STAT406- Methods of Statistical Learning
Lecture 22

Matias Salibian-Barrera

UBC - Sep / Dec 2017
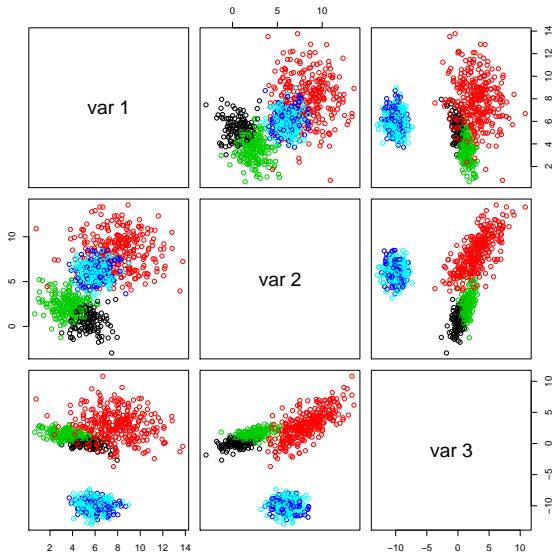
"Noise variables"

- "Noise" (non-cluster-contributing) features can be a problem

- First example: 3 features that define 3 clusters
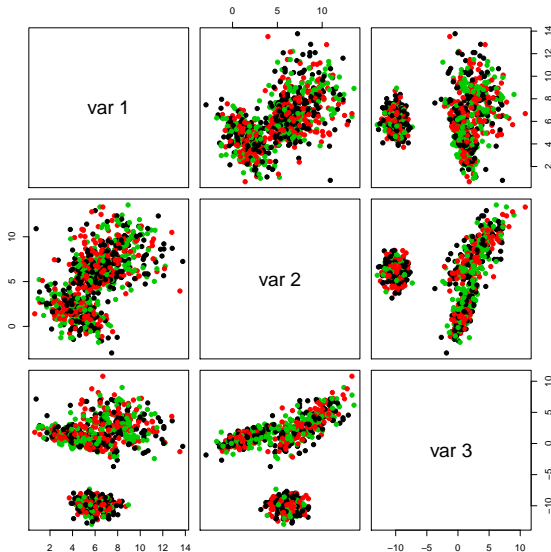
- Three features with uniform random numbers

"Noise variables"

```
set.seed(31)
x1 <- matrix( rnorm(n*3,  mean=3), n, 3) %*% s1.sqrt
mu2 <- c(8, 8, 3)
x2 <- scale( matrix(rnorm(n*3), n, 3) %*% s2.sqrt,
  center=-mu2, scale=FALSE)
mu3 <- c(6, 6, -10)
x3 <- scale( matrix(rnorm(n*3), n, 3), center=-mu3,
  scale=FALSE)
x <- rbind(x1, x2, x3)
x.4 <- cbind(x, matrix(runif(900*3, min=-15, max=15),
  900, 3))
```

"Noise variables"

"Noise variables" - K-means

"Noise variables"

- Second example, same 3 "signal" features

- Plus 3 features with Gaussian random numbers

6

"Noise variables"

```
set.seed(31)
x1 <- matrix( rnorm(n*3,  mean=3), n, 3) %*% s1.sqrt

mu2 <- c(8, 8, 3)
x2 <- scale( matrix(rnorm(n*3), n, 3) %*% s2.sqrt,
  center=-mu2, scale=FALSE)

mu3 <- c(6, 6, -10)
x3 <- scale( matrix(rnorm(n*3), n, 3), center=-mu3,
  scale=FALSE)

x <- rbind(x1, x2, x3)

x.5 <- cbind(x, matrix(rnorm(900*3, sd=10), 900, 3))
```
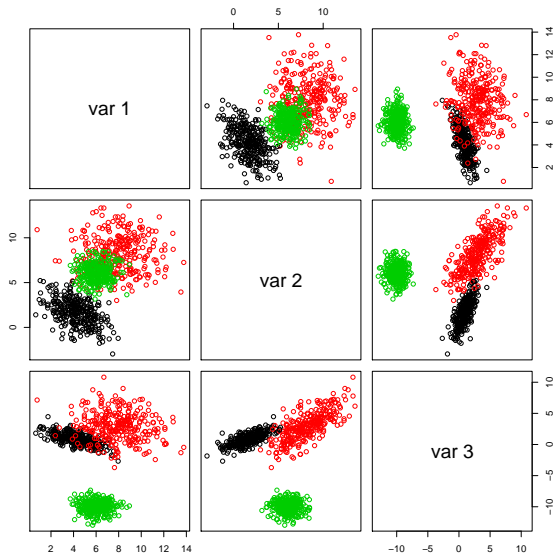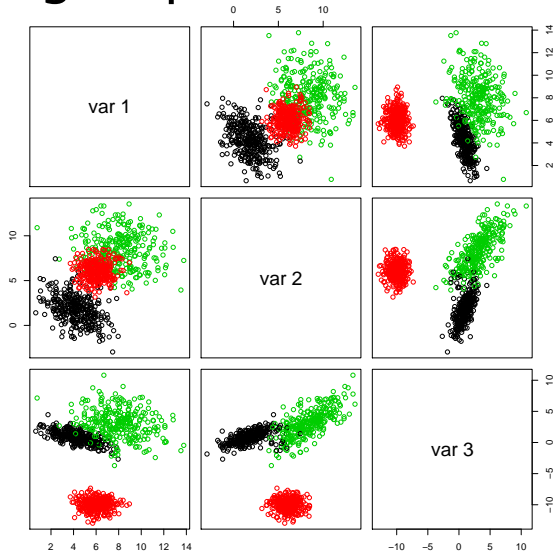
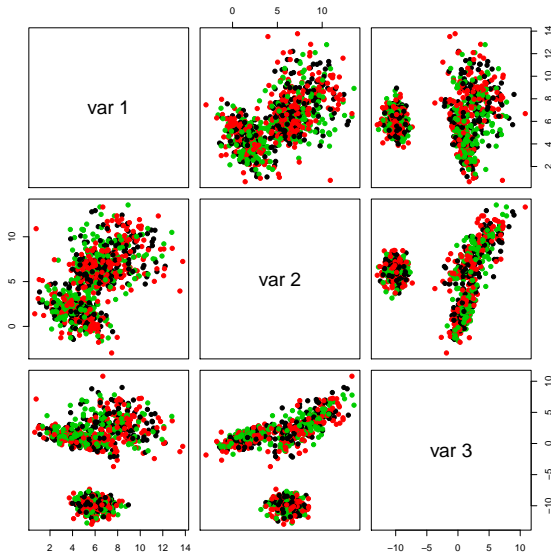# "Noise variables"

8

Force 3 groups

9

"Noise variables" - K-means

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$G
[1] 3
>
> # run model-based clustering with features (dD, dCu65)
> m2 <- Mclust(a[,3:2])
> # no. of clusters found (based on BIC)
> m2$G
[1] 7
```

11
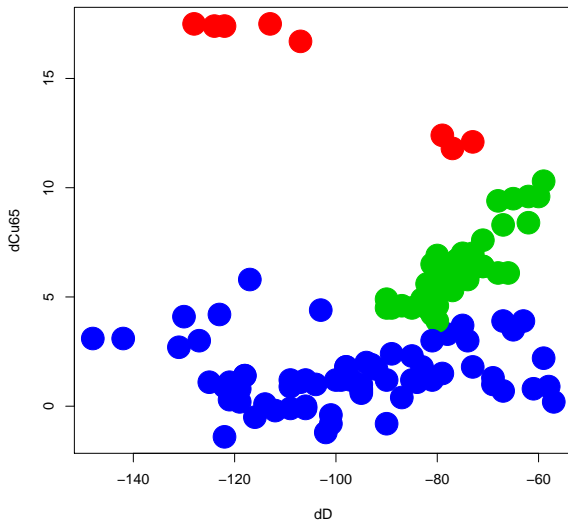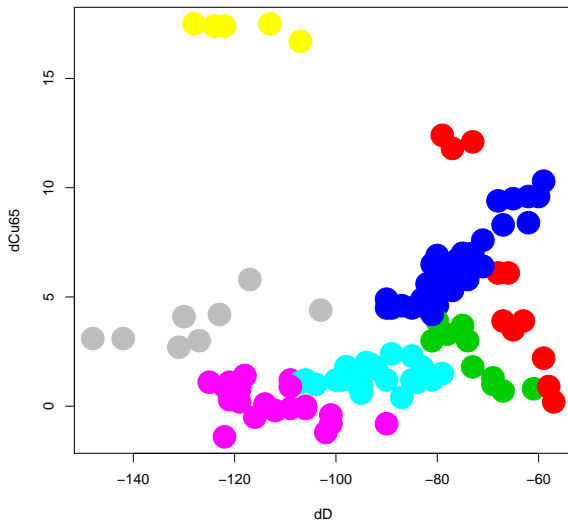
(dCu65, dD)

(dD, dCu65)

Initial (dCu65, dD)
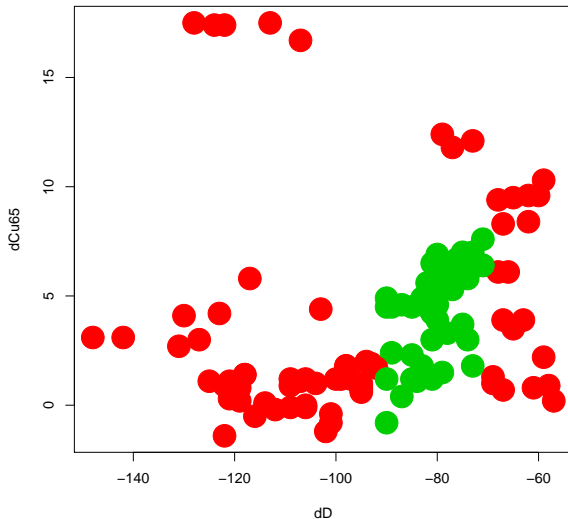
Initial (dD, dCu65)
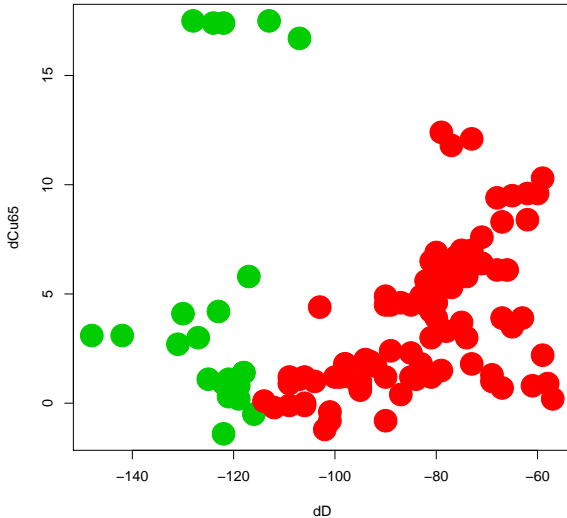
EM algorithm

- Let **X** the observed data, **X**^{m} the missing data

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

1. Initiate with $\hat{\boldsymbol{\theta}}^{(0)}$

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

3. Find $\hat{\boldsymbol{\theta}}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} H(\boldsymbol{\theta})$

4. $j \leftarrow j + 1$ and repeat from step 2.

EM algorithm

Bottlenecks:

- Computing

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

- Maximizing $H(\boldsymbol{\theta})$

Why does EM work?

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

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

and then

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

18

Why does EM work?

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

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

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

- The M-step increases the first term by finding the max over $\boldsymbol{\theta}$

- The second term can only decrease when $\boldsymbol{\theta}$ moves away from $\tilde{\boldsymbol{\theta}}$

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
```
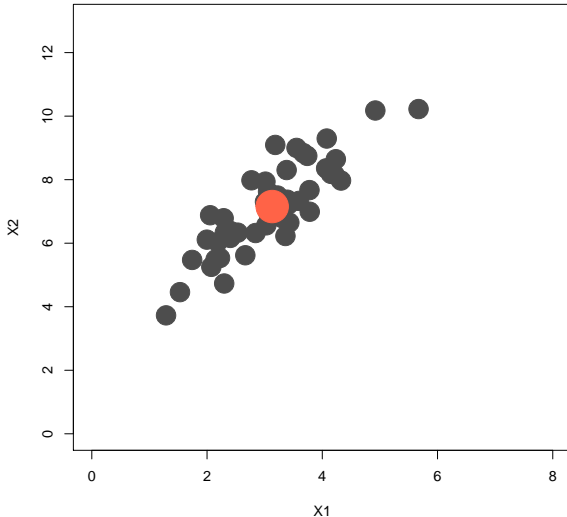
Missing data & EM

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

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

> dim(X)
[1] 368   77
```
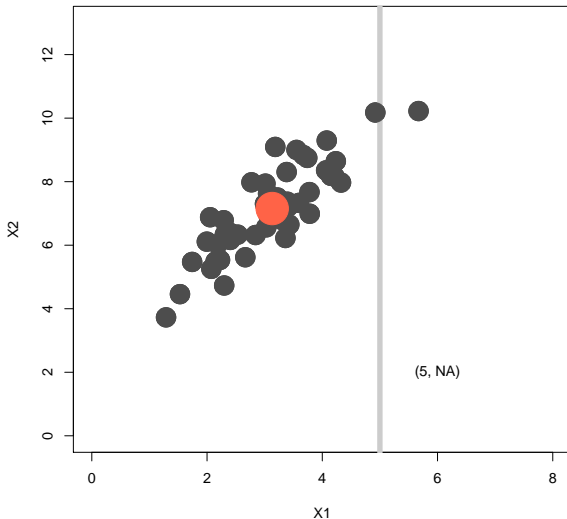
Missing data & EM

- Using only complete records is "sub-optimal"

- Imputation is the process by which one "fills in" missing entries

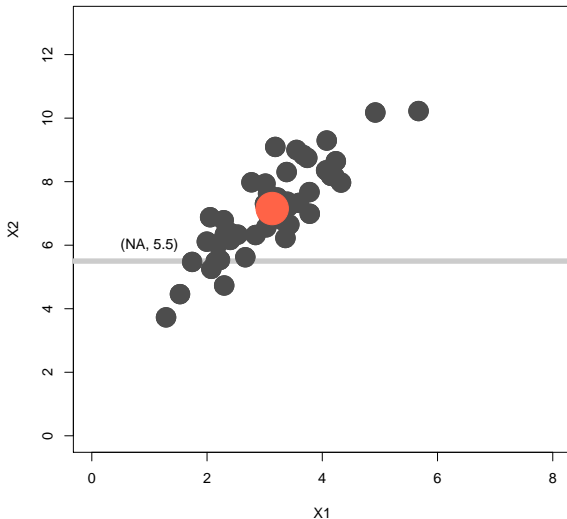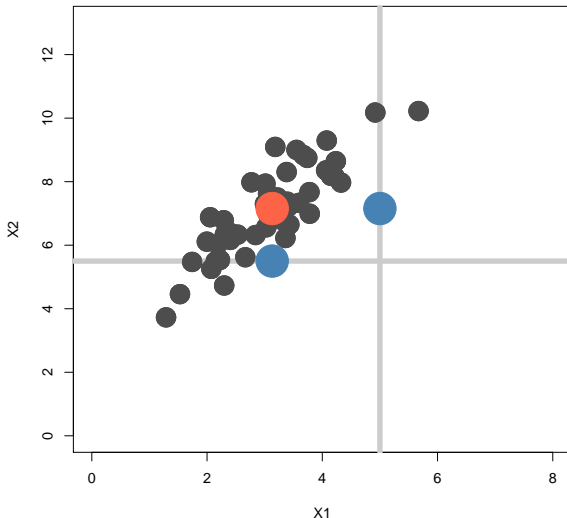- Simplest one is to replace NA's with the average of the observed values

Imputation + EM

Imputation + EM

Imputation + EM

"Marginal imputation"

27

Imputation + EM

- If we assume that **X** is Gaussian

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

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

Imputation + EM

- Suppose that

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

- We need to compute

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

which is not easy, but possible

Imputation + EM

- One can show that

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

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

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

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

is the same as maximizing

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

which is the usual Gaussian MLE for $\boldsymbol{\theta}$, 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