Unsupervised Clustering and Meta-analysis using Gaussian Mixture Copula Models

Anders E. Bilgrau, PhD fellow
abilgrau@math.aau.dk

Supervisors:
Martin Bøgsted and ***Poul Svante Eriksen***

AALBORG UNIVERSITY
DEPARTMENT OF MATHEMATICAL SCIENCES

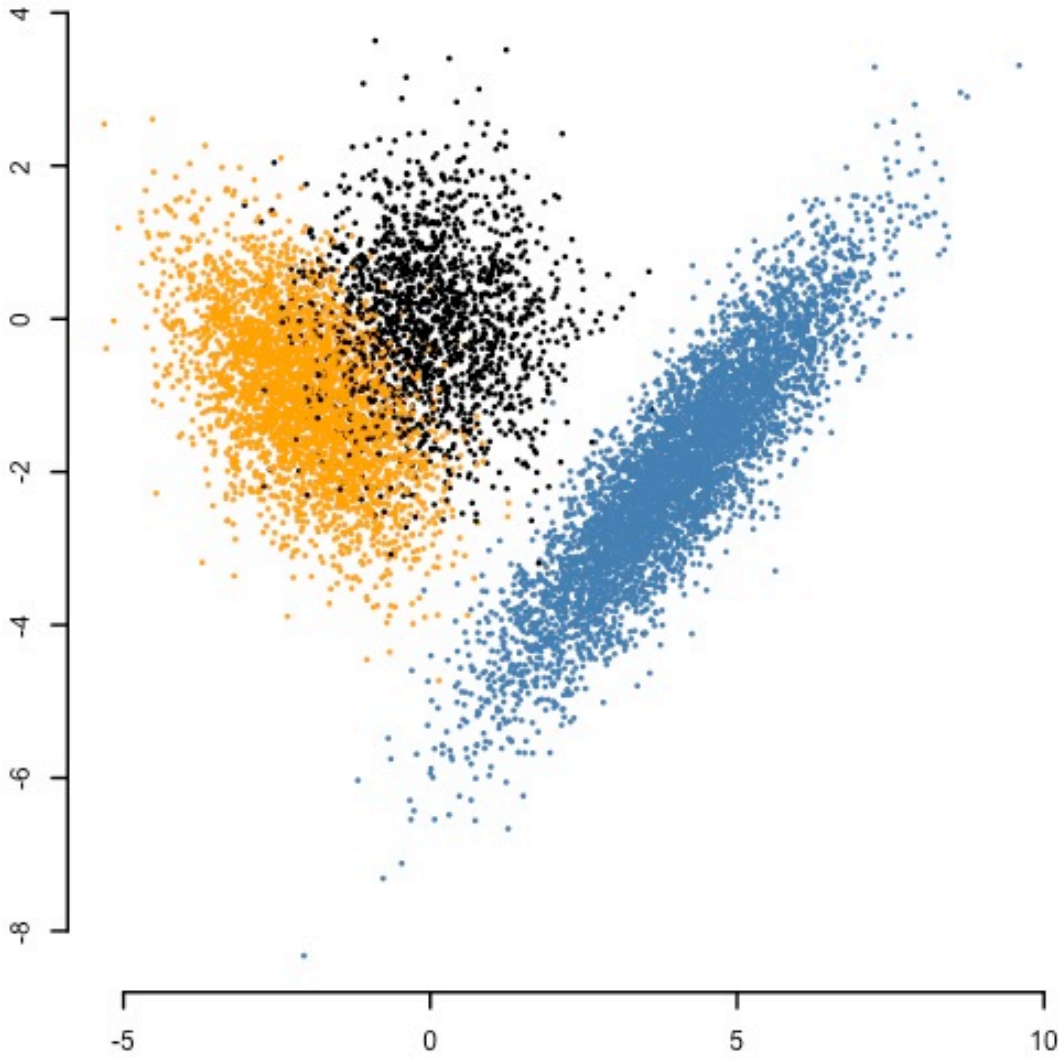
DEPARTMENT OF HAEMATOLOGY
AALBORG UNIVERSITY HOSPITAL

Introduction

- Model based unsupervised clustering
  - Gaussian mixture models (GMM)
  - Gaussian mixture **copula** models (**GMCM**)
    - A semi-parametric version of GMM
    - Example in “regular” clustering
    - Example in “reproducibility analysis”.
- Implemented in package **GMCM** (on CRAN)

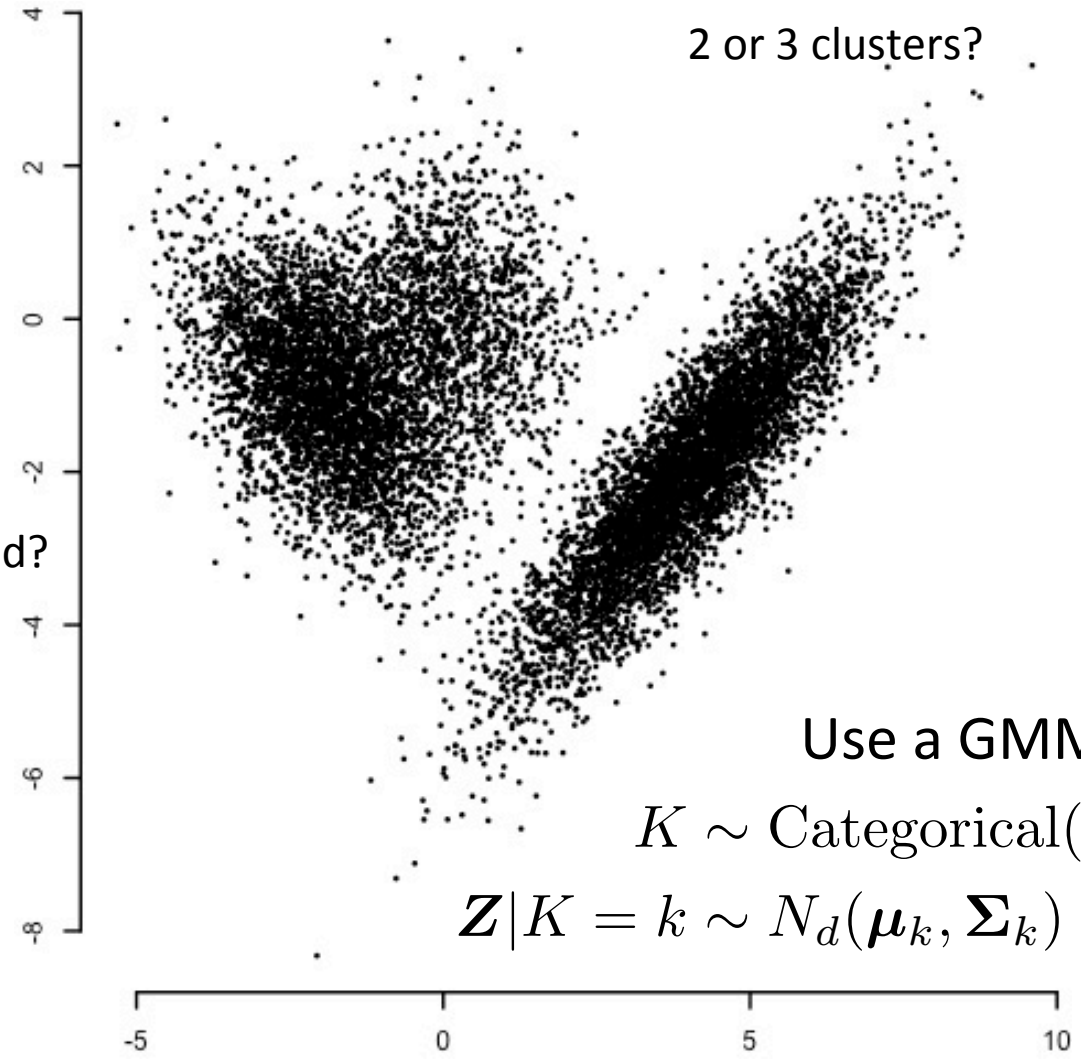
Introduction

- You are presented with some data…



Gaussian Mixture Models

- … but the classes are unknown



2 or 3 clusters?

Normally distributed?

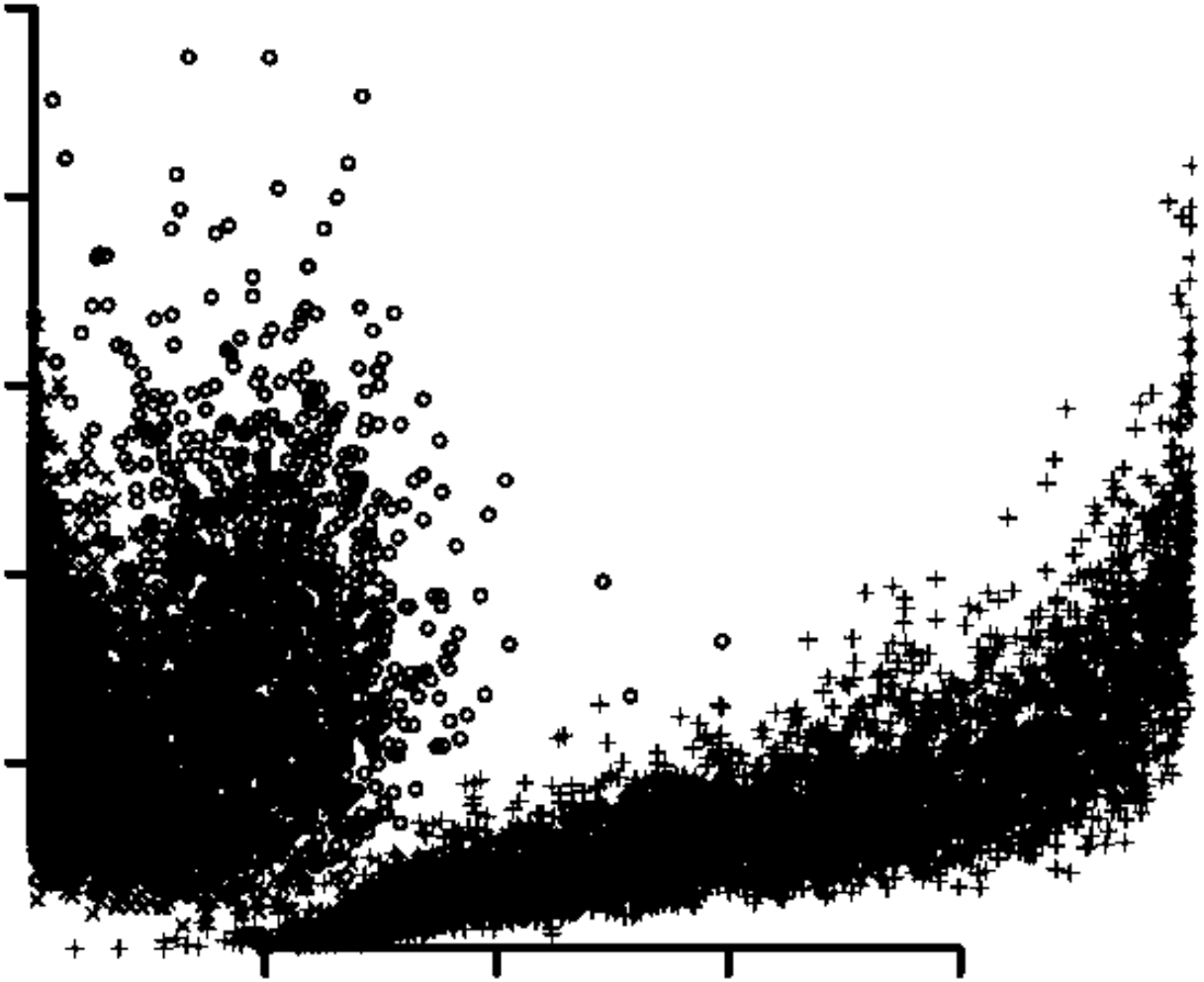
Use a GMM?

$$K \sim \text{Categorical}(\pi_1, \ldots, \pi_m)$$

$$\boldsymbol{Z}|K = k \sim N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian Mixture **Copula** Models

- What if data is clearly non-normal?



But what are GMCMs?

- Model the data with a latent *m*-component GMM:

$$K \sim \text{Categorical}(\pi_1, \ldots, \pi_m)$$
$$\boldsymbol{Z}|K = k \sim N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top$

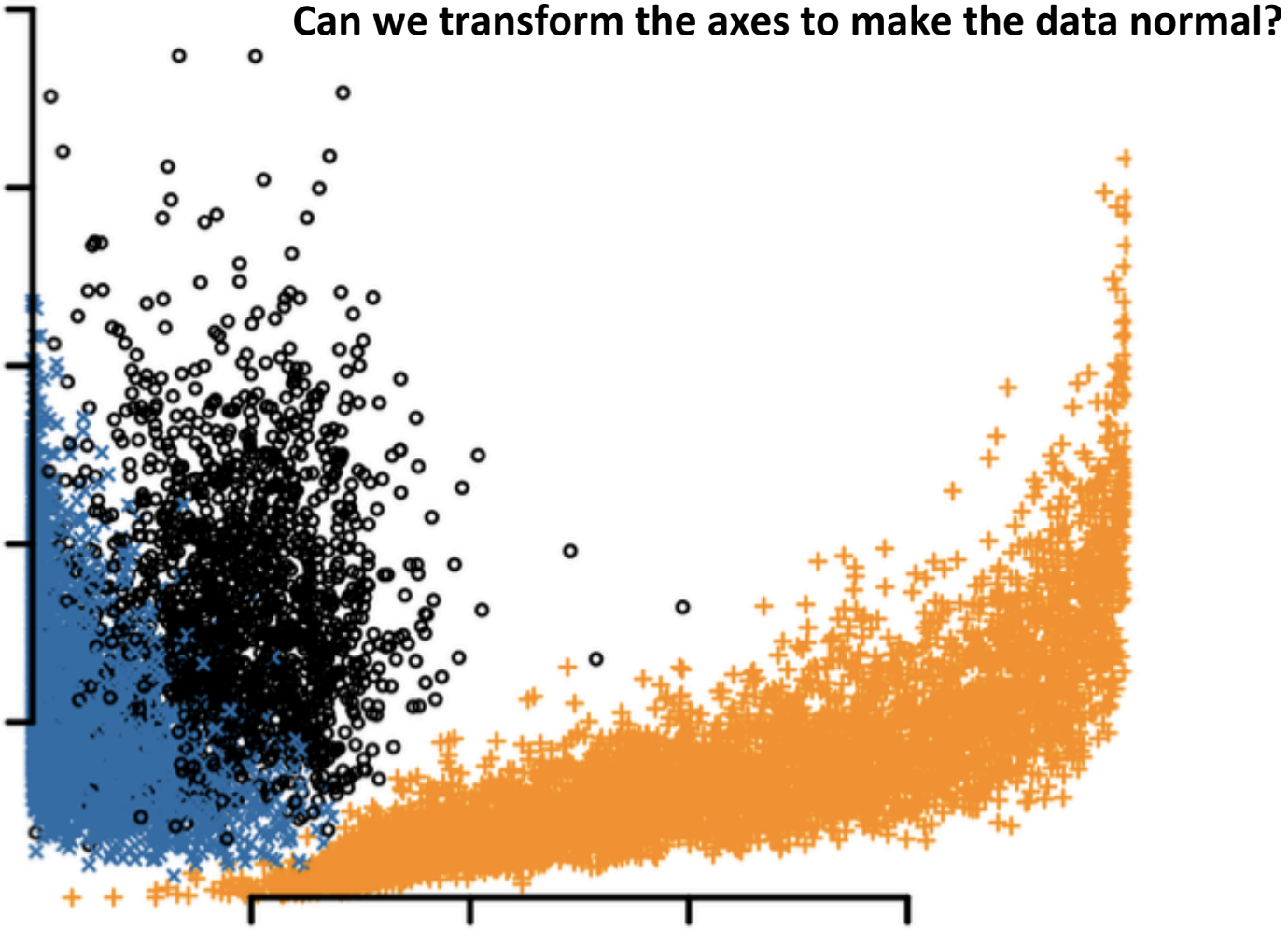
- Observe

$$\boldsymbol{X} = (X_1, \ldots, X_d)^\top = (f_1(Z_1), \ldots, f_d(Z_d))^\top$$

where $f_1, \ldots, f_d$ are *arbitrary* monotone increasing functions

Gaussian Mixture **Copula** Models

- GMCMs are a choice here.

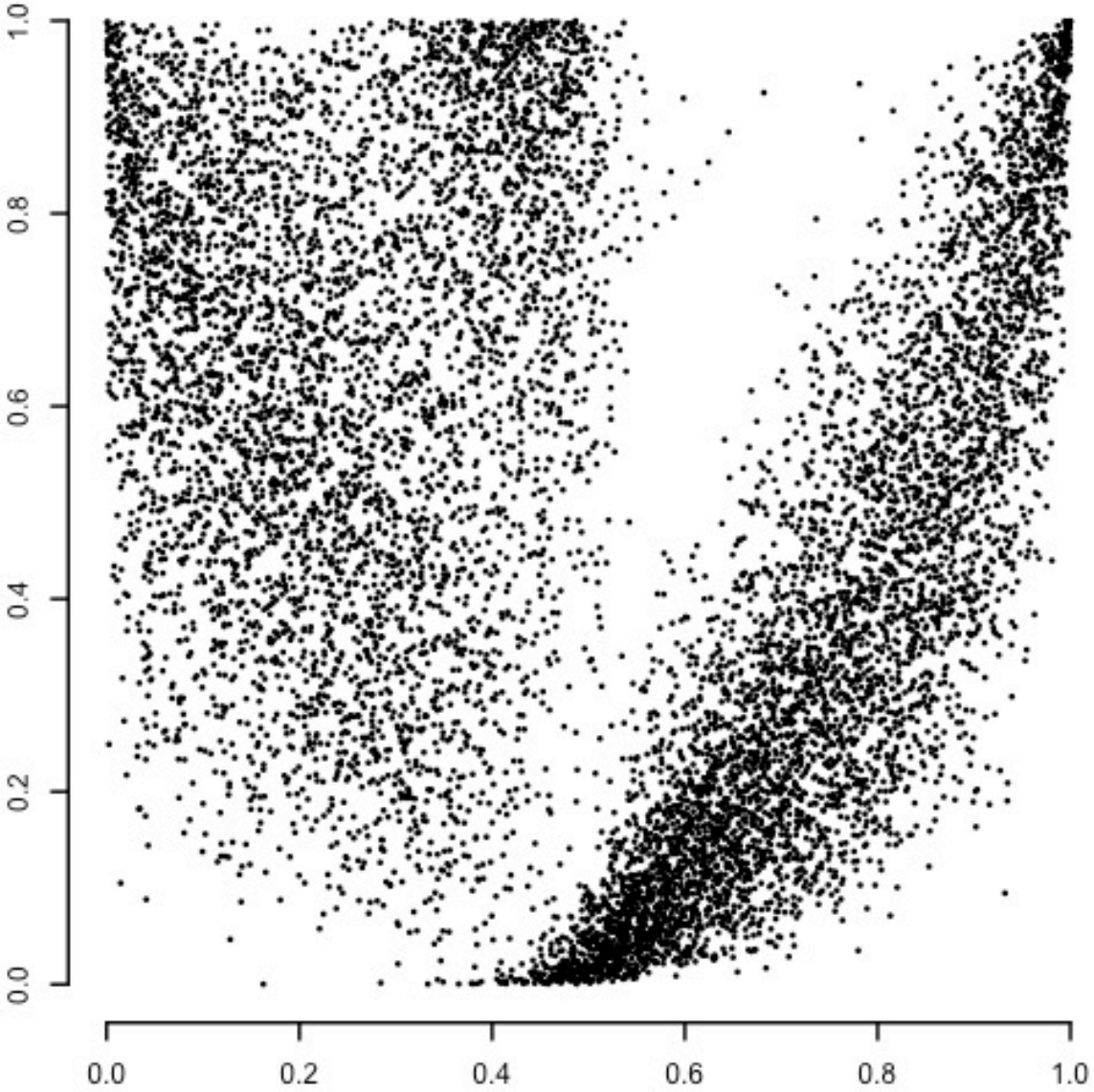


GMCMs in practice

- In practice, we model the ranked observations
- The *copula* of the GMM is a model for the ranked observations.
    - A copula is a distribution function with uniform marginal distributions
- Then optimize the (complicated) likelihood of the ranks

GMCMs in practice

- Plot of the ranks:



Clustering via GMCM

- We can derive the likelihood of the model:

$$K \sim \text{Categorical}(\pi_1, \ldots, \pi_m)$$

$$\boldsymbol{Z}|K = k \sim N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{X} = (X_1, \ldots, X_d)^\top = (f_1(Z_1), \ldots, f_d(Z_d))^\top$$

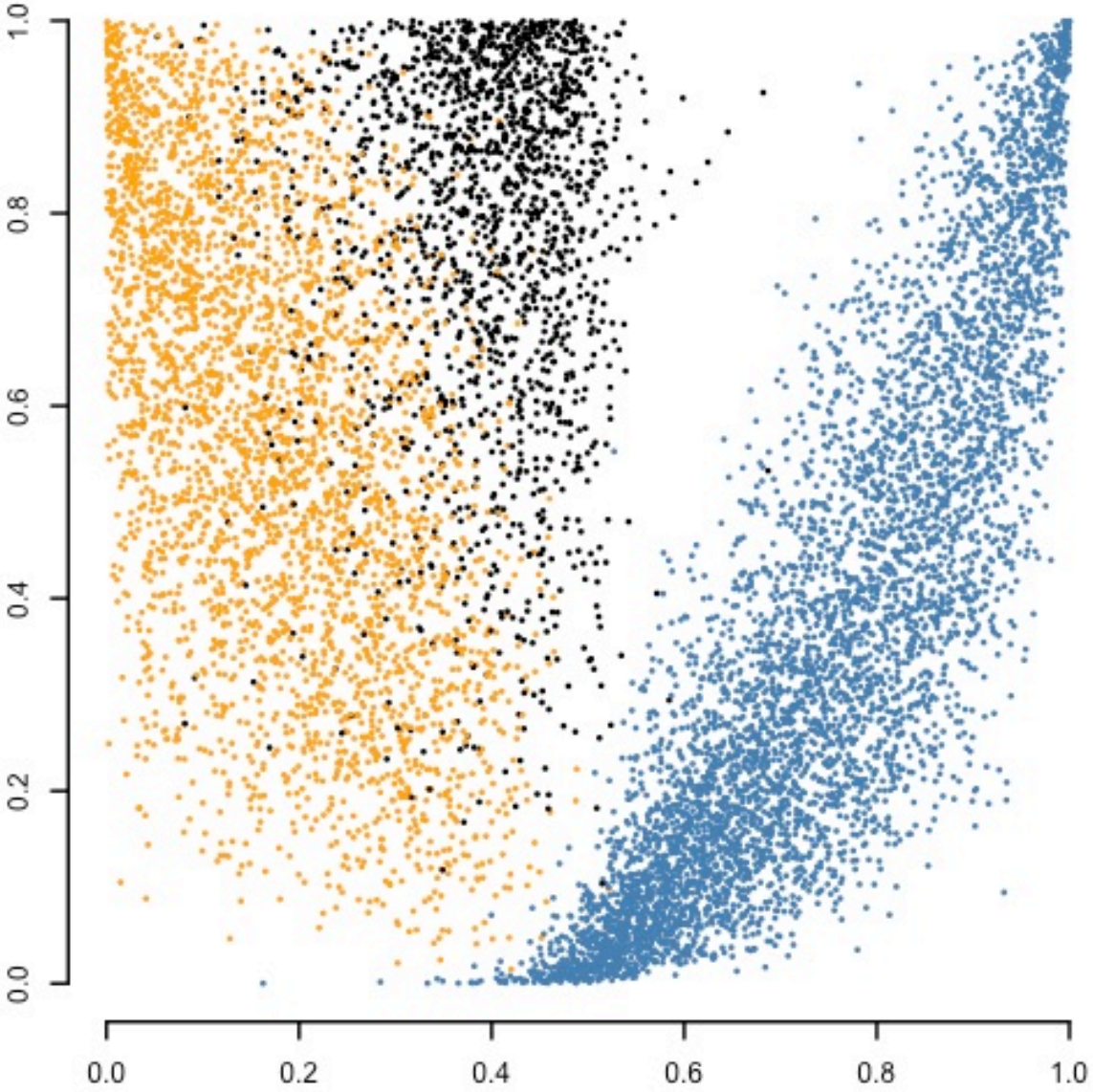
- Estimate parameters with ML `fit.full.GMCM(...)`
  - (Pseudo) EM algorithm
  - Standard numerical optimization methods
- Compute class-probabilities and decide the class; e.g.:

$$\hat{k} = \arg\max_k P(K = k|\boldsymbol{X}, \hat{\theta})$$

`get.prob(...)`

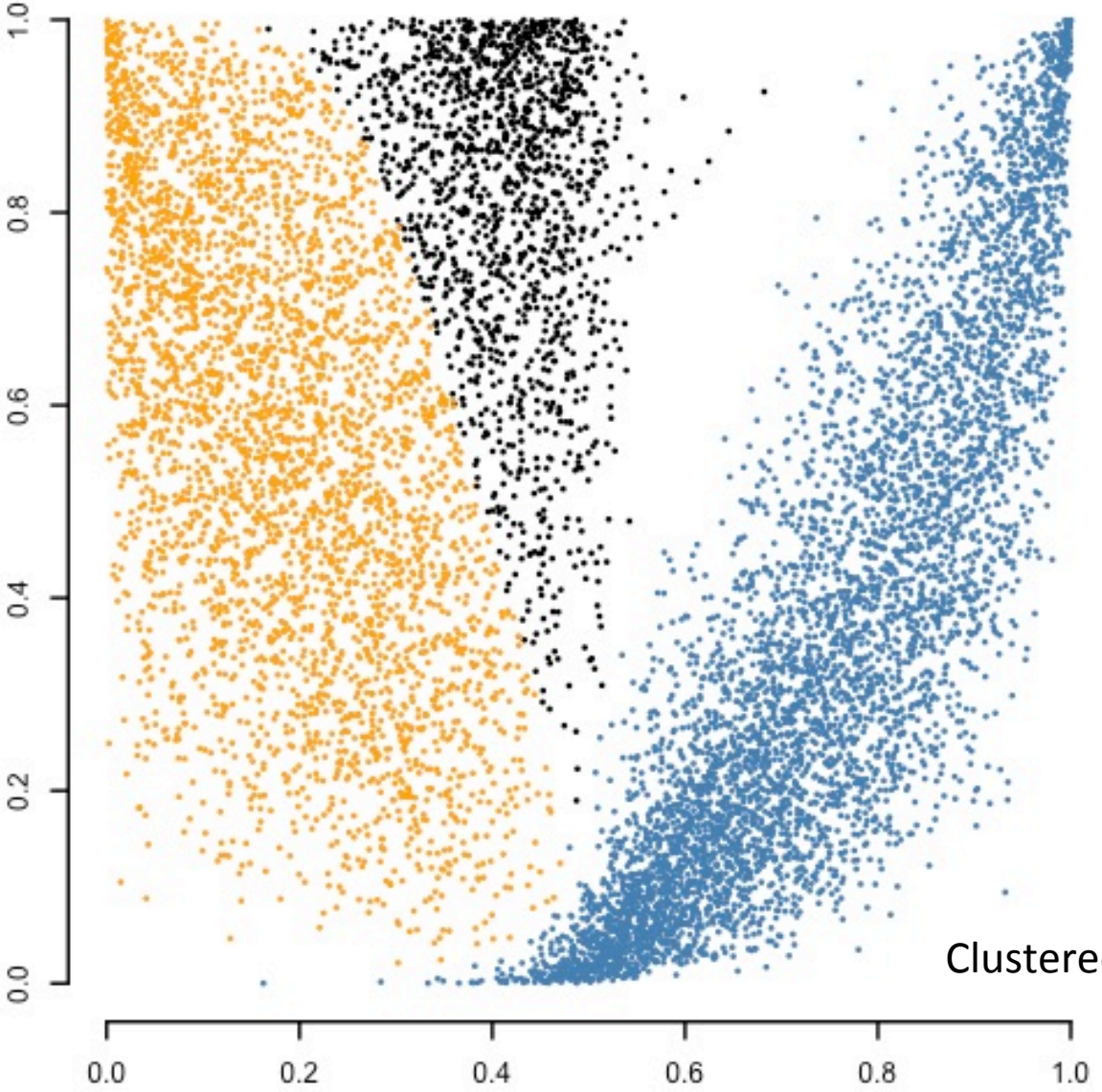
GMCMs in practice

- Plot of the ranks:



GMCMs in practice

- Plot of the ranks:



Clustered with GMCM

Example 1:

- Image segmentation
- An RGB pixel image with *n* pixels
  - can represented as a *n* by 3 table.
  - values in [0, 1]

(non-gaussian)

```
> head(img)
           R         G         B
px1 0.68078594 0.03044214 0.8109695
px2 0.50613594 0.06455260 0.8014441
px3 0.08645027 0.40910793 0.2866232
px4 0.98671489 0.84793493 0.3899271
px5 0.11290949 0.69738184 0.6103914
px6 0.70477680 0.47408089 0.6783339
```

cluster the rows

Example 1:

1500 px × 965 px = 1.4 mpx

Example 1:

Segmented into 11 colours

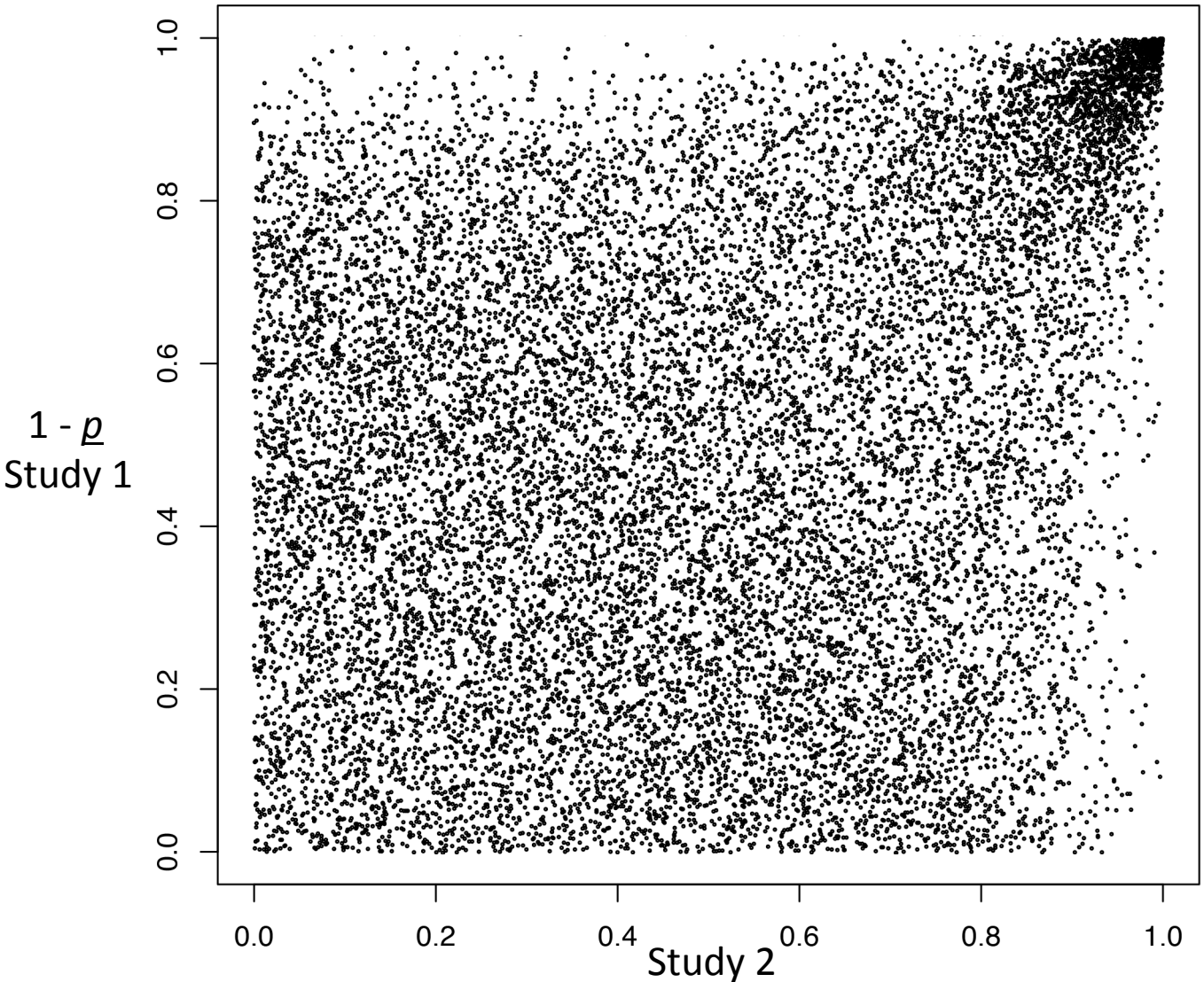
Example 2:

Cross-platform reproducibility of differentially expressed genes

- Two genetic studies investigating the activity of genes in two cell types of lymph nodes.
- The studies were carried out on 2 different microarray types:
  - we test 11,000 genes in two independent studies (think t-tests)

Example 2:

Cross-platform reproducibility of differentially expressed genes



Special case GMCM

[Li et al., 2011]

for reproducibility analysis

- Latent process: (2-component multivariate Gaussian mixture, *d* = 2)

$$K_i \sim \text{Bernoulli}(\pi)$$

$$\begin{pmatrix} z_{i1} \\ z_{i2} \end{pmatrix} \Big| K_i = 0 \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$

$$\begin{pmatrix} z_{i1} \\ z_{i2} \end{pmatrix} \Big| K_i = 1 \sim N_2\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right)$$

$$\theta = (\pi, \mu, \sigma, \rho)$$

- Observed ranks follow marginally uniform data by:

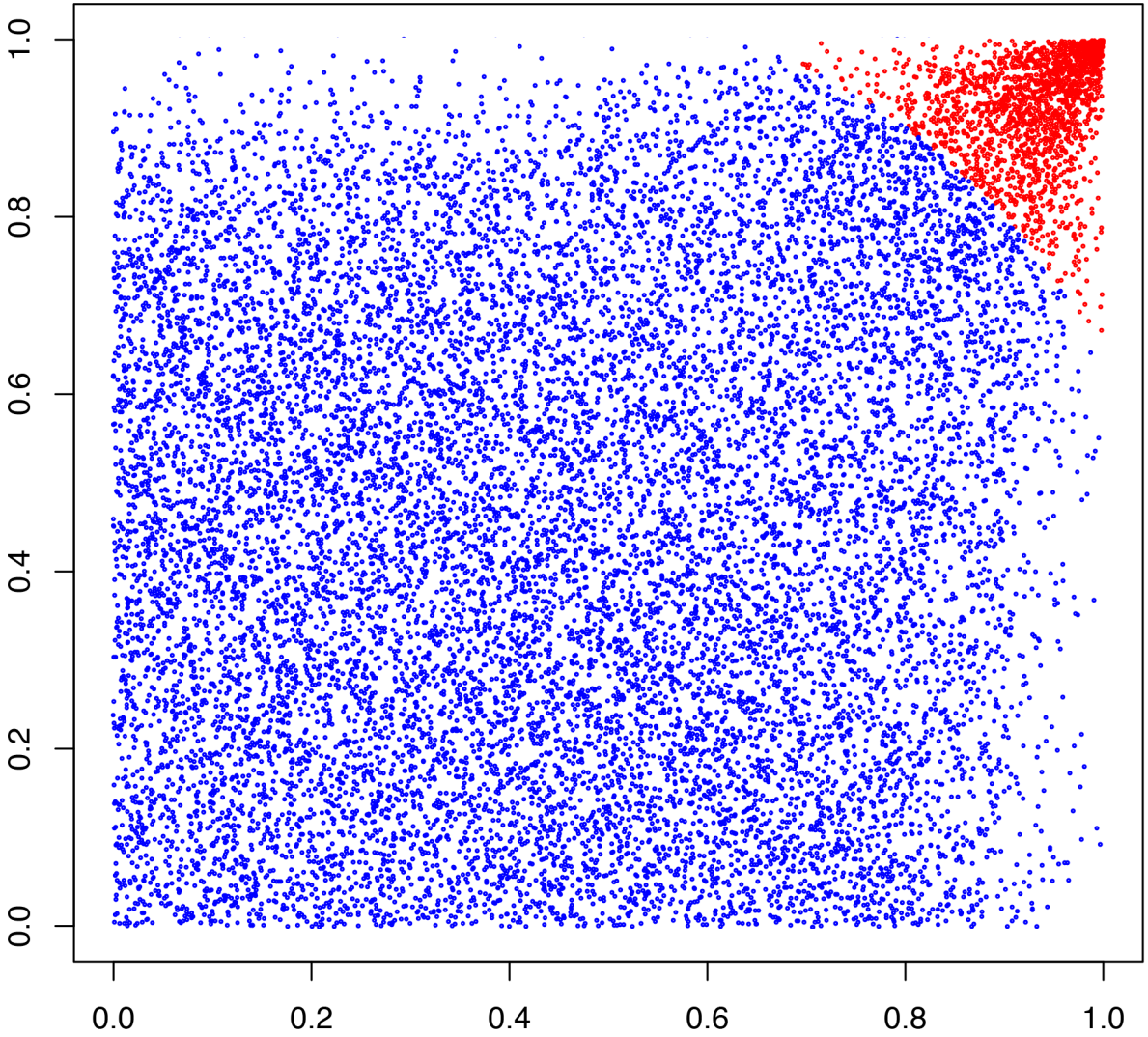
$$u_{i1} = G(z_{i1})$$
$$u_{i2} = G(z_{i2})$$

$$G(z) = (1-\pi)\Phi(z) + \pi\Phi\left(\frac{z-\mu}{\sigma}\right)$$

`GMCM::fit.meta.GMCM(...)`

Example 2:

Cross-platform reproducibility of differentially expressed genes



$P(K = k|\boldsymbol{X}, \hat{\boldsymbol{\theta}})$ is an aggregated level of evidence

14122 irreproducible
1751 reproducible

The GMCM-package

- Fast implementation via **Rcpp/RcppArmadillo**
  - **(> 500x speed-up compared to idr)**
  - Pseudo EM algorithm
    - Suggested by Li et al. (2011) and Tewari et al. (2011)
  - Standard numerical methods via `optim(…)`
- **User friendly** (*I* think)
- Arbitrary number of dimensions, components
  - Special case analysisfor any *d*.

Conclusion & take home

- GMCMs are semi-parametric versions of GMMs
  - potential alternative to GMMs where non-Gaussian clusters are present
- The **GMCM**-package allows simulation from and fitting GMCMs of arbitrary dimension and number of clusters.
- The model has problems of identifiability

Thanks for listening!

abilgrau@math.aau.dk

Further references and information:

- http://cran.r-project.org/package=GMCM (in-depth vignette/JStatSoft paper)

- **[Li et al., 2011]** Li, Q., Brown, J.B., Huang, H. & Bickel, P.J. *Measuring reproducibility of high-throughput experiments*. The Annals of Applied Statistics. 5, 1752-1779 (2011).
- **[Tewari et al. 2011]** Tewari A, Giering MJ, Raghunathan A (2011). "Parametric Characterization of Multimodal Distributions with Non-Gaussian Modes." ICDM 2011 conference, pp. 286–292. doi: 10.1109/ICDMW.2011.135.