STAT406- Methods of Statistical Learning Lecture 20

Matias Salibian-Barrera

UBC - Sep / Dec 2018

K-means / K-medoids

Note that in K-means

- We used $d^2\left(\mathbf{X}_i, \mathbf{X}_j\right) = \left\|\mathbf{X}_i - \mathbf{X}_j\right\|^2$
- The cluster "centers" may not be actual observations
- Need to manipulate the "features" ($\mathbf{X}_i$)
- Can we use different distance measures?
- Can we work with the dissimilarities only?

K-means / K-medoids

A slightly different algorithm is

- Given $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$, for each cluster $\mathcal{C}_r$ find

$$j_r^* = \arg\min_{i \in \mathcal{C}_r} \sum_{j \in \mathcal{C}_r} d\left(\mathbf{X}_i, \mathbf{X}_j\right)$$

  and let $m_r = \mathbf{X}_{j_r^*}$
- Given $m_1, m_2, \ldots, m_K$, assign $\mathbf{X}_i$ to the cluster $\mathcal{C}_j$ with closest centre:

$$\mathbf{X}_i \leftarrow \arg\min_{1 \leq j \leq K} d\left(\mathbf{X}_i, m_j\right)$$

K-means / K-medoids

1. Find K initial cluster centres
2. Given centres m_ℓ, assign points to the cluster $\mathcal{C}_j$ with closest centre:

$$\mathbf{X}_i \leftarrow \arg\min_{1 \leq j \leq K} d\left(\mathbf{X}_i, m_j\right)$$

3. Explore all possible swaps between centres and non-centres.
4. If there's improvement, go to step 2

K-means / K-medoids

Note that now

- We can use any distance – robustness?
- The cluster representatives / prototypes are actual observations
- We do not need the observations, only the dissimilarities

K-means / K-medoids

Beers - 9 beers with 26 attributes

```
> a <- read.table('breweries.dat', header=FALSE)
> a <- t(a)
> a.dis <- dist(a, method='manhattan')
>
> brew.pam <- pam(a.dis, k=3)
>
> brew.pam
Medoids:
     ID
[1,] "7" "V7"
[2,] "2" "V2"
[3,] "6" "V6"
Clustering vector:
V1 V2 V3 V4 V5 V6 V7 V8 V9
 1  2  3  1  2  3  1  3  2
```

Silhouette plot

- For each unit $\mathbf{X}_i \in \mathcal{C}_\ell$

$$a_i = \frac{1}{n_\ell} \sum_{\mathbf{X}_j \in \mathcal{C}_\ell} d\left(\mathbf{X}_i, \mathbf{X}_j\right)$$

- Dissimilarity with other clusters

$$d\left(i, \mathcal{C}_r\right) = \frac{1}{n_r} \sum_{\mathbf{X}_j \in \mathcal{C}_r} d\left(\mathbf{X}_i, \mathbf{X}_j\right)$$

Silhouette plot

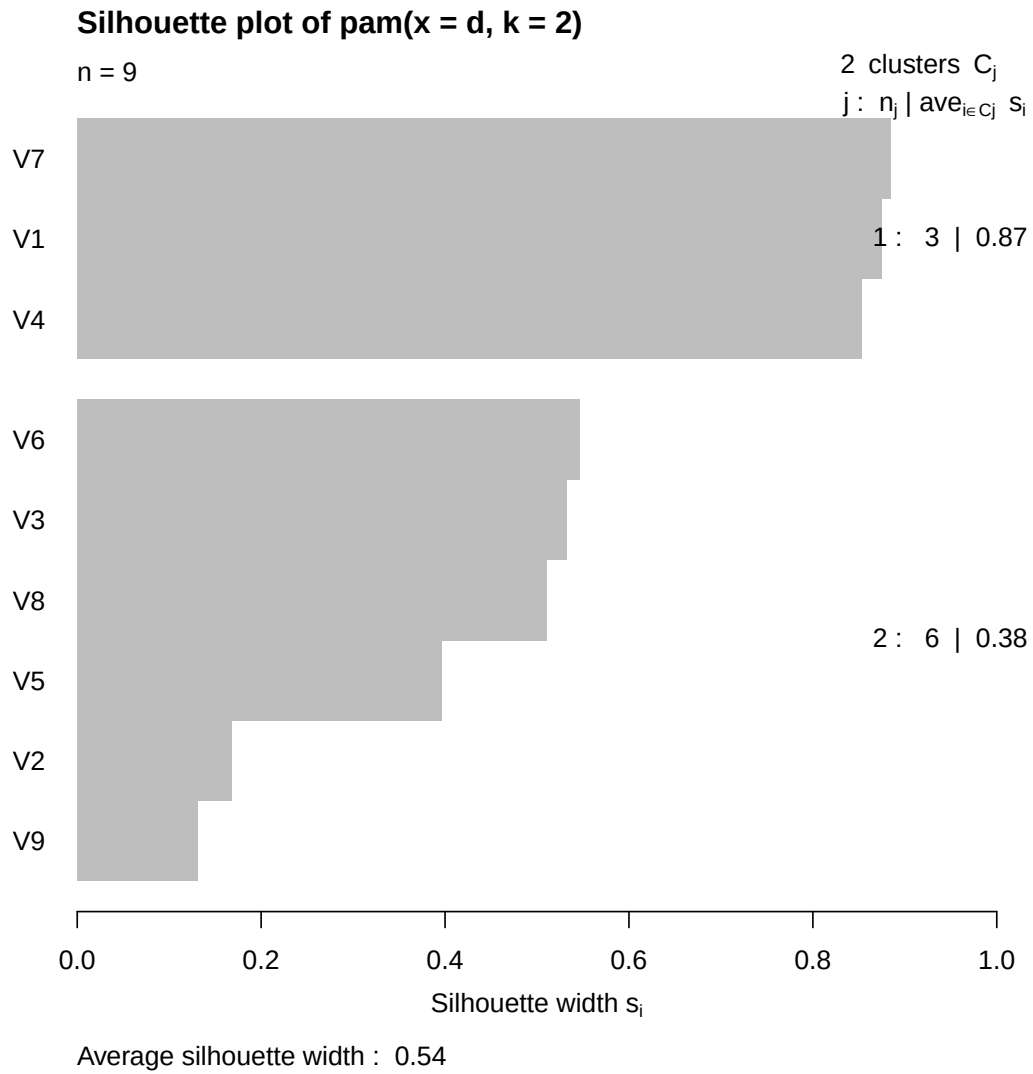
- Then, dissimilarity to closest cluster

$$b_i = \min_{r \neq \ell} d(i, \mathcal{C}_r)$$

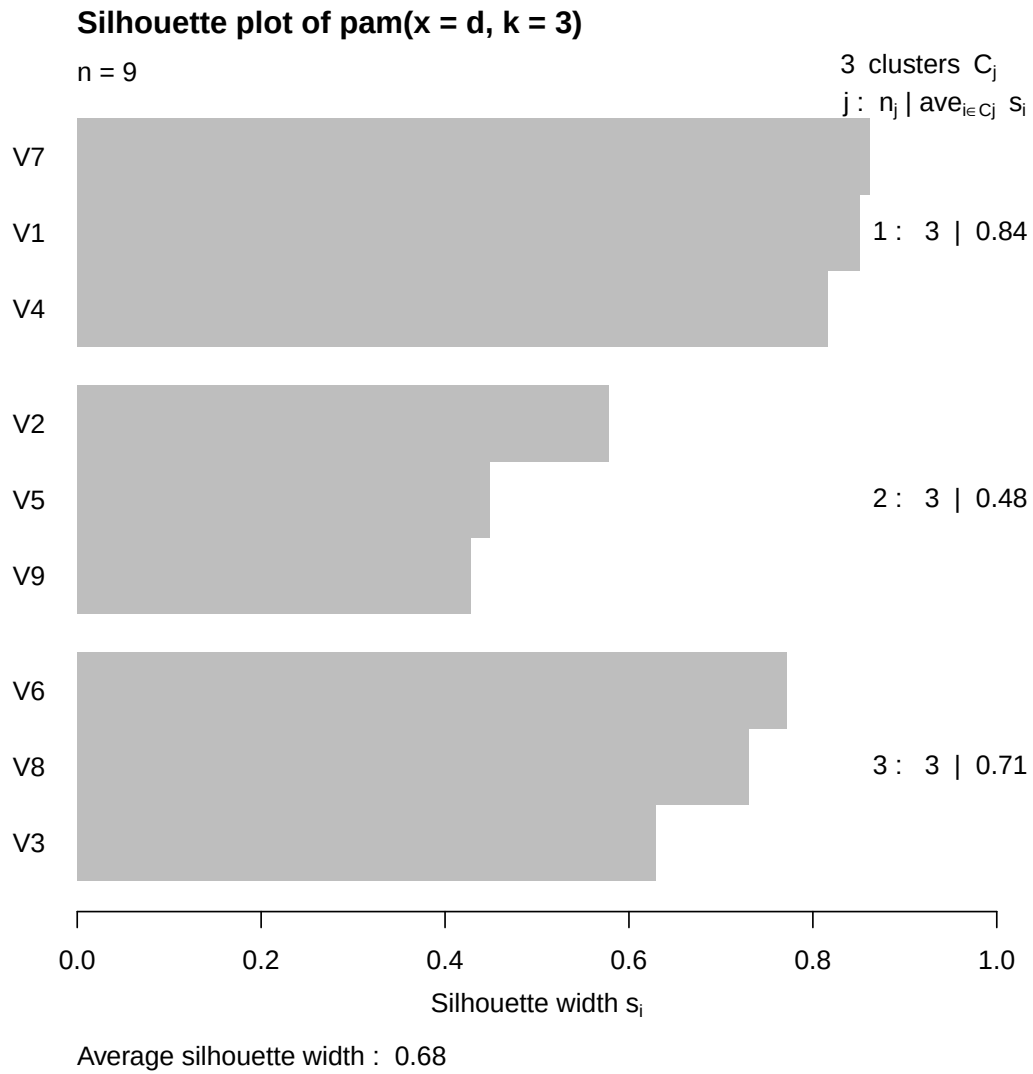
- Silhouette

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

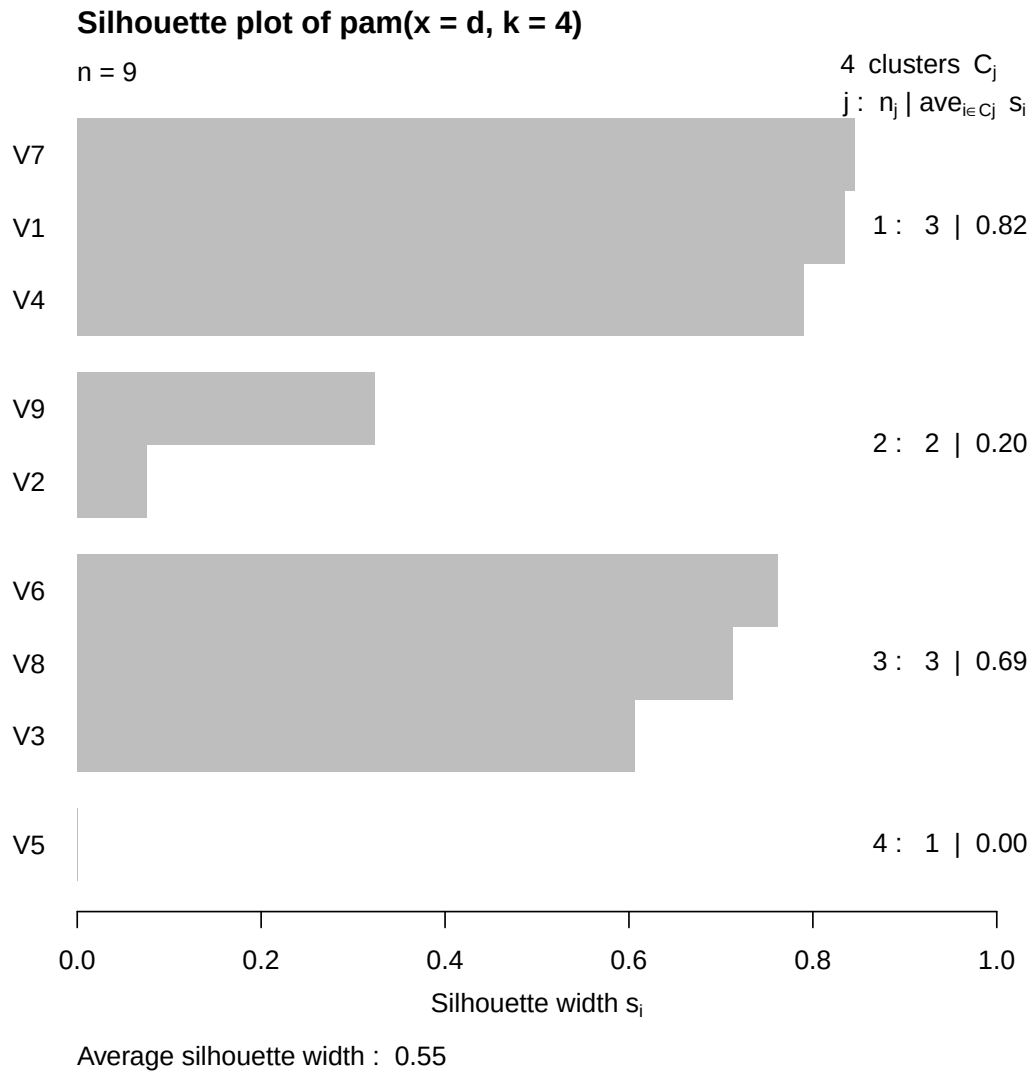
Breweries - K=2



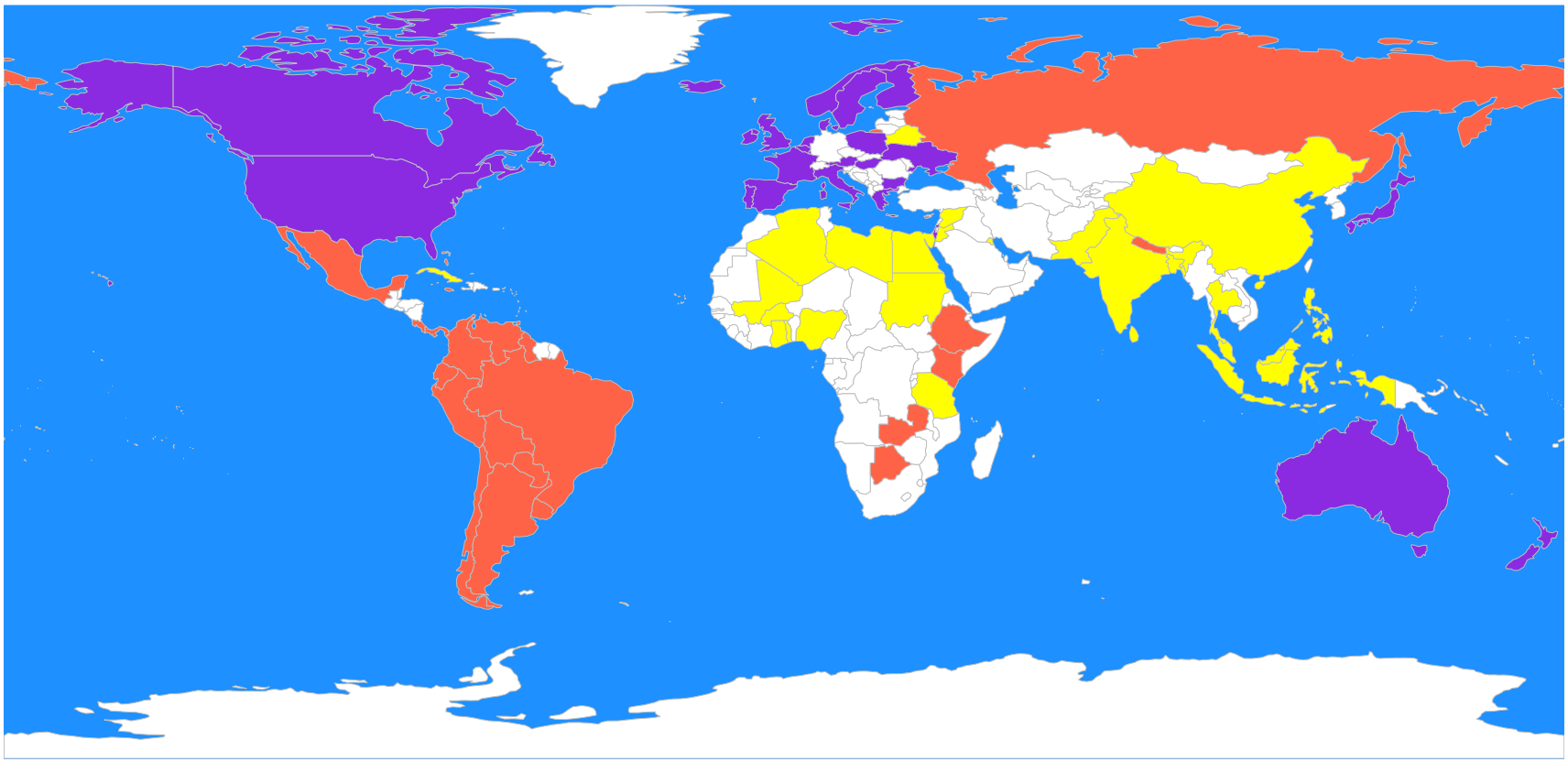
Breweries - K=3



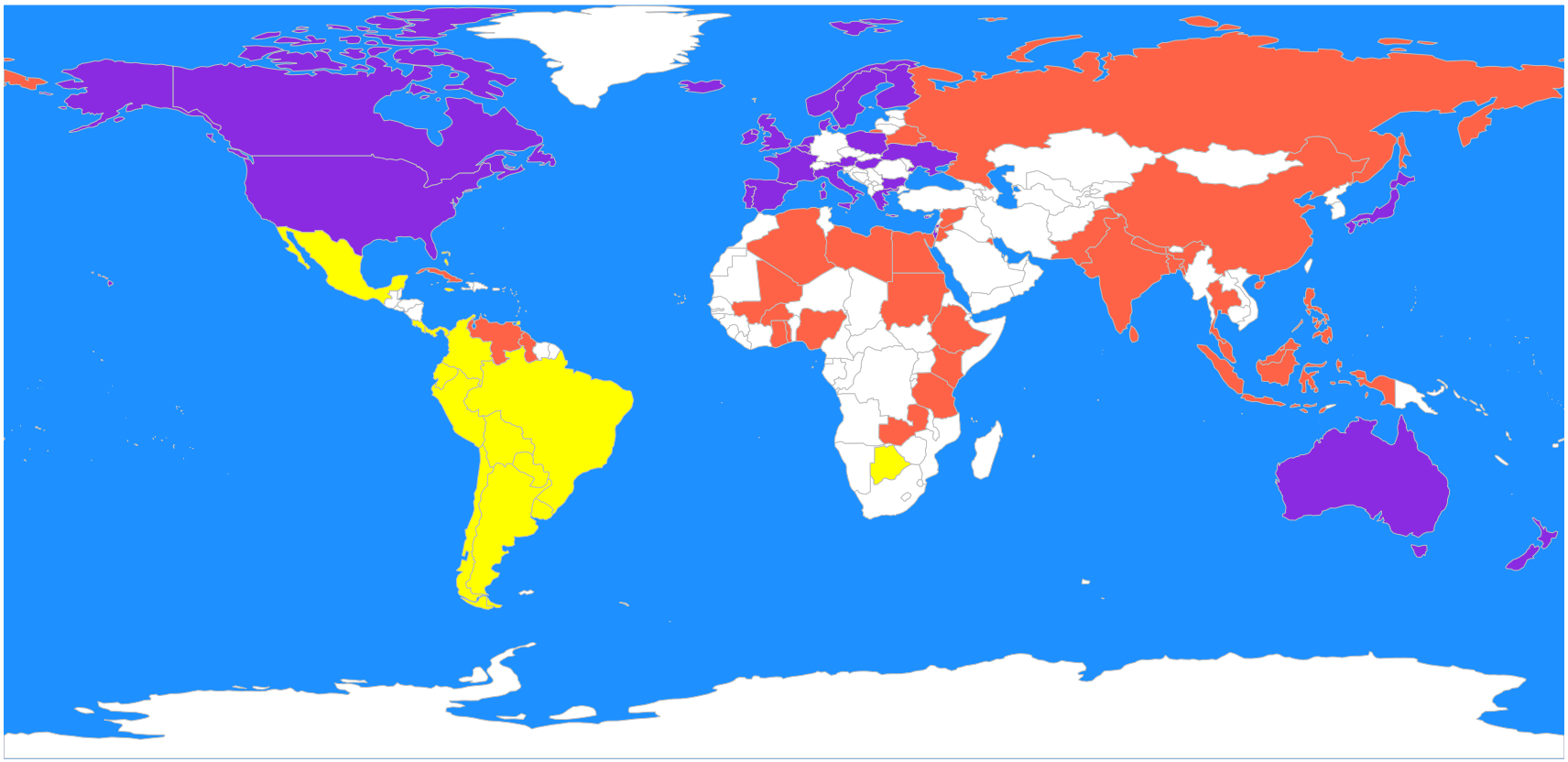
Breweries - K=4



UN Votes PAM - K=3



UN Votes Kmeans - K=3



Mixture models

- **X** denotes the vector of features
- We assume that there are underlying groups / classes
- **g** will denote the class label
- We may consider a model for the distribution of **X** in each class
- i.e. the dist'n of **X** conditional on **g**

Model based clustering

- Model-based clustering - `MCLUST`
- Assume that the random vector **X** and the class label **g** satisfy

$$\mathbf{X} \Big| \mathbf{g} = k \ \sim \ f_k(\theta_k)$$

then

$$f(\mathbf{x}) \ = \ \sum_{k=1}^{K} \sum_{k=1}^{K} \pi_k f_k(\mathbf{X}; \theta_k)$$

where $\pi_\ell = P(\mathbf{g} = \ell)$, $\sum_k \pi_k = 1$

Mixture models

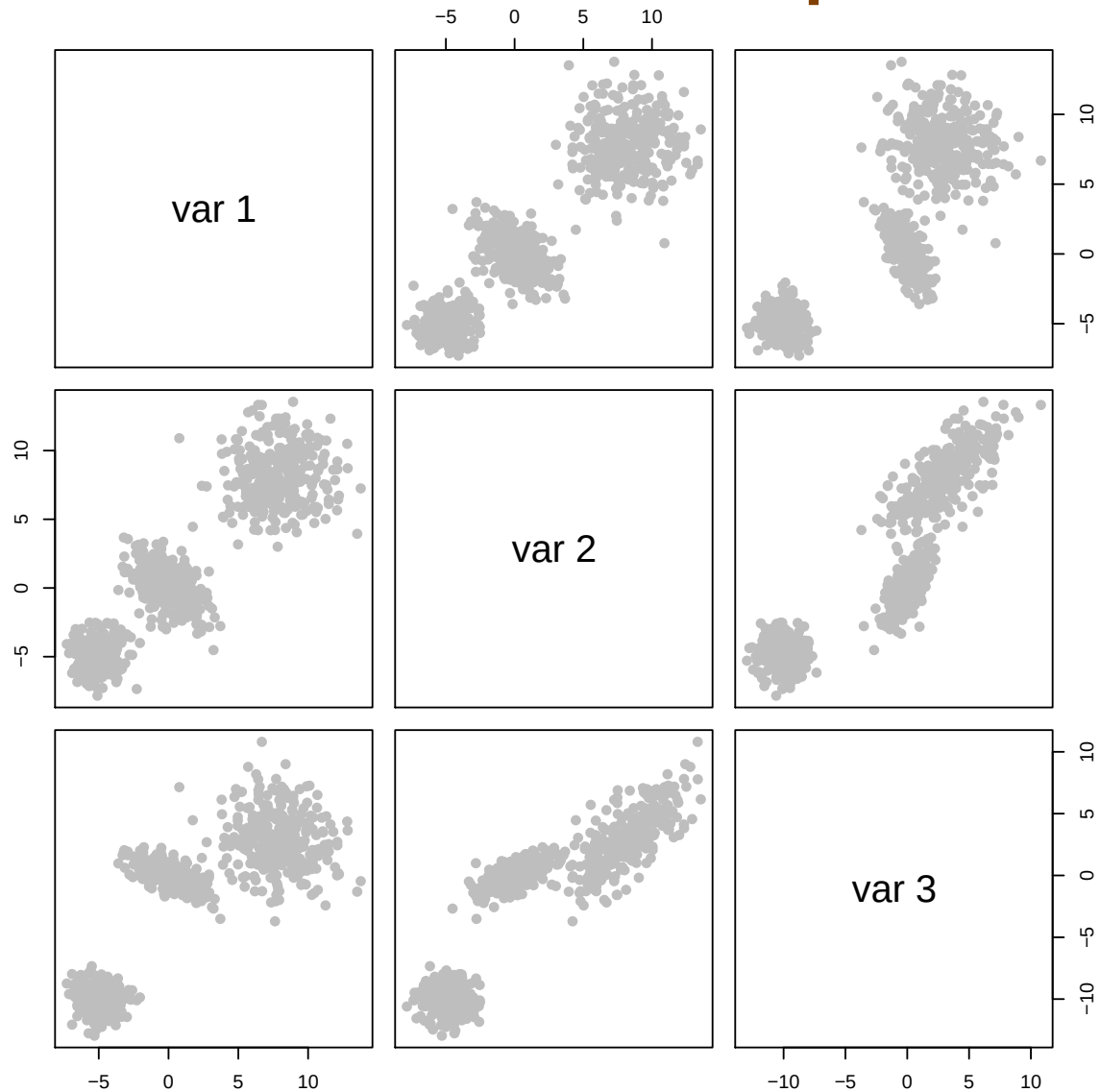
$$\mathbf{X}\Big|_{\mathbf{g}=k} \sim \mathcal{N}_3(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad k = 1, 2, 3$$

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{X};\, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 f_2(\mathbf{X};\, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \pi_3 f_3(\mathbf{X};\, \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

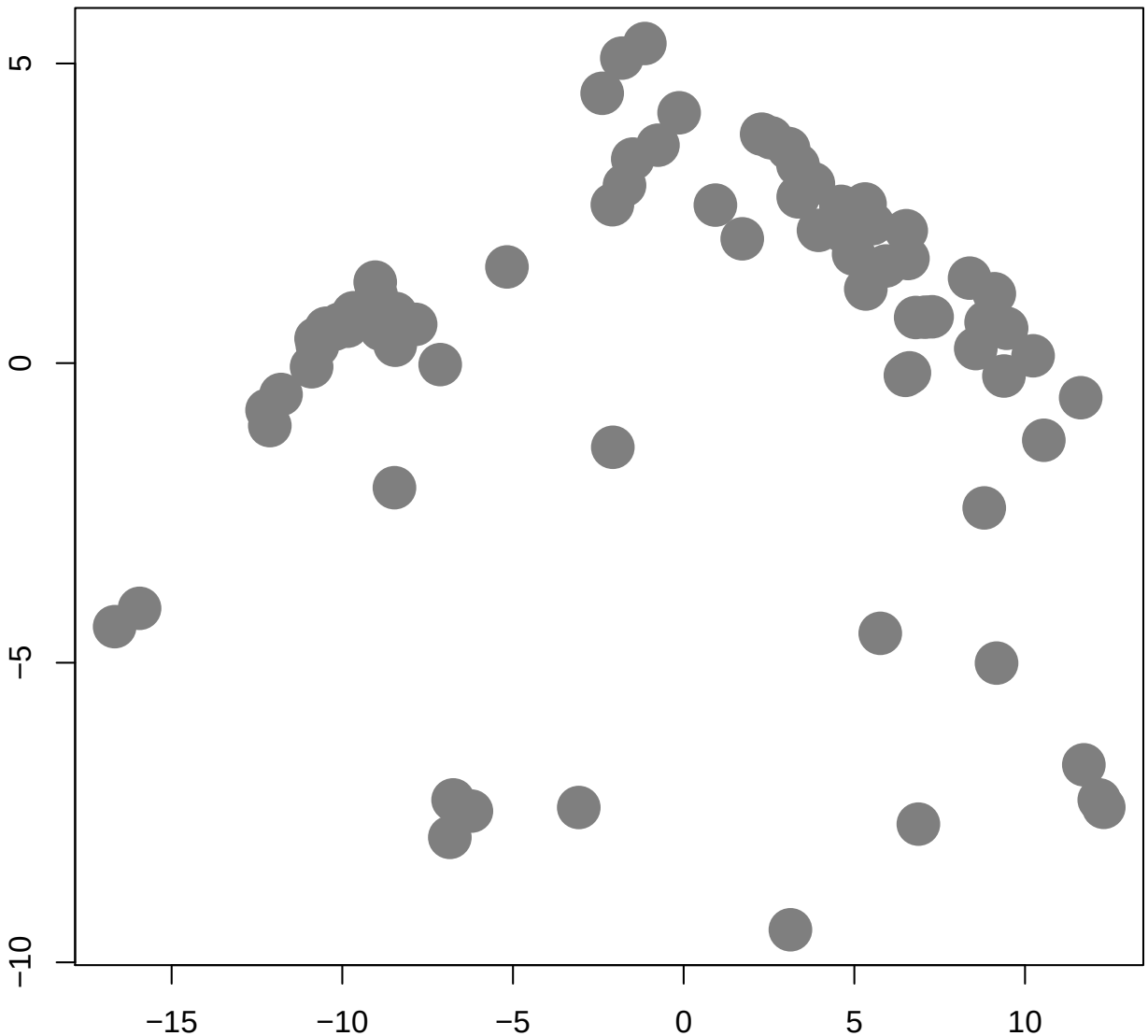
where

$$\pi_1 + \pi_2 + \pi_3 = 1$$

Normal mixture - Simple example



MDS UN Votes



Model based clustering

A two-step (hierarchical) model

If we observed the class lables $\mathbf{g}_1, \ldots, \mathbf{g}_n$ we'd have

$$f(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{g}_1, \ldots, \mathbf{g}_n; \theta, \pi) = \prod_{j=1}^{K} \prod_{\mathbf{g}_i = j} f_j(\mathbf{X}_i; \theta_j) \, \pi_j$$

Model based clustering

$$\text{log-lik} = \sum_{j=1}^{K} \sum_{\mathbf{g}_i=j} \log\left(f_j\left(\mathbf{X}_i;\boldsymbol{\theta}_j\right)\right) + \log\left(\pi_j\right)$$

$$= \sum_{j=1}^{K} \sum_{i=1}^{n} \delta_{j,i}\left[\log\left(f_j\left(\mathbf{X}_i;\boldsymbol{\theta}_j\right)\right) + \log\left(\pi_j\right)\right]$$

$\delta_{j,i} = 1$ if $\mathbf{g}_i = j$ (if $\mathbf{X}_i$ comes from the j-th population), $\delta_{j,i} = 0$ otherwise.

Model based clustering

If the $\delta_{j,i}$'s were available:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_j &= \arg\max_{\boldsymbol{\theta}_j} \sum_{i=1}^{n} \delta_{j,i} \left[\log\left(f_j\left(\mathbf{X}_i;\boldsymbol{\theta}_j\right)\right) + \log\left(\pi_j\right)\right] \\ &= \sum_{\mathbf{g}_i=j} \log\left(f_j\left(\mathbf{X}_i;\boldsymbol{\theta}_j\right)\right) + \log\left(\pi_j\right)\end{aligned}$$

the MLE for the j-th population, and

$$\hat{\pi}_j = \sum_{i=1}^{n} \delta_{j,i}/n$$

Model based clustering

However, the $\delta_{j,i}$'s are not observed.

$$l(\mathbf{X}_1, \ldots, \mathbf{X}_n; \boldsymbol{\delta}; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) = \sum_{j=1}^{K} \sum_{i=1}^{n} \delta_{j,i} \left[\log \left(f_j \left(\mathbf{X}_i; \boldsymbol{\theta}_j \right) \right) + \log \left(\pi_j \right) \right]$$

Model based clustering

For a given set of $\theta_1^{(r)}, \ldots, \theta_k^{(r)}$ and $\hat{\pi}_1^{(r)}, \ldots, \hat{\pi}_k^{(r)}$ we find

$$E\left[l(\mathbf{X}_1, \ldots, \mathbf{X}_n; \delta; \theta_1, \ldots, \theta_k) | \mathbf{X}; \theta^{(r)}\right]$$

$$= \sum_{j=1}^{K} \sum_{i=1}^{n} \gamma_{j,i}^{(r)} \left[\log\left(f_j\left(\mathbf{X}_i; \theta_j^{(r)}\right)\right) + \log(\hat{\pi}_j^{(r)})\right]$$

Model based clustering

$$
\begin{aligned}
\gamma_{j,i}^{(r)} &= P\left(\delta_{j,i}=1 \middle| \mathbf{X}_i;\boldsymbol{\theta}^{(r)}\right) \\
&= \frac{P\left(\delta_{j,i}=1,\mathbf{X}_i;\boldsymbol{\theta}^{(r)}\right)}{\sum_{l=1}^{K}\hat{\pi}_l^{(r)} f_l\left(\mathbf{X}_i;\boldsymbol{\theta}_l^{(r)}\right)} \\
&= \frac{f_j\left(\mathbf{X}_i;\boldsymbol{\theta}^{(r)}\right)\hat{\pi}_j^{(r)}}{\sum_{l=1}^{K}\hat{\pi}_l^{(r)} f_l\left(\mathbf{X}_i;\boldsymbol{\theta}_l^{(r)}\right)}
\end{aligned}
$$

Model based clustering

Now,

$$\boldsymbol{\theta}^{(r+1)}; \hat{\pi}^{(r+1)} \leftarrow$$

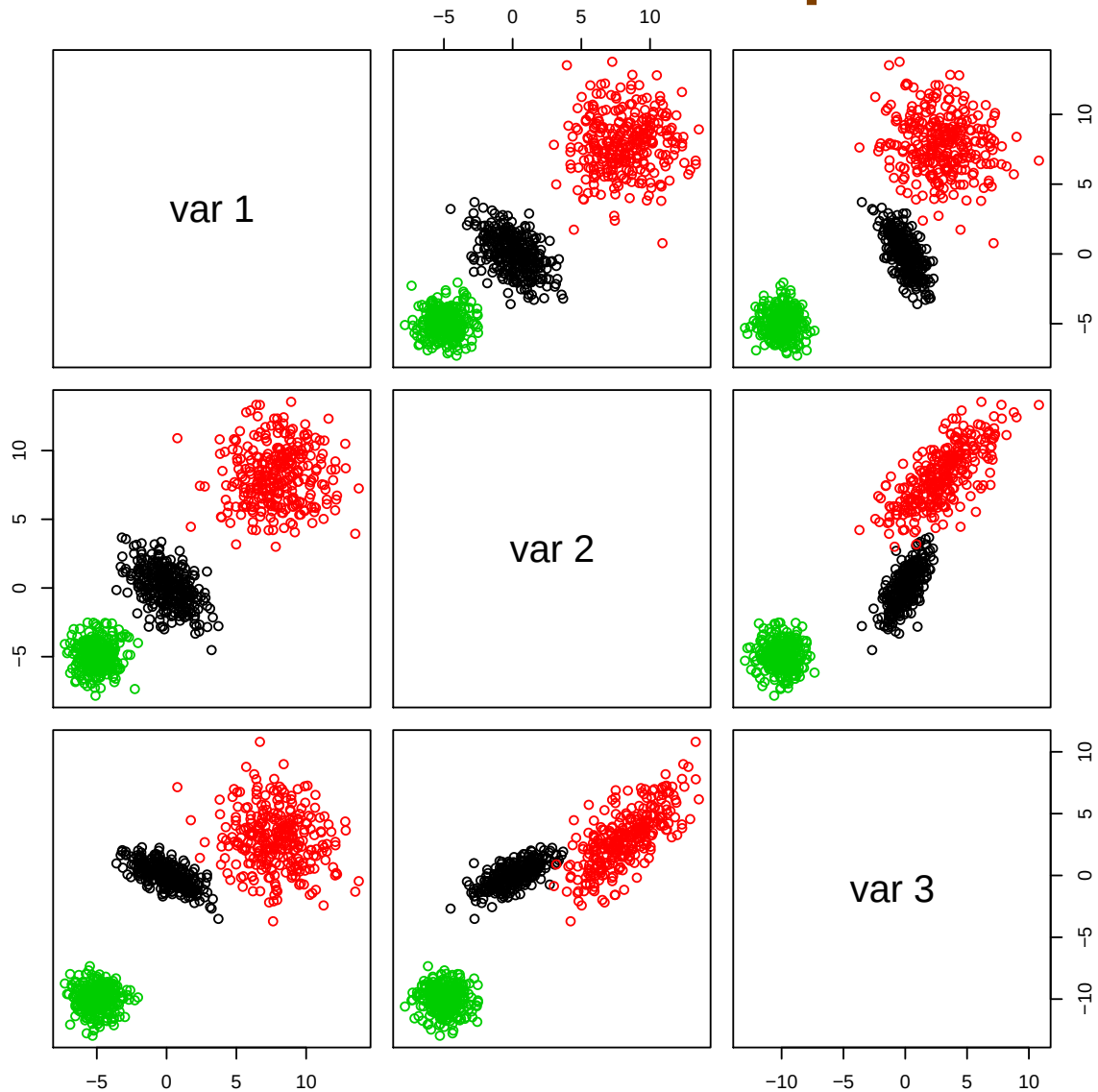
$$\arg\max_{\boldsymbol{\theta}\,\boldsymbol{\pi}} \sum_{j=1}^{K}\sum_{i=1}^{n} \gamma_{j,i}^{(r)} \left[\log\left(f_j\left(\mathbf{X}_i;\boldsymbol{\theta}_j\right)\right) + \log\left(\pi_j\right)\right]$$

- This is the EM algorithm

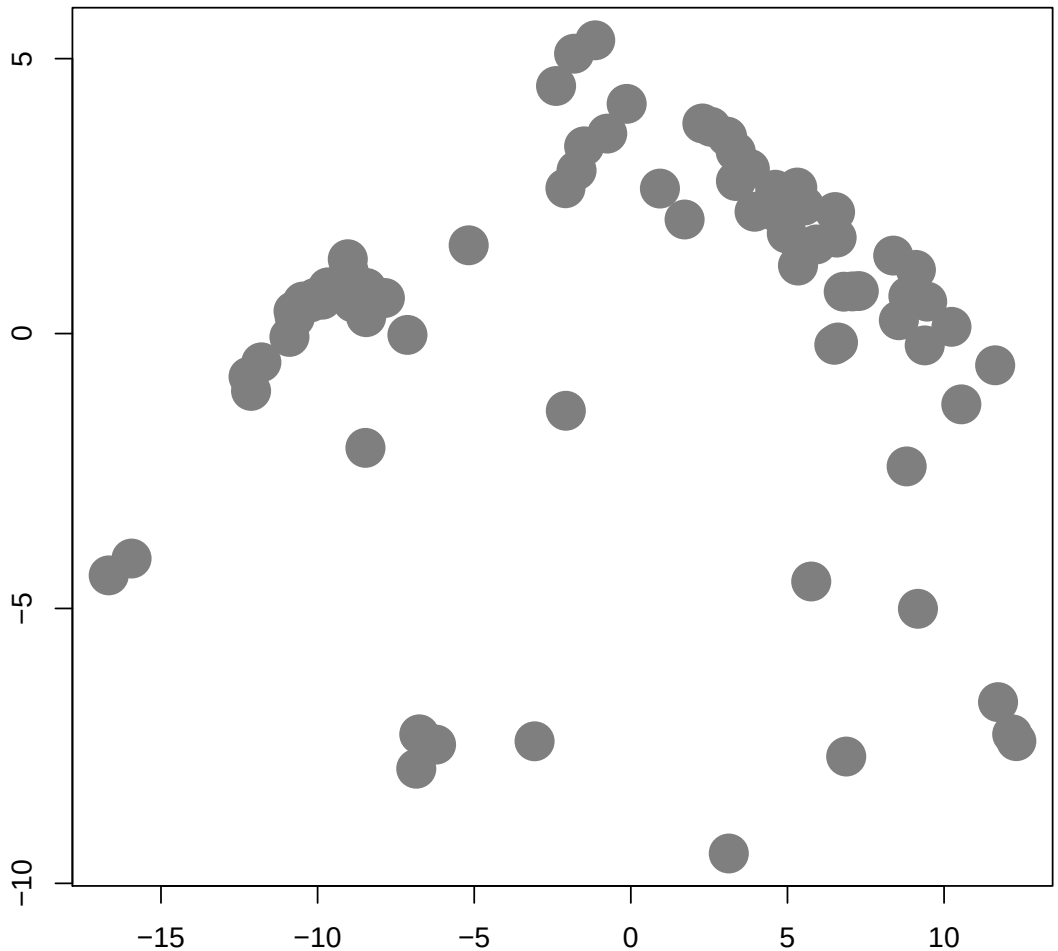
Model based clustering

- It can be shown that the EM algorithm does not decrease the likelihood
- It does not always work well
- The likelihood function for normal mixtures is unbounded
- The EM algorithm only finds local extrema
- Needs to be started from a good initial point, or re-started several times

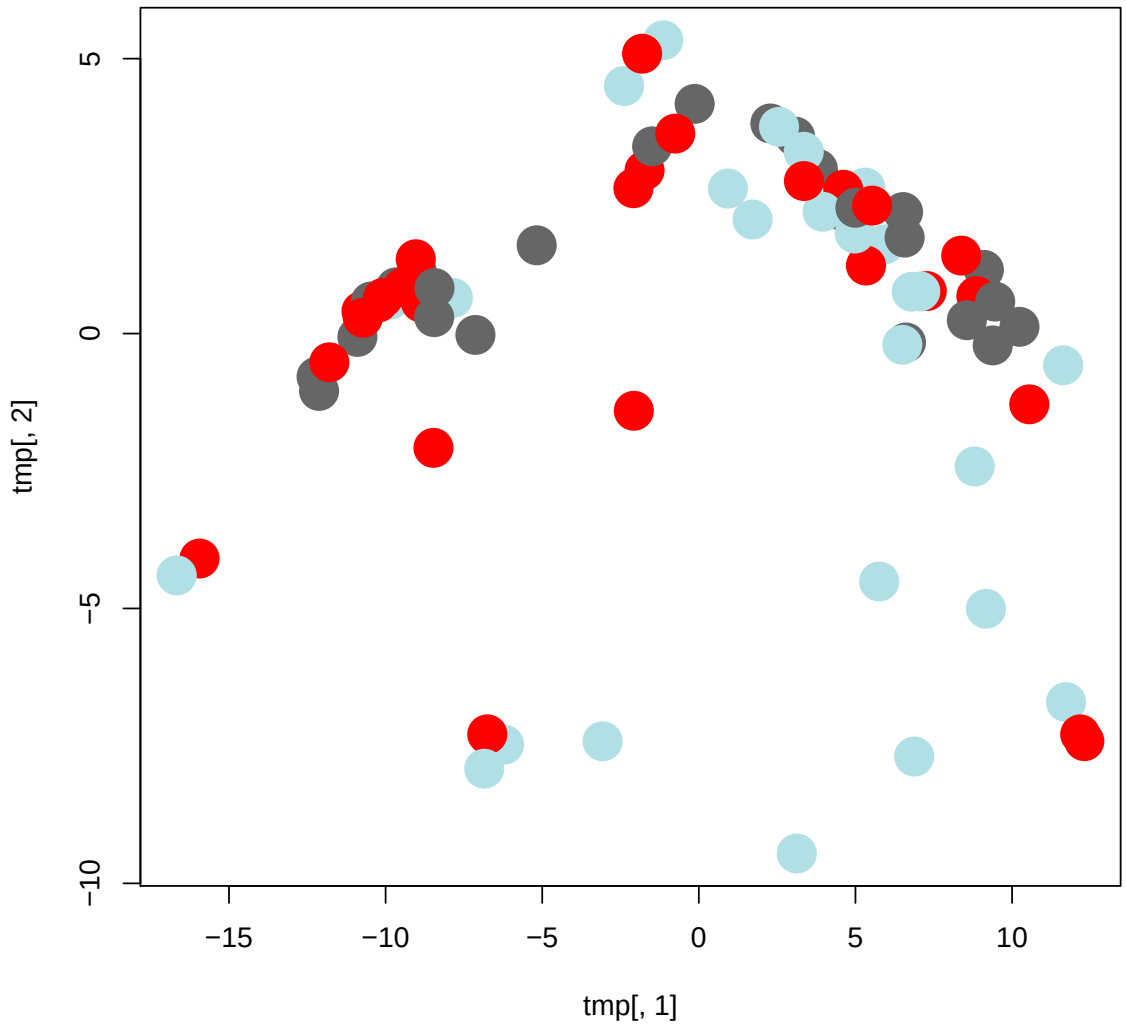
Normal mixture - Simple example



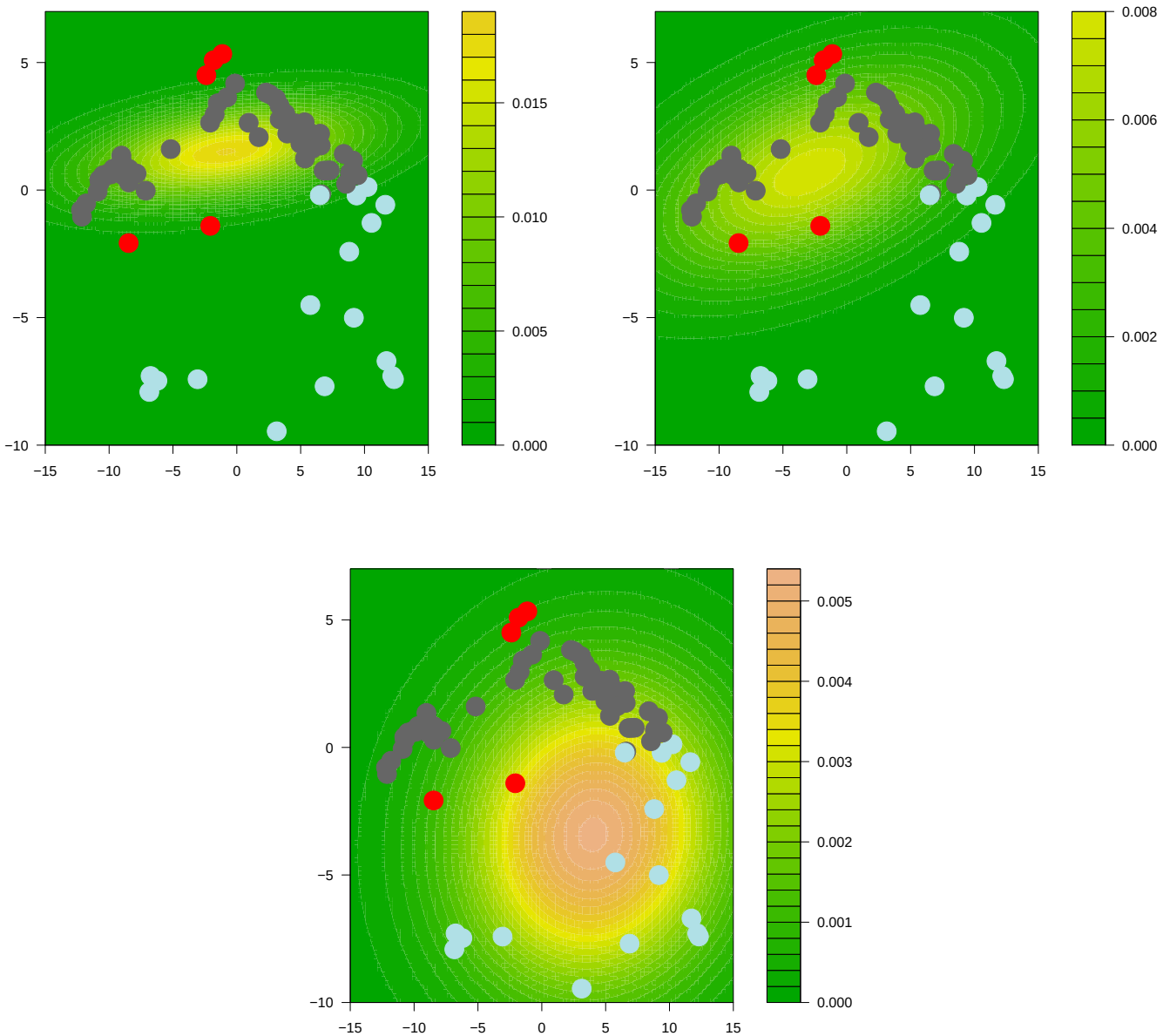
MDS UN Votes



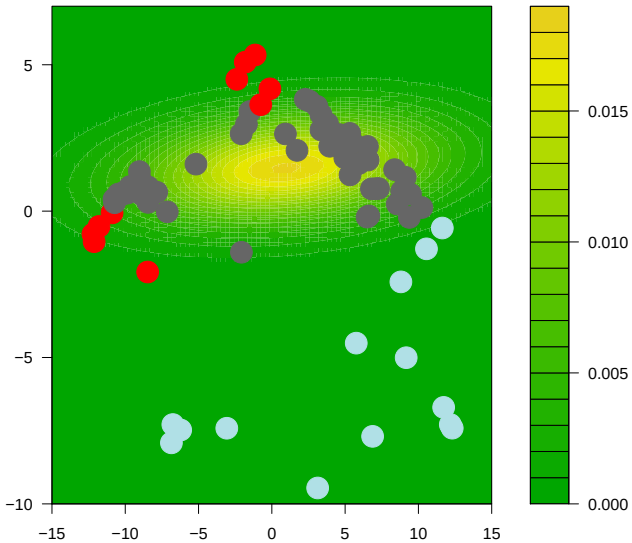
MDS UN Votes - Initial

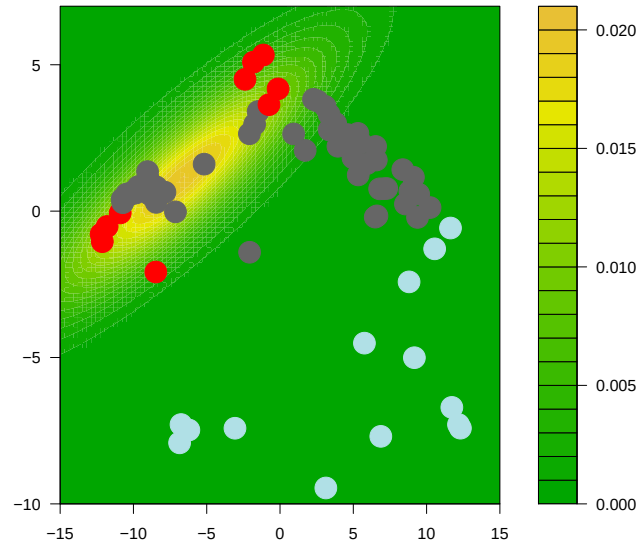


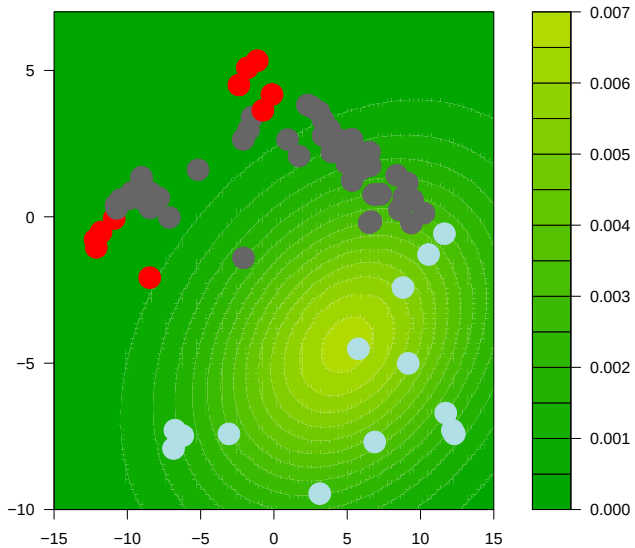
MDS UN Votes - Iter: 5



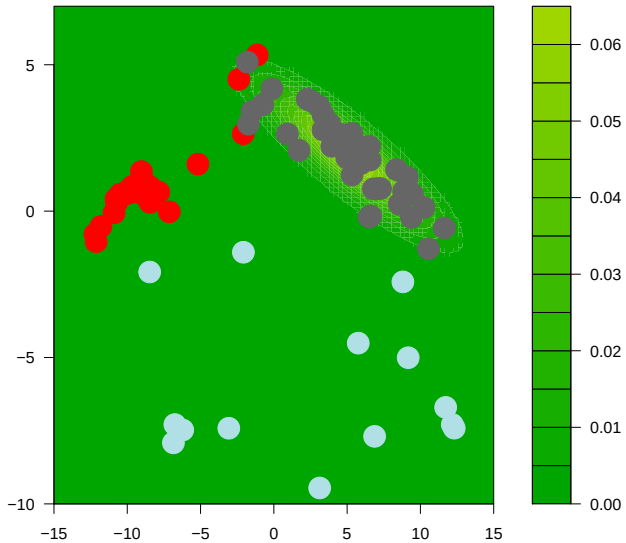
MDS UN Votes - Iter: 10

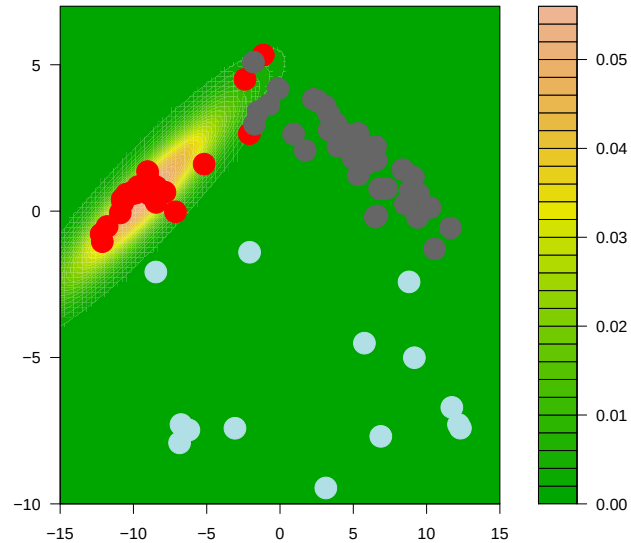


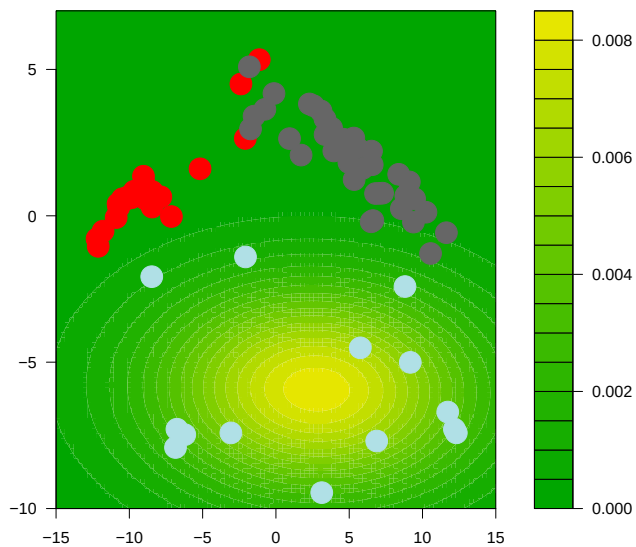




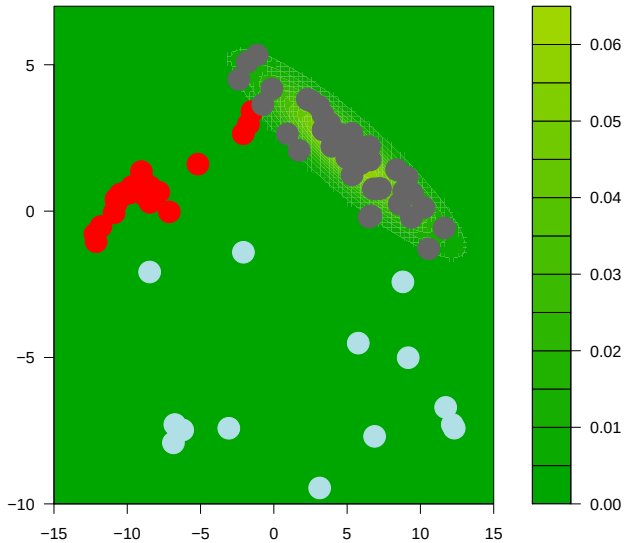
MDS UN Votes - Iter: 15

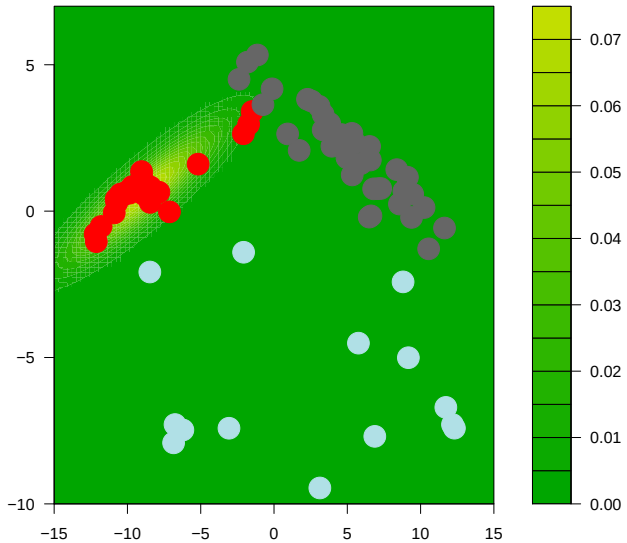


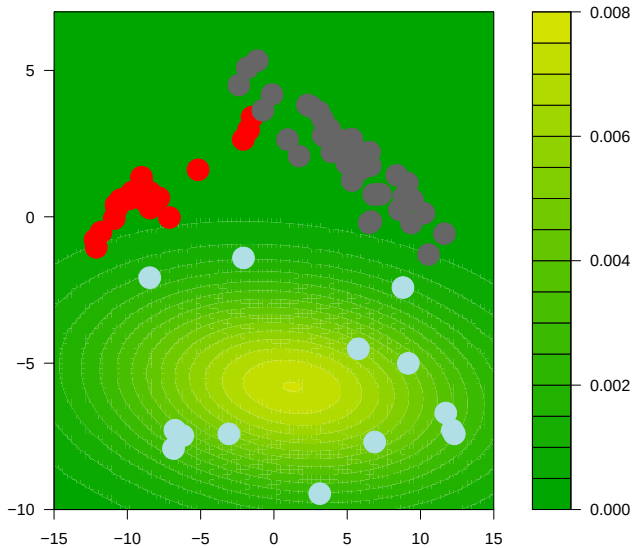




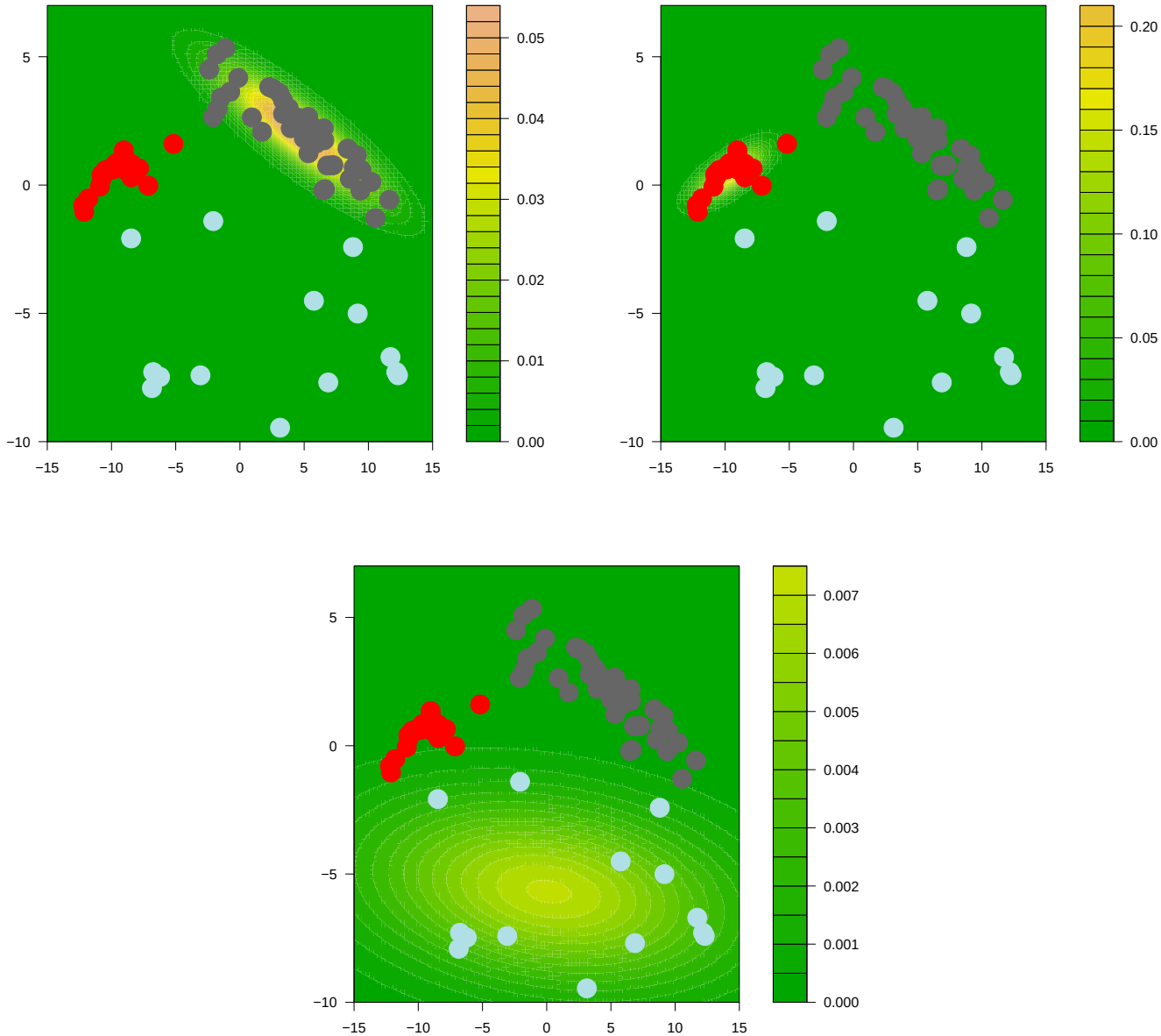
MDS UN Votes - Iter: 20







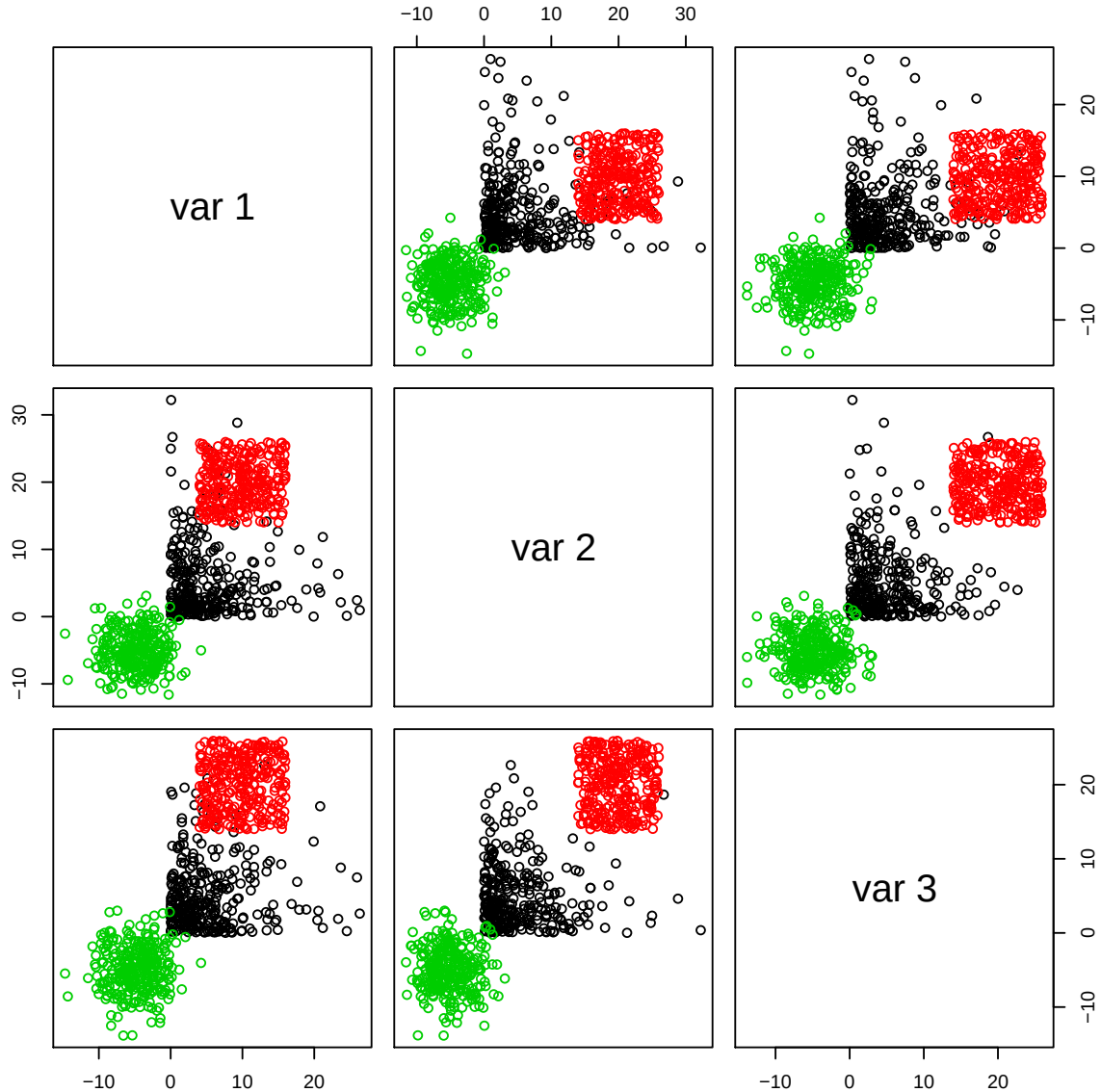
MDS UN Votes - Iter: 120



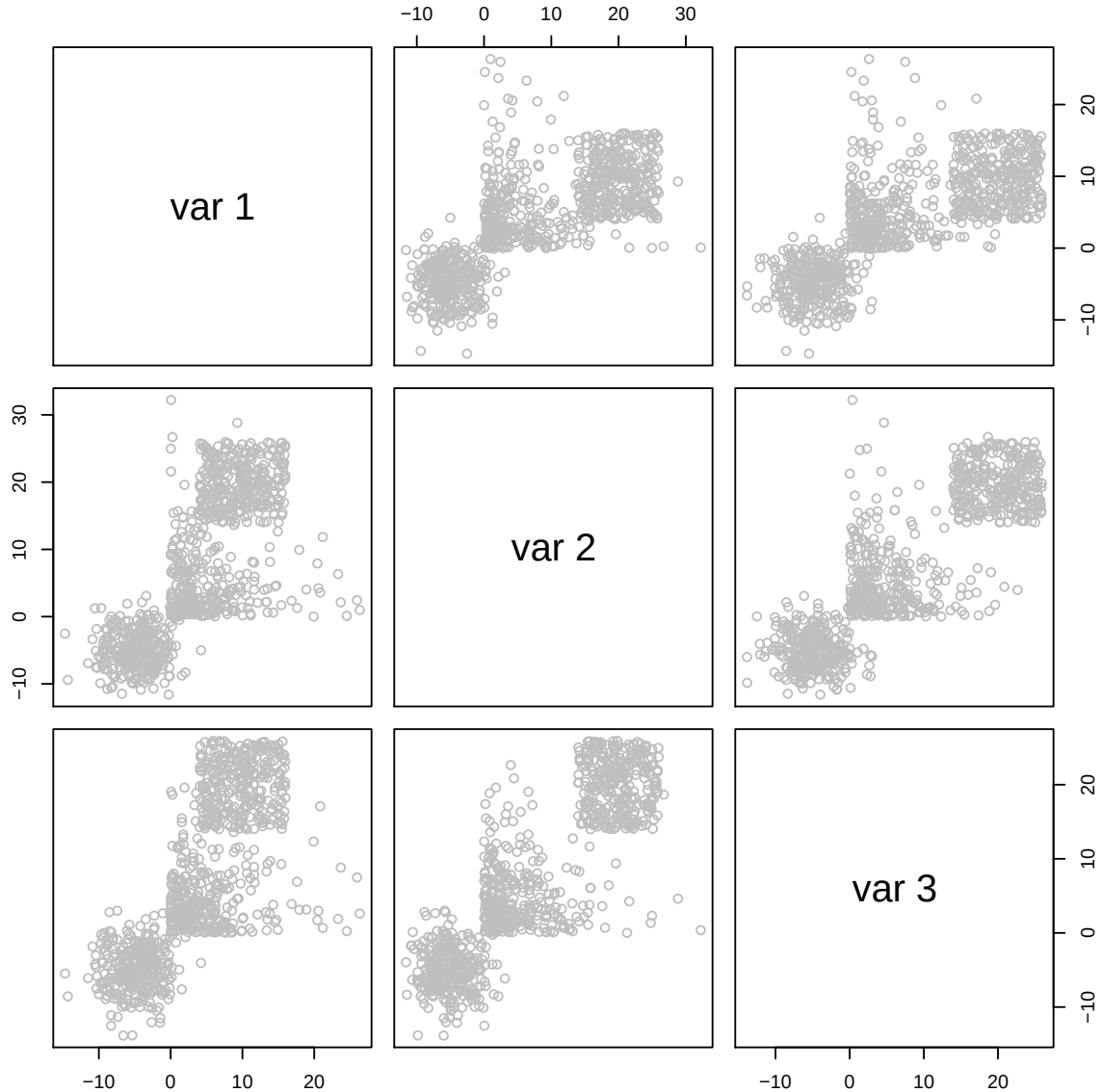
Choosing K

- Having a model, we can use likelihood-based measures to select K
- AIC or BIC, for example
- This works well as long as the model is appropriate

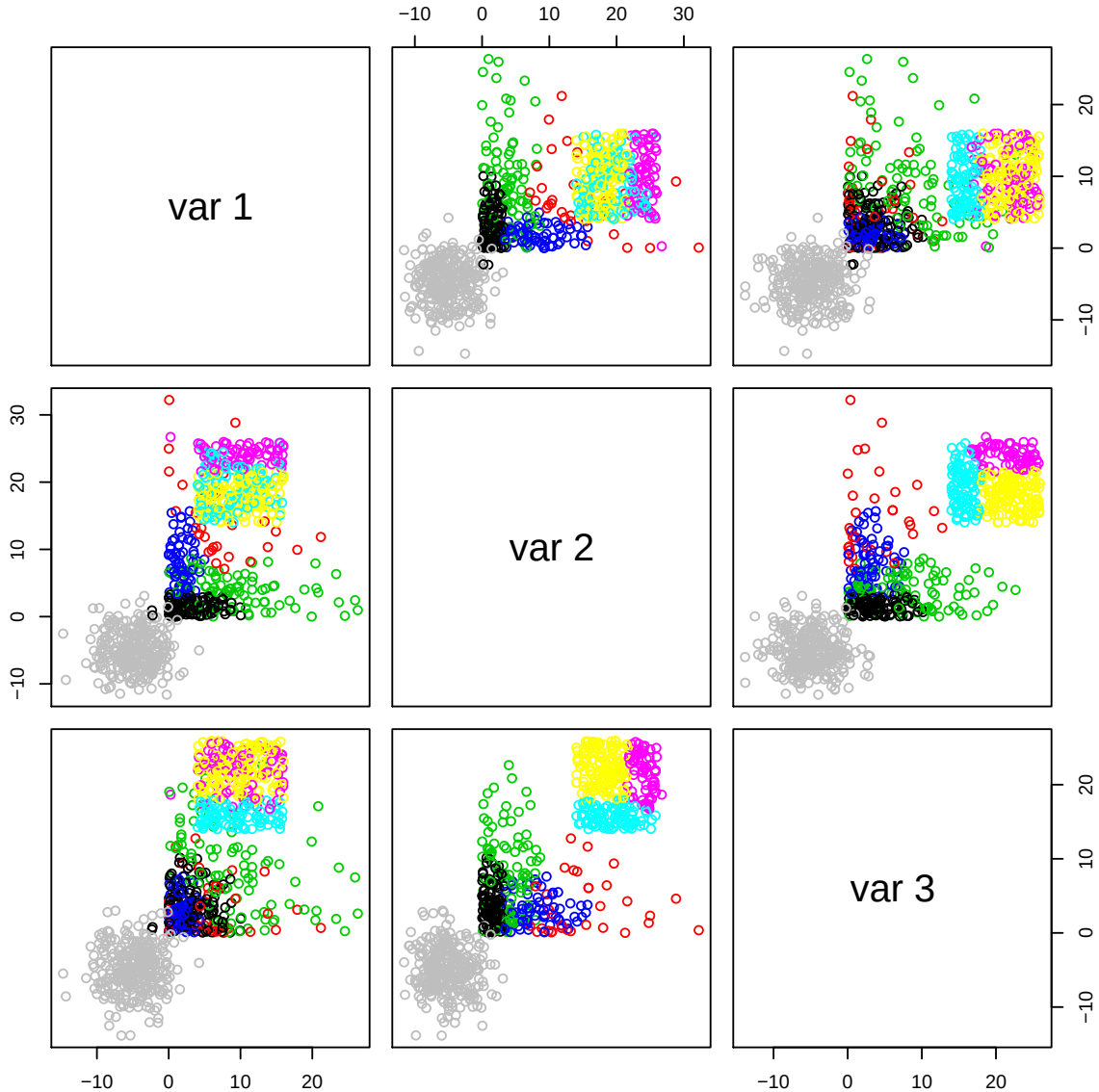
Mixture models - K?



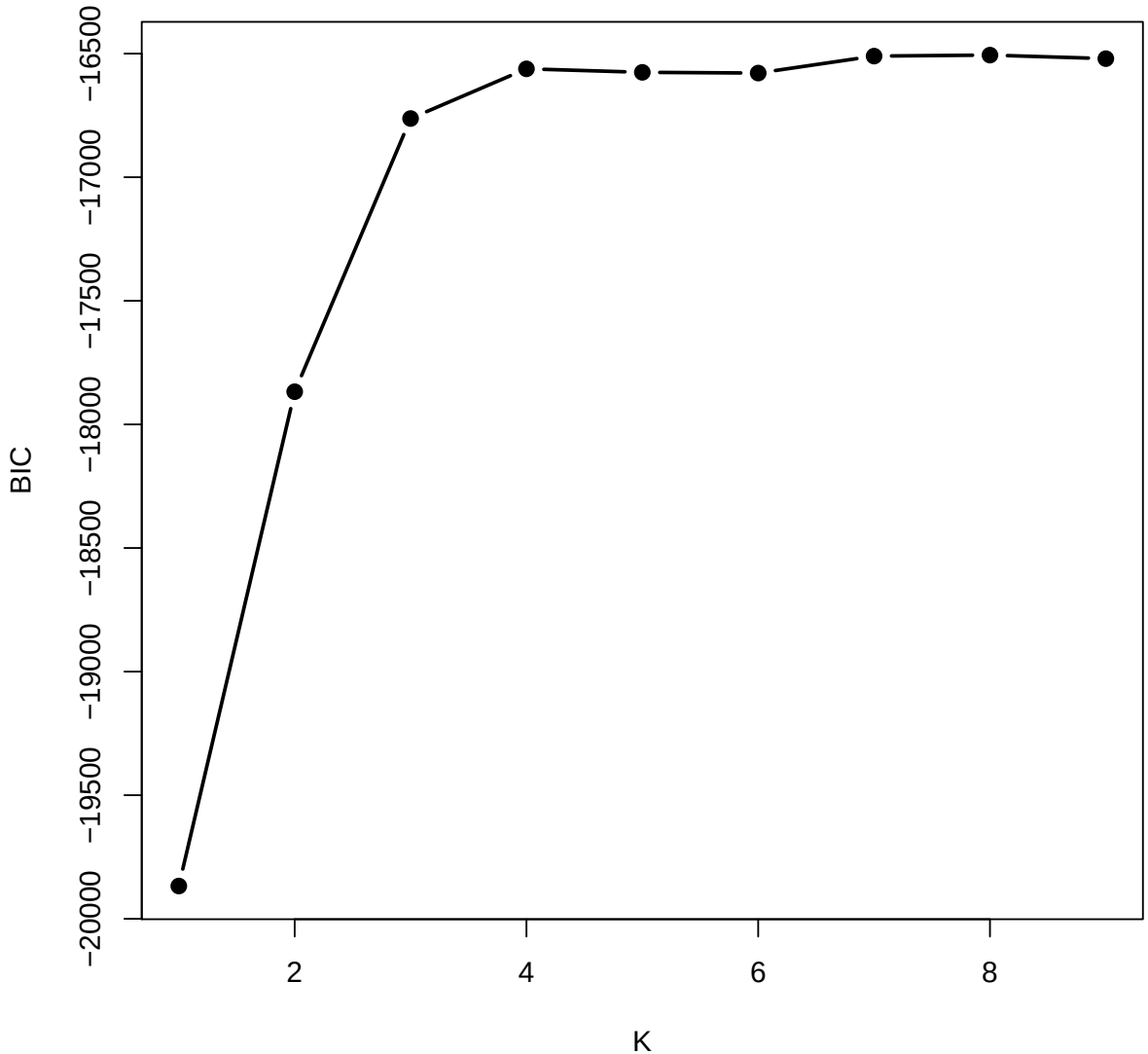
Mixture models - K?



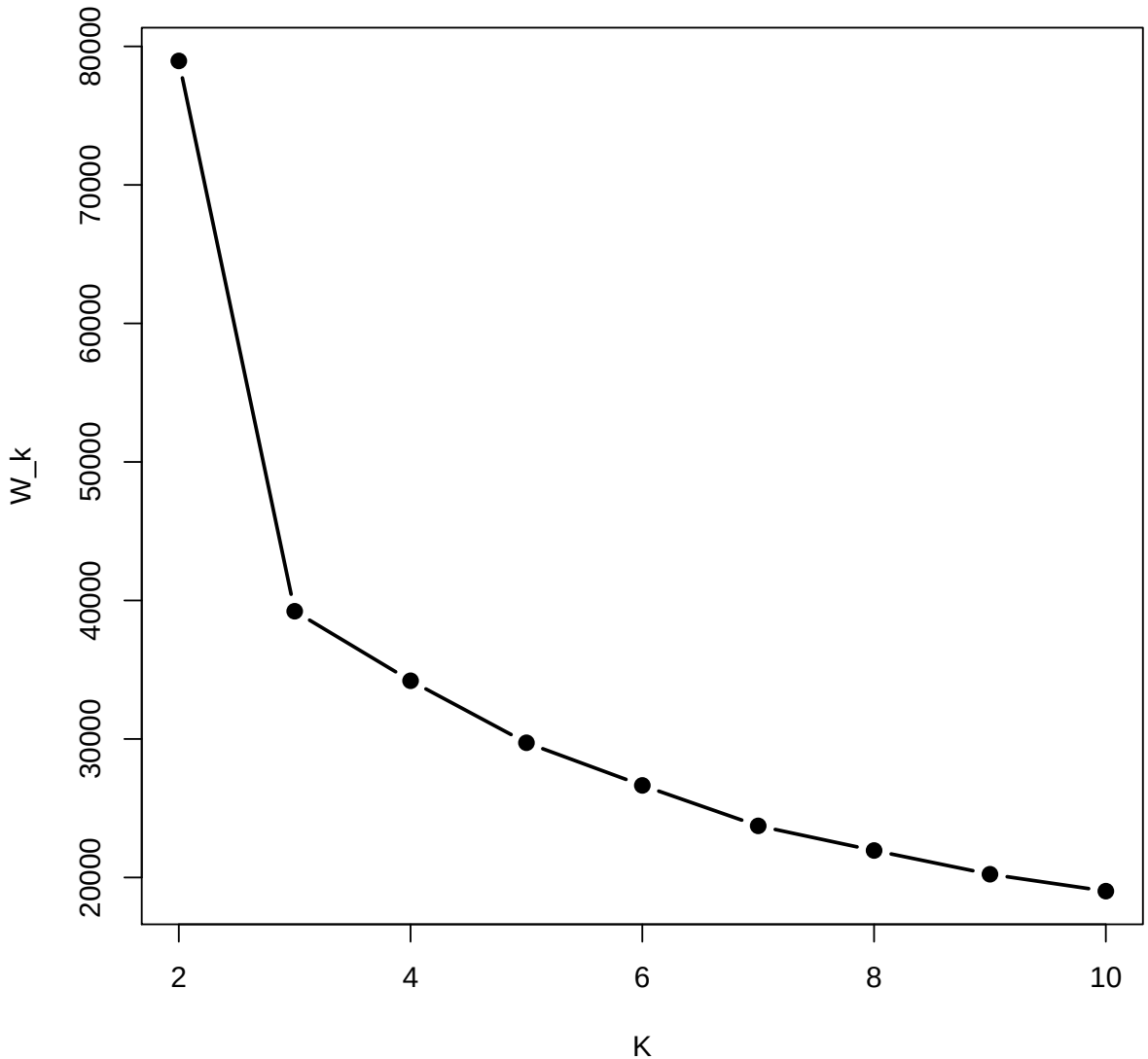
EM solution with best *K*



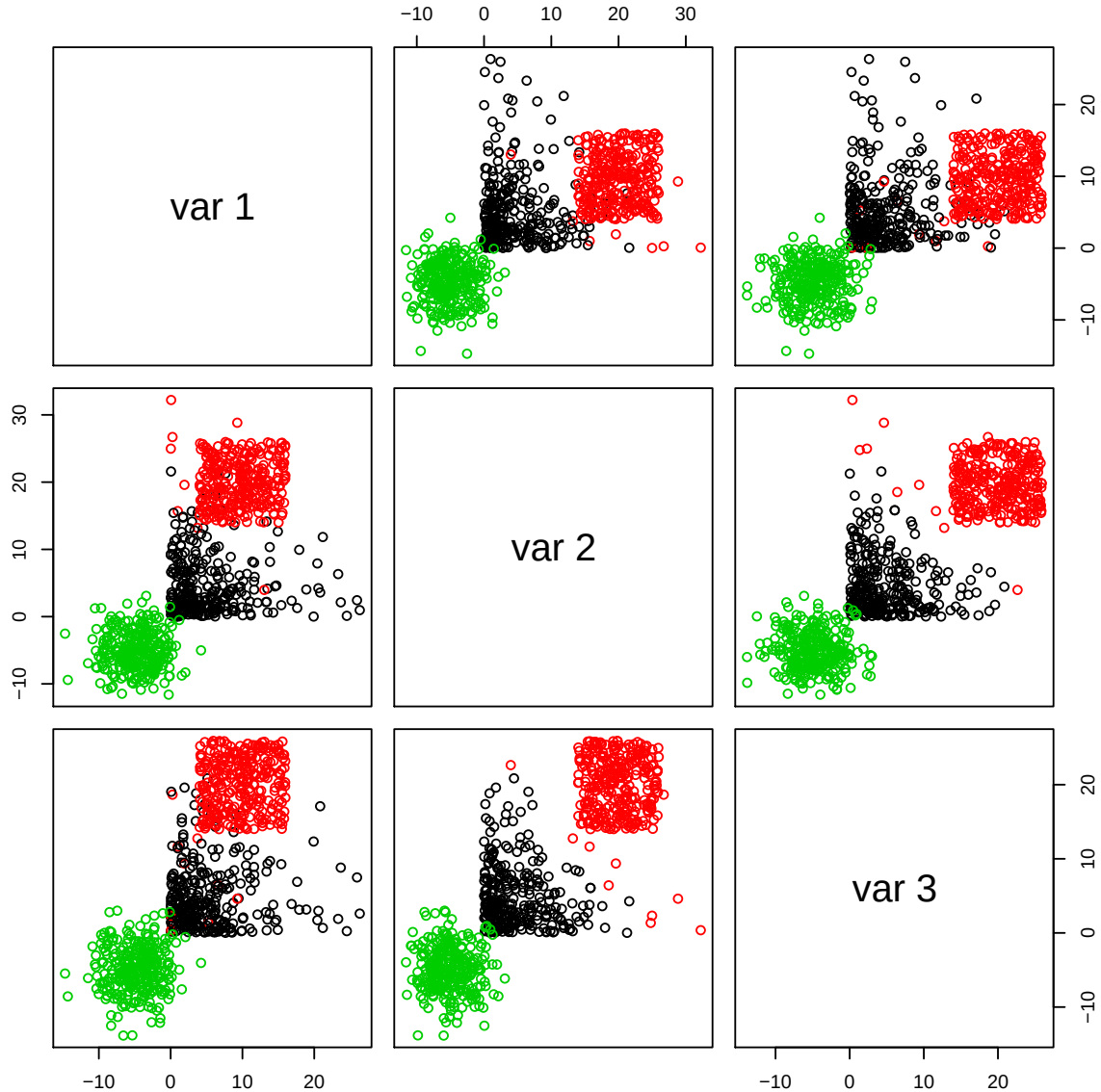
BIC



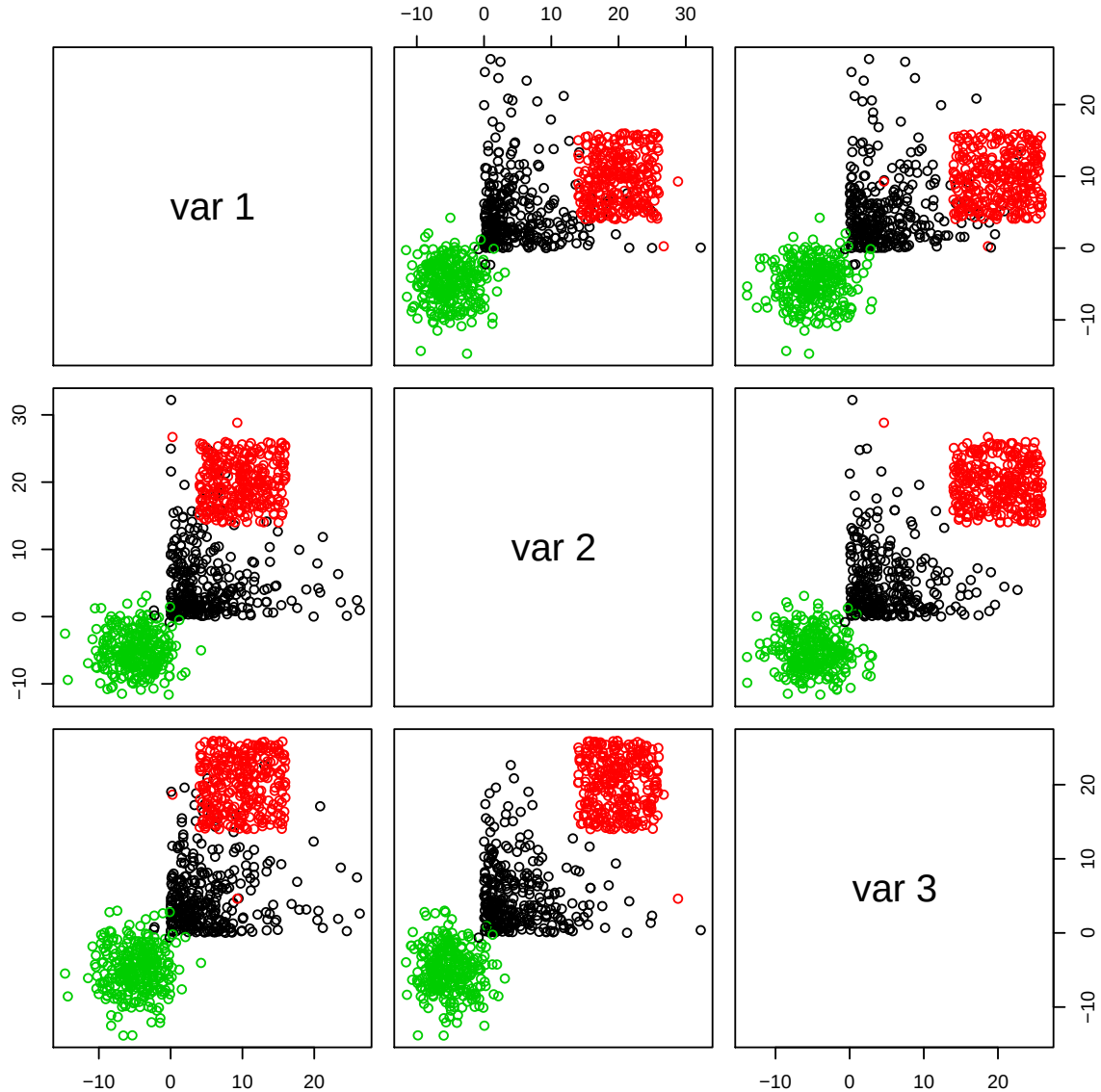
K-means



K-means



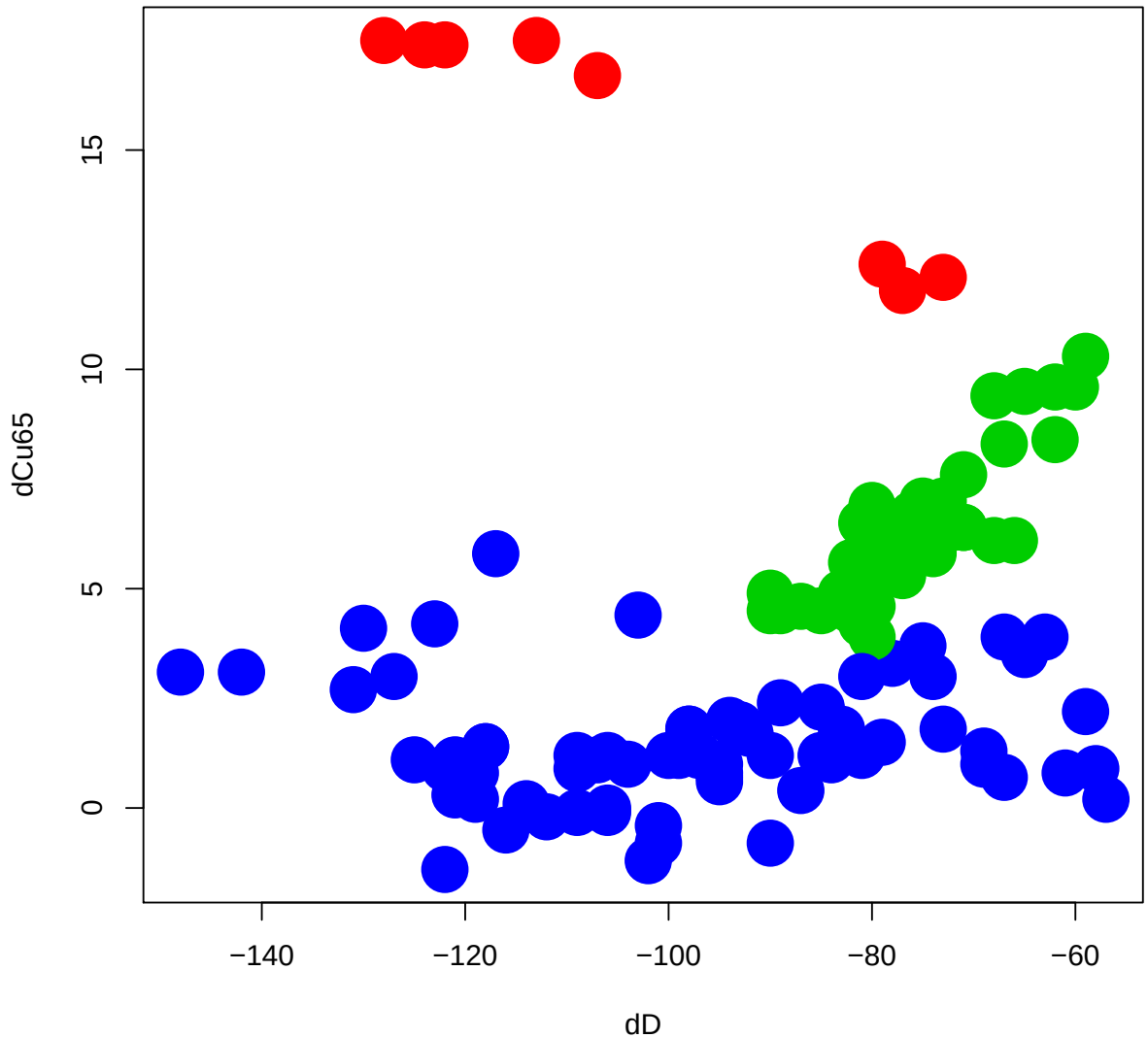
EM with K=3



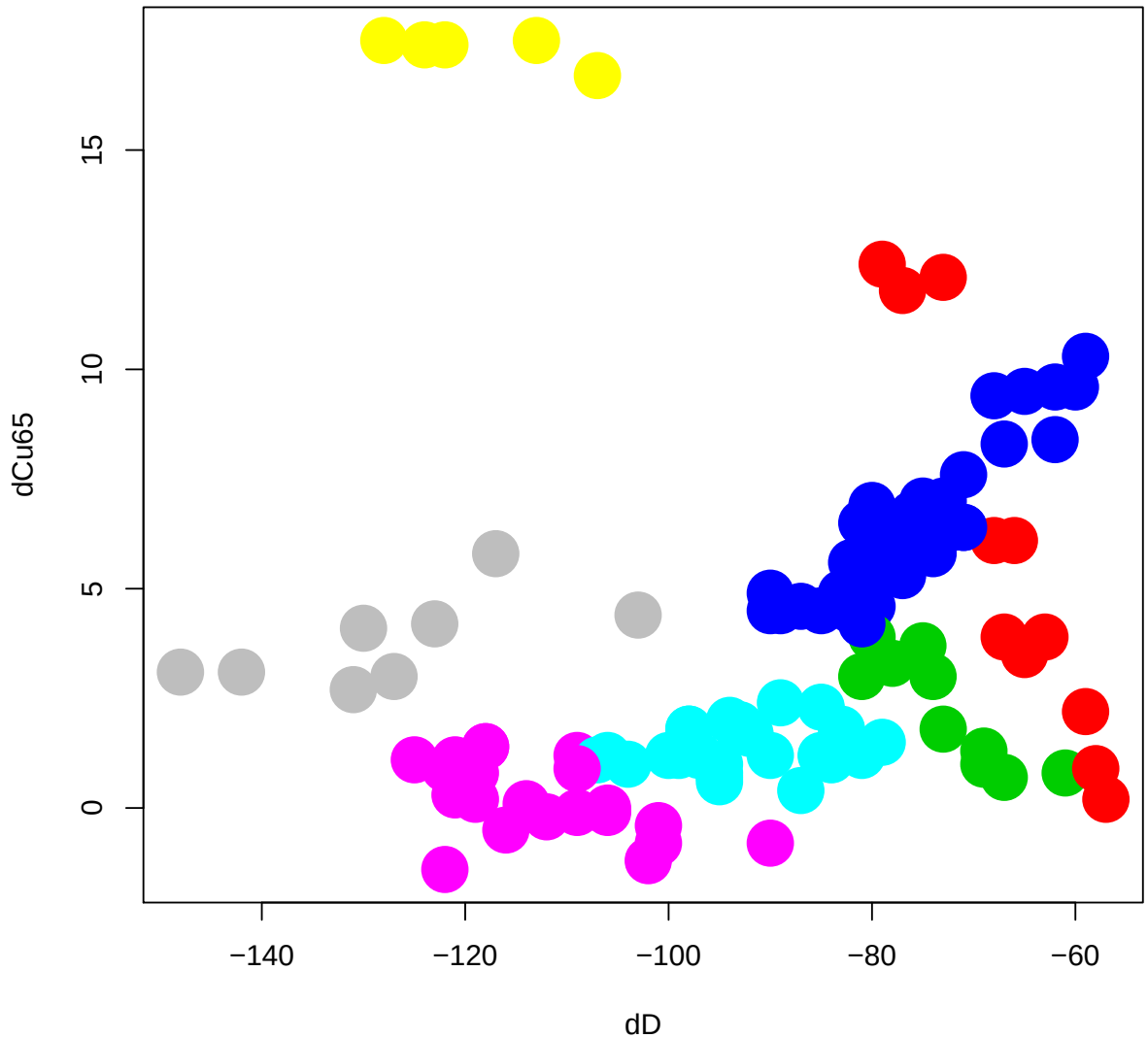
Initialization issues

```
> a <- dget('mclust-fail-dump.txt')
> # n = 130, p = 3 (one is labels, so effectively p = 2)
>
> library(mclust)
>
> # run model-based clustering with features (dCu65, dD)
> m1 <- Mclust(a[,2:3])
> # no. of clusters found (based on BIC)
> m1$G
[1] 3
>
> # run model-based clustering with flipped features (dD, dCu65)
> m2 <- Mclust(a[,3:2])
> # no. of clusters found (based on BIC)
> m2$G
[1] 7
```

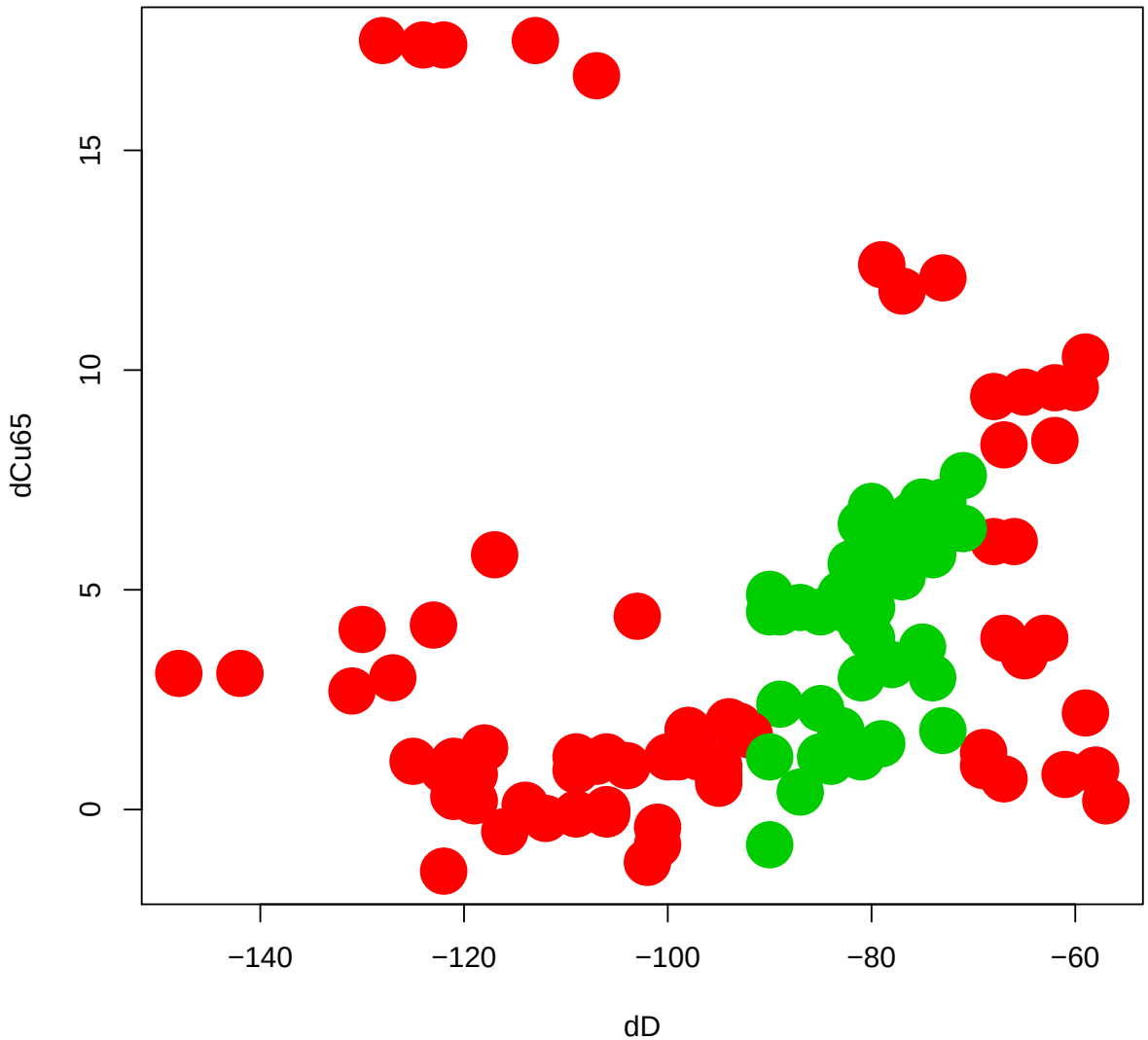
(dCu65, dD)



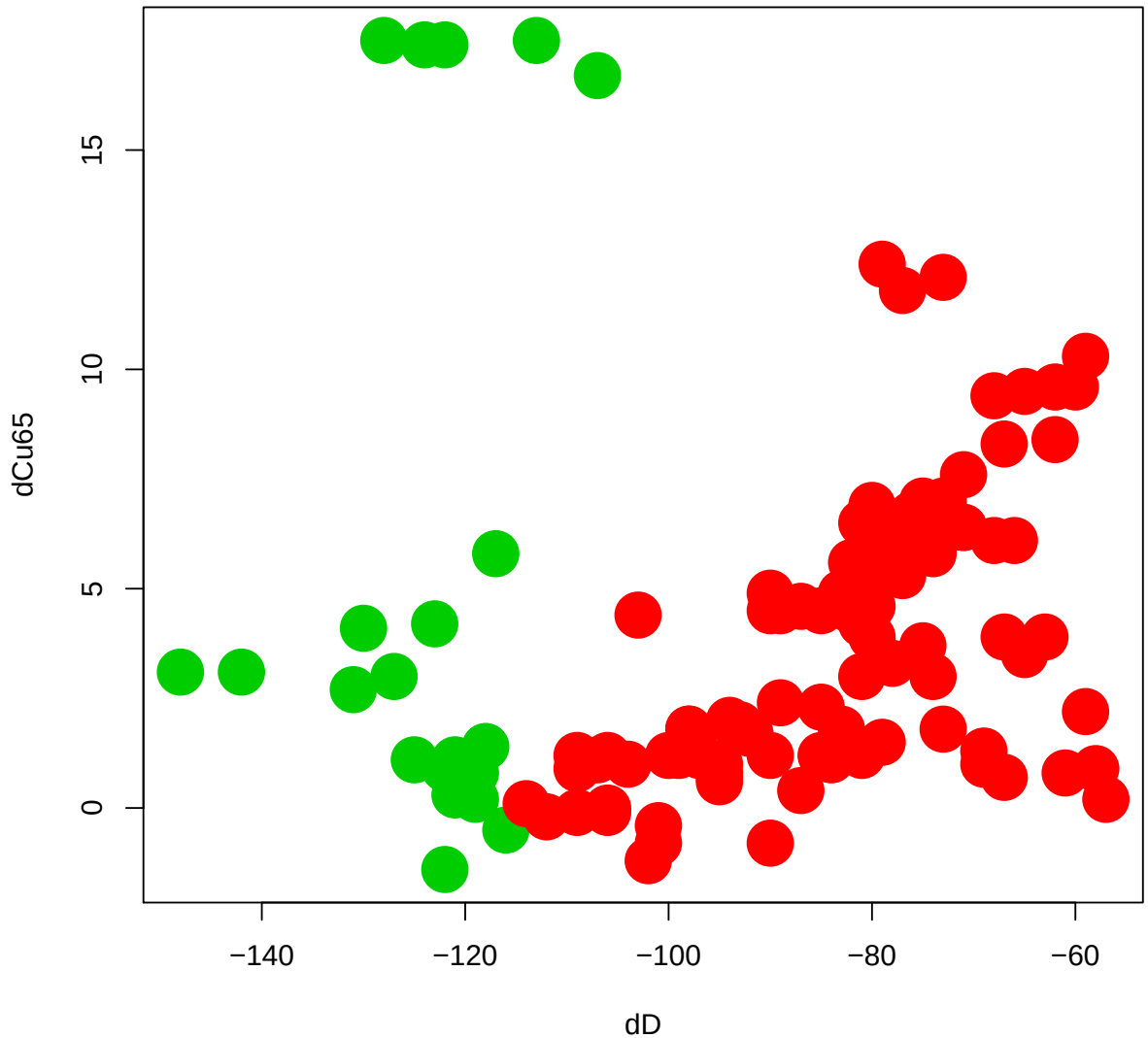
(dD, dCu65)



Initial (dCu65, dD)



Initial (dD, dCu65)



EM algorithm

- Let $\mathbf{X}$ the observed data, $\mathbf{X}^m$ the missing data
- Let $\ell(\mathbf{X}, \mathbf{X}^m; \theta)$ the log-likelihood of the complete data

1. Initiate with $\hat{\theta}^{(0)}$
2. Compute $H(\theta) = E\left(\ell(\mathbf{X}, \mathbf{X}^m; \theta) \middle| \mathbf{X}, \hat{\theta}^{(j)}\right)$
3. Find $\hat{\theta}^{(j+1)} = \arg\max_\theta H(\theta)$
4. $j \leftarrow j + 1$ and repeat from step 2.

EM algorithm

Bottlenecks:

- Computing

$$H(\theta) = E\left(\ell(\mathbf{X}, \mathbf{X}^m; \theta) \middle| \mathbf{X}, \hat{\theta}^{(j)}\right)$$

- Maximizing $H(\theta)$

Why does EM work?

- Data: $(\mathbf{X}, \mathbf{X}^m)$
- Full log-likelihood $\ell_0(\mathbf{X}, \mathbf{X}^m; \boldsymbol{\theta})$

$$P\left(\mathbf{X}^m \middle| \mathbf{X}; \boldsymbol{\theta}\right) = \frac{P\left(\left(\mathbf{X}, \mathbf{X}^m\right); \boldsymbol{\theta}\right)}{P\left(\mathbf{X}; \boldsymbol{\theta}\right)}$$

and then

$$\ell\left(\mathbf{X}; \boldsymbol{\theta}\right) = \ell_0\left(\mathbf{X}, \mathbf{X}^m; \boldsymbol{\theta}\right) - \ell_1\left(\mathbf{X}^m \middle| \mathbf{X}; \boldsymbol{\theta}\right)$$

Why does EM work?

- Hence, for any $\tilde{\theta}$

$$\ell(\mathbf{X};\theta) = E\left[\ell_0(\mathbf{X},\mathbf{X}^m;\theta)\middle|\mathbf{X},\tilde{\theta}\right] -$$

$$E\left[\ell_1(\mathbf{X}^m|\mathbf{X};\theta)\middle|\mathbf{X},\tilde{\theta}\right]$$

- The M-step increases the first term by finding the max over θ
- The second term can only decrease when θ moves away from $\tilde{\theta}$

Missing data & EM

- In general, Gaussian distributions yield closed forms for the maximizers of the expected likelihood
- The method is more general, but requires “specialized software”
- Probably the second most common use of the EM algorithm is **imputation**.

Missing data & EM

	Bahamas	Bangladesh	Belarus	Belgium	Bolivia	Botswana
3249	1	NA	1	2	1	1
3254	1	1	3	1	1	1
3347	1	1	1	2	NA	1
3357	1	3	NA	1	NA	1
3372	2	1	1	2	1	1
3379	NA	1	1	1	1	1

Missing data & EM

- Just using the complete observations might waste a lot of information

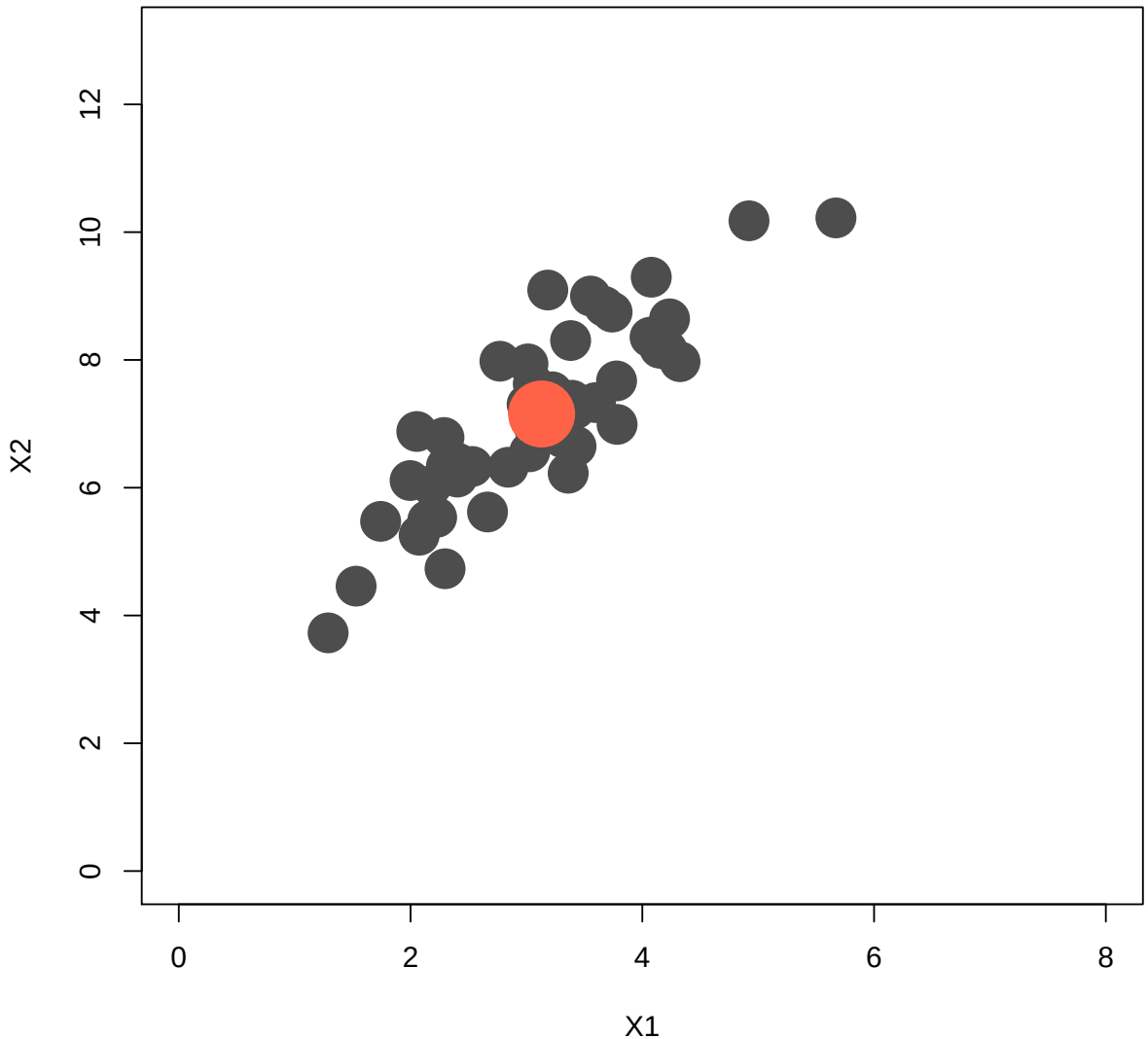
```
> sum(complete.cases(X))
[1] 145

> dim(X)
[1] 368  77
```

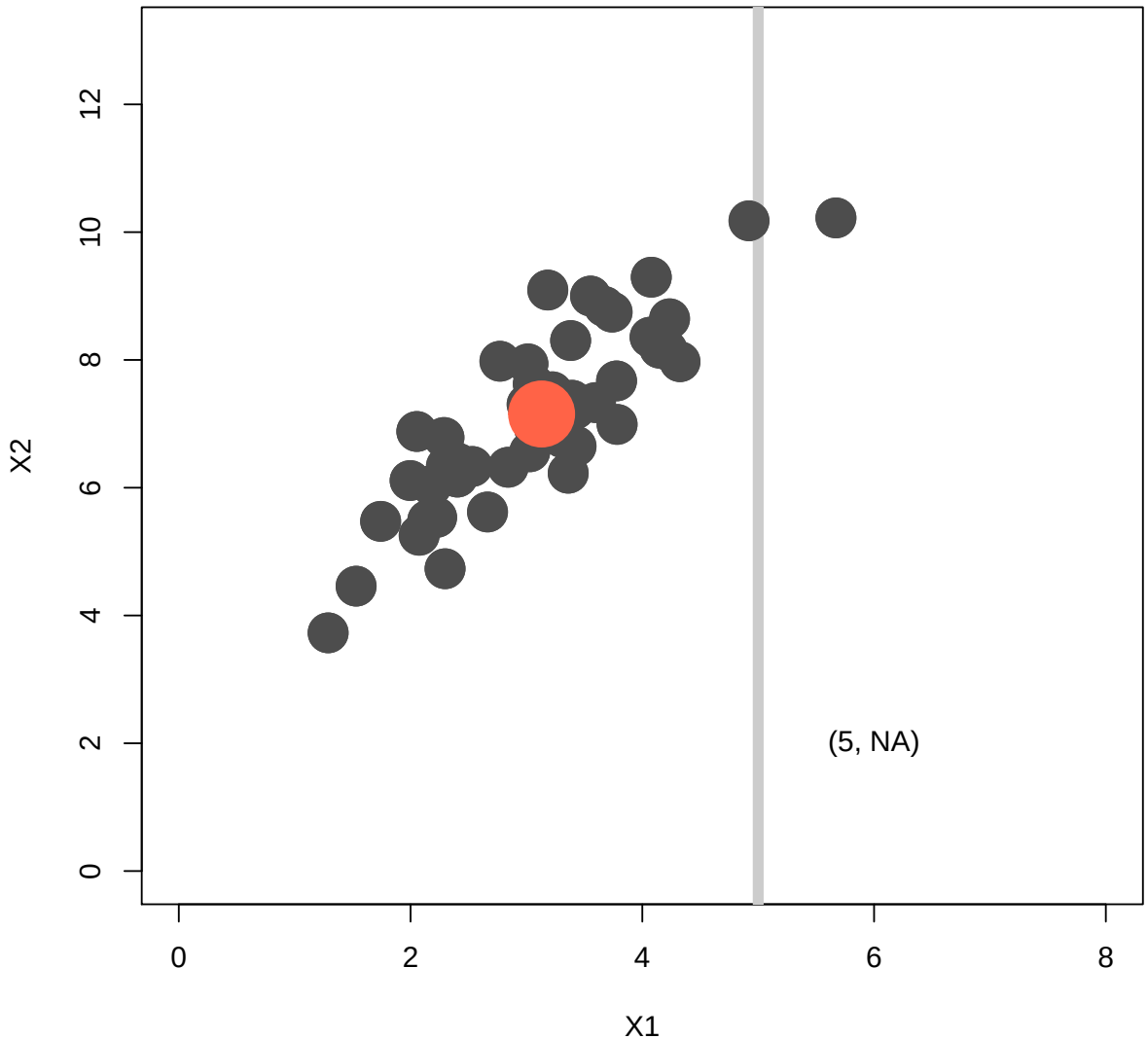
Missing data & EM

- Using only complete records is “sub-optimal”
- Imputation is the process by which one “fills in” missing entries
- Simplest one is to replace `NA`’s with the average of the observed values

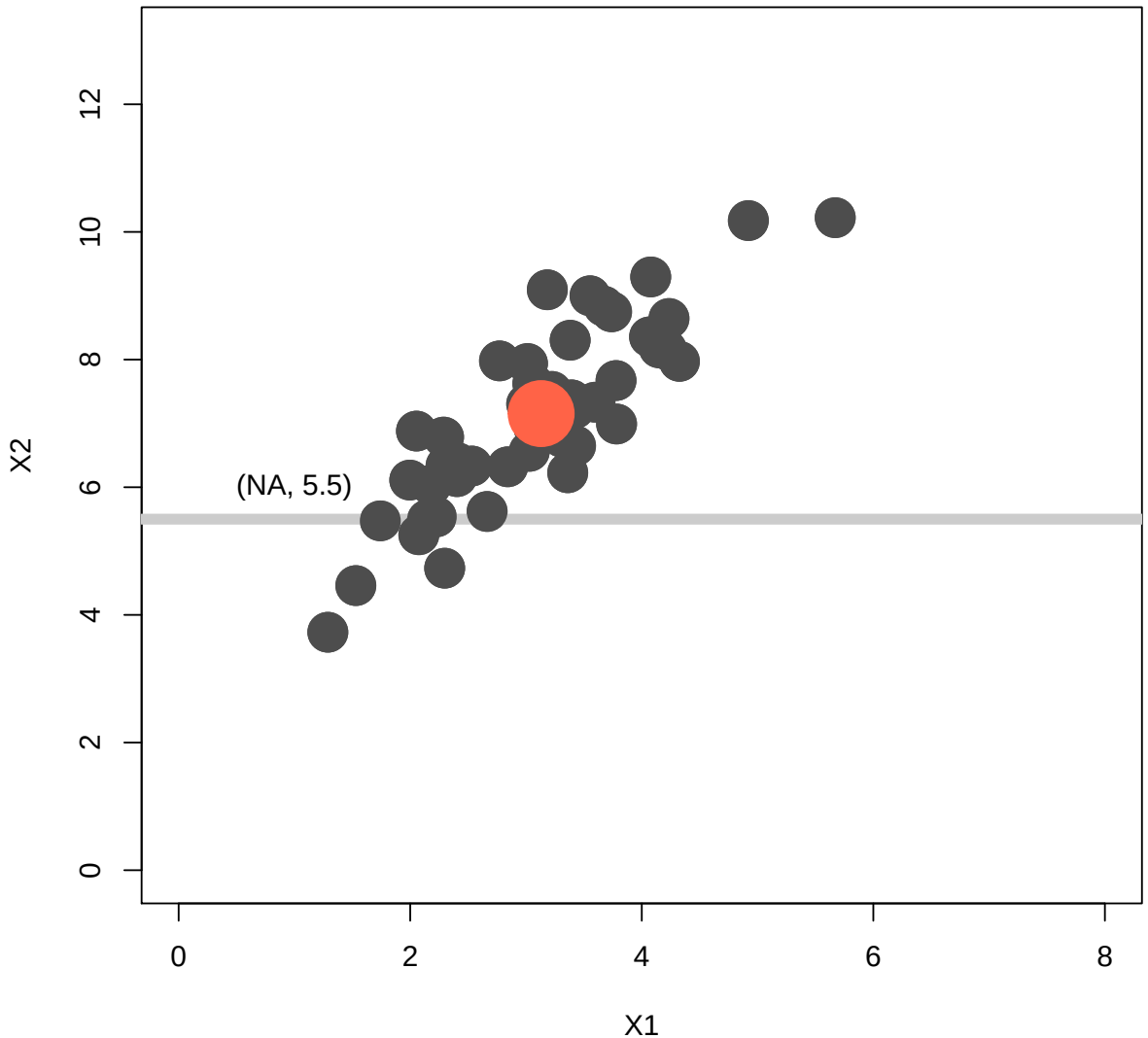
Imputation + EM



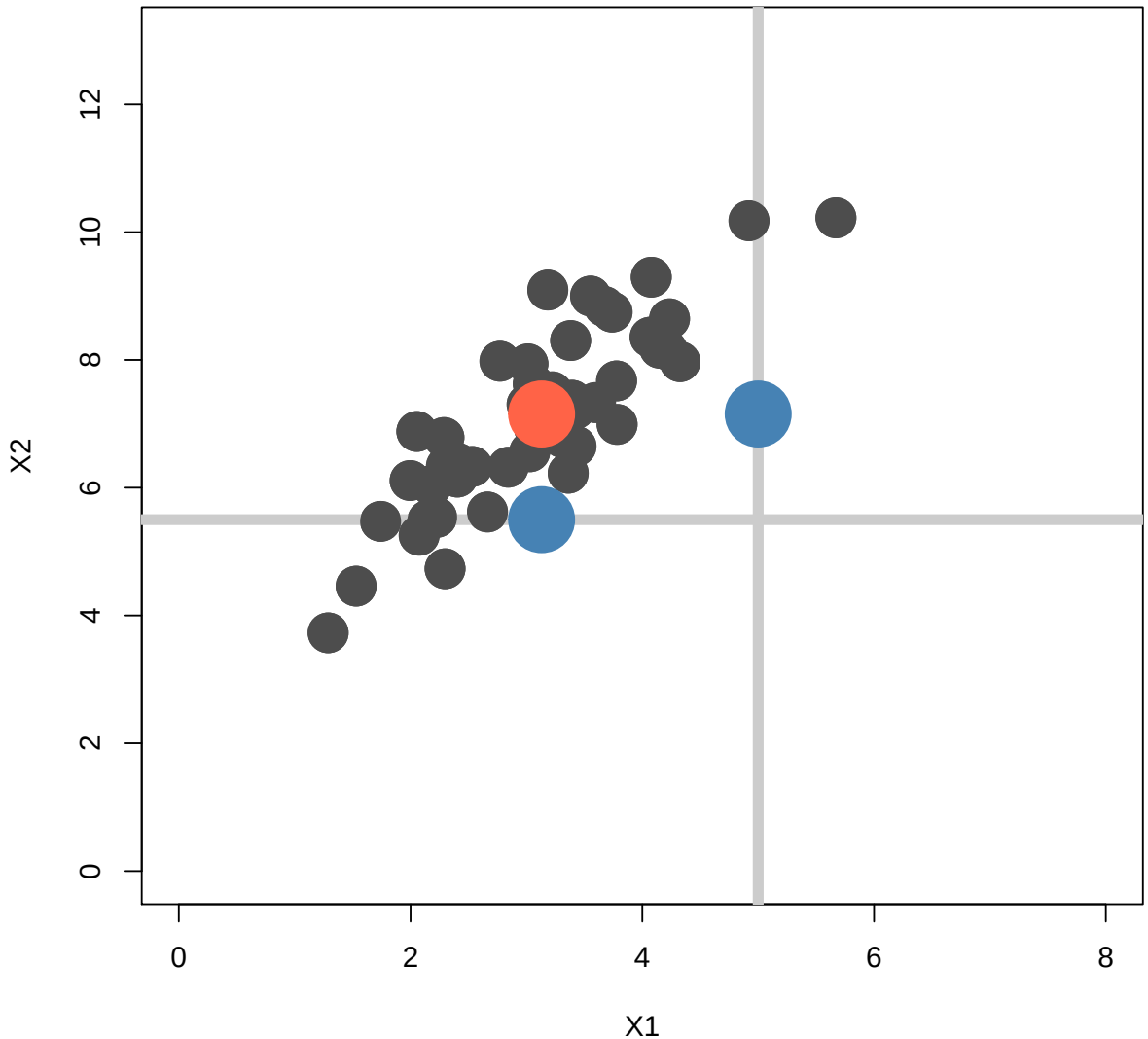
Imputation + EM



Imputation + EM



“Marginal imputation”



Imputation + EM

- If we assume that $\mathbf{X}$ is Gaussian

$$\log f(\mathbf{X};\theta) = -\frac{1}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2}(\mathbf{X}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)$$

$$\ell(\mathbf{X}_1,\ldots,\mathbf{X}_n;\theta) = \sum_{i=1}^{n} \log f(\mathbf{X}_i;\theta)$$

Imputation + EM

- Suppose that

$$\mathbf{X}_i = (X_1, X_2)^\top = (\mathrm{NA}, X)^\top$$

- We need to compute

$$E\left[\log f\left((X_1, X_2)^\top; \theta\right) \middle| X_2, \theta^{(k)}\right]$$

which is not easy, but possible

Imputation + EM

- One can show that

$$E\left[\log f\left((X_1, X_2)^\top; \theta\right) \middle| X_2, \theta^{(k)}\right] =$$

$$C\left(\theta^{(k)}\right) + \log f\left(\left(\tilde{X}_1, X_2\right)^\top; \theta\right)$$

where

$$\tilde{X}_1 = \mu_1^{(k)} + \sigma_{12}^{(k)} \left[\sigma_{22}^{(k)}\right]^{-1} \left(X_2 - \mu_2^{(k)}\right)$$

Imputation + EM

- where

$$\boldsymbol{\theta}^{(k)} = \left(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right)$$

and

$$\boldsymbol{\mu}^{(k)} = \begin{pmatrix} \mu_1^{(k)} \\ \\ \mu_2^{(k)} \end{pmatrix} \quad \boldsymbol{\Sigma}^{(k)} = \begin{pmatrix} \sigma_{11}^{(k)} & \sigma_{12}^{(k)} \\ \\ \sigma_{21}^{(k)} & \sigma_{22}^{(k)} \end{pmatrix}$$

Imputation + EM

- Hence, maximizing

$$H(\theta) = E\left(\ell(\mathbf{X}, \mathbf{X}^m; \theta) \middle| \mathbf{X}, \hat{\theta}^{(j)}\right)$$

is the same as maximizing

$$\ell\left(\mathbf{X}, \tilde{\mathbf{X}}; \theta\right)$$

which is the usual Gaussian MLE for θ, but using $\mathbf{X}$ and $\tilde{\mathbf{X}}$.

Imputation + EM

- Hence, we get

$$\mu^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i$$

and

$$\mathbf{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\mathbf{x}}_i - \mu^{(k+1)}\right)\left(\tilde{\mathbf{x}}_i - \mu^{(k+1)}\right)^\top$$

Imputation + EM