Topics in Model-based Clustering and Classification

Irene Vrbik

Department of Statistics, University of British Columbia Okanagan

April 5, 2017

Outline

1 Introduction
- Model-based clustering via finite mixture models

2 Finite Mixture Models (FMM)
- Definition
- Skew-t mixture FMM
- A quick demonstration

3 Fractionally-supervised classification (FSC)
- A quick demonstration

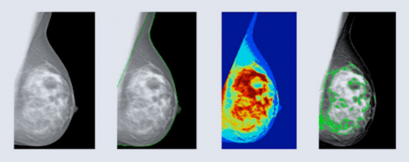
An introduction to clustering

What is clustering?

- Clustering is an unsupervised learning method for finding hidden patterns or grouping in data.
- Clustering algorithms are performed on datasets consisting of input data without labeled responses.
- The goal of clustering is to categorize data into meaningful groups where the similarity within groups and the dissimilarity between groups are maximized.

Some applications of clustering

- Image processing

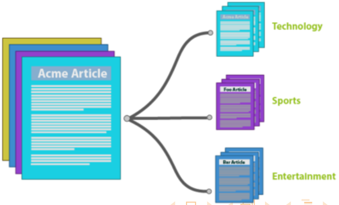


[1]

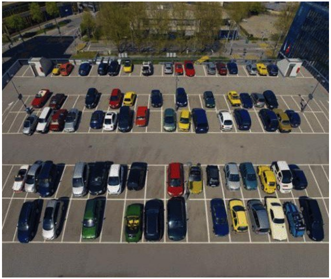
- Market segmentation

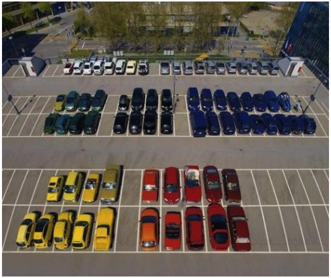
Source: Mahmoud Reza Saremi DBA, MBA/MKT

- Text mining Source:analyticbridge



An introduction to clustering





How do we determine clusters?

- Distance-based clustering
  - Based on distance metric such as euclidean or manhattan
  - Some methods include: hierarchical clustering, partitioning methods (e.g. k-means)
- Model-based clustering
  - Based on probability models
  - Most popular example: `mclust`

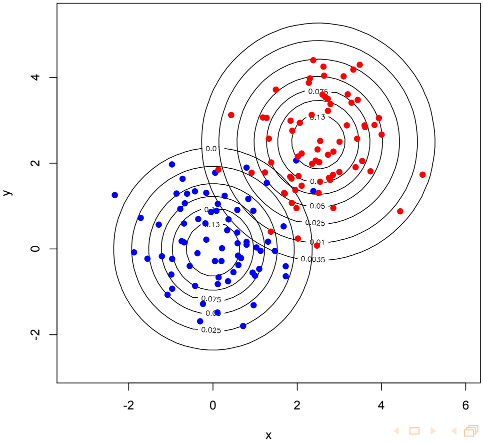
Finite Mixture Models (FMM)

- Assumes the population is composed of a collection of G sub-populations called groups or components each following its own distribution.
- The density of a G-component mixture model is of the form

$$f(\mathbf{x} \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g), \tag{1}$$

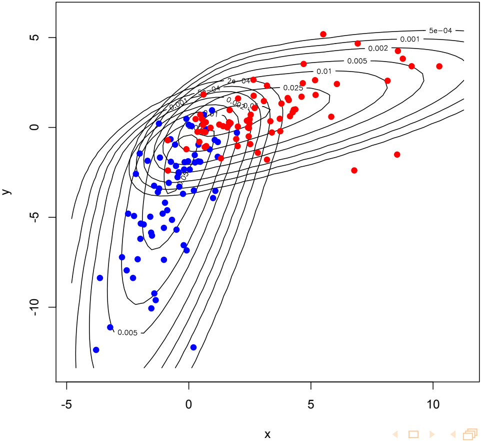
where π_g are the nonnegative mixing proportions such that $\sum_{g=1}^{G} \pi_g = 1$, and f_g is the density for the gth group with parameters $\boldsymbol{\theta}_g$.

f_g has predominantly taken to be Gaussian which is a very strong assumption on the structure of the data.



Alternatives to Gaussian FMM

There has been increasing interest in non-gaussian assumptions on the sub-populations.



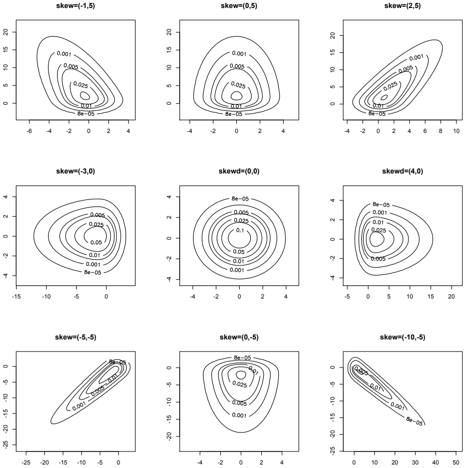


Figure: Density contours from bivariate skew-normal distributions with $\boldsymbol{\xi} = \mathbf{0}$, $\boldsymbol{\Sigma} = \mathbf{I}$, and varying values for the skewness parameter $(\boldsymbol{\lambda})$.

Multivariate skew-t mixture model

A formulation of a finite mixture of skew-t distributions assumes f_g follows a skew-t distribution with location $\boldsymbol{\xi}_g$, skewness vector $\boldsymbol{\lambda}_g$, scale matrix $\boldsymbol{\Sigma}_g$ and degrees of freedom ν_g given by

$$\varphi(\mathbf{x} \mid \boldsymbol{\xi}_g, \boldsymbol{\lambda}_g, \boldsymbol{\Sigma}_g, \nu_g) = 2t_p(\mathbf{x} \mid \boldsymbol{\xi}_g, \boldsymbol{\Omega}_g)\, T_{p+\nu_g}\left(\frac{\alpha_g}{\beta_g} \sqrt{\frac{\nu_g + p}{\nu_g + \delta_g}} \right),$$

where

$$\begin{aligned}
\boldsymbol{\Omega}_g &= \boldsymbol{\Sigma}_g + \boldsymbol{\lambda}_g \boldsymbol{\lambda}_g^\top, & p &= \text{dimension of the data} \\
\alpha_g &= \boldsymbol{\lambda}_g^\top \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\xi}_g), & t_p &= \text{PDF of a } p\text{-variate } t \\
\beta_g^2 &= (1 - \boldsymbol{\lambda}_g^\top \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\lambda}_g), & T_{p+\nu} &= \text{CDF of a univariate } t \\
\delta_g &= (\mathbf{x} - \boldsymbol{\xi}_g)^\top \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\xi}_g), & & \text{with degrees of freedom } p + \nu.
\end{aligned}$$

The EM Algorithm

- Parameter estimates are found using the expectation-maximization (EM) algorithm (Dempster et al., 1977).

- The EM algorithm is an iterative method for computing the maximum likelihood (ML) estimates when the data are incomplete.

- One source of missing data is the group labels $\mathbb{Z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$ where $\mathbf{z}_j = (z_{j1}, \ldots_{jG})$ with

$$z_{jg} = \begin{cases} 1, & \text{if } \boldsymbol{x}_j \text{ arises from group } g \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

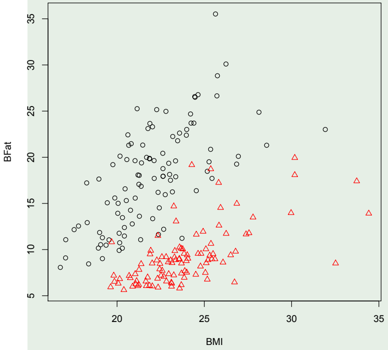
From a model based clustering standpoint, our interest lies mainly in the "fuzzy zeds", i.e. the estimates given by the posterior probability of observation x_j belonging to component g,

$$\hat{z}_{jg} = \frac{\pi_g f_g(x_j \mid \boldsymbol{\theta}_g)}{\sum_{g=1}^{G} \pi_g f_g(y_j \mid \boldsymbol{\theta}_g)}$$

A "hard zed" can be given by,

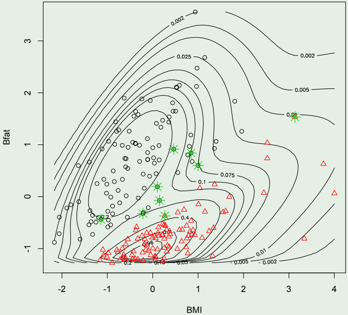
$$\widetilde{z}_{jg} = \mathrm{MAP}(\hat{z}_{jg}) = \begin{cases} 1, & \text{if } \hat{z}_{jg} > \hat{z}_{jk} \text{ for all } k \neq g \\ 0, & \text{otherwise.} \end{cases}$$

Example (Australian Institute of Sport (AIS) data)



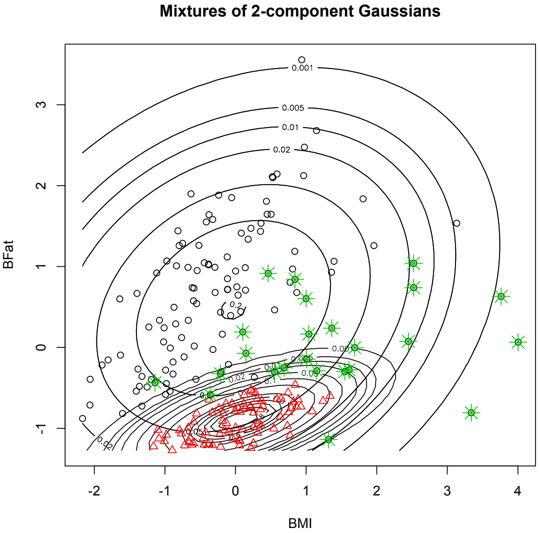
- Variables:
  body mass index (BMI) and body fat (Bfat)
- n=202 athletes
- genders are taken to be unknown
- ○ = female (100 obs),
  △= male (102 obs)

Example (Australian Institute of Sport (AIS) data)



- Contour plot of the two-component MST mixture model fitted to the scaled data
- MST fits well to asymmetry data
- Misclassifies 9 observations

The 2-component solution for a Gaussian FMM

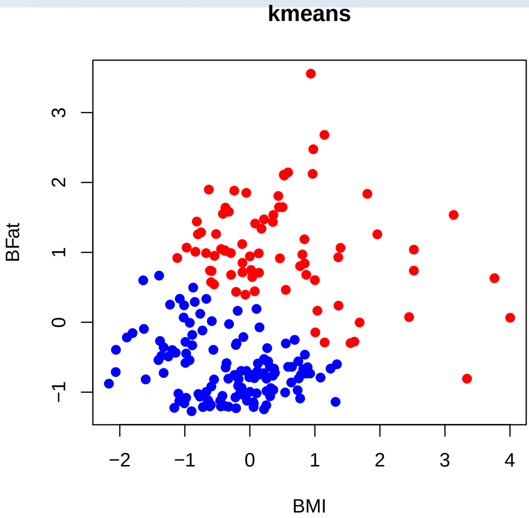


A note of FMM

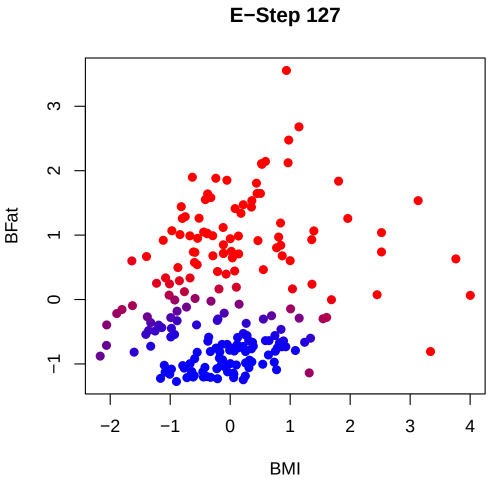
WARNING: this algorithm is susceptible to converge to local maximum. To help avoid local maxima one could use

- Multiple random starts
- Deterministic annealing EM (or DAEM) as proposed by Ueda and Nakano (1998)
- A conservative stopping criterion

A fuzzy 2-component Gaussian solution

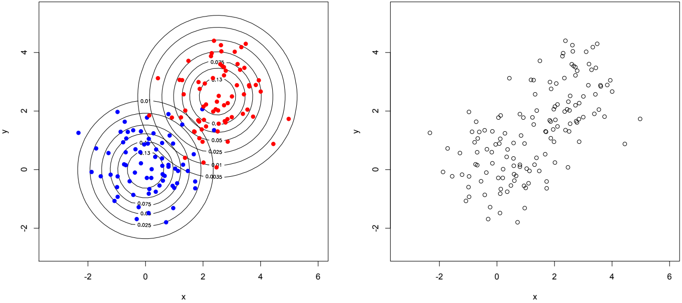


A Gaussian FMM movie

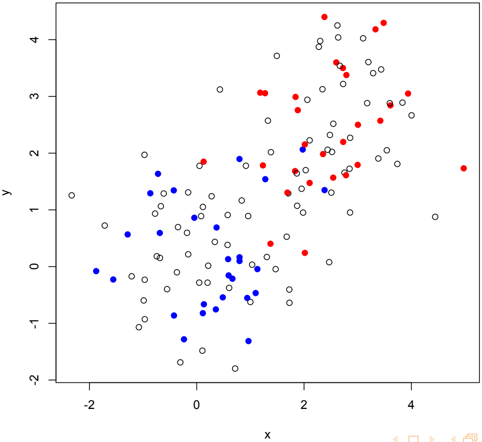


Unsupervised learning

So far all of our examples have assumed no information about the group labels (unsupervised learning).



In practice, it is quite possible that we would have the labels for some of the observations



Species of Classification

Model-based classification can be divided into 3 major designations:

- **Supervised Classification (discriminant analysis or DA)**: uses labelled data to build a classifier for labelling unlabelled observations.
- **Unsupervised Classification (Clustering)**: Aims at assigning labels to unlabelled data (no labelled data used).
- **Semi-supervised Classification (Classification)**: Uses both labelled and unlabelled data to infer the labels un unclassified data.

Species of Classification

- **Supervised Classification (discriminant analysis or DA)**: uses only labelled data $\mathcal{X}_1$, and their labels $\mathcal{Z}_1$, to infer the unknown labels ($\mathcal{Z}_2$) of the unlabelled observations $\mathcal{X}_2$.

- **Unsupervised Classification (Clustering)**: Aims at assigning labels ($\mathcal{Z}_2$) based solely on the feature data $\mathcal{X}_2$.

- **Semi-supervised Classification (Classification)**: Uses both labelled ($\mathcal{X}_1, \mathcal{Z}_1$) and unlabelled data ($\mathcal{X}_2$) to infer the group membership of the unclassified data ($\mathcal{Z}_2$).

The respective complete data log-likelihood for DA, classification and clustering is given by,

$$\begin{aligned}
{}_{c}\mathcal{L}_{DA}(\boldsymbol{\Theta}|\mathcal{X}_1,\mathcal{Z}_1) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{1j}\mid\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}} \\
{}_{c}\mathcal{L}_{Clsf}(\boldsymbol{\Theta}|\mathcal{X},\mathcal{Z}) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{1j}\mid\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}} \qquad \prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{2k}\mid\theta_g)]^{z_{kg}^{(2)}} \\
{}_{c}\mathcal{L}_{Clst}(\boldsymbol{\Theta}|\mathcal{X}_2,\mathcal{Z}_2) &= \qquad \prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{2k}\mid\theta_g)]^{z_{kg}^{(2)}} \\
{}_{c}\mathcal{L}_{FSC}(\boldsymbol{\Theta}|\mathcal{X}_2,\mathcal{Z}_2) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{1j}\mid\boldsymbol{\theta}_g)]^{\alpha_c z_{jg}^{(1)}} \qquad \prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{2k}\mid\theta_g)]^{(1-\alpha_c)z_{kg}^{(2)}}
\end{aligned} \tag{3}$$

The respective complete data log-likelihood for DA, classification and clustering is given by,

$$
\begin{aligned}
{}_{\mathrm{c}}\mathcal{L}_{DA}(\boldsymbol{\Theta}|\mathcal{X}_1,\mathcal{Z}_1) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{1j}\mid\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}} & & \\
{}_{\mathrm{c}}\mathcal{L}_{Clsf}(\boldsymbol{\Theta}|\mathcal{X},\mathcal{Z}) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{1j}\mid\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}} & & \prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{2k}\mid\theta_g)]^{z_{kg}^{(2)}} \\
{}_{\mathrm{c}}\mathcal{L}_{Clst}(\boldsymbol{\Theta}|\mathcal{X}_2,\mathcal{Z}_2) &= & & \prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{2k}\mid\theta_g)]^{z_{kg}^{(2)}} \\
{}_{\mathrm{c}}\mathcal{L}_{FSC}(\boldsymbol{\Theta}|\mathcal{X}_2,\mathcal{Z}_2) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{1j}\mid\boldsymbol{\theta}_g)]^{\alpha_c z_{jg}^{(1)}} & & \prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\mathbf{x}_{2k}\mid\theta_g)]^{(1-\alpha_c)z_{kg}^{(2)}}
\end{aligned}
\tag{3}
$$

This generalized the three approaches and offers any level of supervision between unsupervised and supervised

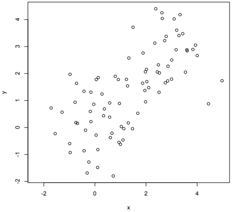
$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\Theta}|\mathcal{X},\mathcal{Z}) &= \prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\boldsymbol{x}_{1j}\mid\boldsymbol{\theta}_g)]^{\alpha_c z_{jg}^{(1)}}\prod_{k=1}^{n_2}\prod_{g=1}^{G}[\pi_g\phi(\boldsymbol{x}_{2k}\mid\theta_g)]^{(1-\alpha_c)z_{kg}^{(2)}}\\ &= \prod_{i=1}^{m}\prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g\phi(\boldsymbol{x}_{ij}\mid\boldsymbol{\theta}_g)]^{\alpha_i z_{jg}^{(1)}}\end{aligned}$$

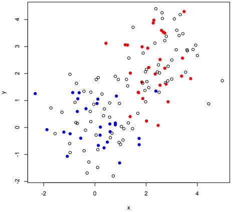
where $\alpha_1 = \alpha_c$ and $\alpha_2 = (1-\alpha_c)$ and

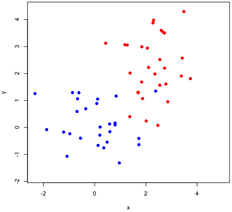
- $\alpha_c = 1/2$ corresponds to model-based classification
- $\alpha_c = 0$ corresponds to model-based clustering
- $\alpha_c = 1$ corresponds to DA.

Clst ClsF DA

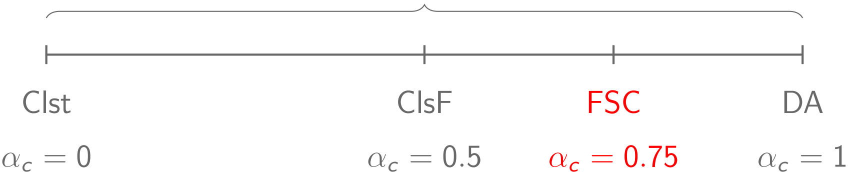
$\alpha_c = 0$ $\alpha_c = 0.5$ $\alpha_c = 1$







Fractionally Supervised Classification (FSC)



Clst

$\alpha_c = 0$

ClsF

$\alpha_c = 0.5$

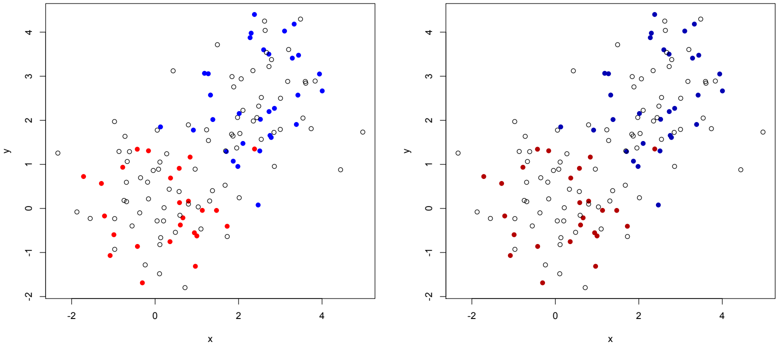
FSC

$\alpha_c = 0.75$

DA

$\alpha_c = 1$

FSC softens known classes



- This model is a special case of weighted likelihood (WL).
- WL used to combine information from multiple populations at varying degree.
- Herein, we use Gaussian mixture model-based approaches to illustrate our FSC approach, and we fix the weights.
- FSC is very flexible and can easily be extended to non-Gaussian mixtures and/or different weights.

Choice of weight

- Some guidance towards their construction have been presented in a number of articles [2] [3]
- In practice, there has been no definitive way of defining the weights of the WL.
- Hu has suggested that the weights should reflect the similarity between the pdf of the *m*th population and the target population.

- We first tested *static* weights

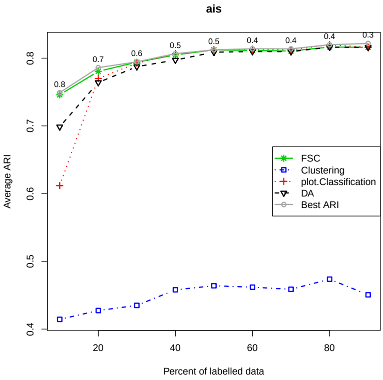
$$\alpha_c = \frac{\max\{n_1, n_2\}}{n_1 + n_2},$$

where n_1 is the number of labelled observations and n_2 is the number of unlabelled.

Application

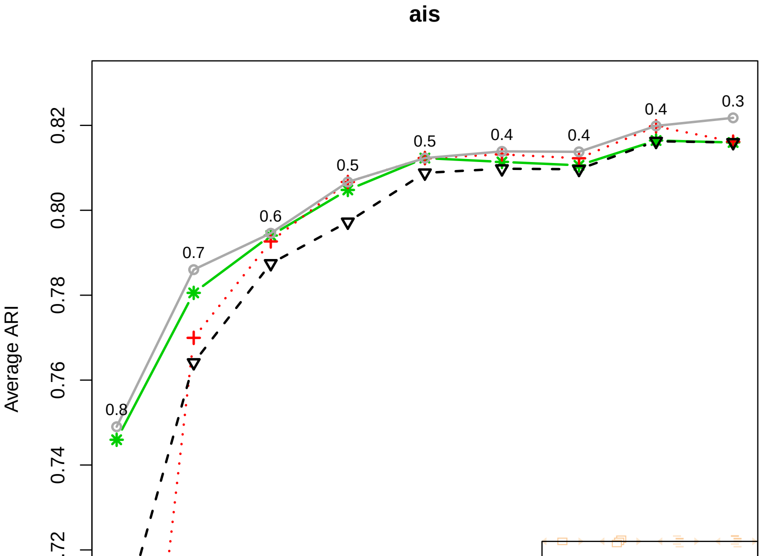
- We demonstrate how FSC compares with the the three species of classification when applied to the clustering data sets.
- For each data set, 100 random subsets are considered with varying proportions labelled ranging from 0.2, 0.3, . . . , 0.9 (800 runs in total).
- We initialize our algorithm at $\hat{\Theta}_{\text{DA}}$, where $\hat{\Theta}_{\text{DA}}$ is the fitted parameters produced by performing DA.
- The Adjusted Rand Index (ARI) is used to assess the clustering results
  - An ARI of 1 corresponds to perfect clustering
  - An ARI of 0 corresponds to no better than random clustering

AIS Data



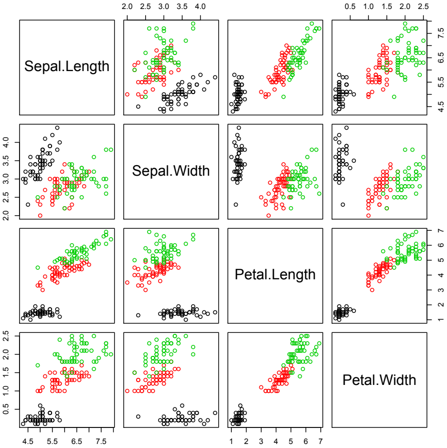
- Clustering produced inferior results compared to the other species.
- FSC produces results comparable to best possible ARI obtainable on average.
- With less than 20% labelled, DA performs better on average than model based classification.

AIS Datat

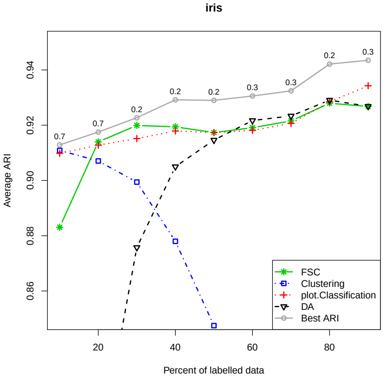


Iris Data Fisher (1936); Anderson (1935).

- Measure sepal length and width and petal length and width
- Consists of 150 observations
- Contain 3 species: Iris *setosa*, *versicolor*, and *virginica*


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
7.5 6.5 5.5 4.5
4.0 3.5 3.0 2.5 2.0
7 6 5 4 3 2 1
2.5 2.0 1.5 1.0 0.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7

Iris Data

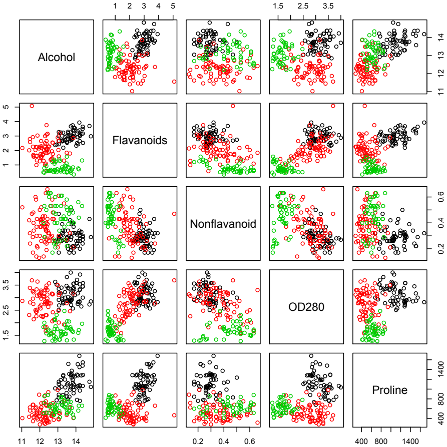


- Favours smaller values for α_c for $>$20% known.
- There is some room for improvement in terms of optimal ARI with static weights

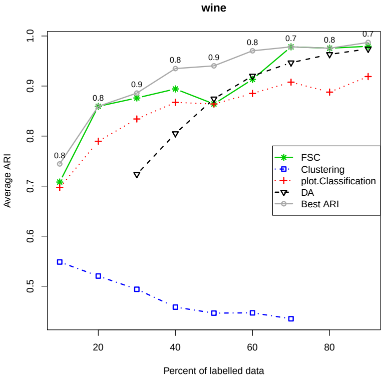
Wine Data (Forina et al., 1988)

- Measures 13 chemical and physical properties of wines
- Consists of 178 from 3 different grape cultivars
- This data set can be accessed through the gclus package (Hurley, 2012) for R (R Core Team, 2017)

Wine Data (Forina et al., 1988)



Wine Data



- Using FSC consistently over all analyses leeds to a smaller overall classification error than sticking with any one species.
- There is some room for improvement in terms of optimal ARI with static weights

Conclusion

Introduced parsimonious multivariate skew-t mixture models.

- The added skewness parameter accounts of asymmetry in the data
- Incorporating heavy tails to account of atypical observations via the robustness parameter ν
- Able to capture the Gaussian and multivariate t distribution as special cases

Future Work

Introduced FSC: a flexible paradigm that allows any level of supervision ranging from unsupervised and supervised.

- Investigate alternative choices for (α_1, α_2), e.g. *dynamic* weights that change at each iteration.
- Extend to $m > 2$.
- Investigate mixtures of MSN and MST in the FSC framework.

Acknowledgements

This work was supported by a Queen Elizabeth II Graduate Scholarship in Science and Technology (Vrbik), Ontario Graduate Scholarship (Vrbik), and by an Early Researcher Award from the Ontario Ministry of Research and Innovation (McNicholas).

Reference I

[1] Jayani Withanawasam. *Apache Mahout Essentials*. Packt Publishing Ltd, 2015.

[2] Feifang Hu and James V Zidek. The relevance weighted likelihood with applications. In *Empirical Bayes and Likelihood Inference*, pages 211–235. Springer, 2001.

[3] Xiaogang Wang, James V Zidek, et al. Selecting likelihood weights by cross-validation. *The Annals of Statistics*, 33(2):463–500, 2005.

Questions ?

Irene Vrbik
vrbiki@gmail.com