gross ist der Anteil der ersten beiden Komponenten an der Gesamtvarianz (d.h. wie gross ist der Anteil an der Streuung, der durch diese beiden Komponenten beschrieben wird).

g) Die ursprünglichen Variablen haben unterschiedliche Einheiten mit unterschiedlichen Grössenordnungen. In solchen Fällen werden die Variablen für die Hauptkomponentenanalyse in der Regel zuerst mit der Standardabweichung **standardisiert**. Standardisieren Sie die ursprünglichen Daten so, dass die transformierten Daten den Mittelwert 0 und eine Standardabweichung von 1 haben (Z-Transformation $Z = \frac{X-\mu}{\sigma}$). D.h. Mittelwert abziehen und durch die Standardabweichung der Daten teilen. Führen Sie die Schritte der Teilaufgabe a-e) danach noch einmal durch. Was ändert sich? (Zum Standardisieren können Sie die Funktion `scale` verwenden oder noch einfacher, bei der Funktion `prcomp` einfach das Argument `scale` auf `TRUE` setzen.)

h) Jede PC ist eine linear Kombination der Original-Variablen und den Koeffizienten der Rotationsmatrix (auch Ladungen genannt). Betrachten Sie die Koeffizienten für den standardisierten und den nicht standardisierten Fall (`...$rotation`). Inwiefern unterscheiden sich diese? Würden Sie hier mit den standardisierten oder den nicht standardisierten Daten arbeiten?

Aufgabe 3: Abstimmungsverhalten der Kantone

Laden Sie den Datensatz *abst.Rdata*. Er enthält die Resultate (Anteil Ja-Stimmen in Prozent) der letzten 64 eidgenössischen Abstimmungen aufgeteilt nach Kantonen. Es handelt sich um aufbereitete Daten von der Webseite: https://www.bfs.admin.ch/bfs/de/home/statistiken/politik/abstimmungen.assetdetail.3362357.html. Machen Sie sich mit dem Datensatz vertraut. Genauere Informationen zu den einzelnen Abstimmungen finden Sie im Netz.

a) Gewinnen Sie zuerst einen Überblick über die Daten. Was fällt Ihnen auf? Kommentieren Sie!

b) Führen Sie eine Hauptkomponenten-Analyse durch (auf welchen Daten?) und stellen Sie die Daten in den ersten beiden Hauptkomponenten dar. Beschreiben Sie die Struktur der Daten in dieser Darstellung. Benutzen Sie als die Kantonsnamen als Label (z.B. `text(..., labels=row.names(...))`).

c) Genügen die beiden ersten Hauptkomponenten, um die Variabilität der Daten sinnvoll zu approximieren?

d) Benutzen Sie für die Beurteilung der Approximation das Scree-Diagramm. Kommen Sie zum gleichen Schluss wie in c)?

Aufgabe 4: Zehnkampf

Die folgende Tabelle zeigt die Resultate vom Zehnkampf der Olympischen Spiele in Rio de Janeiro 2016. Die Daten sind im File **zehnkampf.csv** abgespeichert und enthalten folgende Variablen:

`m100`	100m Lauf (s)	`m400`	400m Lauf (sek.)	`speer`	Speerwurf (m)
`weit`	Weitsprung (m)	`hurd`	110m Hürdenlauf (s)	`m1500`	1500m Lauf (s)
`kugel`	Kugelstoss (m)	`disc`	Diskuswurf (m)	`punkte`	Totale Punktzahl
`hoch`	Hochsprung (cm)	`stab`	Stabhochsprung (cm)		

a) Beurteilen Sie anhand von Boxplots, ob sich die Resultate einer PCA, basierend auf unskalierten und auf skalierten Variablen, unterscheiden werden. Begründen Sie die Antwort.

b) Führen Sie eine PCA durch. Lassen Sie die Variable Punkte dabei weg, da es sich dabei nicht um eine Diziplin handelt. Wie viele Hauptkomponenten sind nötig, um 80% der Varianz zu erklären?

c) Stellen Sie die Ergebnisse der PCA in einem Biplot `biplot()` dar. Können Sie die Principal Components (PCs) im Hinblick auf die ursprünglichen Variablen interpretieren? Welche Disziplinen korrelieren miteinander? Welche Disziplinen sind im 2D Score-Plot nicht optimal wiedergegeben. Gibt es Gruppen von Athleten?

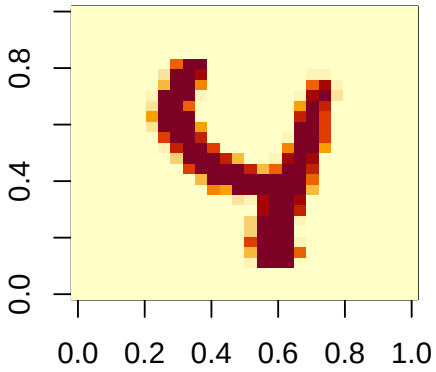
Aufgabe 5: Ausreisserdetektion mit PCA

Laden Sie den Datensatz `mnist20x4and1000x0.Rdata`. Er erhält zwei Objekte. `x` ist eine Matrix mit Pixel-Grauwert-Daten von handgeschriebenen Zahlen, wobei jede Zeile zu einem "linearisierten" Bild gehört. Im Vektor `xlabel` sind die entsprechenden Labels zu finden. Der Datensatz enthält 1000 Nullen die mit 20 Vierern kontaminiert sind. Das Ziel dieser Aufgabe ist es, die Ausreisser mittels PCA zu identifizieren.

a) Machen Sie sich mit dem Datensatz vertraut. Einzelne Bilder können Sie wie folgt anschauen. Visualisieren Sie einige Elemente.

```
load("Daten/mnist20x4and1000x0.RData")
image_nummer <- 7 # hier können Sie ein Bild von 1 bis 1020 auswählen
```

```
bild <- matrix( x[image_nummer, ], ncol=28, byrow = TRUE)
image(t(bild[28:1,1:28]))
```



b) Führen Sie mit der Funktion `prcomp` eine Hauptkomponentenanalyse auf die Matrix `x` durch. Sollten die Variablen standardisiert werden oder nicht? Warum? Was genau macht der `prcomp` Befehl?

c) Wieviel Hauptkomponenten braucht es, um 80% der Varianz in den Daten zu erklären? Greifen Sie mit `$sdev` auf die Standardabweichung der Hauptkomponenten zu.

d) Erstellen Sie einen Scatterplot der ersten beiden Hauptkomponenten. Wählen Sie andere Farben für 4-er und 0-er Beobachtungen. Können Sie einige Ausreisser identifizieren?

e) Laden Sie jetzt die Pakete `MASS` und `rrcov` und führen Sie mit folgendem Befehl eine robuste Hauptkomponentenanalyse durch: `pca.rob = PcaHubert(x, k = 2)`. Wofür steht das Argument k? Schauen Sie sich die Struktur des pca.rob Objekts an. Das Objekt ist von der Klasse S4, d.h. Sie können verschiedene Attributen mit `@` ansprechen, ähnlich zu `$` in einem data.frame Objekte. Plotten Sie auch hier die ersten beiden Hauptkomponenten (Tipp: `plot(pca.rob@scores, ...)`).

f) Nun plotten Sie die orthogonale Distanzen `pca.rob@od` der Beobachtung zur 2D-Ebene gegen die Beobachtungsnummer. Mit `abline(h = pca.rob@cutoff.od)` können Sie in den Plot noch eine vordefinierte Cutoff-Linie einzeichnen. Wählen Sie wieder andere Farben für 4-er und 0-er Beobachtungen. Welche Beobachtungen sind Ausreiser?

Aufgabe 6: Leistungsnachweis

Versuchen Sie einen eigenen Datensatz für den Leistungsnachweis zu finden.

Anforderungen an Datensatz

- Mindestens 4 quantitative Variablen
- Mindestens eine qualitative Variable, welche als Zielvariable für ein Klassifikationsproblem verwendet werden kann (nicht zu viele Stufen)

- Mindestens 100 Beobachtungen

Erste Schritte

- Datensatz einlesen
- Kurze Beschreibung des Datensatz
- allenfalls Datenaufbereitung
- Kurze summary-Statistik