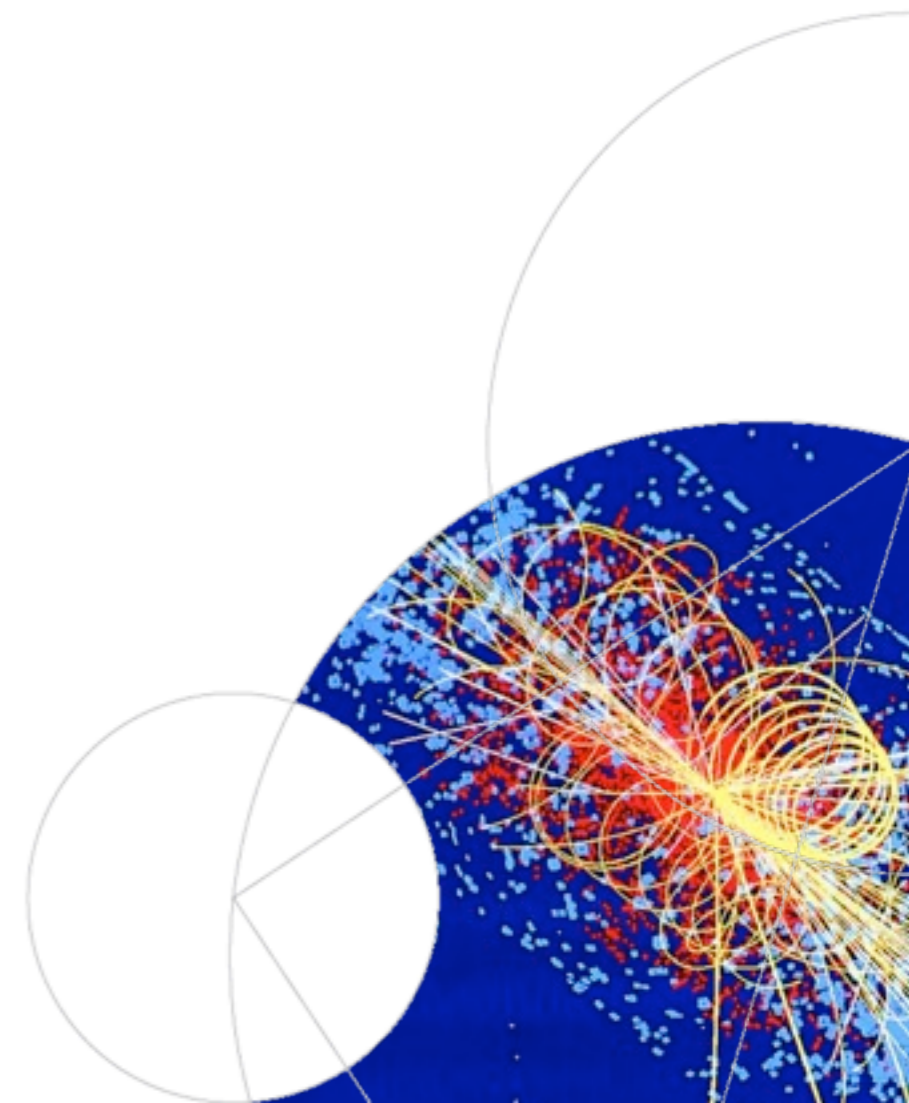




Eksplorativ dataanalyse

Eksperimentel Fysik, blok 4 - 2011

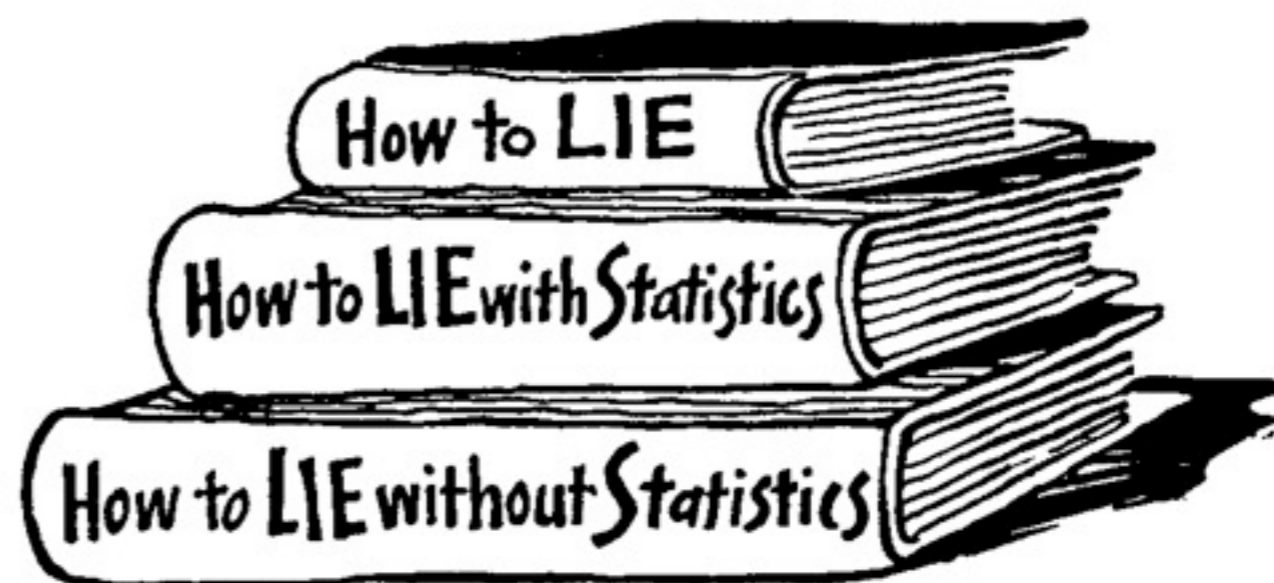
Morten Dam Jørgensen
mdj@mdj.dk





Oversigt

- Projekter
- Motivation
- Korrelation
- Principal Component Analysis
- Opgaver
- MATLAB hjælp





Projekter

Partikelfysik

- "The Large Hadron Collider" har nu kørt et år, og vi er i fuld gang med at analysere data. LHC projektet går ud på at kigge på simuleret data for en mulig ny type partikel kaldet "R-Hadron". Ved at kombinere målinger fra forskellige dele af ATLAS detektoren, er det muligt at komme med et præcist masse estimat for partiklen, hvis den eksisterer...

Kosmisk stråling

- På loftet af Rockefeller, HCØ og her på Blegdamsvej, sidder en serie målestationer, der opfanger stråling skabt i atmosfæren af energi-rige partikler af kosmisk oprindelse. Projektet omhandler dataanalyse af virkelig data fra disse stationer, samt lignende data fra udlandet. Er det muligt at se om der er korrelationer mellem antallet af målte partikler og atmosfærens tryk?

Dopplereffekten og Fourier transformationer

- Ved hjælp af Fourier analyse, en mikrofon, og MATLAB er det muligt at måle hastigheden af en forbipasserende ambulance, men det er lidt svært...

Nearest neighbour clustering
"old faithful geyser"



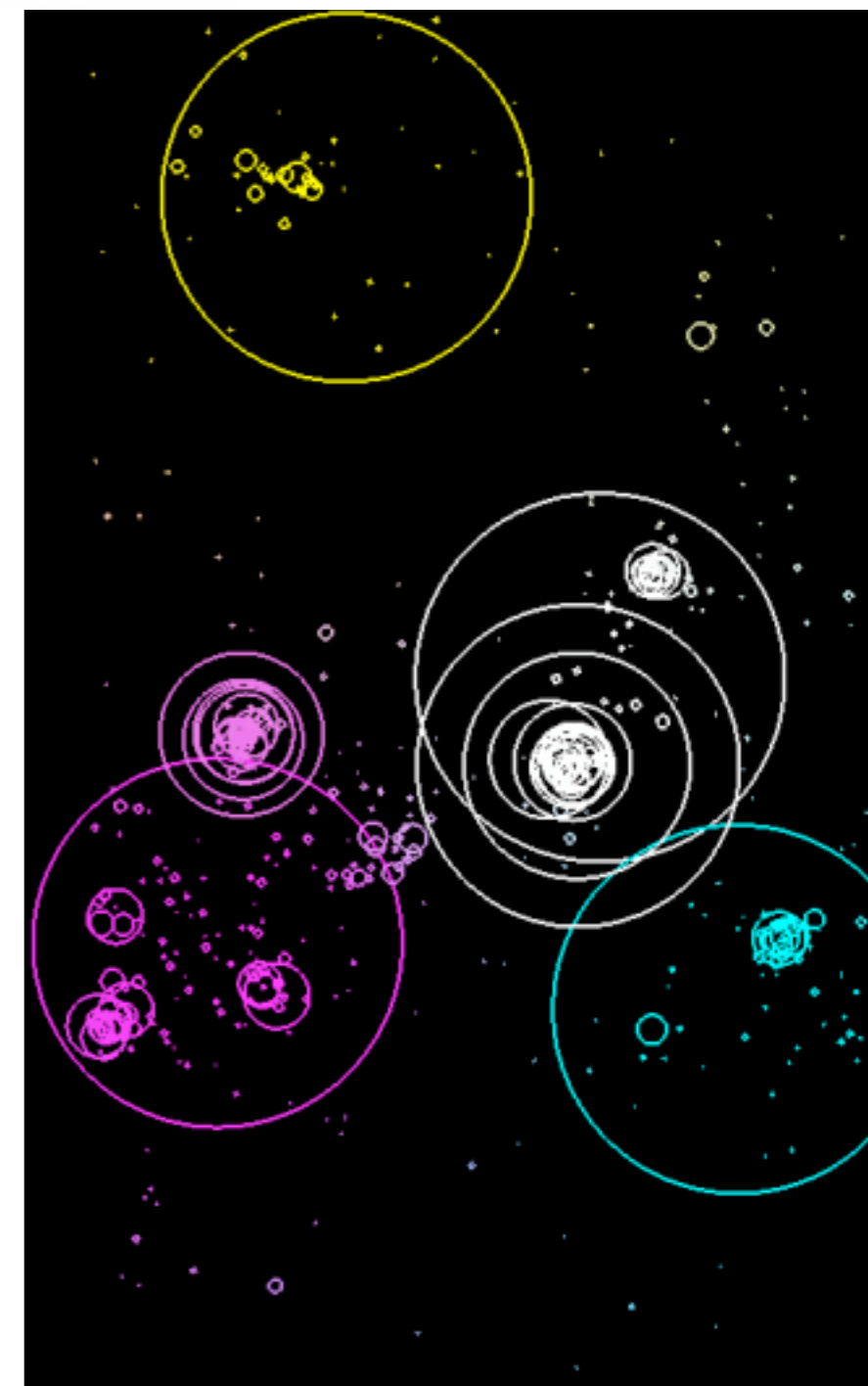
Find grupperinger af data
C-Means clustering algoritme
(Partikel jets i i LHC data, pythia simulation)

Motivation

Idé – Forsøg – Observation – Teori ←

- ikke altid så nemt...
- Typisk: mange forskellige målinger på det samme system, hvad er sammenhængen mellem dem?
- Er nogle målinger redundante?
- Er der simple relationer mellem komplicerede observationer?
- Separation af signal og støj.
- Klassifikation af observationer
- Interpolering (bestem manglende information ud fra målinger)

Statistiske metoder kan bruges til andet end kvantisering af måleresultater, de kan give os et nyt perspektiv, der gør det nemmere at finde nye sammenhænge.



Korrelationer

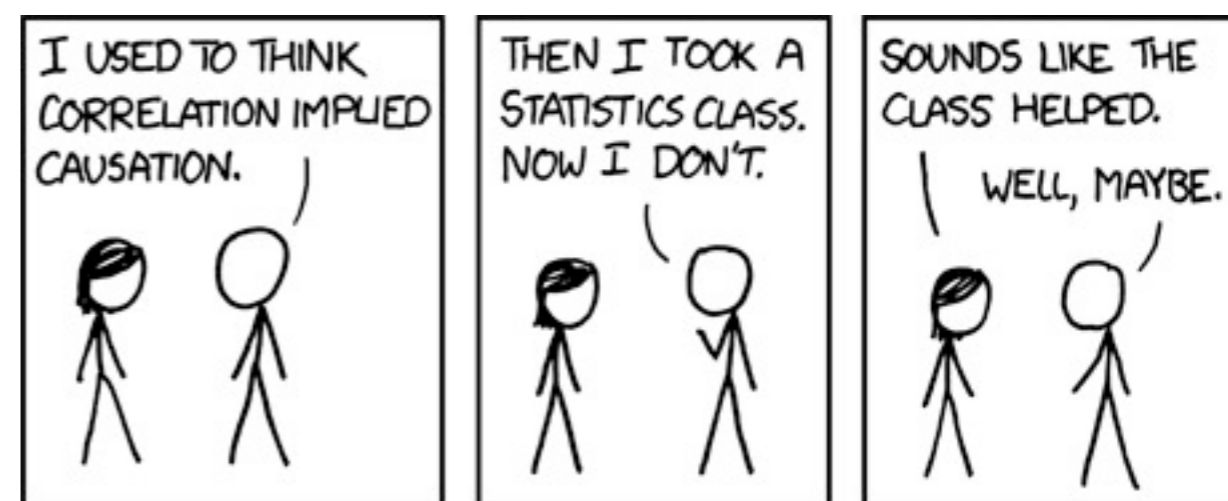
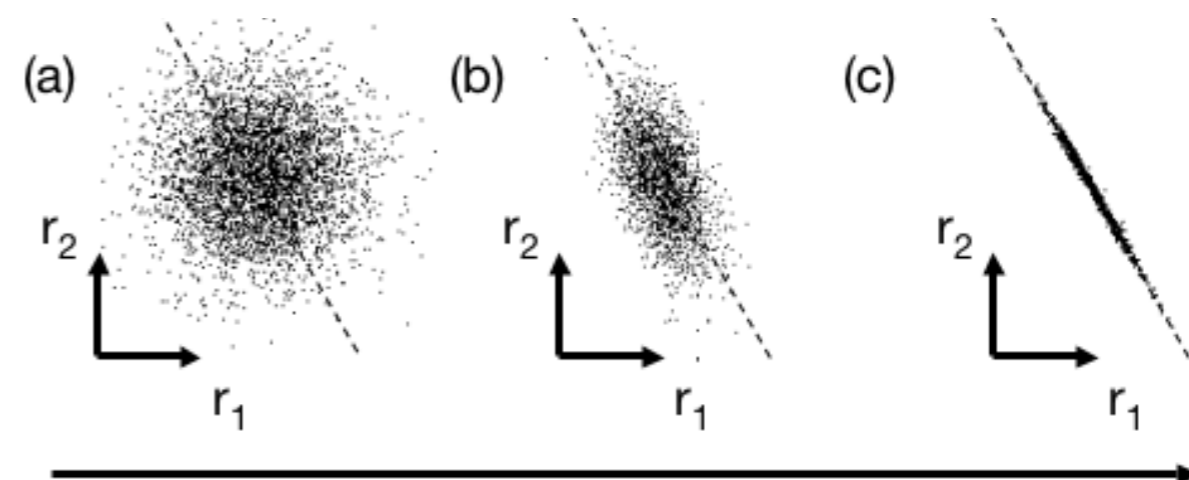
- Korrelation = hvis noget sker for x sker det i nogen grad også for y. graden af korrelation beskriver hvor meget de to variable varierer i forhold til hinanden.

$$\rho = \frac{\text{COV}(x, y)}{\sigma_x \sigma_y}$$

- "Normaliseret" kovarians.
- At noget er korreleret er ikke ensbetydende med at der er et kausalt forhold.
- At noget ikke er (lineært) korreleret betyder ikke nødvendigvis at der ikke er en sammenhæng!

Kovarians:

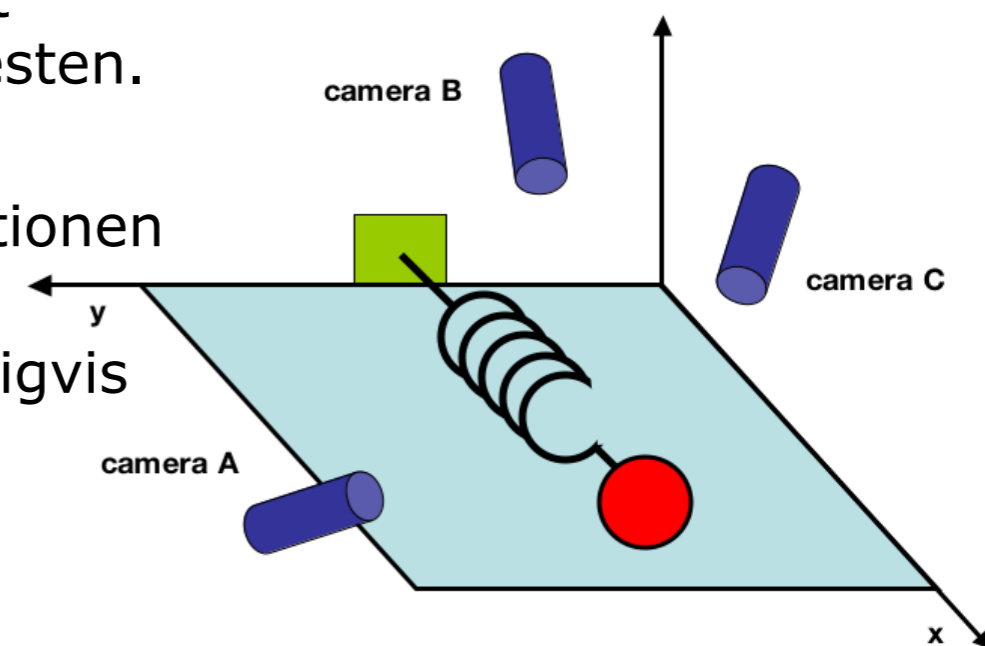
$$\text{COV}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

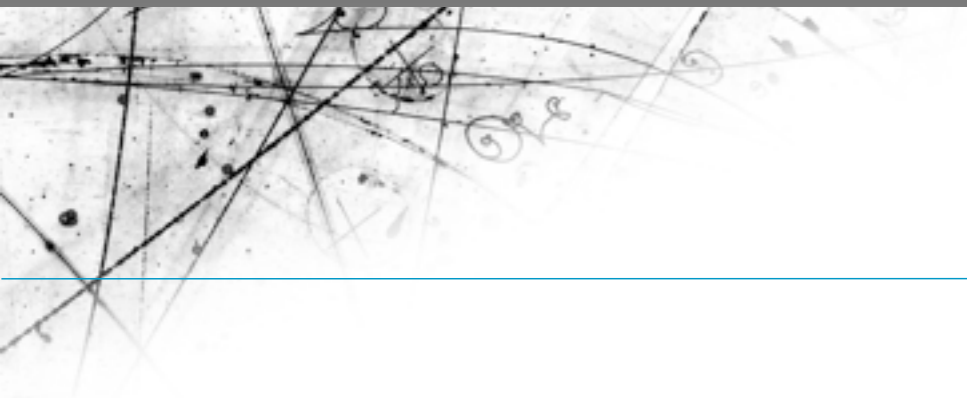




Principal Component Analysis

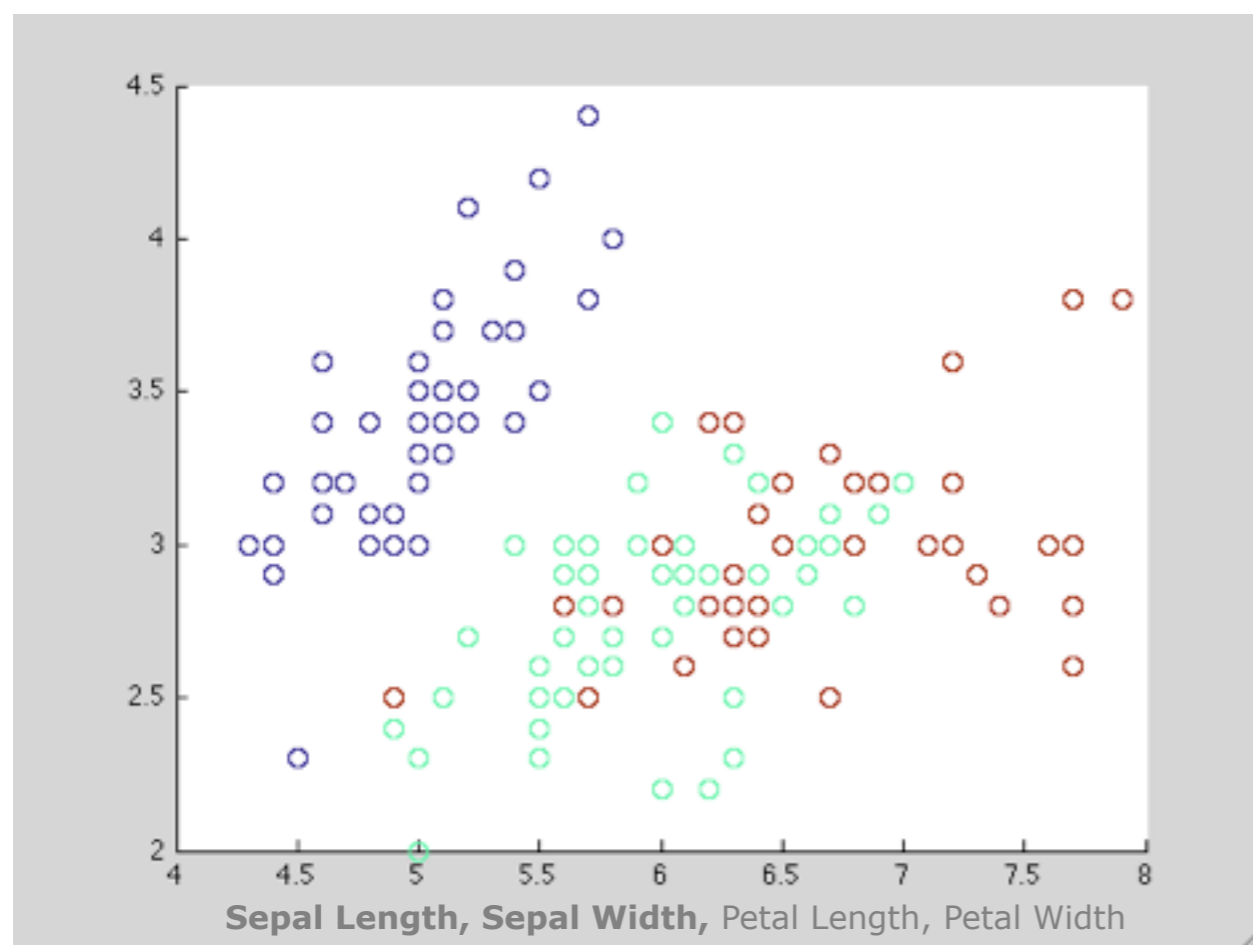
- Simple metode til at bestemme de mest relevante komponenter (kombinationer af variable) i et forsøg.
- Lineær Algebra (i troede det var slut?... hehe)
 - Find en basis for målingerne der minimerer korrelationen mellem komponenterne.
 - Sorter komponenterne sådan at de mest dominerede kommer først, og ignorerer resten.
- Eksempel: en kugle på en fjeder, hvor positionen er målt i $\{x, y\}$ af tre kameraer med forskellige perspektiver (ikke nødvendigvis vinkelret på hinanden eller fjederen!).
 - Mål: find den grundlæggende dynamik (Hooke's lov) i et ikke-ideelt system.



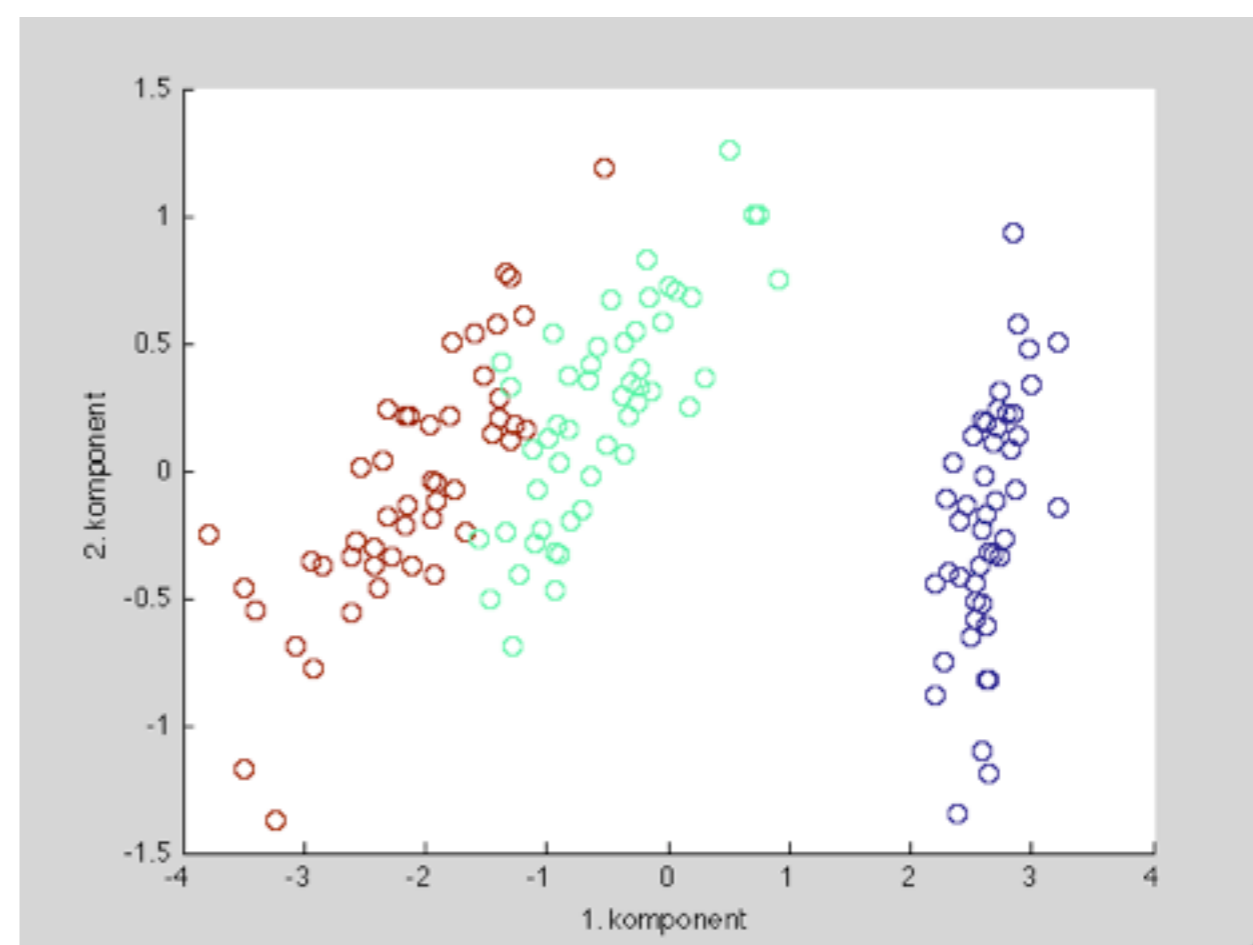


PCA Eksempel "iris" datasættet

Før PCA



Efter PCA

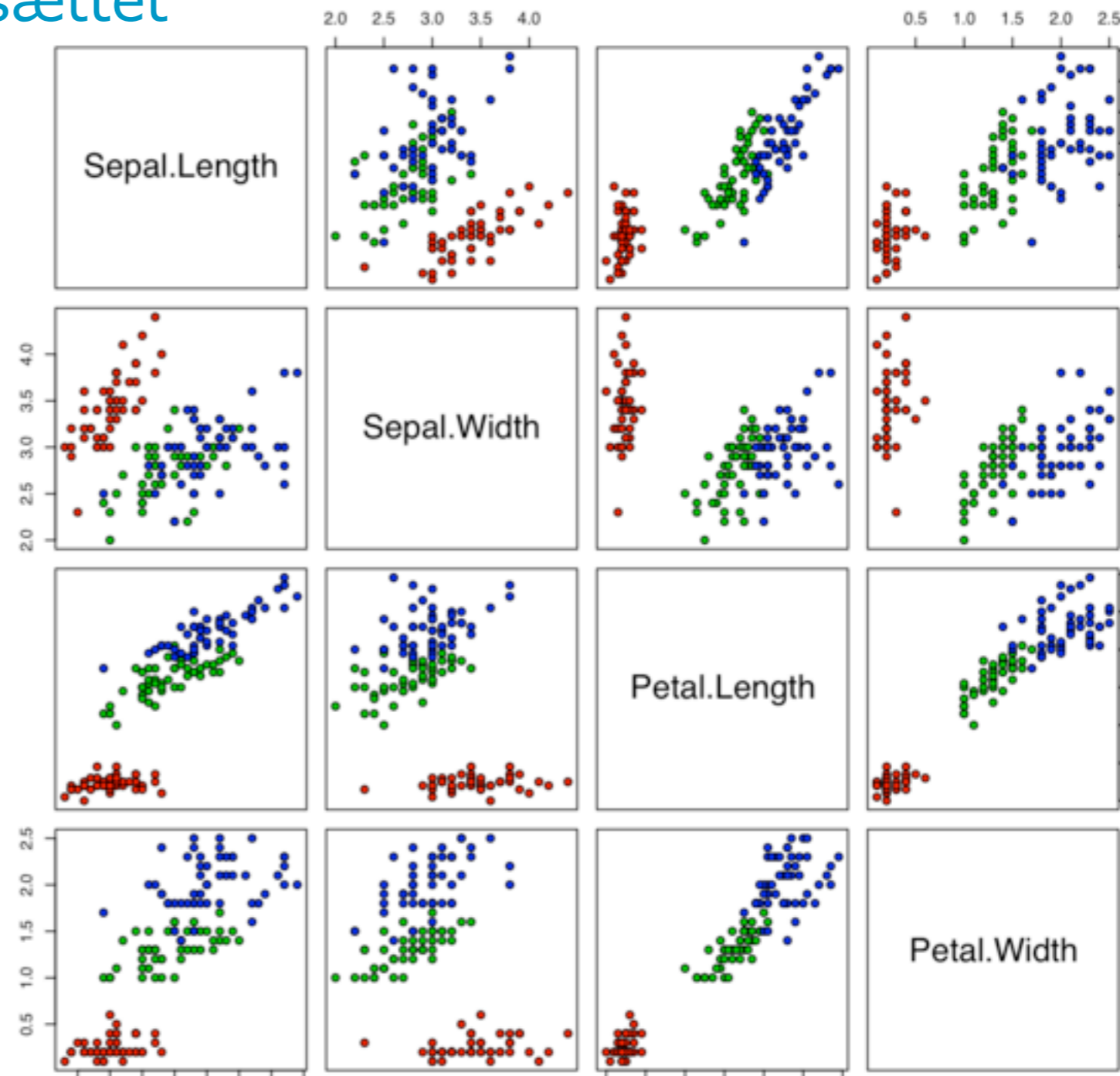
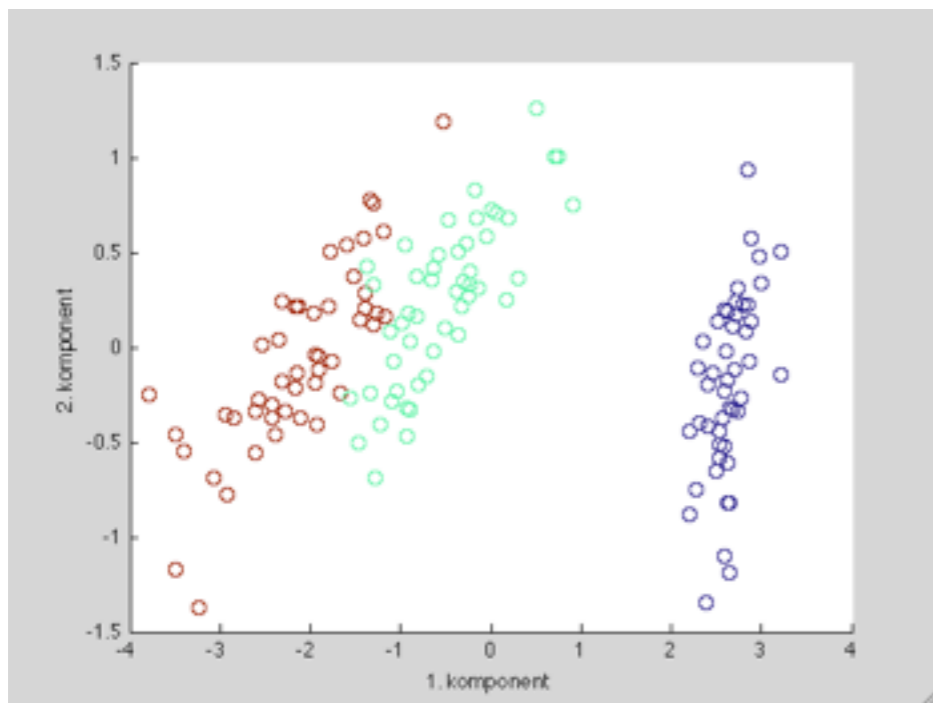


http://en.wikipedia.org/wiki/Iris_flower_data_set



PCA Eksempel "iris" datasættet

Iris Data (red=setosa, green=versicolor, blue=virginica)





Algoritme

- Forbered datasættet med N datapunkter af $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i}, \dots, \mathbf{x}_{ni})$ som en matrix \mathbf{X} med dimensionerne $N \times M$.
- Fratræk middelværdien af hver søjle (observabel), $\mathbf{B} = \mathbf{X}_i - \mu_i$
- Beregn kovarians matricen, $\mathbf{C} = \frac{1}{N} \mathbf{B}^T \mathbf{B}$
- Beregn egenværdierne og egenvektorerne af \mathbf{C} , sådan at $\mathbf{V}^{-1} \mathbf{C} \mathbf{V} = \mathbf{D}$, hvor \mathbf{V} er egenvektorerne af \mathbf{C} og \mathbf{D} er en $M \times M$ matrice med egenværdierne af \mathbf{C} langs diagonalen.
- Sorter søjlerne af \mathbf{D} med de største egenværdier først. Benyt samme sortering på egenvektorerne i \mathbf{V} .
- Fravælg egenværdierne som falder under en grænseværdi η , der sættes sådan at de tilbageværende komponenter beskriver mest muligt (typisk 90%).
- Projekter det oprindelige datasæt med egenvektorerne: $\mathbf{X}' = \mathbf{V}^T \mathbf{B}$.



Øvelser

På Absalon findes tre zip filer med MATLAB opgaver (*.m filerne) og datasæt, hent filerne og åben opgaverne.

Opgave 0: (opvarmning, spring bare over hvis det er for nemt)

- Cavendish' målinger af Jordens massefylde.
 - Middelværdien, standardafvigelsen, outliers.

Opgave 1:

- Forholdet mellem landes areal og deres indbyggertal.
 - Plot forholdet
 - Find kovariansen og korrelationen.

Opgave 2:

- Et ubehageligt datasæt...
 - Find korrelationen mellem to variable...

Opgave 3:

- Principal Component Analyse af iPod data
 - Hvilke effekter dominerer i virkeligheden målingerne af accelerationen i tre retninger?



MATLAB cheat sheet

Data import

```
data = dataimport('data.txt');
```

Scatter plots

```
scatter(data(:,1), data(:,2))
```

Histogrammer

```
hist(data(:,1), nbins)
```

Egenværdier og vektorer

```
[V e] = eig(data)
```

Statistik

```
mean(⋄), std(⋄), var(⋄), cov(⋄), corr  
(⋄), pcacov(⋄)
```

Fitting

Curve fitting toolbox eller `fit(⋄)`

Programmering

```
for(⋄), sort(⋄)
```

Matricer

```
diag(⋄), repmat(⋄), size(⋄)
```

Løsningerne til opgaverne findes på absalon.



MATLAB cheat sheet (når du virkelig har problemer)

Kovariansmatrix

```
kovarians = 1./size(data.data,1) .* data.data' * data.data;
```

Korrelationsmatrix

```
for i=1:size(kovarians,1)
    for j=1:size(kovarians,2)
        corrl(i,j) = kovarians(i, j)./(sqrt(kovarians(i,i)) .* sqrt(kovarians(j,j)));
    end
end
```