Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

HS 2016

Statistisches Data Mining (StDM)

Woche 2

Aufgabe 1 Metric MDS vs PCA

Der Datensatz `stiffness.da` enthält für $n = 30$ Holzbretter je vier verschiedenartige Messungen S1, S2, S3 und S4, die alle irgendwie die Festigkeit dieser Bretter messen. Die erste Messart besteht darin, dass eine Schockwelle durch das Brett gesendet wird. Die zweite wird während das Brett vibriert bestimmt, und die letzten beiden Messungen werden aus einem statischen Test ermittelt.

a) Verschaffen Sie sich einen Überblick über die und stellen Sie die Daten in einer Streudiagramm-Matrix dar. Was stellen Sie fest? Fallen Ihnen noch andere grafische Darstellungsmöglichkeiten für diese vierdimensionalen Daten ein?

b) Soll die PCA auf skalierten Variablen durchgeführt werden.

c) Führen Sie eine PCA auf den geeigneten Daten durch und stellen Sie die Daten in den ersten beiden Hauptkomponenten dar, verwenden Sie dazu die function `eqscplot`. Beschreiben Sie die Struktur der Daten in dieser Darstellung.

d) Stellen Sie eine Tabelle mit den prozentualen Beiträgen zur totalen Varianz jedes Eigenwertes zusammen. Genügen die beiden ersten Hauptkomponenten, um die Variabilität der Daten sinnvoll zu approximieren?

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

f) Wie könnte man die Hauptkomponenten interpretieren?

Aufgabe 2 Metric MDS vs PCA

a) Create a data matrix with dimension 5x2 (5 examples, 2 features) by drawing random numbers from a Gaussian `X <- matrix(rnorm(10), nrow = 5)` and plot the data matrix.

b) Calculate all pairwise Euklidean distances with (`dist`) and do a metric MDS (`cmdscale`)

c) Do a principal component analysis and compare to a) and b).

d) Metric MDS can also used (for slightly) non-Euklidean Distances. The file `airdist.Rmd` contains the distance between some US airports (in miles). These distances can't be Euclidean since we live on a sphere.

e) Using the `Shepard` function in the `MASS` package, plot the distances before and after the MDS against each other. What is the largest difference?