# Flexible EM Clustering beyond the i.i.d paradigm

Matthieu Jonckheere

joint work with Frédéric Pascal and Violeta Roizman

DIG Seminar January 2021

# Some challenges for clustering

## Heterogeneous datasets

- Datasets with outliers/noise.
- Heavy tailed distributions.
- Different scales/distributions.
- Continuous and discrete data.

# Some challenges for clustering

## Heterogeneous datasets

- Datasets with outliers/noise.
- Heavy tailed distributions.
- Different scales/distributions.
- Continuous and discrete data.

## Lots of data ($n \gg$)

- high computational cost.
- need of parallelization / batch versions

# Some challenges for clustering

## Heterogeneous datasets

- Datasets with outliers/noise.
- Heavy tailed distributions.
- Different scales/distributions.
- Continuous and discrete data.

## Lots of data ($n \gg$)

- high computational cost.
- need of parallelization / batch versions

## High dimensional context ($m \gg$)

- ill-posed problems
- data on manifolds
- $\Rightarrow$ regularization, dimensionality reduction

# Some challenges for clustering

Focus here on:

## Heterogeneous datasets

- Datasets with outliers/noise.
- Different scales/distributions.

# Some challenges for clustering

Focus here on:

**Heterogeneous datasets**

- Datasets with outliers/noise.
- Different scales/distributions.

We address ”not too high dimensions” regimes (say 30-100).

# Outline

# Introduction and State of the art

# K-means

Given $\{\mathbf{x}_i\}_{i=1}^{n}$, find $\hat{\mathbf{C}} = \{C_1, \ldots, C_K\}$ with $\boldsymbol{\mu}_k = \frac{1}{\#(C_k)} \sum_{\mathbf{x} \in C_k} \mathbf{x}$ such that

$$\hat{\mathbf{C}} = \underset{\mathbf{C}=\{C_1,\ldots,C_K\}}{argmin} \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2$$

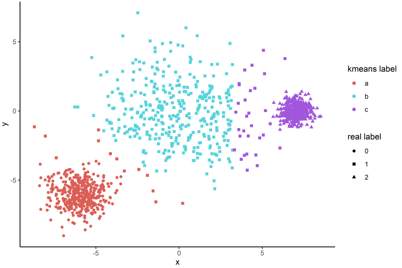
Plain optimization problem.

Simple idea. ✓

Very fast. ✓

Works well only when: ✗

- round-shaped clusters,
- with similar variance, and
- well-separated.

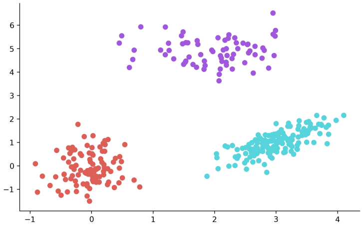


# Gaussian Mixture Model (GMM)

We model data as a mixture of Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_k, \mathsf{M}_k)$:

$$f(\mathsf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathsf{x}),$$

with $\pi_k$ the proportion of cluster $k$ and $f_k$ the normal p.d.f.




$$f_k(\mathsf{x}) = \frac{1}{(2\pi)^{m/2}|\mathsf{M}_k|^{1/2}} \exp\left[-\frac{(\mathsf{x}-\boldsymbol{\mu}_k)^T \mathsf{M}_k^{-1}(\mathsf{x}-\boldsymbol{\mu}_k)}{2}\right]$$

# Expectation-Maximization (EM) algorithm

Statistical algorithm to estimate parameters based on a likelihood.

In the GMM case, we would need the labels of the data points to estimate the parameters. Labels $\rightarrow$ Latent variables

# Expectation-Maximization (EM) algorithm

Statistical algorithm to estimate parameters based on a likelihood.

In the GMM case, we would need the labels of the data points to estimate the parameters. Labels $\rightarrow$ Latent variables

## E-STEP

Computation of the membership a posteriori probabilities

$$p_{ik} = P(Z_i = k | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x}_i)}$$

with $f_k$ the Gaussian p.d.f.

# Expectation-Maximization (EM) algorithm

Statistical algorithm to estimate parameters based on a likelihood.

In the GMM case, we would need the labels of the data points to estimate the parameters. Labels $\rightarrow$ Latent variables

## E-STEP

Computation of the membership a posteriori probabilities

$$p_{ik} = P(Z_i = k | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x}_i)}$$

with $f_k$ the Gaussian p.d.f.

## M-STEP

Estimation of the parameters

$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} p_{ik}$$

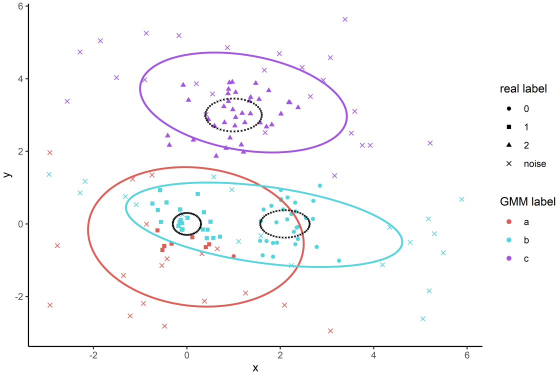
$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{n\widehat{\pi}_k} \sum_{i=1}^{n} p_{ik} \mathbf{x}_i$$

$$\widehat{\mathbf{M}}_k = \frac{1}{n\widehat{\pi}_k} \sum_{i=1}^{n} p_{ik} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T$$

# What happens to GMM when the data has some noise or non Gaussian data?

The GMM has problems to cluster and estimate parameters for data with noise, different distribution shapes and outliers.

Result with data contaminated:



# What happens to GMM when the data has some noise or is non Gaussian?

**Why?**

- The **estimators** are **not robust**.
- **Mismatch** between the **model** and the data.
- **No outlier rejection**.

# How to address the robustness challenge?

There are mainly two directions to robustify clustering methods in the literature:

- model generalizations
  - Extra uniform cluster [Banfield and Raftery, 1993]
  - Model low density areas (RIMLE and OTRIMLE) [Coretto and Hennig, 2016]
  - Mixture of $t-$distributions (t-EM) [Peel and McLachlan, 2000]
- models that introduce classical robust techniques in the estimation
  - Trimming methods (TCLUST) [García-Escudero et al., 2008]
  - k-tau [Gonzalez et al., 2019] and Spatial-EM [Yu et al., 2015]

# Some drawbacks

Some drawbacks of the state of the art robust clustering methods:

- No closed equations on the M-step, reliance on non-linear optimizers (t-EM).
- Extra parameters difficult to be tuned (RIMLE, TCLUST).
  e.g. if we misspecify the proportion of noise in the TCLUST algorithm [Gonzalez et al., 2019].
- Models are too specific.

# Some drawbacks

Some **drawbacks** of the state of the art robust clustering methods:

- No closed equations on the M-step, reliance on non-linear optimizers (t-EM).
- Extra parameters difficult to be tuned (RIMLE, TCLUST).
  e.g. if we misspecify the proportion of noise in the TCLUST algorithm [Gonzalez et al., 2019].
- Models are too specific.

**Our goal:**

- flexibility to very general models
- no extra parameters

# F-EM: Model, derivation and properties

# Model

We consider $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ independent vectors.

These vectors belong to some clusters $C_1, \ldots, C_K$.

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ ARE NOT i.i.d. !

**Cluster characterization**

$\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to $C_k$ if they are drawn from a distribution with the same features

$$\boldsymbol{\mu}_k \quad \text{and} \quad \boldsymbol{\Sigma}_k$$

The location and the scatter matrix are the features that characterize the clusters and not a particular distribution as in GMM or t-EM.

# F-EM: A flexible algorithm relying on a very general model

F-EM is based on a model where the $\mathbf{x}_1, \ldots, \mathbf{x}_n$ independent vectors are characterized by

## Stochastic representation

$$\mathbf{x}_i \in C_k \Rightarrow \mathbf{x}_i \overset{d}{=} \boldsymbol{\mu}_k + \sqrt{\mathcal{Q}_{ik}}\,\sqrt{\tau_{ik}}\,\mathbf{A}_k\,\mathbf{u}_i$$

- $\boldsymbol{\mu}_k$ is the mean of the cluster $k$.
- $\mathcal{Q}_{ik}$ is an independent positive random variable.
- $\tau_{ik}$ **are scale (nuisance) parameters that increase the flexibility of the model.**
- $\mathbf{A}_k$ is such that $\mathbf{A}_k^T\mathbf{A}_k = \boldsymbol{\Sigma}_k$ (the scatter matrix of the cluster $k$).
- $\mathbf{u}_i$ is a uniform vector on the unit hyper-sphere.

# Elliptical Symmetric family

The stochastic characterization [Cambanis et al., 1981] represents vectors of the Elliptical Symmetric family [Kelker, 1970].

The density can be written as

$$\boxed{f_{\mathbf{x}_i}(\mathbf{x}) = A_m |\tau_{ik} \boldsymbol{\Sigma}_k|^{-1/2} \; \mathbf{g}_{ik}\left(\tau_{ik}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)}$$

for some function $\mathbf{g}_{ik}$ called the **density generator**. We denote it as $\mathbf{x} \sim \mathrm{ES}(\boldsymbol{\mu}_k, \tau_{ik}\boldsymbol{\Sigma}_k, g_{ik})$.

$g_{ik}$ characterizes $\mathcal{Q}_{ik}$ and gives the **shape** of the distributions

This family includes Gaussian, $t-$distribution, Generalized Gaussian distribution. Heavier and lighter (than Gaussian) tails.

# Different scenarios

We consider different scenarios based on the nature of the density generator functions:

$$g_{ik} = \begin{cases} g_i, & \text{each point might come from different shaped distributions} \\ & \text{BUT shapes do not depend on the cluster} \\ g, & \text{the density generator function is} \\ & \text{always the same (e.g. Gaussian case)} \\ g_k, & \text{cluster dependent shapes} \end{cases}$$

# F-EM: A flexible algorithm relying on a very general model

## Parameter space

Given $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^m$ we have to estimate the usual parameters

$$\boldsymbol{\Theta} = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1,..,K}$$

AND we now have a lot of (nuisance) parameters $\tau$

$$\widetilde{\boldsymbol{\Theta}} = \{\tau_{ik}\}_{\substack{k=1,..,K \\ i=1,..,n}}$$

## MLE

We derive the two-step (E-M) algorithm based on the likelihood of the model (using the trick of [Ollila and Tyler, 2012]).

# E-step: First Miracle

**Proposition**

*Assume $g_{ik} = g_i$, then the membership probabilities MLE are*

$$\widehat{p}_{ik} = \frac{\widehat{\pi}_k \left( (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k) \right)^{-m/2} |\widehat{\boldsymbol{\Sigma}}_k|^{-1/2}}{\sum_{j=1}^{K} \widehat{\pi}_j \left( (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_j)^T \widehat{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_j) \right)^{-m/2} |\widehat{\boldsymbol{\Sigma}}_j|^{-1/2}}.$$

Insensitivity: the expression of the membership does not depend on the particular density $g_i$ that generates each data point

**Proposition (Location and scatter matrix estimators)**

*We almost obtain Tyler's estimators.*

$$\widehat{\boldsymbol{\mu}}_k = \frac{\displaystyle\sum_{i=1}^{n} \frac{\hat{p}_{ik}\mathbf{x}_i}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}_k{}^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}}{\displaystyle\sum_{i=1}^{n} \frac{\hat{p}_{ik}}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}}$$

$$\widehat{\boldsymbol{\Sigma}}_k = m \sum_{i=1}^{n} \frac{w_{ik}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}_k{}^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}, \quad \text{with} \quad w_{ik} = \hat{p}_{ik} / \sum_i \hat{p}_{ik}$$

# And the taus?

Furthermore,

$$\widehat{\tau}_{ik} = \frac{(\mathrm{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} (\mathrm{x}_i - \widehat{\boldsymbol{\mu}}_k)}{a_{ik}},$$

where $a_{ik}$ depends only on $g_{ik}$, for example for the Gaussian case $a_{ik} = m$.

# Estimators intuitively

$\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\boldsymbol{\Sigma}}_k$ are like usual sample estimators with small weights for outlying points

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} \gamma_i \mathbf{x}_i$$

$$\frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^T \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} \gamma_i (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^T$$

with $\gamma_i = C \frac{\hat{p}_{ik}}{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)}$

**Tyler** estimators [Tyler, 1987] (classical robust estimator [Maronna, 1976]) fulfill very similar equations. HINT about robustness of the model.

# About F-EM

## Properties

- The random vectors that represent the data points are independent but not necessarily i.i.d.
- Generalizes GMM. (Gaussian $\in$ ES)
- If $g_{ik} = g_i$, the membership probabilities do not depend on the shape of the distributions!
- If $g_{ik} = g_k$, we can derive extra estimators to be computed on the M-Step.
- The model leads to estimators that are similar to classical robust estimators (Tyler) [Ollila and Tyler, 2012].

# If the dimension grows... Some hints

When the dimension grows we can better estimate the parameters $\tau_{ik}$.

**Convergence of $\hat{\tau}$ when $g$ is the Gaussian density generator**

Let $\mathbf{x} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{\tau}\mathbf{A}\mathbf{q}$, with $\mathbf{q}$ a standard Gaussian. Under some assumptions, for any $a \in \mathbb{R}$, $\forall \varepsilon > 0$ and $\mathbf{y} \sim \mathcal{N}(\tau, 2\tau^2/m)$, then

$$|\mathbb{P}(\{\hat{\tau} \leq a\}) - \mathbb{P}(\mathbf{y} \leq a)| < \varepsilon, \text{ if } n \text{ and } m \text{ are large enough}$$

This is in agreement with previous RMT results [Couillet et al., 2014].

We can combine this result with parsimonious restrictions on the covariance matrix to avoid issues in the case of **very large $m$**.

# Implementation details

- The trace of the scatter matrix estimator is fixed.
- Possible centers initialization: quick run of k-means.
- Code available: *`github.com/violetr/fem`*

# F-EM: Experimental results

# Measuring the performance

## We compare our algorithm to

- k-means
- GMM-EM
- Spectral Clustering
- Mixture of Student's $t$ ($t$-EM or EMMIX)
- TClust
- RIMLE

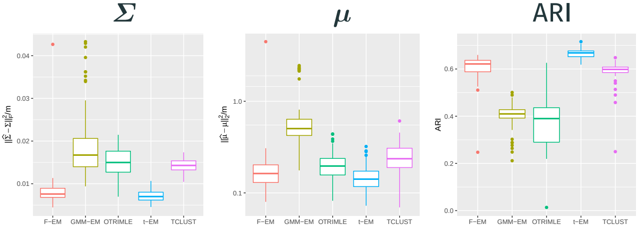
## Metrics

- Adjusted Mutual Information (AMI),
- Adjusted Rand Index (AR).
- Estimation error of the parameters (only for simulations).

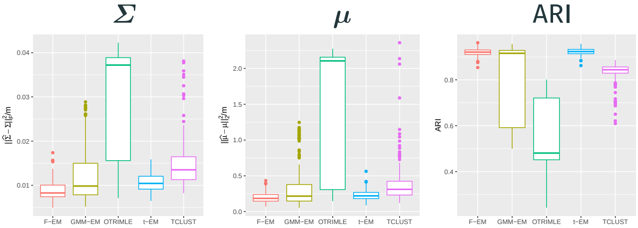
# Some simulation results

Mixtures of t-distributions with different degrees of freedom and covariance matrix classes, mixtures of more general distributions, clusters with different $g_i$.

**Setup 1:**
$t-$distributions
$\nu$ small



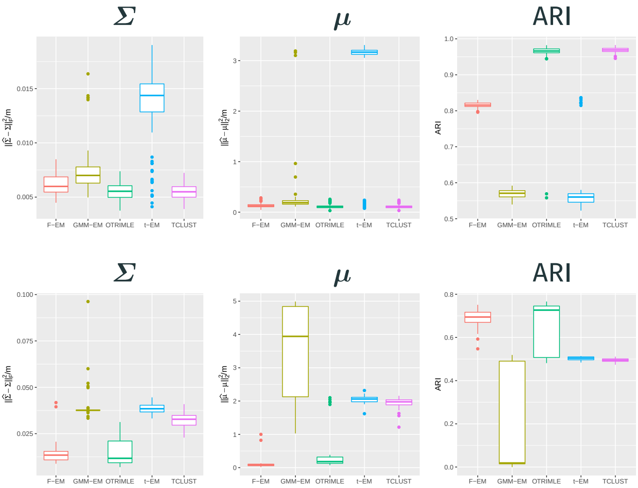

**Setup 2:**
$t-$distributions
$\nu = 10$



# Some simulation results

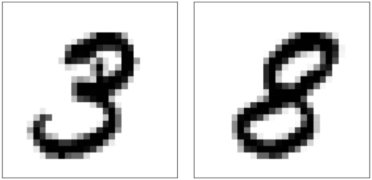
**Setup 3:** Gaussian + uniform noise

**Setup 4:** Elliptical different $g_i$

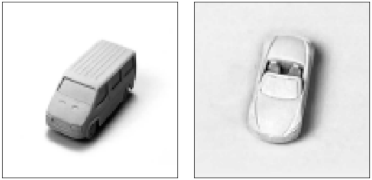



F-EM performs well even in the situations that do not match the model.

# Real data clustering results



MNIST (LeCun, 1998)



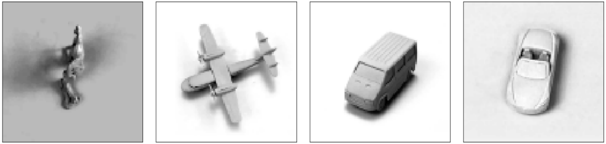
NORB (LeCun, 2004)

| Set | k-means | GMM | $t$-EM | F-EM | spectral | TCLUST | RIMLE |
|---|---|---|---|---|---|---|---|
| MNIST38 | 0.2884 | 0.5716 | 0.6397 | **0.6887** | 0.6866 | 0.6847 | 0.2494 |
| MNIST71 | 0.8486 | 0.8905 | **0.9432** | 0.9360 | 0.9384 | 0.6885 | 0.2493 |
| MNIST386 | 0.6338 | 0.7332 | 0.8262 | 0.8306 | **0.8542** | 0.8366 | 0.4274 |
| MNIST386+n | 0.4475 | 0.4909 | 0.5296 | 0.5548 | 0.3115 | **0.6908** | 0.1498 |
| smallNORB | 0.0015 | 0.0468 | 0.4223 | **0.5067** | $\sim 0$ | 0.1330 | 0.1472 |
| 20news | 0.1883 | 0.2739 | 0.4426 | **0.5114** | 0.0987 | 0.2664 | 0.0026 |

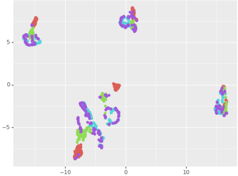
**Table 1:** Median AMI

# Real data clustering results - The NORB case

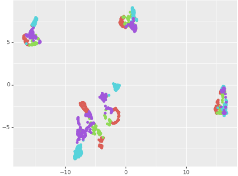
| Dataset | kmeans | GMM-EM | $t$-EM | F-EM | spectral | TCLUST | RIMLE |
|---|---|---|---|---|---|---|---|
| small NORB | 0.0015 | 0.0468 | 0.4223 | **0.5067** | $\sim 0$ | 0.1330 | 0.1472 |



t-SNE embedding of the dataset colored with labels:




**real** labels




**F-EM** labels


**GMM-EM** labels

# References i


Banfield, J. D. and Raftery, A. E. (1993).
Model-based gaussian and non-gaussian clustering.
*Biometrics*, 49(3):803–821.

Cambanis, S., Huang, S., and Simons, G. (1981).
On the theory of elliptically contoured distributions.
*Journal of Multivariate Analysis*, 11(3):368–385.

Coretto, P. and Hennig, C. (2016).
Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust gaussian clustering.
*Journal of the American Statistical Association*, 111(516):1648–1659.

Couillet, R., Pascal, F., and Silverstein, J. W. (2014).
Robust estimates of covariance matrices in the large dimensional regime.
*IEEE Transactions on Information Theory*, 60(11):7269–7278.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008).
A general trimming approach to robust cluster analysis.
*Ann. Statist.*, 36(3):1324–1345.

# References ii


Gonzalez, J. D., Yohai, V. J., and Zamar, R. H. (2019).
Robust Clustering Using Tau-Scales.
*arXiv e-prints*, page arXiv:1906.08198.

Kelker, D. (1970).
Distribution theory of spherical distributions and a location-scale parameter generalization.
32(4):419–430.

Maronna, R. A. (1976).
Robust M-Estimators of multivariate location and scatter.
*The Annals of Statistics*, 4(1):51–67.

Ollila, E. and Tyler, D. E. (2012).
Distribution-free detection under complex elliptically symmetric clutter distribution.
In *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 413–416.

Peel, D. and McLachlan, G. J. (2000).
Robust mixture modelling using the t distribution.
*Statistics and Computing*, 10(4):339–348.

Tyler, D. E. (1987).
A distribution-free $M$-estimator of multivariate scatter.
*The Annals of Statistics*, 15(1):234–251.

Yu, K., Dang, X., Bart, H., and Chen, Y. (2015).
Robust model-based learning via spatial-em algorithm.
*IEEE Transactions on Knowledge and Data Engineering*, 27(6):1670–1682.

# THANKS!