MIXTURE MODELS IN R

Univariate Gaussian Mixture Models

Victor Medina
Researcher at SBIF

Gender dataset

```
> gender %>% head()
```

```
  Gender   Height   Weight        BMI
1   Male 73.84702 241.8936 0.04435662
2   Male 68.78190 162.3105 0.03430822
3   Male 74.11011 212.7409 0.03873433
4   Male 71.73098 220.0425 0.04276545
5   Male 69.88180 206.3498 0.04225479
6   Male 67.25302 152.2122 0.03365316
```

```
> gender %>% select(-Gender) %>% head()
```

```
    Height   Weight        BMI
1 73.84702 241.8936 0.04435662
2 68.78190 162.3105 0.03430822
3 74.11011 212.7409 0.03873433
4 71.73098 220.0425 0.04276545
5 69.88180 206.3498 0.04225479
6 67.25302 152.2122 0.03365316
```

Modeling with Mixture Models

1. Which is the suitable probability distribution?
2. How many sub-populations should we consider?
3. Which are the parameters and their estimations?

Clustering with one variable

```
> head(gender %>% select(-Gender))
```

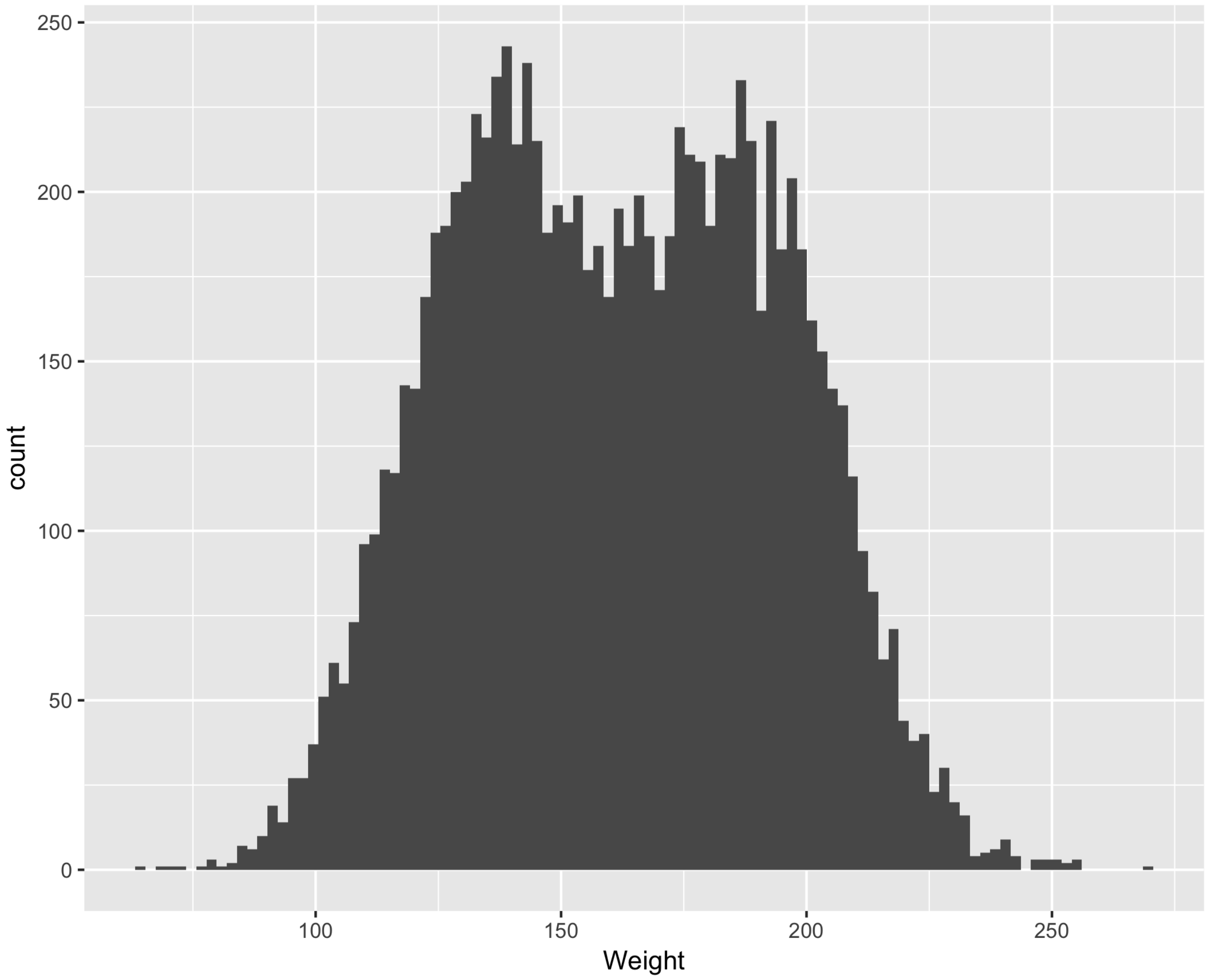
```
    Height   Weight        BMI
1 73.84702 241.8936 0.04435662
2 68.78190 162.3105 0.03430822
3 74.11011 212.7409 0.03873433
4 71.73098 220.0425 0.04276545
5 69.88180 206.3498 0.04225479
6 67.25302 152.2122 0.03365316
```

```
> head(gender %>% select(Weight))
```

```
    Weight
1 241.8936
2 162.3105
3 212.7409
4 220.0425
5 206.3498
6 152.2122
```

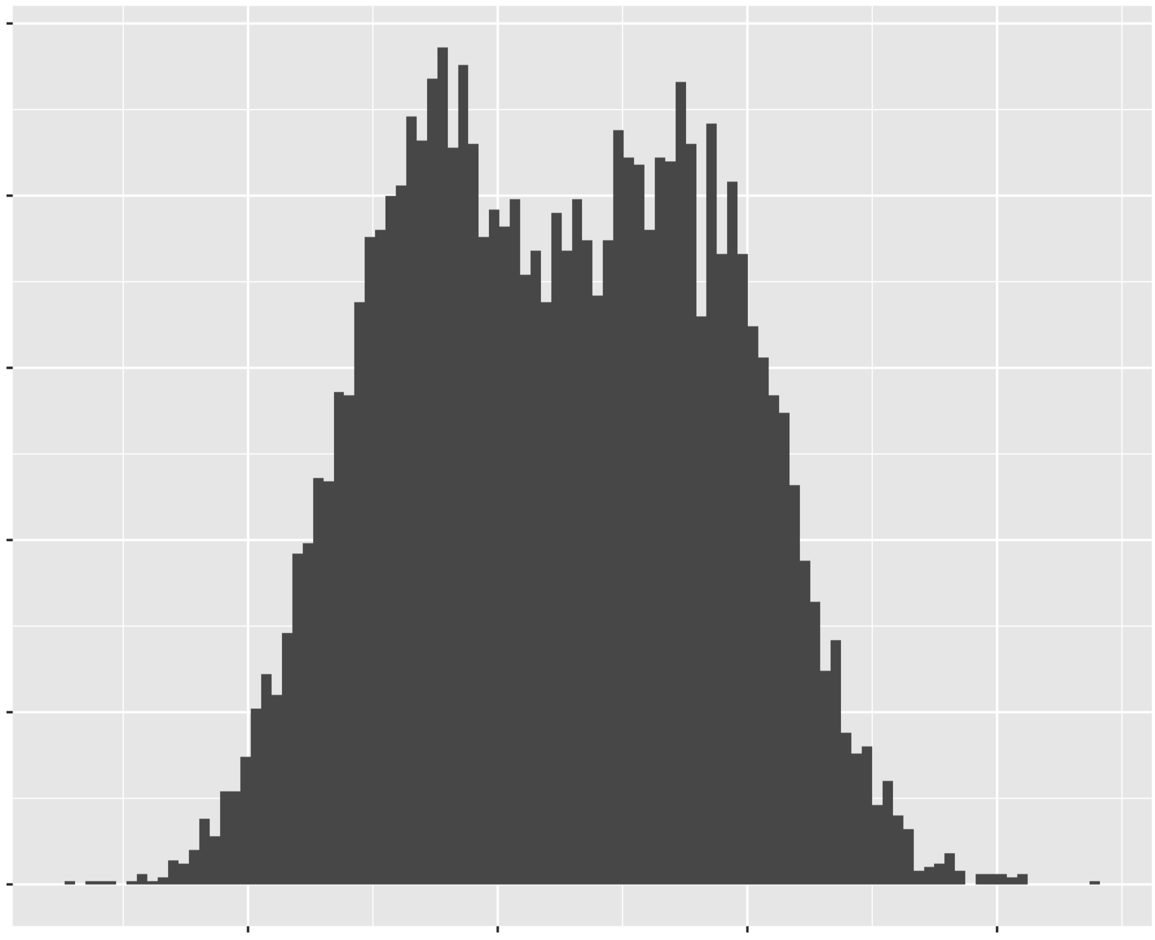
Exploratory data analysis

```
gender %>%
  ggplot(aes(x = Weight)) + geom_histogram(bins = 100)
```

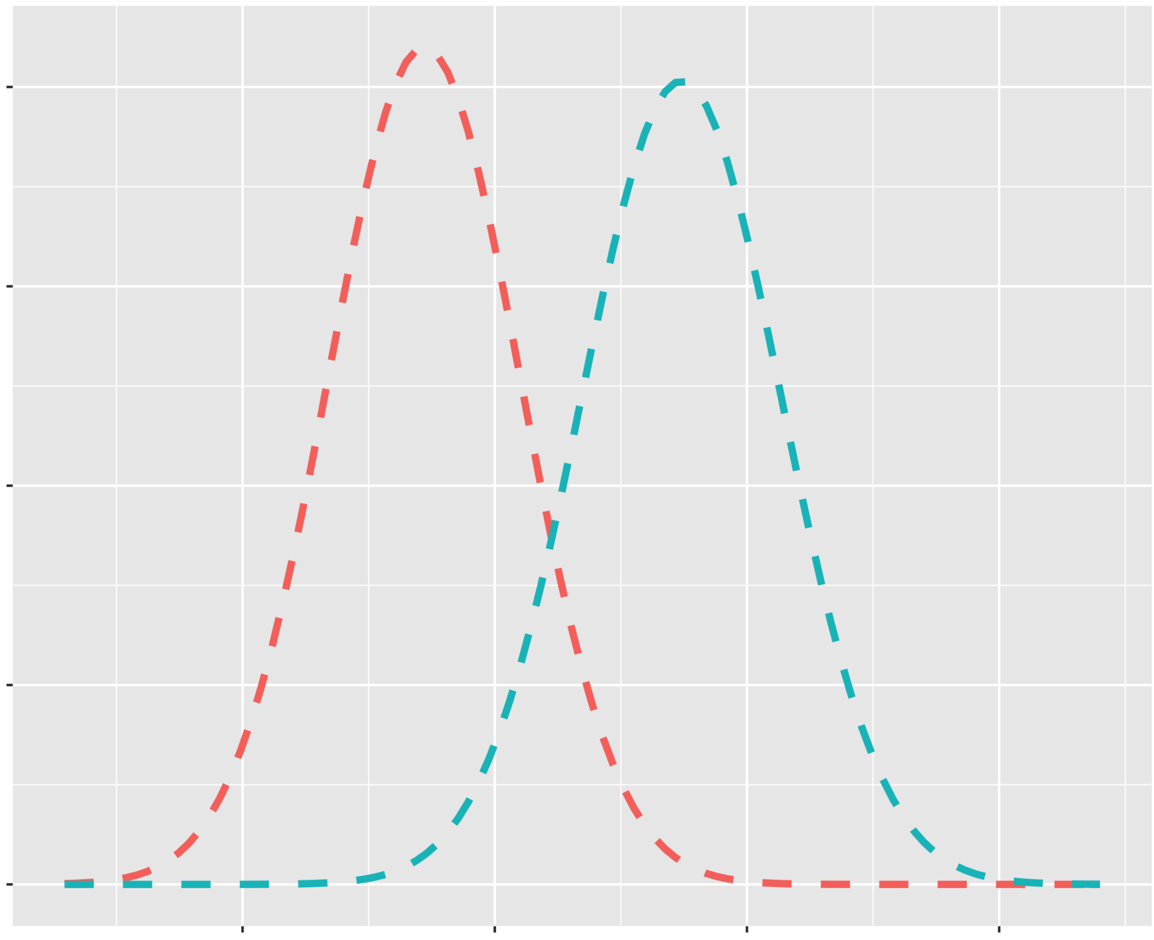


Which distribution?

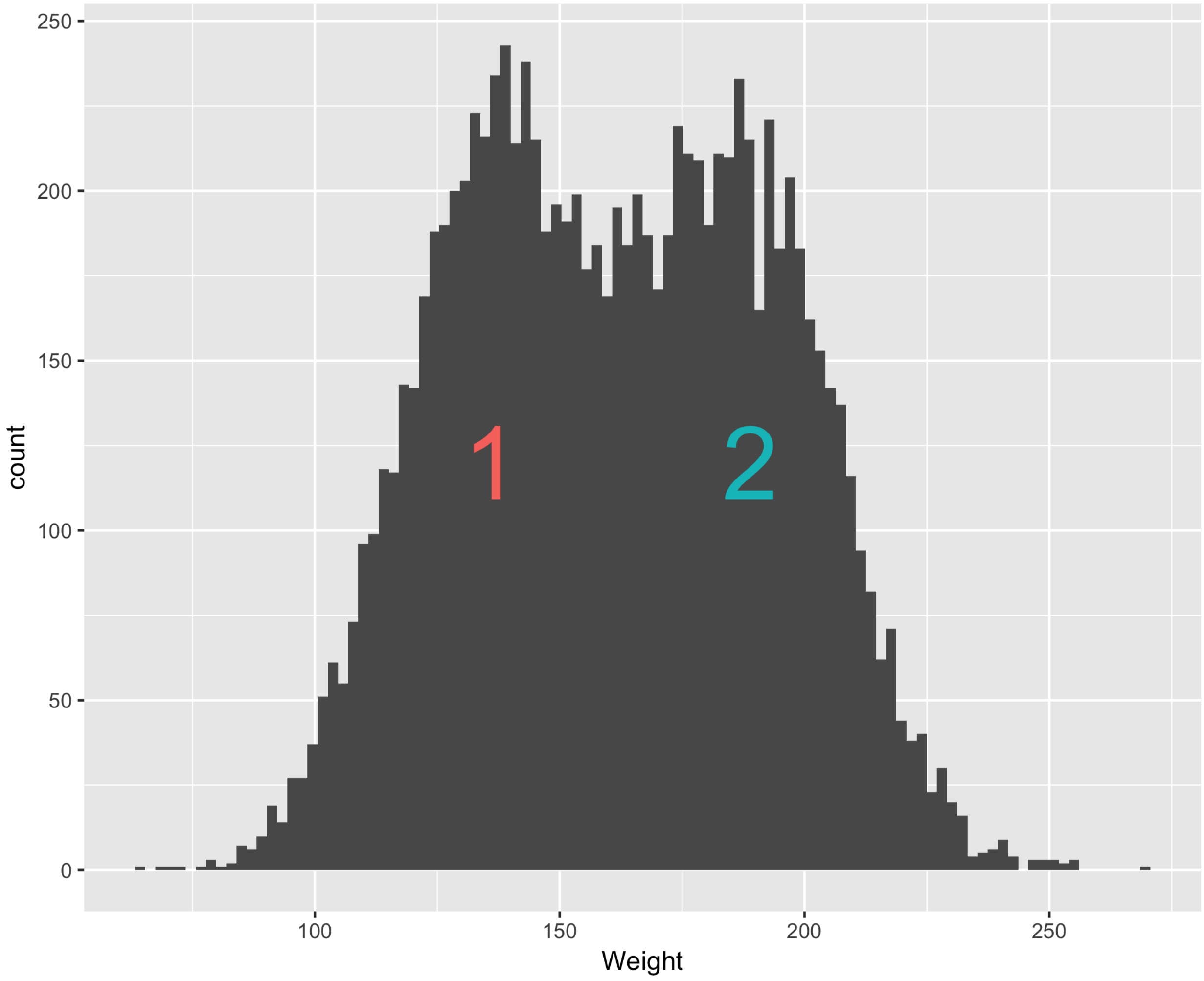
Histogram



Gaussian distributions



How many clusters?



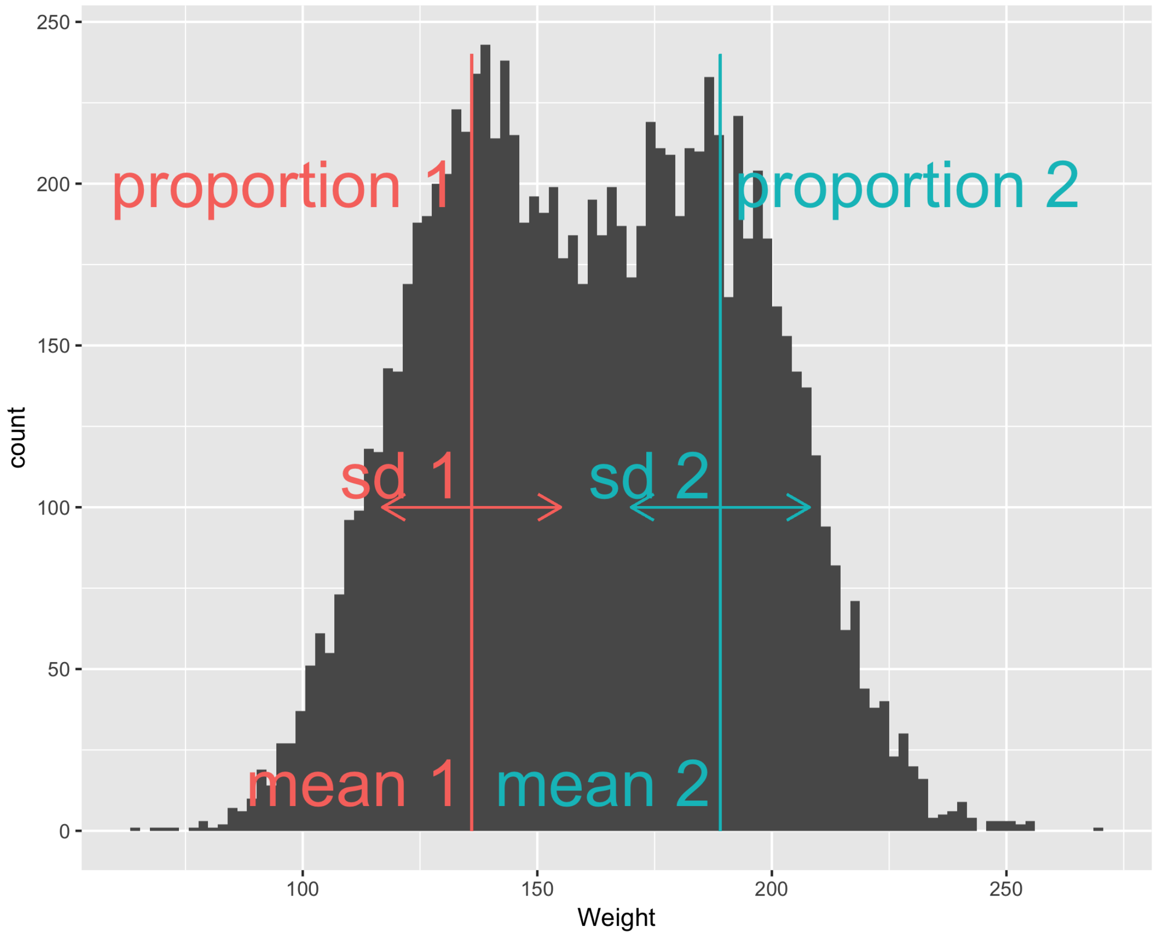
Which parameters and how to estimate them?

Which parameters?

- Two means
- Two standard deviations
- Two proportions

How to estimate them?

- EM algorithm implemented in `flexmix`



MIXTURE MODELS IN R

Let's practice!

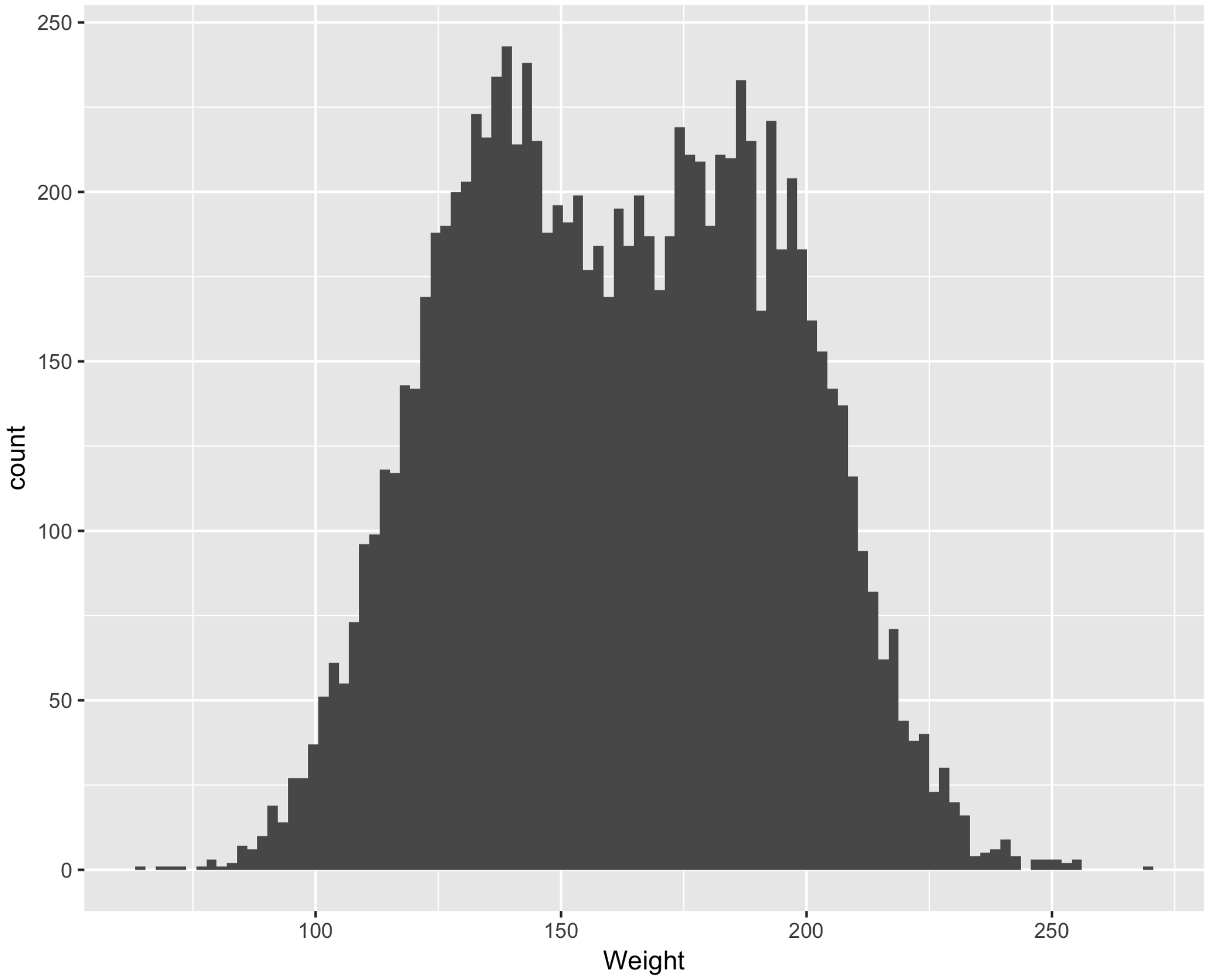
MIXTURE MODELS IN R

Univariate Gaussian Mixture Models with flexmix

Victor Medina
Researcher at SBIF

Gender dataset

```
gender %>%
  ggplot(aes(x = Weight)) + geom_histogram(bins = 100)
```



Modeling with Mixture Models

1. Which is the suitable probability distribution?
   - Univariate Gaussian distributions
2. How many sub-pupulations shoud we consider?
   - 2 clusters
3. Which are the parameters and their estimations?
   - EM algorithm implemented in `flexmix` to estimate the **means**, the **standard deviations** and the **proportions**

flexmix function

```
flexmix(formula, data, k, model, control, ...)
```

- **formula**: description of the model to be fit ($variable \sim 1$)
- **data**: data frame
- **k**: number of clusters
- **model**: specifies the distribution (`FLXMCnorm1`, `FLXMCmvnorm`, `FLXMCmvbinary`, `FLXMRglm`, `FLXMCmvpois`)
- **control**: specifies the max number of iterations, the tolerance, etc.

Fit univariate Gaussian mixture model

```
> fit_mixture <- flexmix(Weight ~ 1, # the means and sds are constant
+                     data = gender, # the data frame
+                     k = 2, # the number of clusters,
+                     model = FLXMCnorm1(), # univariate Gaussian
+                     control = list(tol = 1e-15, # tolerance for EM stop
+                                    verbose = 1, # show partial results
+                                    iter = 1e4)) # max number of iterations
```

```
Classification: weighted
   1 Log-likelihood :   -48880.0782
   2 Log-likelihood :   -48880.0745
   3 Log-likelihood :   -48880.0732
   4 Log-likelihood :   -48880.0727
   .     .    .    .    .    .    .    .
   .     .    .    .    .    .    .    .
   .     .    .    .    .    .    .    .
3454 Log-likelihood :   -48518.3717
3455 Log-likelihood :   -48518.3717
3456 Log-likelihood :   -48518.3717
3457 Log-likelihood :   -48518.3717
converged
```

The proportions: prior function

```
> proportions <- prior(fit_mixture)
> proportions
```

```
[1] 0.4929668 0.5070332
```

The means and the sds: parameters function

Both distributions

```
> parameters(fit_mixture)
```

```
                      Comp.1     Comp.2
coef.(Intercept) 135.54652 186.61583
sigma             18.94726   19.96097
```

Each of them

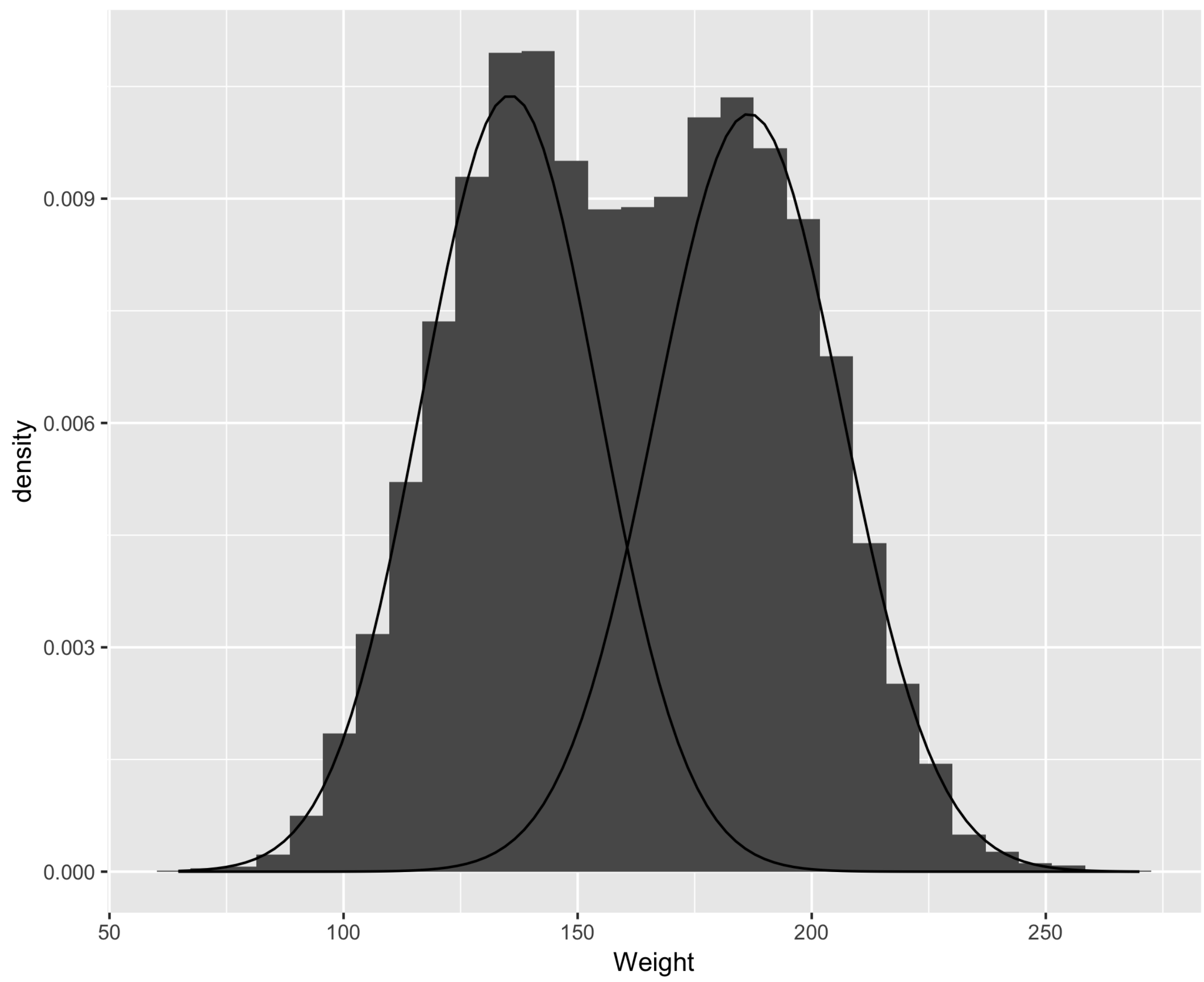
```
> comp_1 <- parameters(fit_mixture, component = 1)
> comp_2 <- parameters(fit_mixture, component = 2)
> comp_2
```

```
                      Comp.2
coef.(Intercept) 186.61583
sigma             19.96097
```

Visualize the resulting distributions

```
> gender %>%
+   ggplot() + geom_histogram(aes(x = Weight, y = ..density..)) +
+   stat_function(geom = "line", fun = fun_prop,
+                 args = list(mean = comp_1[1],
+                             sd = comp_1[2],
+                             proportion = proportions[1])) +
+   stat_function(geom = "line", fun = fun_prop,
+                 args = list(mean = comp_2[1],
+                             sd = comp_2[2],
+                             proportion = proportions[2]))
```

Visualize the resulting distributions



The probabilities and assignments

`posterior` function

```
> posterior(fit_mixture) %>% head()
```

```
             [,1]      [,2]
[1,] 6.836341e-06 0.9999932
[2,] 4.421760e-01 0.5578240
[3,] 5.994160e-04 0.9994006
[4,] 1.998798e-04 0.9998001
[5,] 1.547774e-03 0.9984522
[6,] 7.544450e-01 0.2455550
```

`clusters` function

```
> clusters(fit_mixture) %>% head()
```

```
[1] 2 2 2 2 2 1
```

Assignments comparison

```
> table(gender$Gender, clusters(fit_mixture))
```

```
          1    2
Female 4500  500
Male    444 4556
```

MIXTURE MODELS IN R

Let's practice!

MIXTURE MODELS IN R

Bivariate Gaussian Mixture Models

Victor Medina
Researcher at SBIF

Gender data

One variable

```
> gender %>%
    select(Weight) %>%
    head()
```

```
    Weight
1 241.8936
2 162.3105
3 212.7409
4 220.0425
5 206.3498
6 152.2122
```

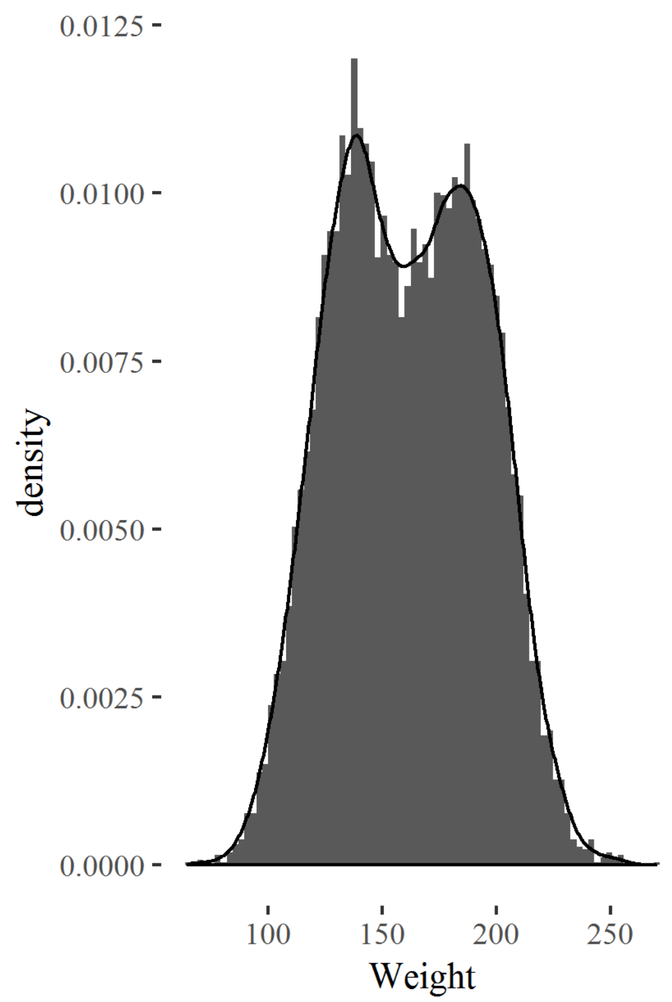
Two variables

```
> gender %>%
    select(Weight, BMI) %>%
    head()
```

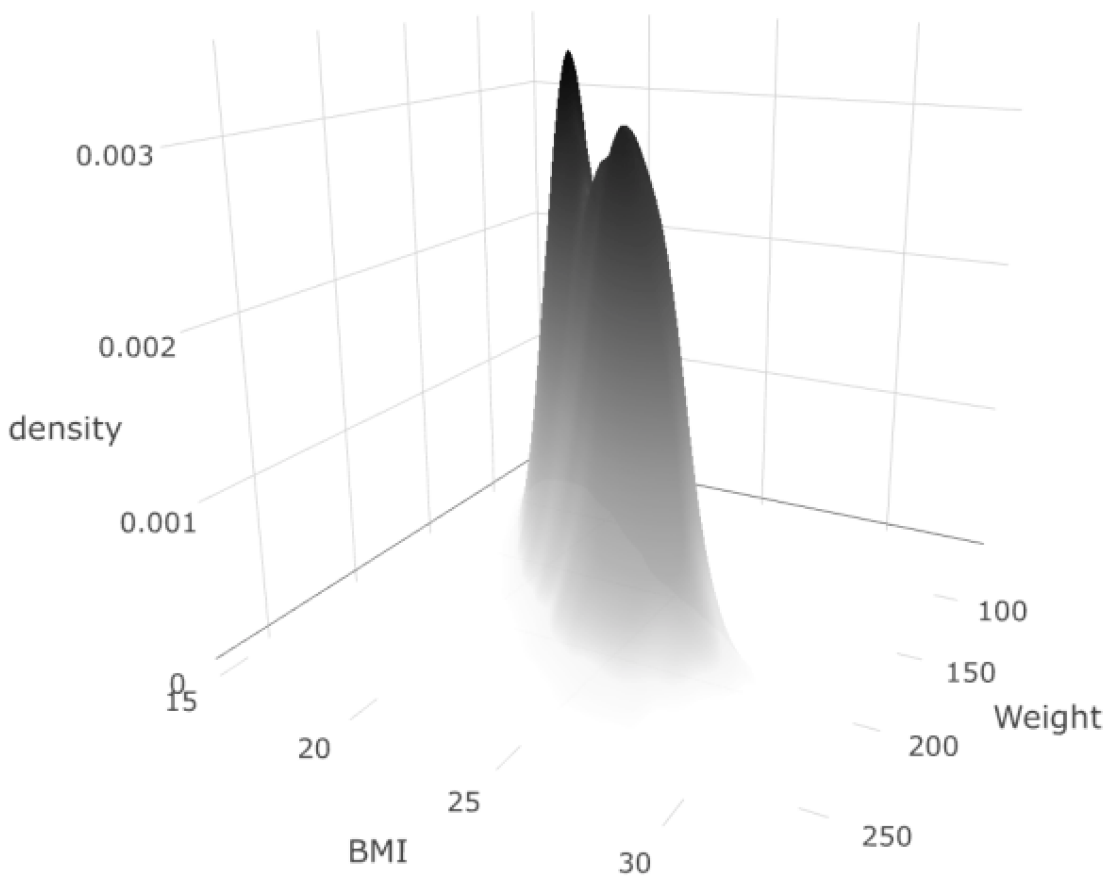
```
    Weight      BMI
1 241.8936 31.18576
2 162.3105 24.12104
3 212.7409 27.23291
4 220.0425 30.06706
5 206.3498 29.70803
6 152.2122 23.66049
```

Exploratory data analysis

One variable



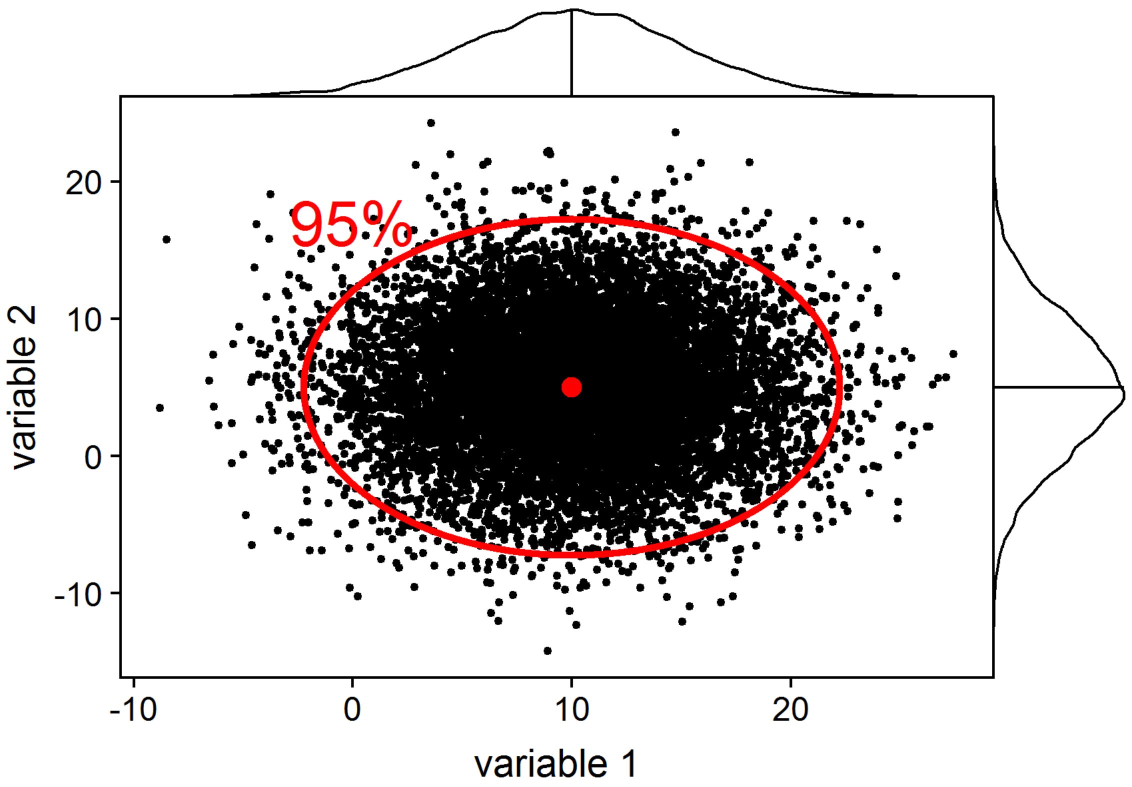
Two variables



Modeling with Mixture Models

1. Which is the suitable probability distribution?
   - Bivariate Gaussian distribution
2. How many sub-populations should we consider?
   - Two clusters
3. Which are the parameters and their estimations?
   - The means (now in 2 dimension), the "standard deviation" (now a matrix) and the proportions
   - `flexmix` for the estimations

Bivariate Gaussian distribution




```
> mean
```

```
[1] 10  5
```

```
> covariance_matrix
```

```
     [,1] [,2]
[1,]   25    0
[2,]    0   25
```

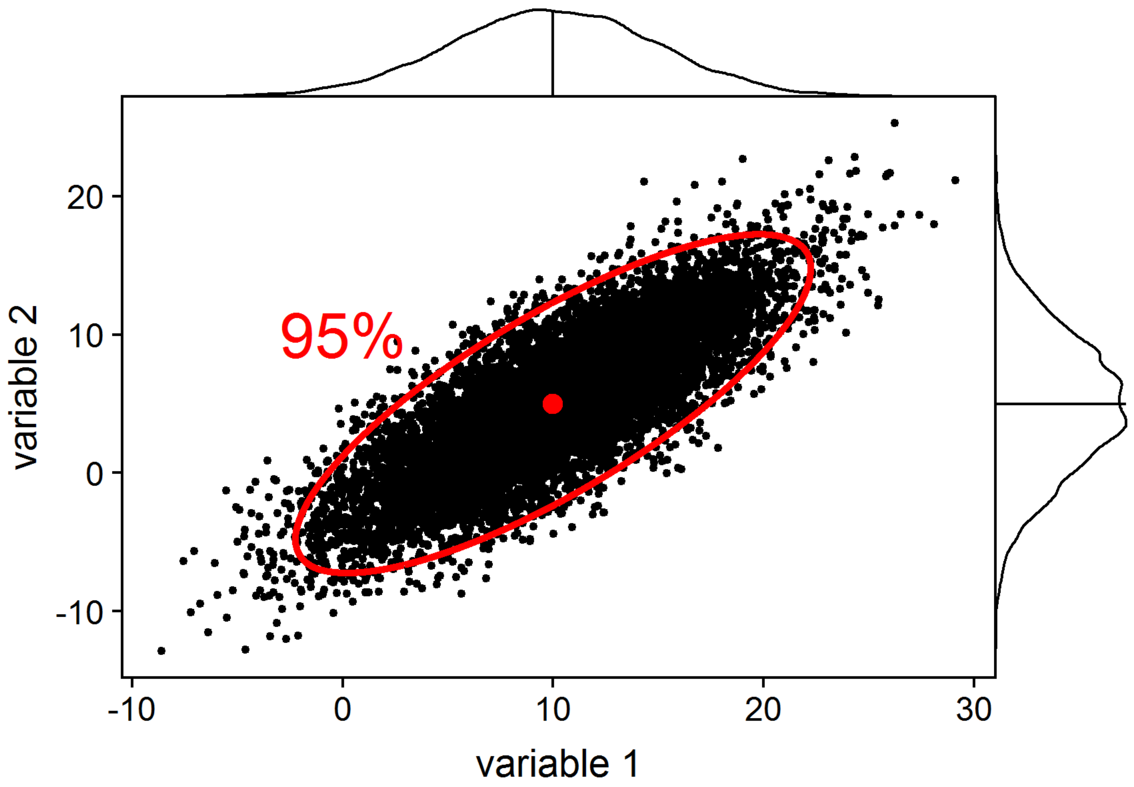
Bivariate Gaussian distribution

```
> mean
```

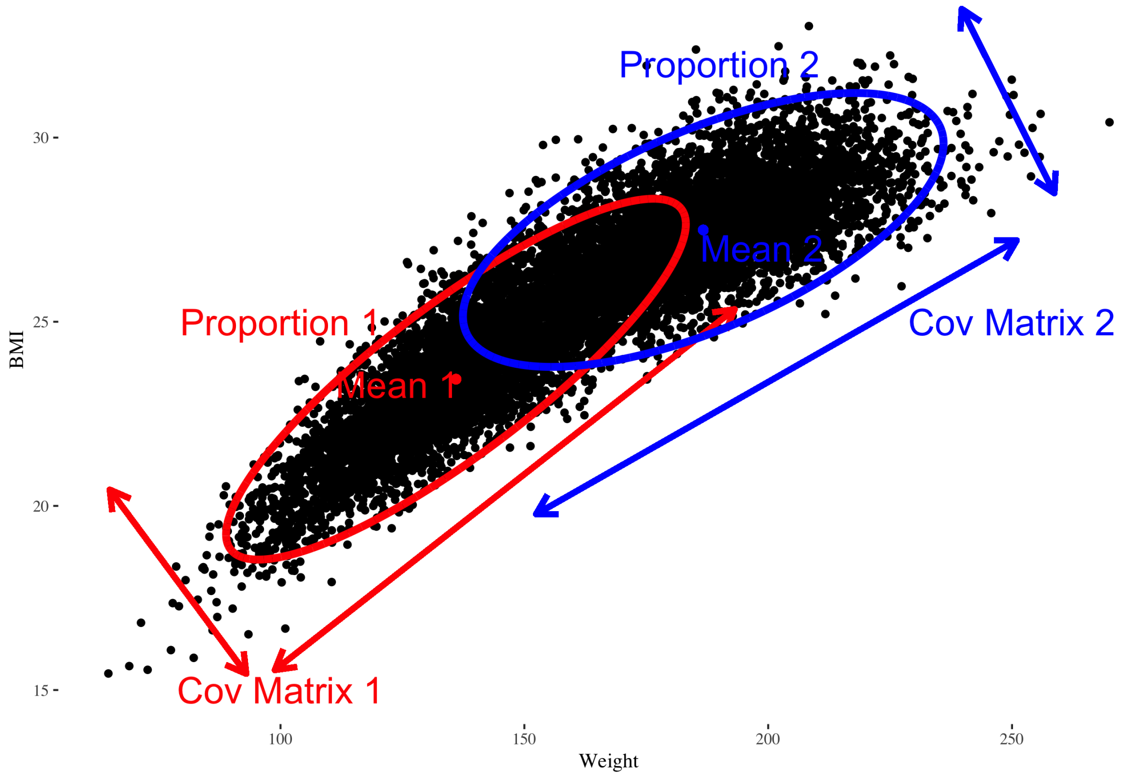
```
[1] 10  5
```

```
> covariance_matrix
```

```
     [,1] [,2]
[1,]   25   20
[2,]   20   25
```



Coming back to the Gender data



1. Which distribution?
    - Bivariate Gaussian distribution
2. How many clusters?
    - Two
3. Which parameters?
    - The proportions
    - The means
    - The covariance matrices

MIXTURE MODELS IN R

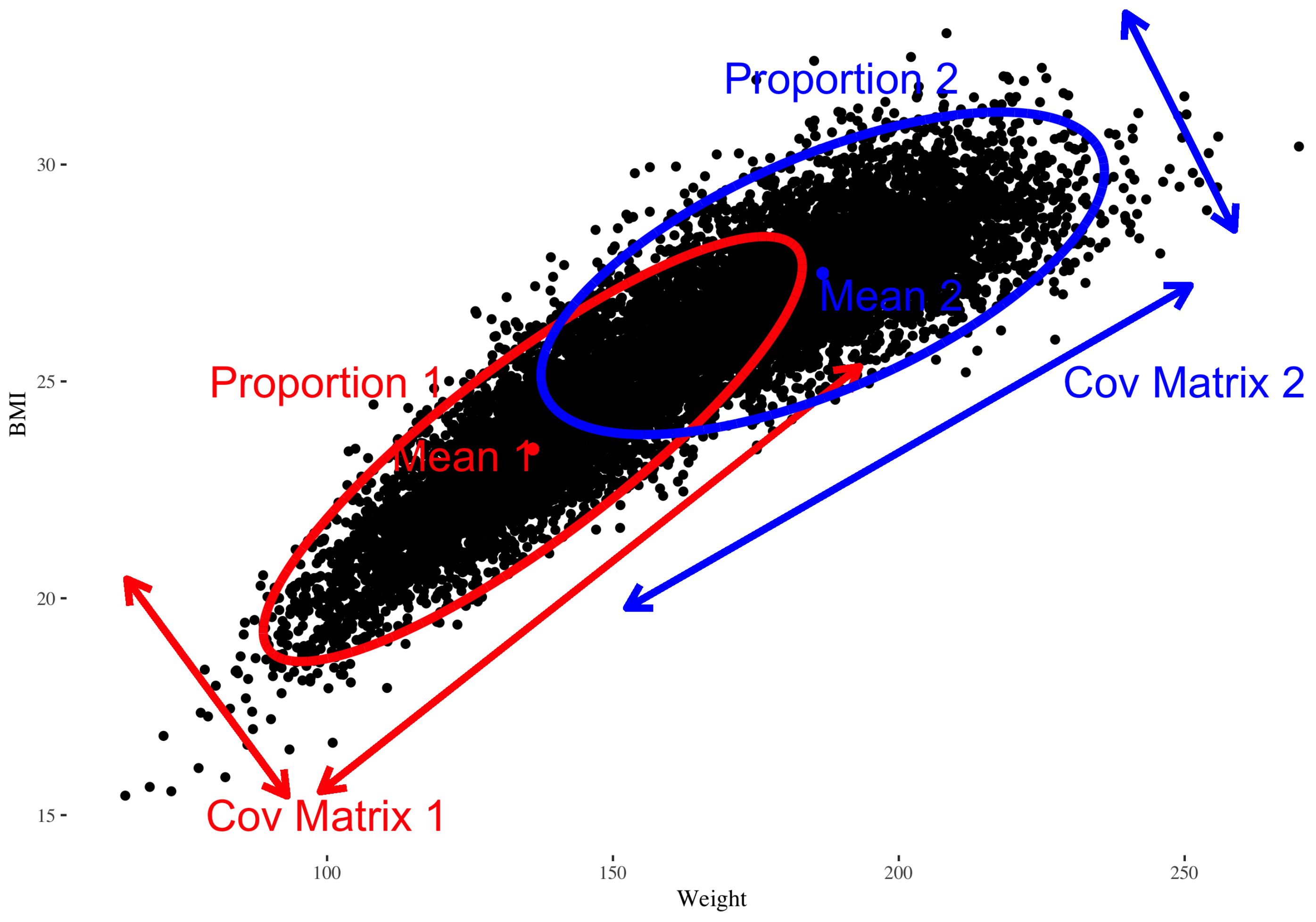
Let's practice!

MIXTURE MODELS IN R

Bivariate Gaussian Mixture Models with flexmix

Victor Medina
Researcher at SBIF

Bivariate Gaussian Mixture Model



Fit bivariate Gaussian Mixture Model

Covariance matrices without cross-terms

```
> fit_without_cov <- flexmix(cbind(Weight, BMI) ~ 1,
+                    k = 2,
+                    data = gender,
+                    model = FLXMCmvnorm(diag = TRUE),
+                    control = list(tolerance = 1e-15, iter.max = 1000))
```

- Formula from `Weight ~ 1` to `cbind(Weight, BMI) ~ 1`
- Model from `FLXMCnorm1()` to `FLXMCmvnorm(diag = TRUE)`

The proportions: prior function

```
> proportions <- prior(fit_without_cov)
> proportions
```

```
[1] 0.5314674 0.4685326
```

parameters function

```
> parameters(fit_without_cov)
```

```
                 Comp.1     Comp.2
center.Weight 186.309154 133.231102
center.BMI     27.521840  23.154197
cov1          366.830490 286.899357
cov2            0.000000   0.000000
cov3            0.000000   0.000000
cov4            2.012768   3.065863
```

Extract the means

```
> # Extract each component
> comp_1 <- parameters(fit_without_corr, component=1)
> comp_2 <- parameters(fit_without_corr, component=2)

> # Extract the means
> mean_comp_1 <- comp_1[1:2]
> mean_comp_2 <- comp_2[1:2]
```

```
> mean_comp_1
```

```
[1] 186.30915  27.52184
```

```
> mean_comp_2
```

```
[1] 133.2311  23.1542
```

Extract the diagonal covariance matrices

```
> # Extract the covariance matrices
> covariance_comp_1 <- matrix(comp_1[3:6], nrow=2)
> covariance_comp_2 <- matrix(comp_2[3:6], nrow=2)
> covariance_comp_1
```

```
         [,1]     [,2]
[1,] 366.8305 0.000000
[2,]   0.0000 2.012768
```

```
> covariance_comp_2
```

```
         [,1]     [,2]
[1,] 286.8994 0.000000
[2,]   0.0000 3.065863
```

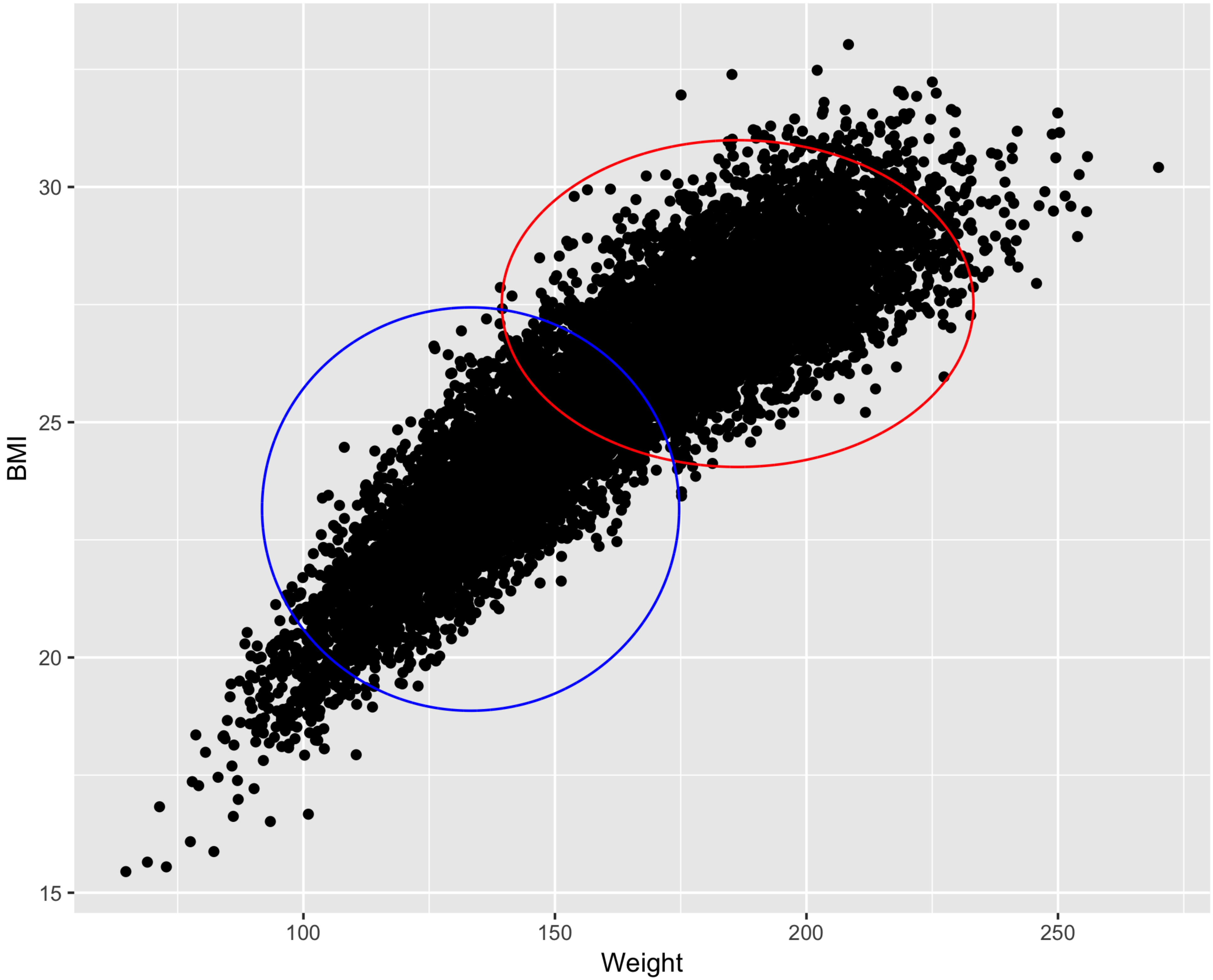
Ellipse curves

```
> library(ellipse)
> # ellipse curve for component 1
> ellipse_comp_1 <- ellipse(x = covariance_comp_1,
+                           centre = mean_comp_1,
+                           npoints = nrow(gender))
> # ellipse curve for component 2
> ellipse_comp_2 <- ellipse(x = covariance_comp_2,
+                           centre = mean_comp_2,
+                           npoints = nrow(gender))
> head(ellipse_comp_1)
```

```
            x        y
[1,] 219.4592 29.97739
[2,] 219.4384 29.97893
[3,] 219.4175 29.98047
[4,] 219.3967 29.98201
[5,] 219.3758 29.98355
[6,] 219.3549 29.98509
```

Visualize the resulting distributions

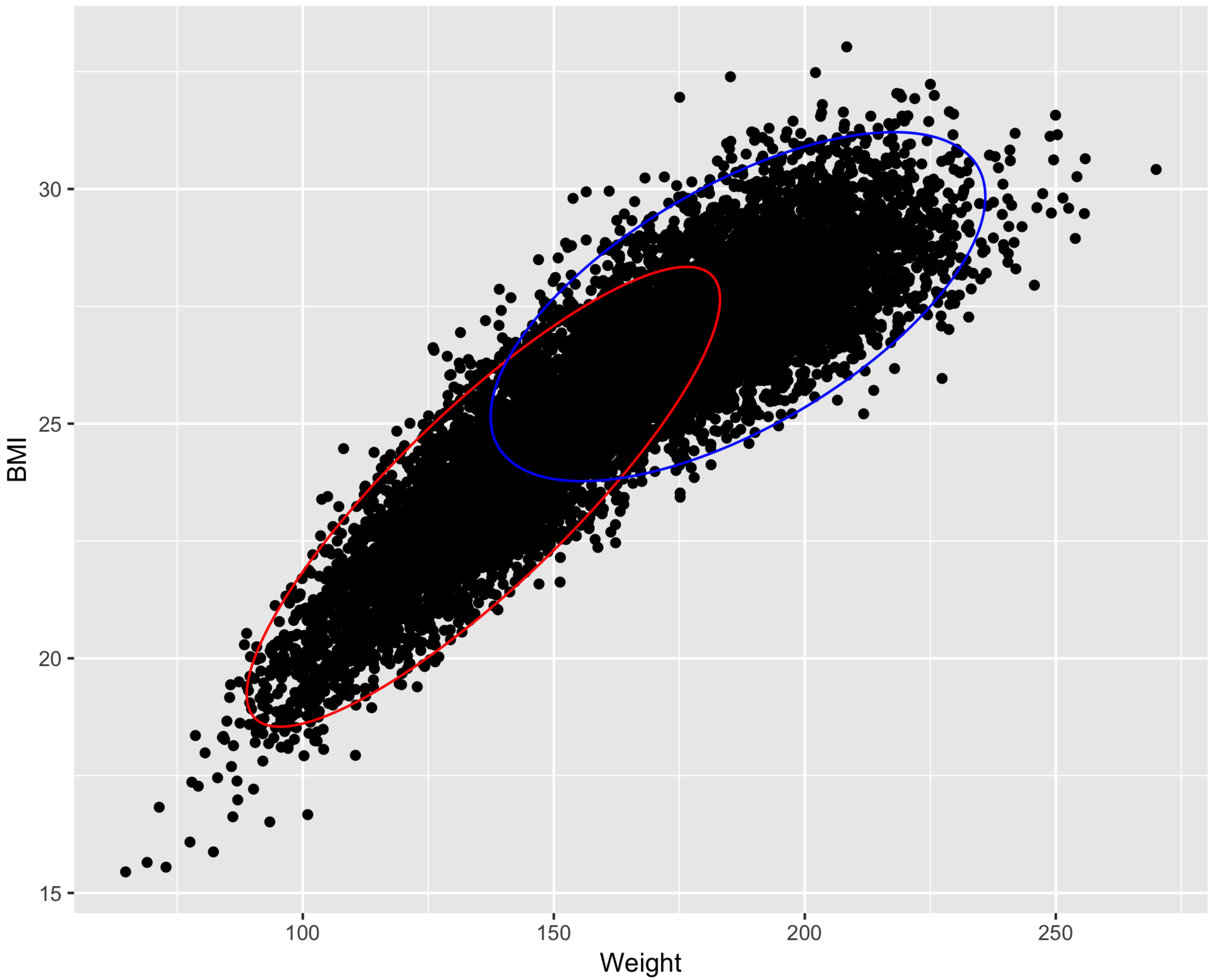
```
> gender %>%
+   ggplot(aes(x = Weight, y = BMI)) + geom_point() +
+   geom_path(data = data.frame(ellipse_comp_1), aes(x=x,y=y), col = "red") +
+   geom_path(data = data.frame(ellipse_comp_2), aes(x=x,y=y), col = "blue")
```


BMI
30
25
20
15
100
150
200
250
Weight

Fit bivariate Gaussian mixture model

Covariance matrices with cross-terms

```
> fit_with_corr <- flexmix(cbind(Weight,BMI) ~ 1,
+               k = 2,
+               data = gender,
+               model = FLXMCmvnorm(diag = FALSE),
+               control = list(tolerance = 1e-15, iter.max = 1000))
```


BMI
30
25
20
15
100
150
200
250
Weight

MIXTURE MODELS IN R

Let's practice!