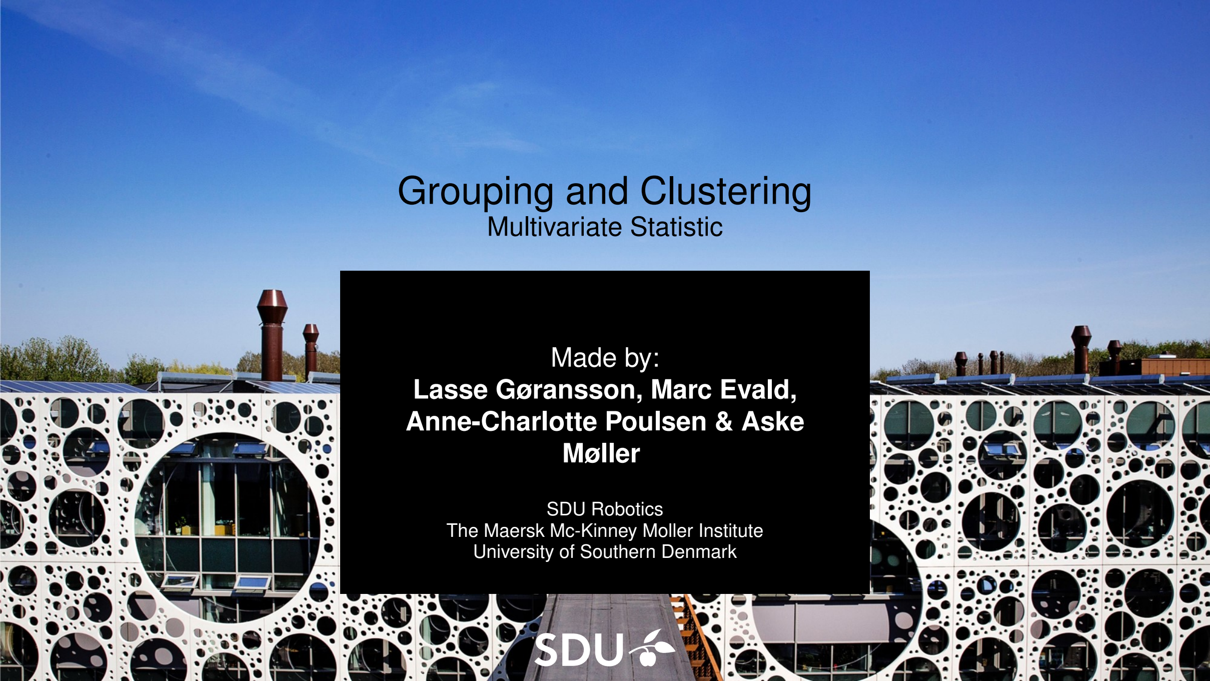


Grouping and Clustering

Multivariate Statistic

Made by:

Lasse Gøransson, Marc Evald, Anne-Charlotte Poulsen & Aske Møller

SDU Robotics
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark

Agenda

- Motivation
- Distance Measure
- Search Methods
  - Hierarchical Grouping
  - K-Means Clustering
  - Gaussian Mixture

Motivation

Split observation into K clusters

Distance Measure

Between Observations

Properties

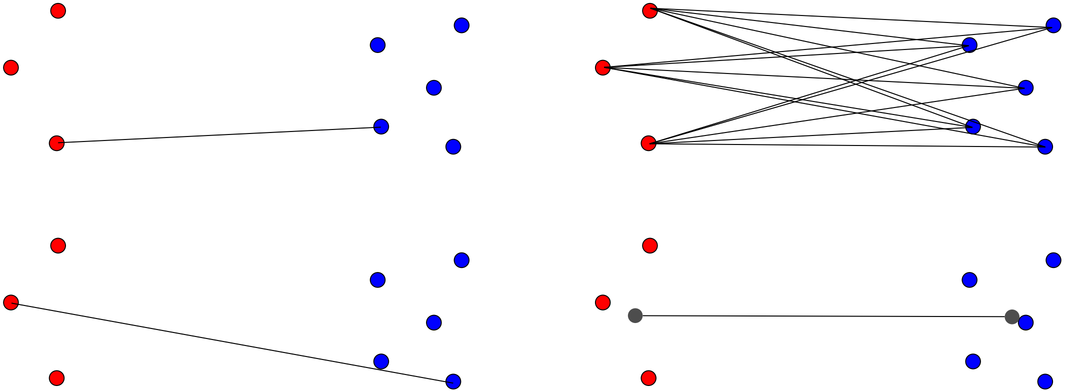
$$d\left(x,y\right)=\left(\sum_{i=1}^{p}\left|x_i-y_i\right|^m\right)^{\frac{1}{m}}$$

$$d\left(x,y\right)=\sqrt{\sum_{i=1}^{p}\left(x_i-y_i\right)^2}d\left(x,y\right)$$

$$d\left(x,y\right)=\sum_{i=1}^{p}\left|x_i-y_i\right|$$

Distance Measure

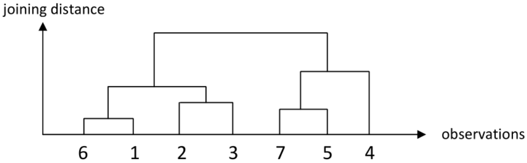
Between Clusters



Search Methods

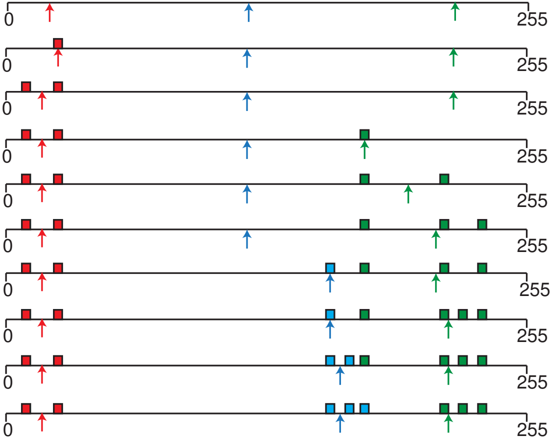
Hierarchical

$$D = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ & \ddots & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$



Search Methods

K-means



Search Methods

Gaussian Mixture

- Model

$$f_X\left(X\right) = \sum_{k=1}^{K} p_k f_{Y_k}\left(x\right), \qquad x = \left[x_1, \cdots, x_p\right]^T$$

Search Methods

Gaussian Mixture

- Model

$$f_X\left(X\right) = \sum_{k=1}^{K} p_k f_{Y_k}\left(x\right), \qquad x = \left[x_1, \cdots, x_p\right]^T$$

- Estimate

$$L\left(\{p_k\}_{k=1}^{K}, \{\mu_k\}_{k=1}^{K}, \{\Sigma_k\}_{k=1}^{K}\right)$$

Search Methods

Gaussian Mixture

- Model

$$f_X\left(X\right) = \sum_{k=1}^{K} p_k f_{Y_k}\left(x\right), \qquad x = \left[x_1, \cdots, x_p\right]^T$$

- Estimate

$$L\left(\{p_k\}_{k=1}^{K}, \{\mu_k\}_{k=1}^{K}, \{\Sigma_k\}_{k=1}^{K}\right)$$

- Evaluation

$$AIC = 2\log L_{\max} - 2\left(K \cdot \left[1 + \frac{p\left(p+3\right)}{2}\right] - 1\right)$$

$$BIC = 2\log L_{\max} - \left(K \cdot \left[1 + \frac{p\left(p+3\right)}{2}\right] - 1\right)\log n$$

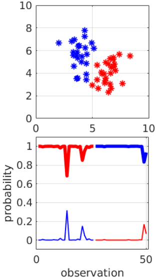
- Assign Clusters

Search Methods

Gaussian Mixture

K = 2

clustering and posteriors



$2*\log L_{max} = -314.2$

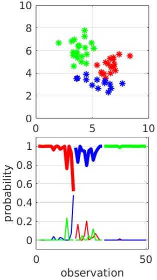
$N_{par} = 11$

AIC = -336.2

BIC = -357.2 (K=2 best BIC)

K = 3

clustering and posteriors



$2*\log L_{max} = -301.7$

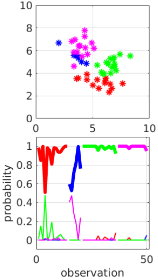
$N_{par} = 17$

AIC = -335.7 (K=3 best AIC)

BIC = -368.2

K = 4

clustering and posteriors



$2*\log L_{max} = -291.7$

$N_{par} = 23$

AIC = -337.7

BIC = -381.7