# Data Mining Tutorial

## Session 5: Clustering

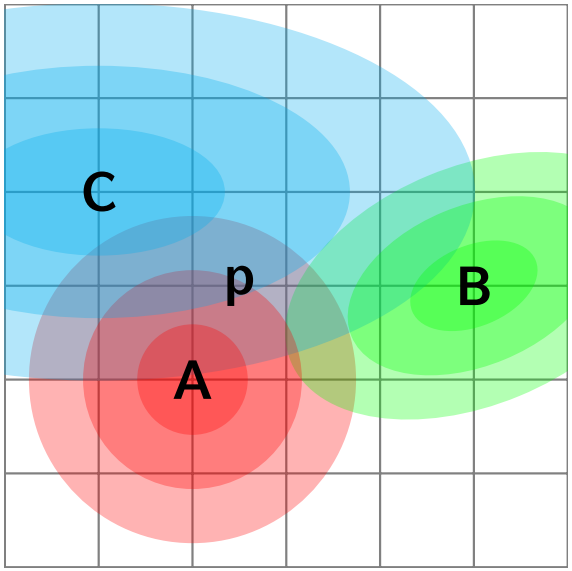
Erich Schubert, Eirini Ntoutsi

Ludwig-Maximilians-Universität München

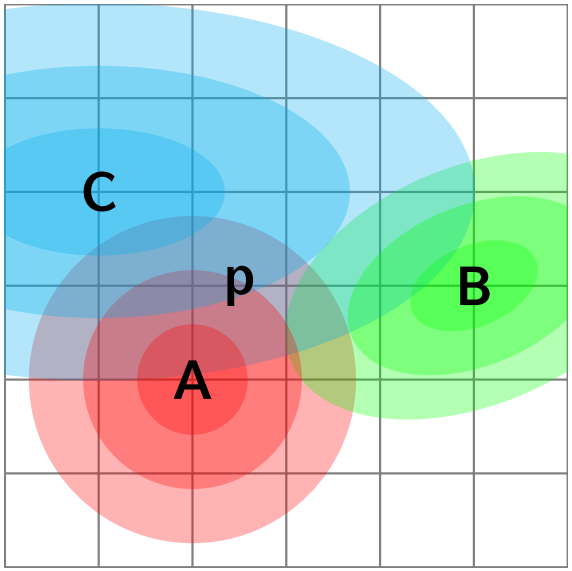
2012-06-28 — KDD class tutorial

# Multivariate normal distribution





# Multivariate normal distribution

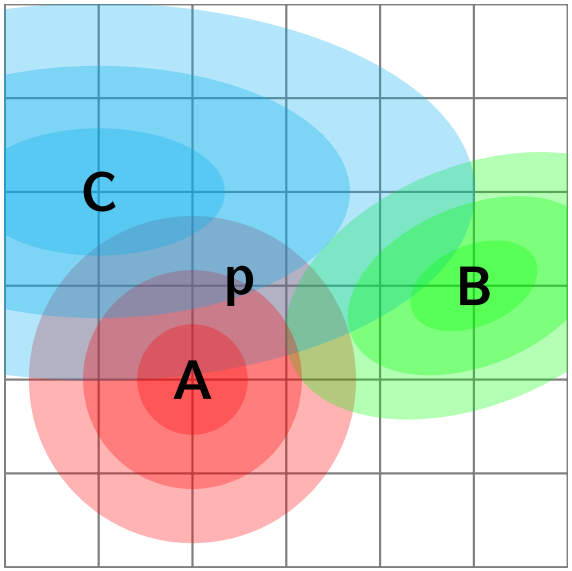



Clusters can be:

- Spherical (A)
- Ellipsoid (C)
- Rotated ellipsoid (B)

But: same formula for each situation!

# Multivariate normal distribution

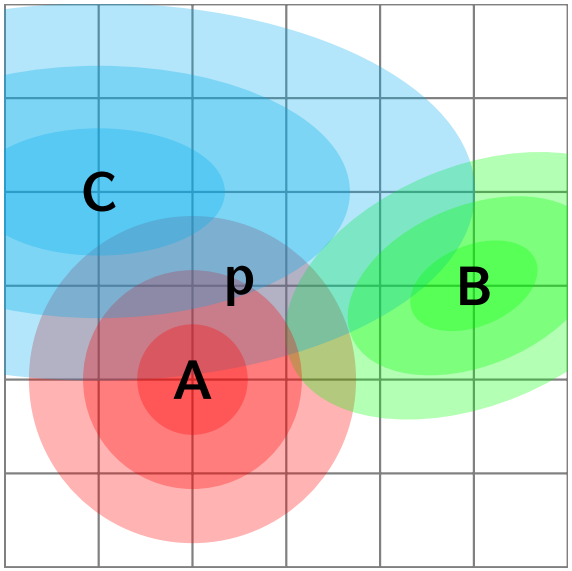



Probability Density Function of multivariate normal distribution:

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

The most essential multi-dimensional distribution!

# Multivariate normal distribution




Let us look at this in more detail!

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

normalization and squared distance from mean

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

normalization and squared distance from mean

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis distance:

$$d_{Mahalanobis}(x, \mu, \Sigma) := \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis distance:

$$d_{Mahalanobis}(x, \mu, \Sigma)^2 := (x-\mu)^T \Sigma^{-1} (x-\mu)$$

# Multivariate normal distribution

## Dissecting the formula

$$pdf(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

One-dimensional normal distribution

$$pdf(x, \mu, \sigma) := \frac{1}{\sqrt{(2\pi)\sigma^2}} \cdot e^{-\frac{1}{2}\left((x-\mu)\sigma^{-2}(x-\mu)\right)}$$

Mahalanobis distance:

$$d_{Mahalanobis}(x, \mu, \Sigma)^2 := (x-\mu)^T \Sigma^{-1} (x-\mu)$$

What is the role of $\Sigma^{-1}$?

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Covariance matrix is symmetric, and non-negative on diagonal, therefore can usually be inverted (ignore constant dimensions).

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Covariance matrix is symmetric, and non-negative on diagonal, therefore can usually be inverted (ignore constant dimensions).
They can be decomposed (equivalently):

$$V\Lambda V^{-1} = \Sigma \quad \equiv \quad V\Lambda^{-1}V^{-1} = \Sigma^{-1}$$

where $V$ contains the Eigenvectors and $\Lambda$ diagonal containing the Eigenvalues.

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Covariance matrix is symmetric, and non-negative on diagonal, therefore can usually be inverted (ignore constant dimensions).
They can be decomposed (equivalently):

$$V\Lambda V^{-1} = \Sigma \quad \equiv \quad V\Lambda^{-1}V^{-1} = \Sigma^{-1}$$

where $V$ contains the Eigenvectors and $\Lambda$ diagonal containing the Eigenvalues. $V \cong$ rotation, $\Lambda \cong$ scaling!
(This is the key idea of Principal Component Analysis!)

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$d^2_{\text{Mahalanobis}} = (x-\mu)^T\Sigma^{-1}(x-\mu)$$

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$d^2_{\text{Mahalanobis}} = (x-\mu)^T V\Omega(V\Omega)^T(x-\mu)$$

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \end{aligned}$$

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \\ &= L_2((V\Omega)^T(x-\mu))^2 \end{aligned}$$

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \\ &= L_2((V\Omega)^T(x-\mu))^2 \end{aligned}$$

$L_2$ is the $L_2$ norm (Euclidean distance $d(x,y) = L_2(x-y)$!)

# Multivariate normal distribution

## Inverse of covariance matrix – $\Sigma^{-1}$

Build $\Omega$ using $\omega_i = 1/\sqrt{\lambda_i} = \lambda_i^{-\frac{1}{2}}$. Then $\Omega\Omega = \Lambda^{-1}$.

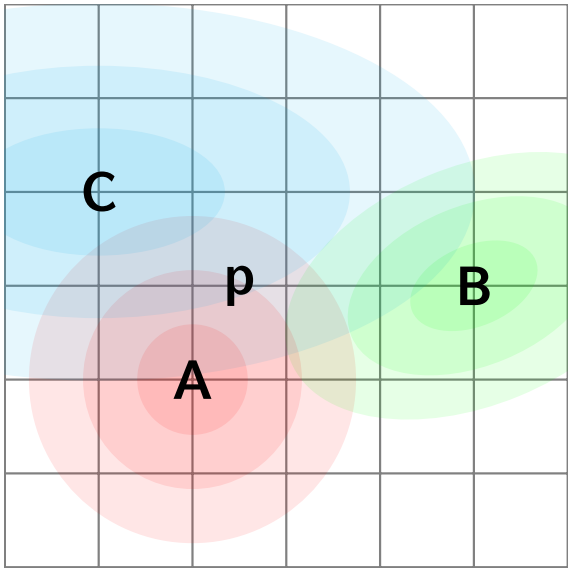
$$\Sigma^{-1} = V\Lambda^{-1}V^{-1} = V\Omega\Omega^T V^T = V\Omega(V\Omega)^T$$

$$\begin{aligned} d^2_{\text{Mahalanobis}} &= (x-\mu)^T V\Omega(V\Omega)^T(x-\mu) \\ &= \left\langle (V\Omega)^T(x-\mu), (V\Omega)^T(x-\mu) \right\rangle \\ &= L_2((V\Omega)^T(x-\mu))^2 \end{aligned}$$

$L_2$ is the $L_2$ norm (Euclidean distance $d(x,y) = L_2(x-y)$!)
Mahalanobis $\approx$ Euclidean distance after PCA!

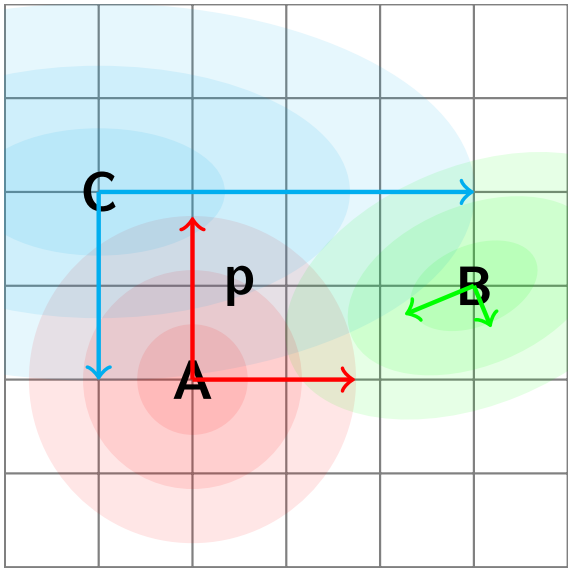
# EM homework



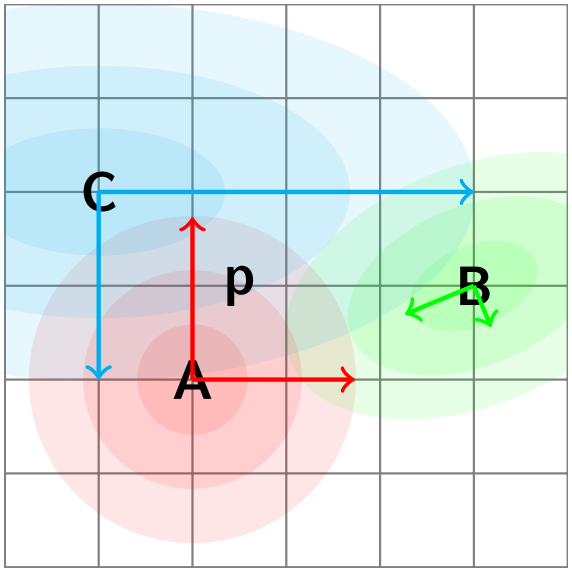


# EM homework





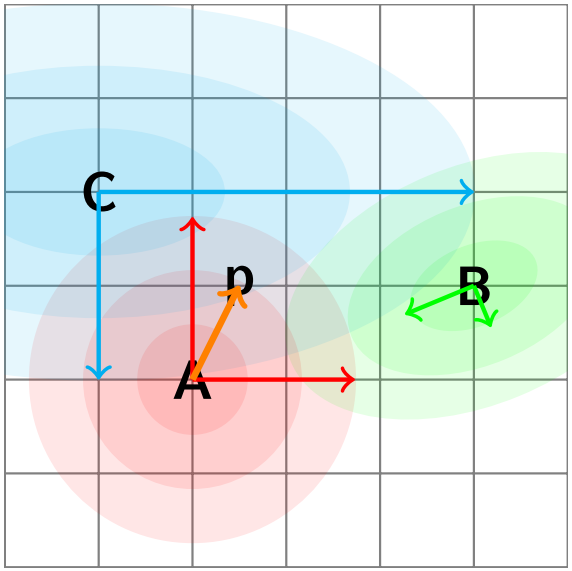
# EM homework




$$\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$$

# EM homework

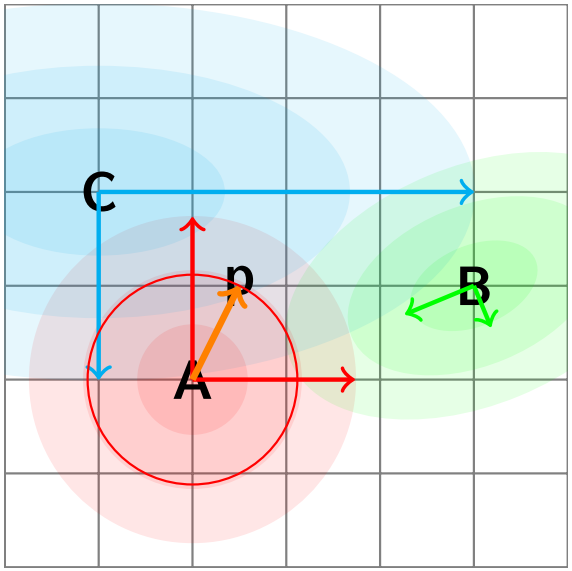





$$\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$$

$$p - \mu_A = (0.5, 1)$$

# EM homework

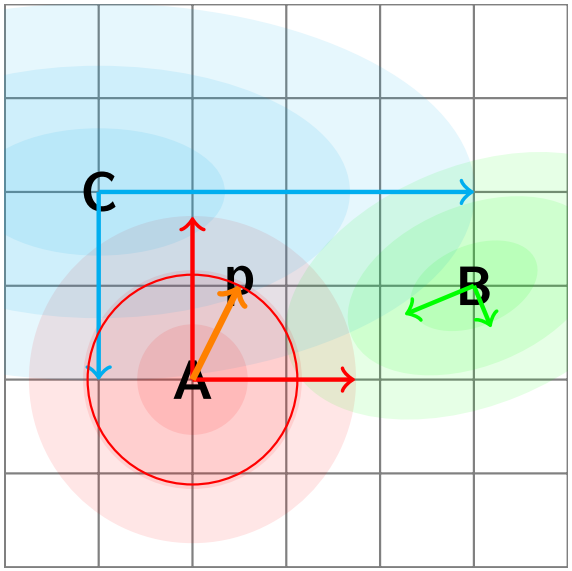



$$\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$$

$$p - \mu_A = (0.5, 1)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 0.41666$$

# EM homework



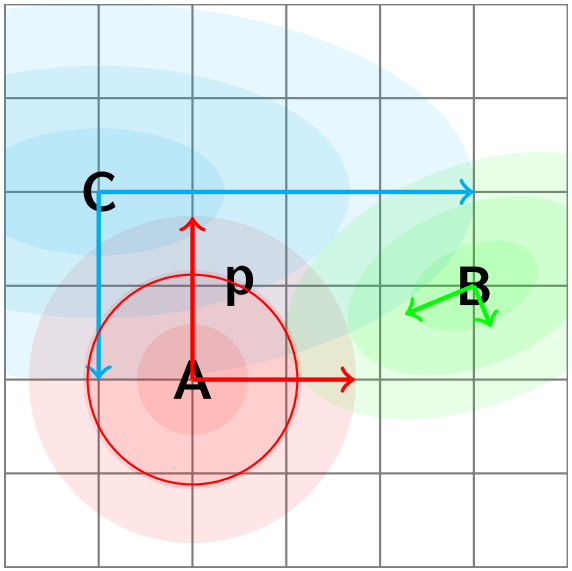

$$\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$$

$$p - \mu_A = (0.5, 1)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 0.41666$$

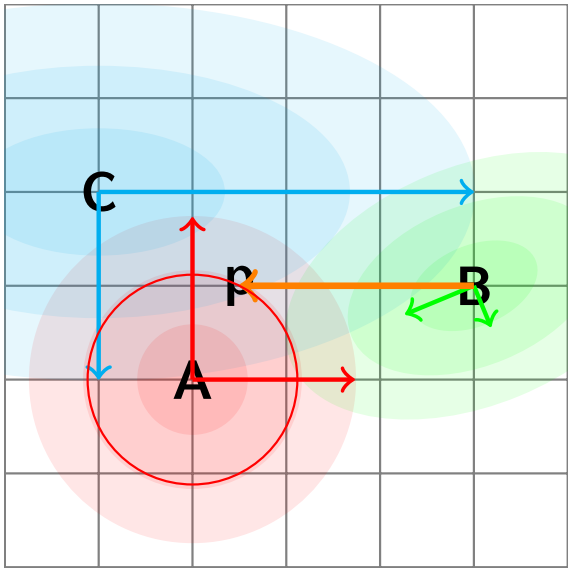
$$prob_A \approx \frac{1}{\sqrt{(2\pi)^2 9}} e^{-\frac{1}{2}0.41666}$$
$$\approx 0.04307456$$

# EM homework




$$\Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix}$$

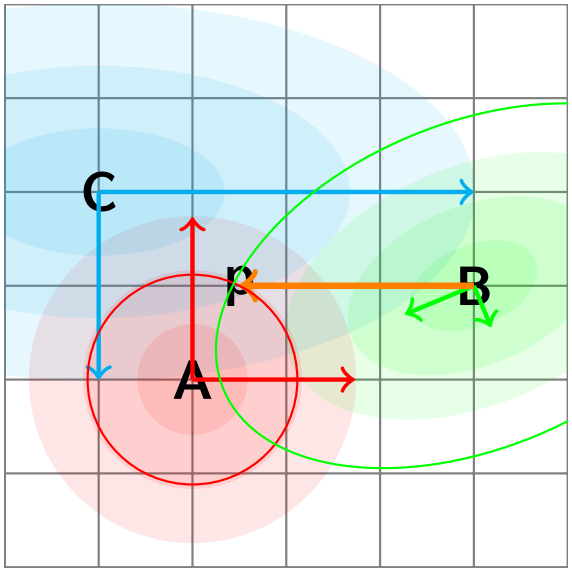
# EM homework




$$\Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix}$$

$$p - \mu_B = (-2.5, 0)$$

# EM homework

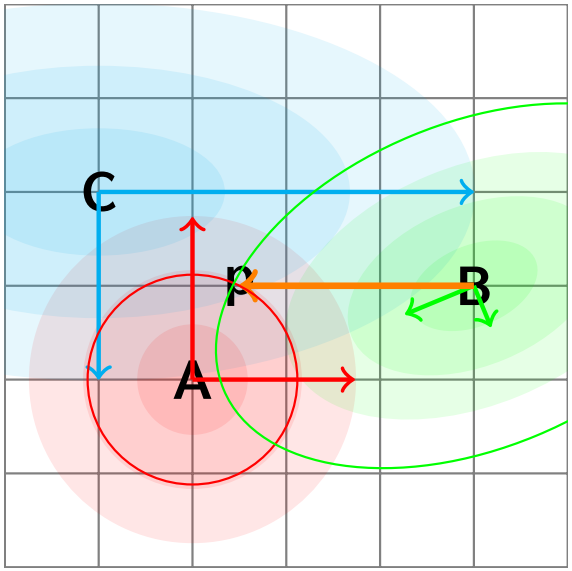



$$\Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix}$$

$$p - \mu_B = (-2.5, 0)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 3.5714285$$

# EM homework



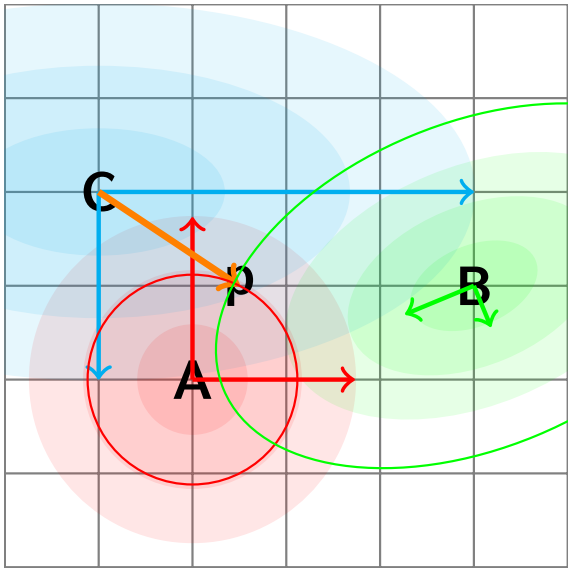

$$\Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix}$$

$$p - \mu_B = (-2.5, 0)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 3.5714285$$

$$prob_B \approx \frac{1}{\sqrt{(2\pi)^2 7}} e^{-\frac{1}{2}3.5714285}$$
$$\approx 0.01008661$$

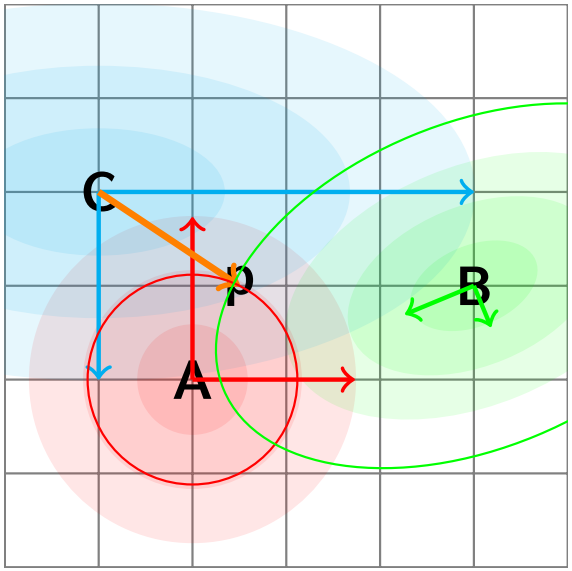
# EM homework




$$\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix} \quad \Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

# EM homework

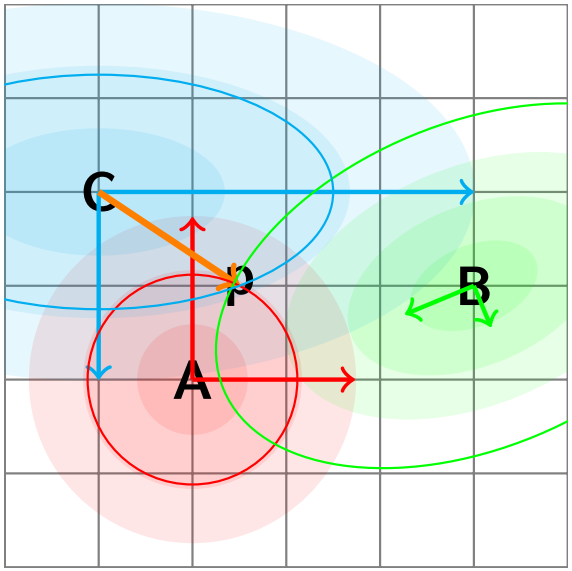





$$\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix} \quad \Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

$$p - \mu_C = (1.5, -1)$$

# EM homework

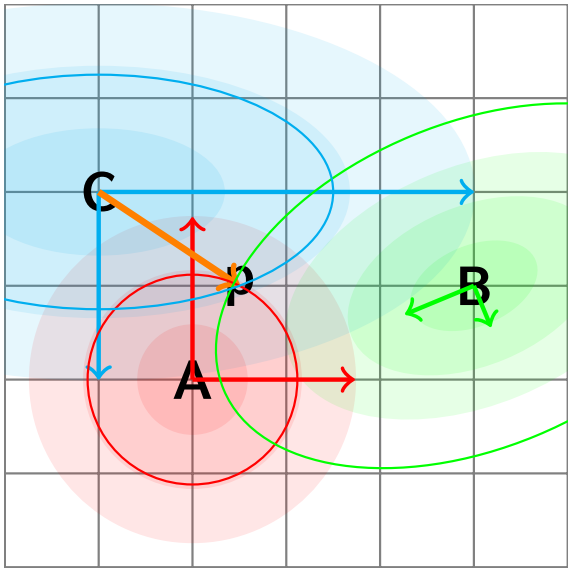



$$\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix} \quad \Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

$$p - \mu_C = (1.5, -1)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 0.390625$$

# EM homework




$$\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix} \quad \Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

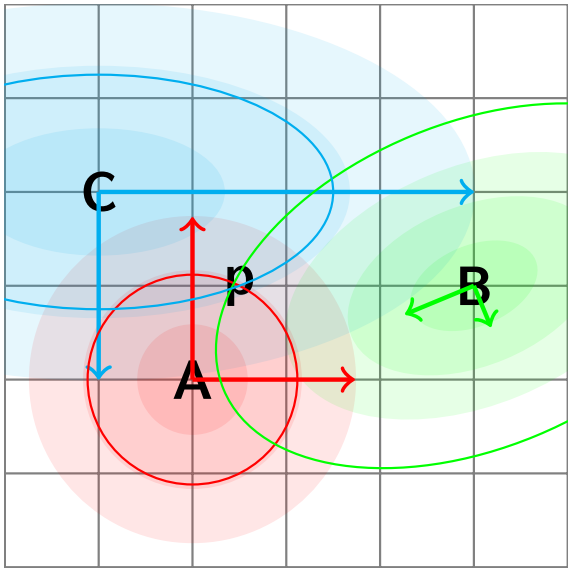
$$p - \mu_C = (1.5, -1)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 0.390625$$

$$prob_C \approx \frac{1}{\sqrt{(2\pi)^2 64}} e^{-\frac{1}{2}0.390625}$$
$$\approx 0.01636466$$

# EM homework

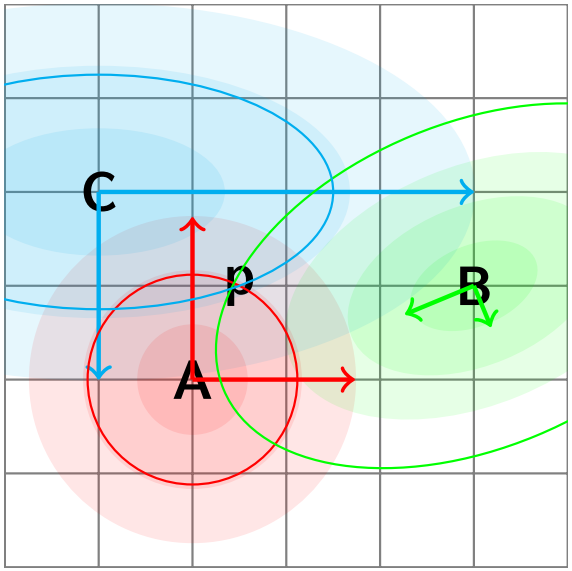





| | $A$ | $B$ | $C$ |
|---|---|---|---|
| *prob* | 0.043075 | 0.010087 | 0.016365 |
| *size* | 30% | 20% | 50% |

# EM homework

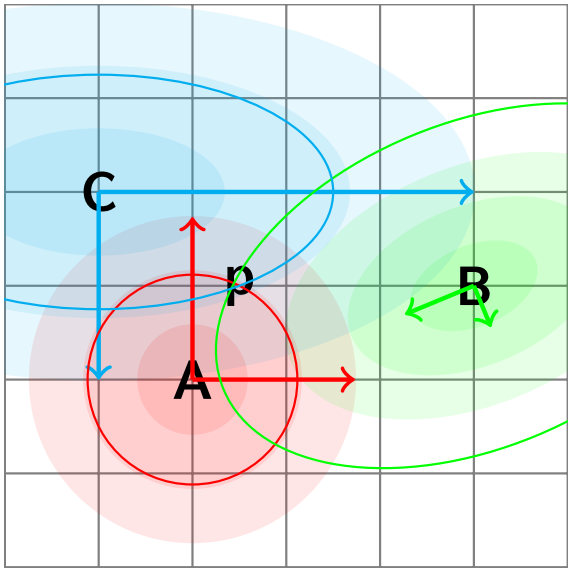





| | $A$ | $B$ | $C$ |
|---|---|---|---|
| *prob* | 0.043075 | 0.010087 | 0.016365 |
| *size* | 30% | 20% | 50% |
| *score* | 0.012922 | 0.002017 | 0.008182 |

# EM homework

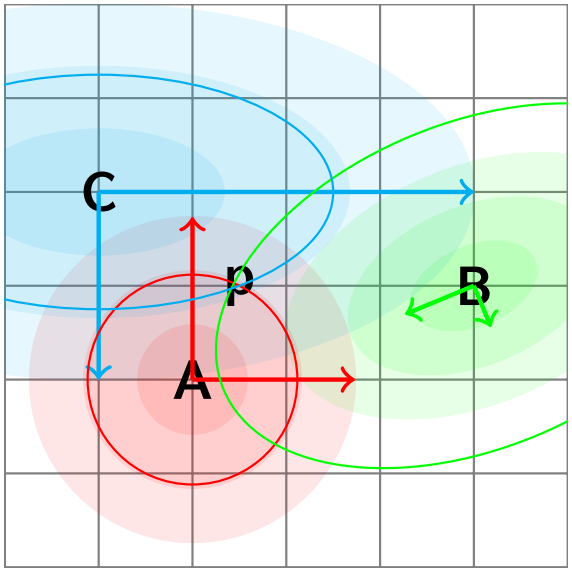





| | $A$ | $B$ | $C$ |
|---|---|---|---|
| *prob* | 0.043075 | 0.010087 | 0.016365 |
| *size* | 30% | 20% | 50% |
| *score* | 0.012922 | 0.002017 | 0.008182 |
| *sum* | divide by 0.023122 | | |

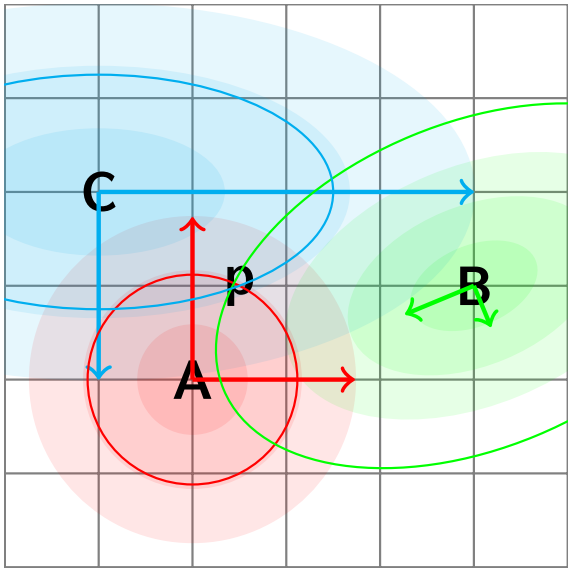
# EM homework






| | $A$ | $B$ | $C$ |
|---|---|---|---|
| *prob* | 0.043075 | 0.010087 | 0.016365 |
| *size* | 30% | 20% | 50% |
| *score* | 0.012922 | 0.002017 | 0.008182 |
| *sum* | divide by 0.023122 | | |
| *weight* | $\approx 55.9\%$ | $\approx 8.2\%$ | $\approx 35.4\%$ |

# EM homework



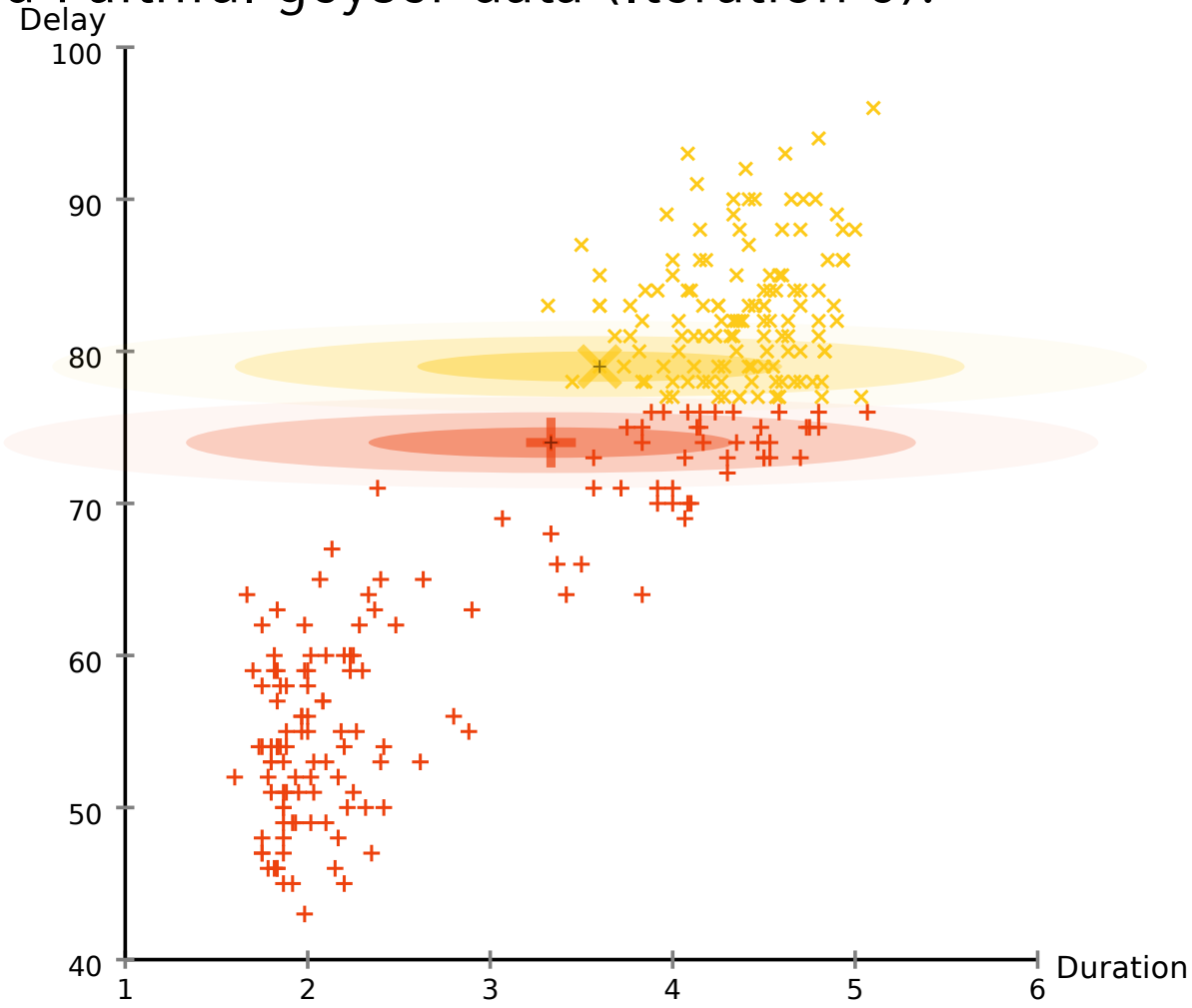

| | $A$ | $B$ | $C$ |
|---|---|---|---|
| *prob* | 0.043075 | 0.010087 | 0.016365 |
| *size* | 30% | 20% | 50% |
| *score* | 0.012922 | 0.002017 | 0.008182 |
| *sum* | divide by 0.023122 | | |
| *weight* | $\approx 55.9\%$ | $\approx 8.2\%$ | $\approx 35.4\%$ |

Point $p$ belongs mostly to cluster $A$!

# EM demonstration

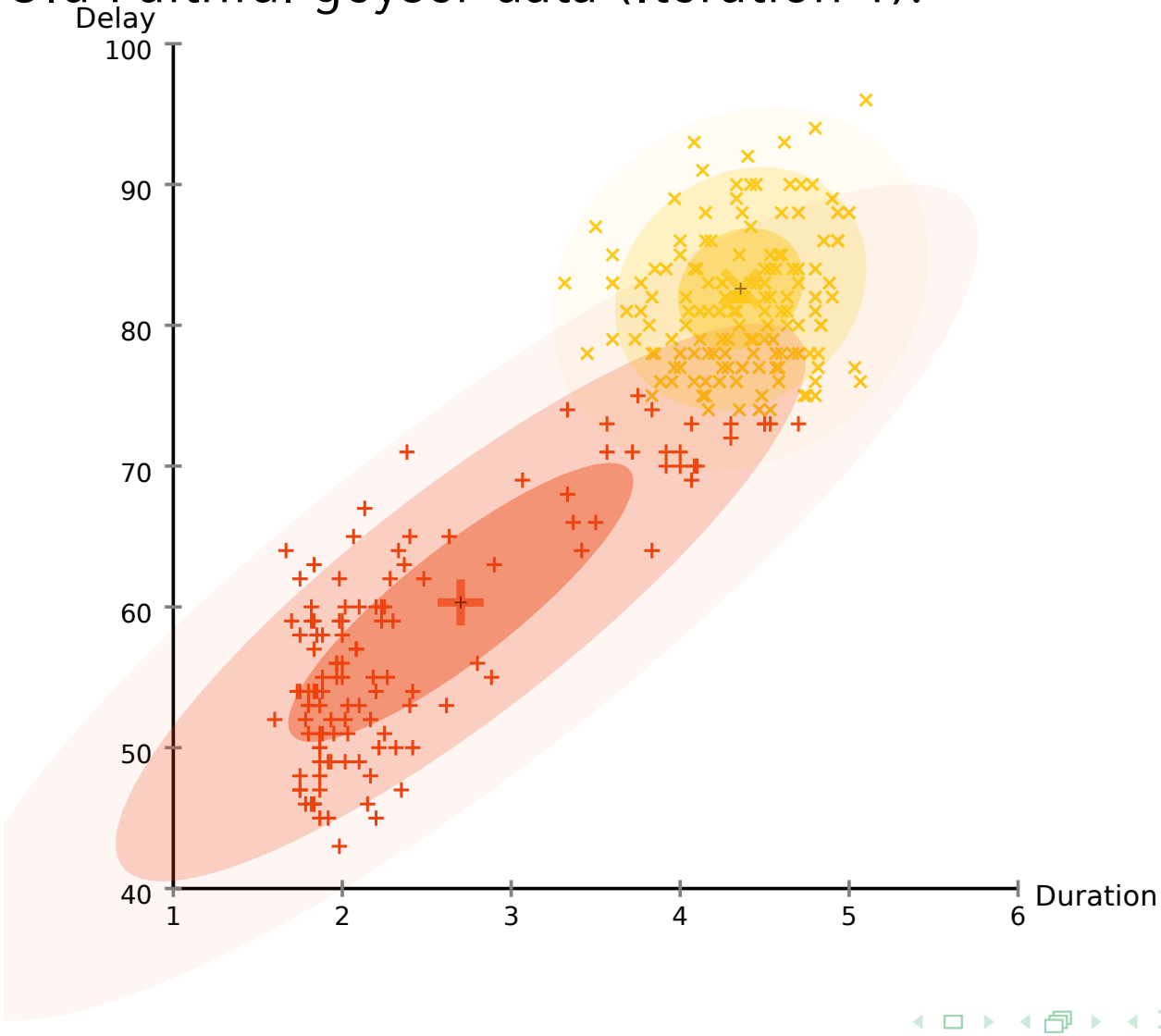


## Old Faithful geyser data (Iteration 0):



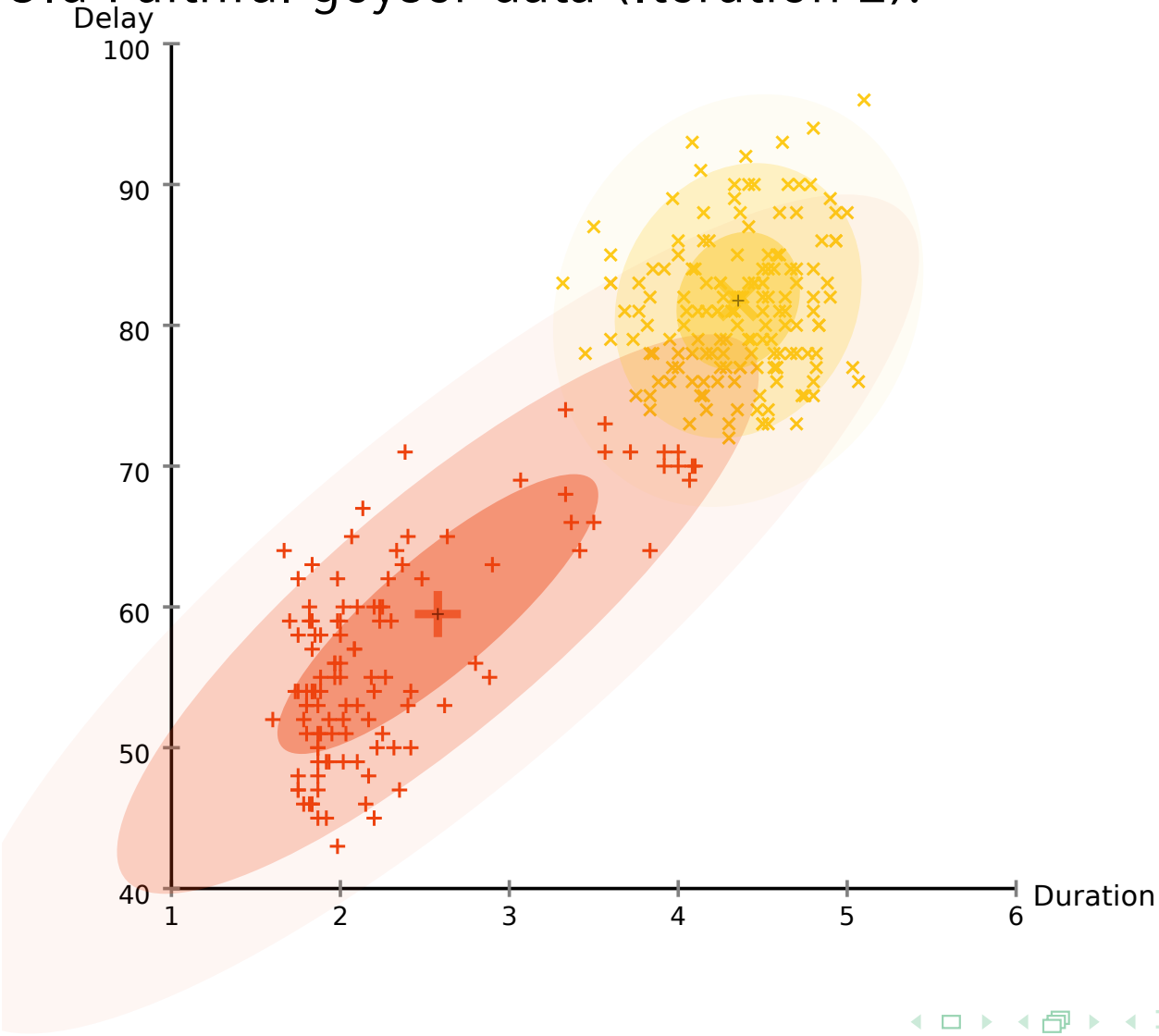
# EM demonstration

Old Faithful geyser data (Iteration 1):



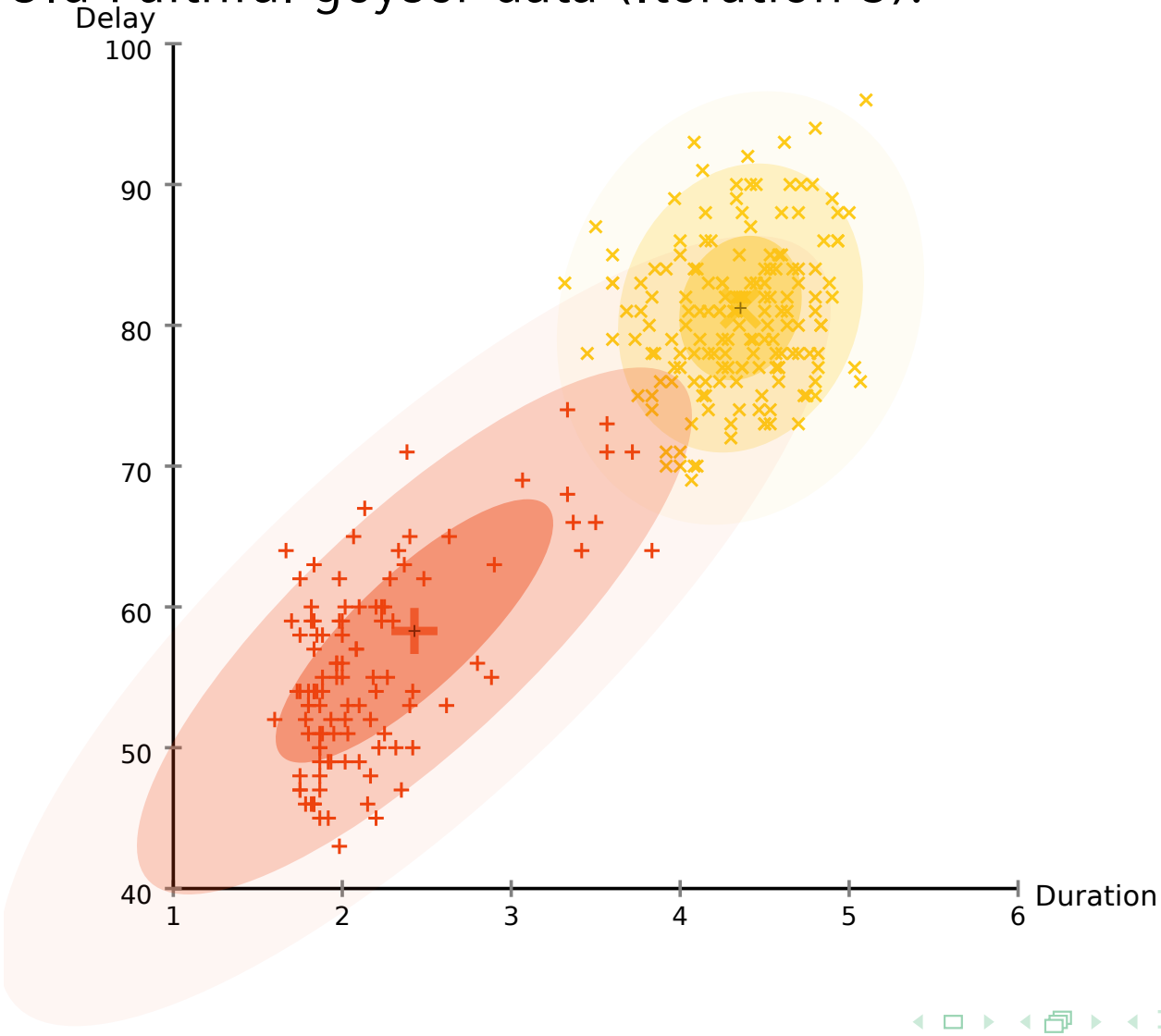
# EM demonstration

Old Faithful geyser data (Iteration 2):



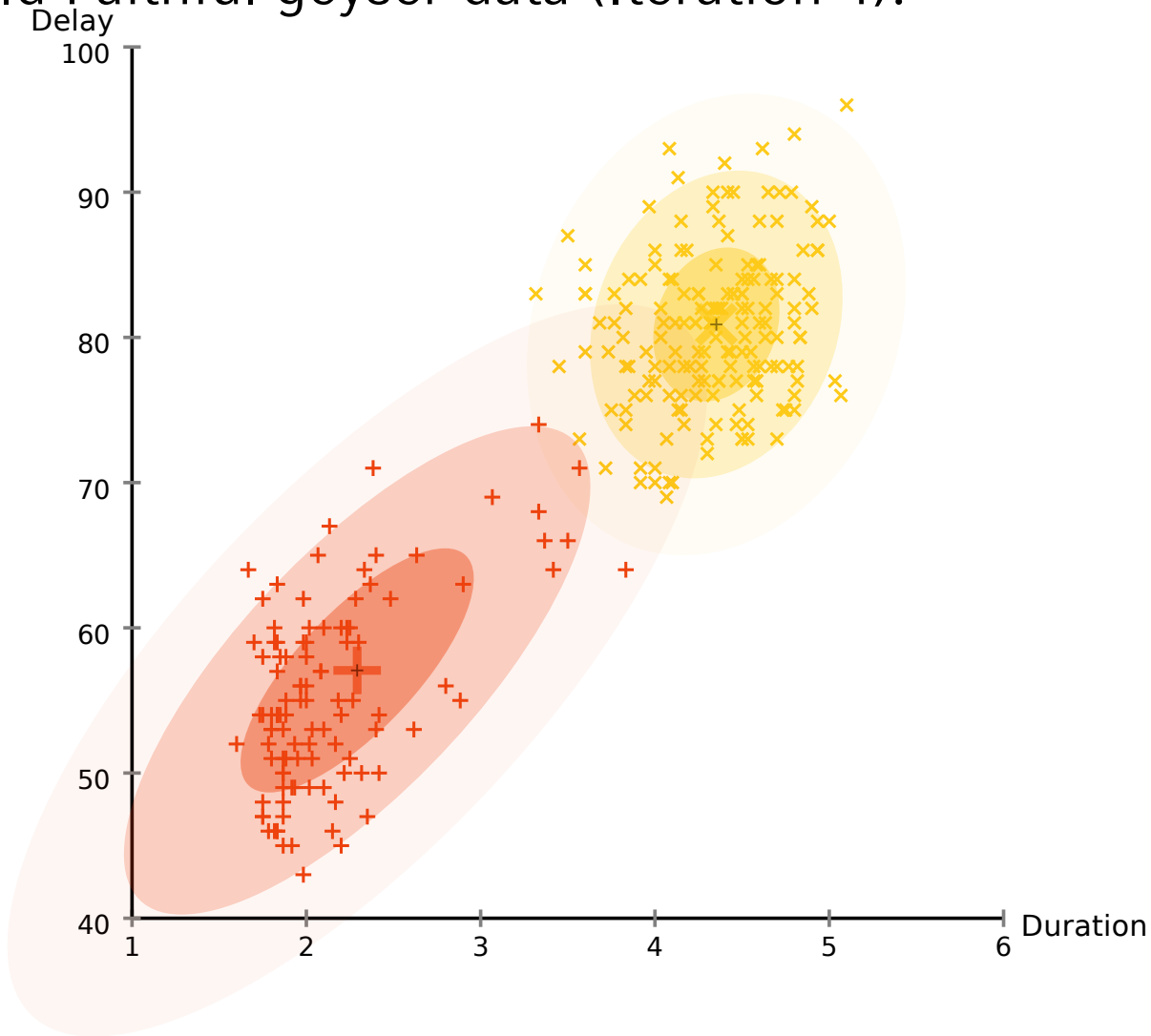
# EM demonstration

Old Faithful geyser data (Iteration 3):



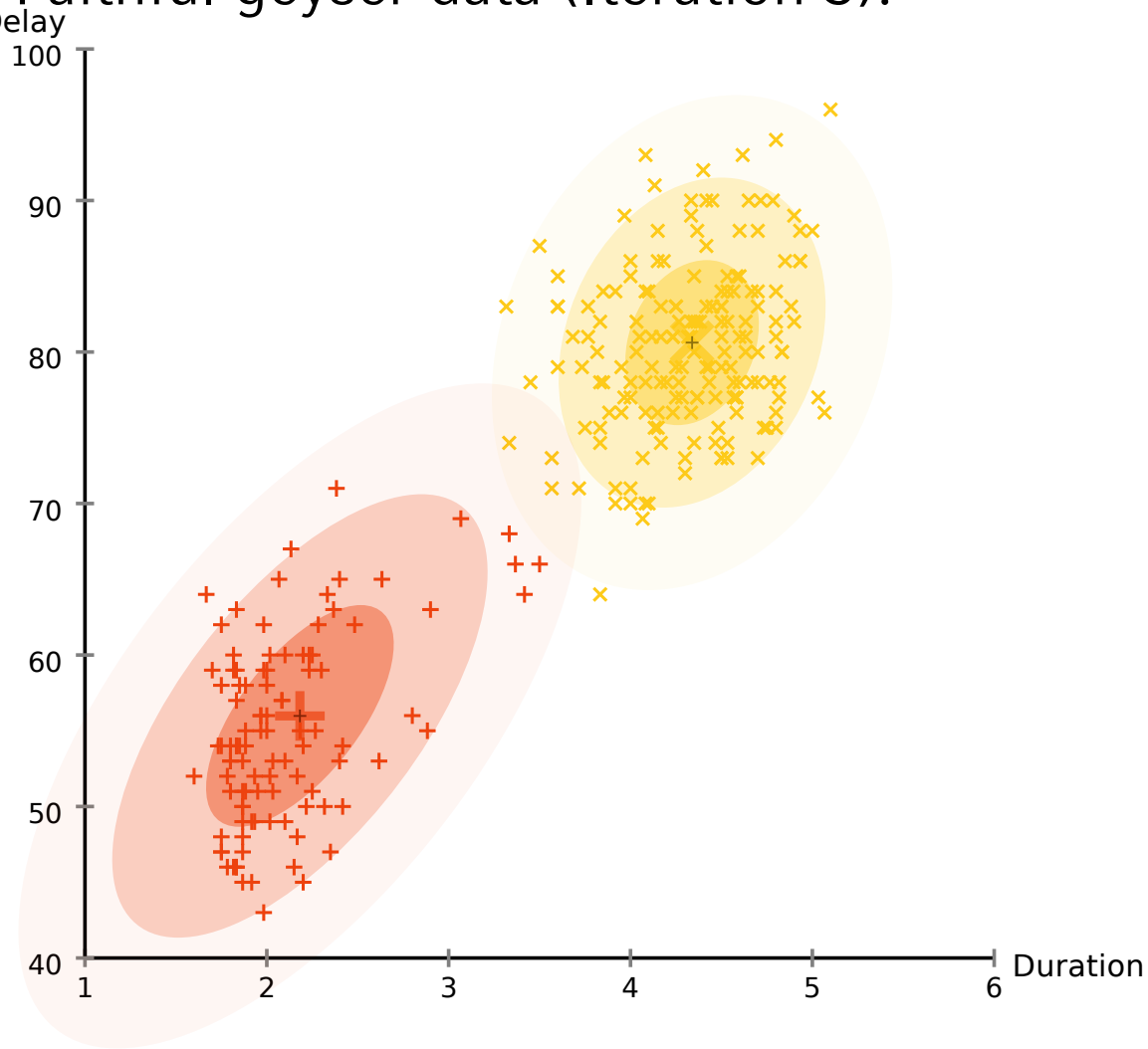
# EM demonstration

Old Faithful geyser data (Iteration 4):



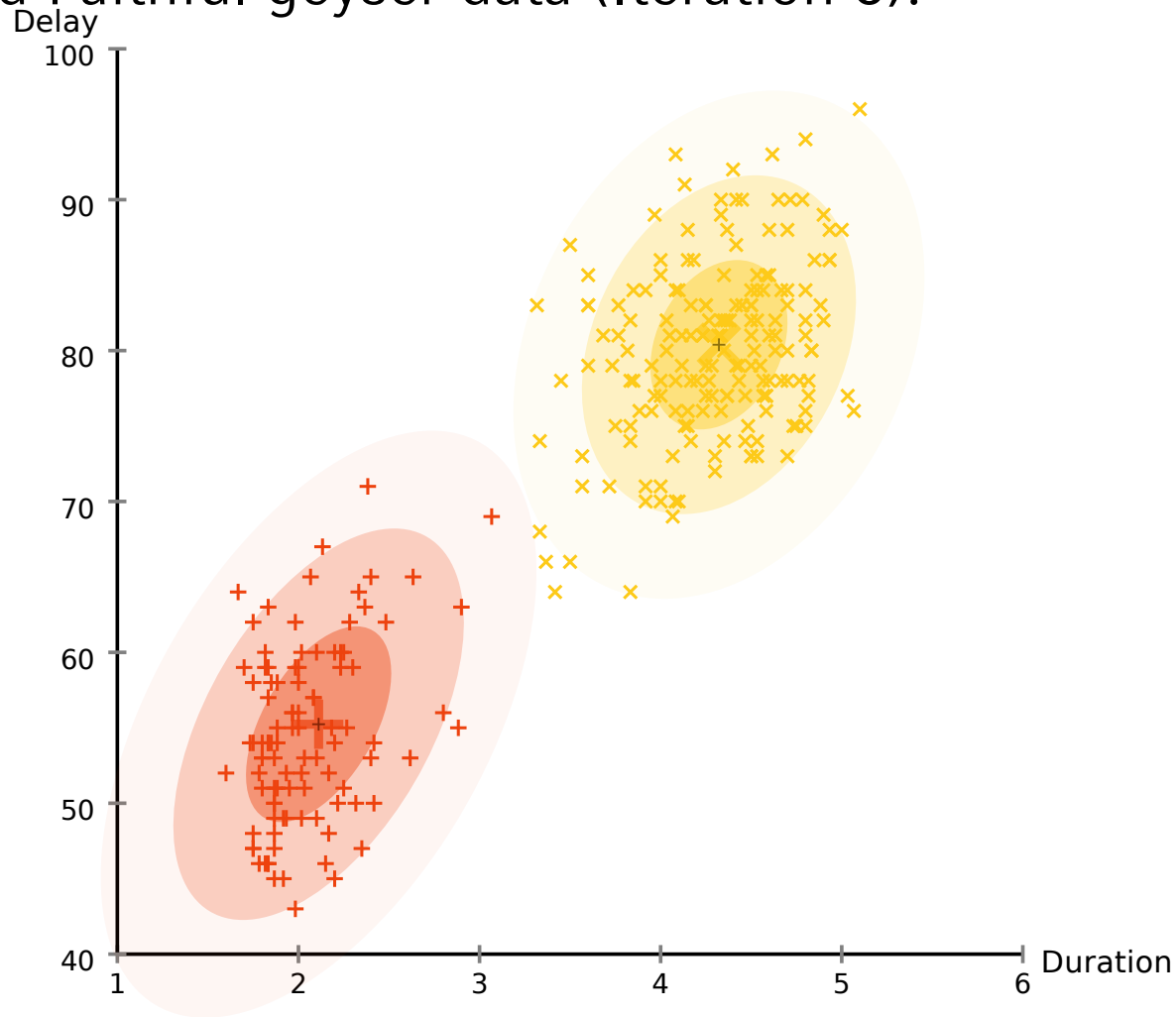
# EM demonstration

## Old Faithful geyser data (Iteration 5):



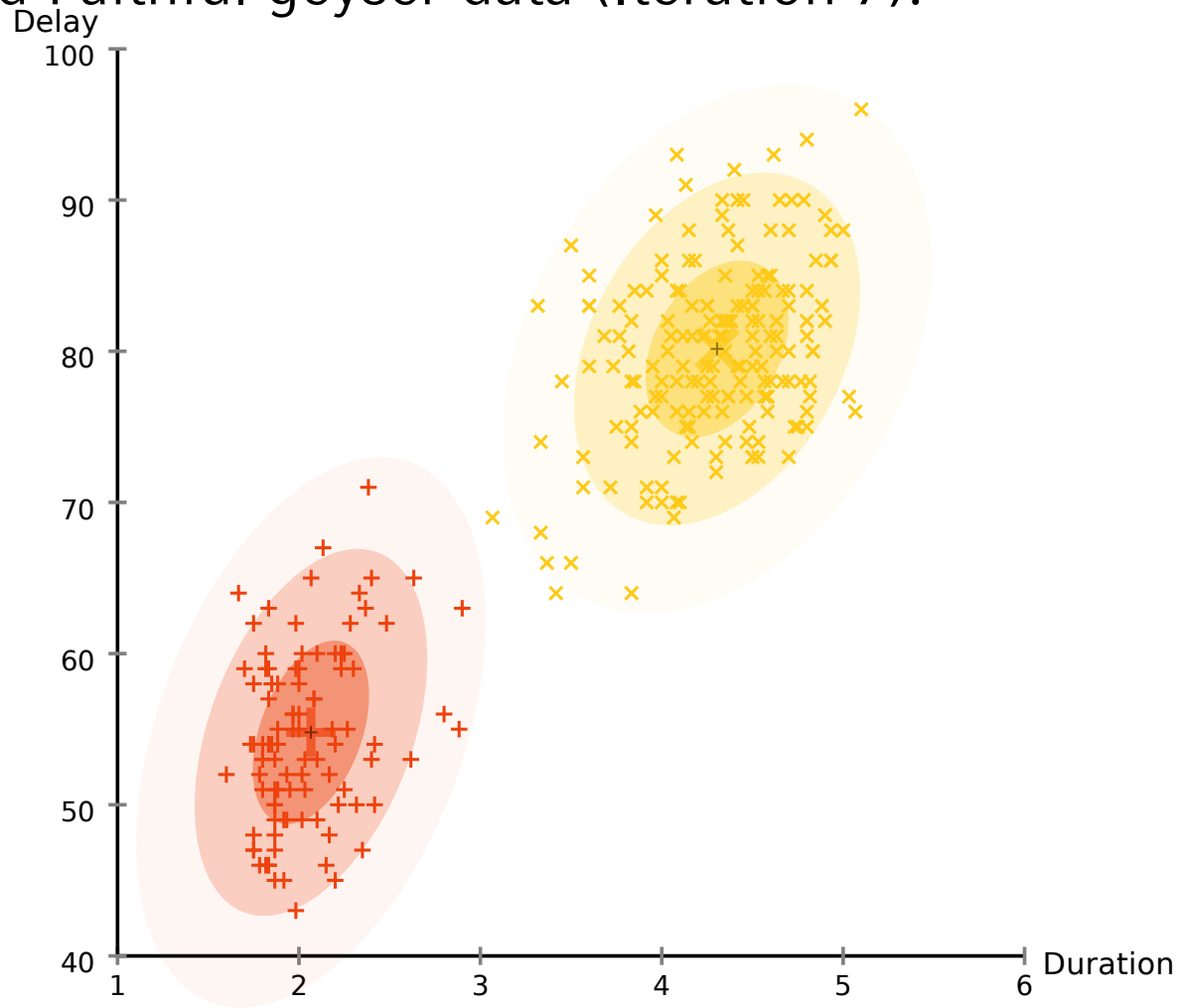
# EM demonstration

Old Faithful geyser data (Iteration 6):



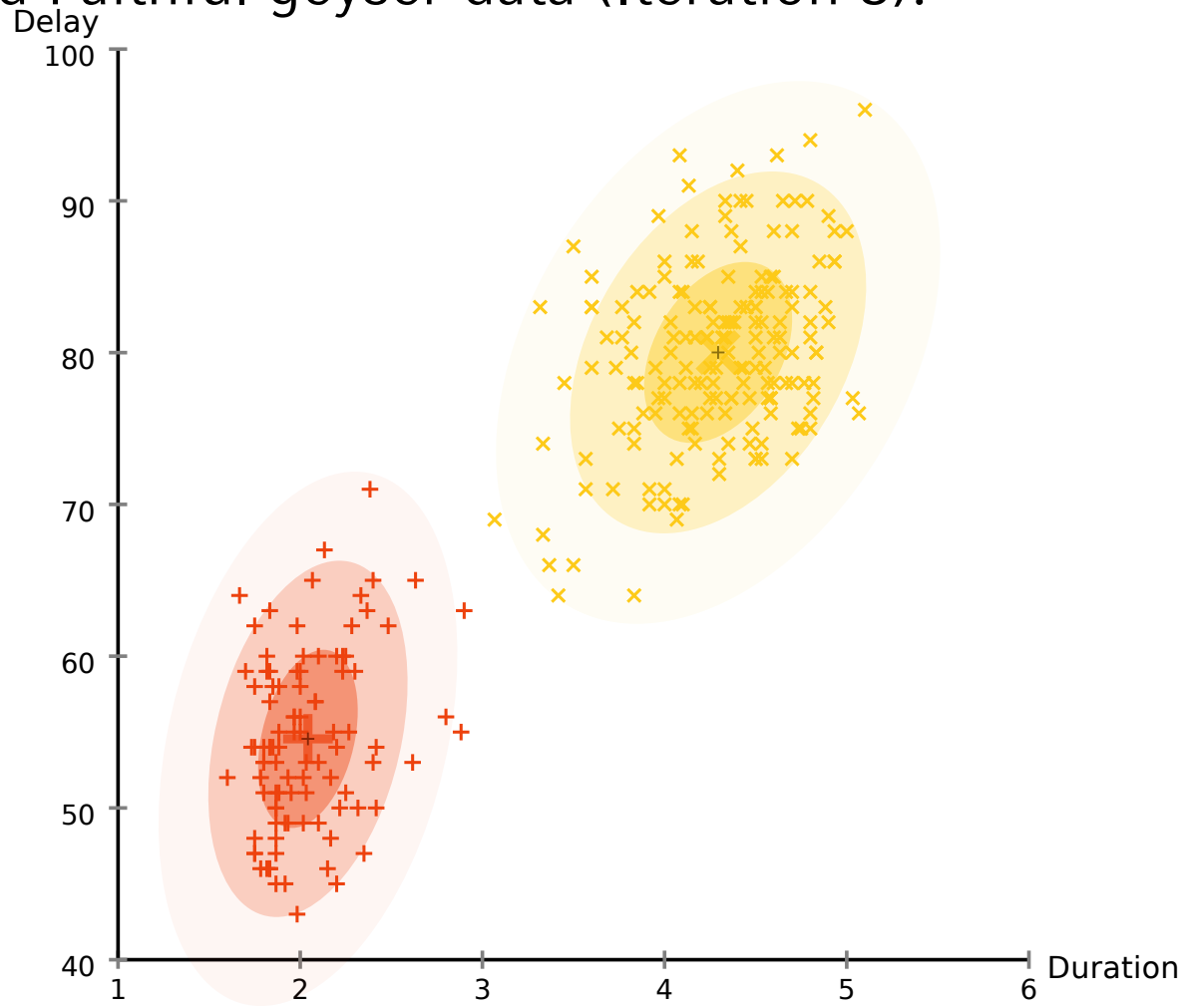
# EM demonstration

Old Faithful geyser data (Iteration 7):



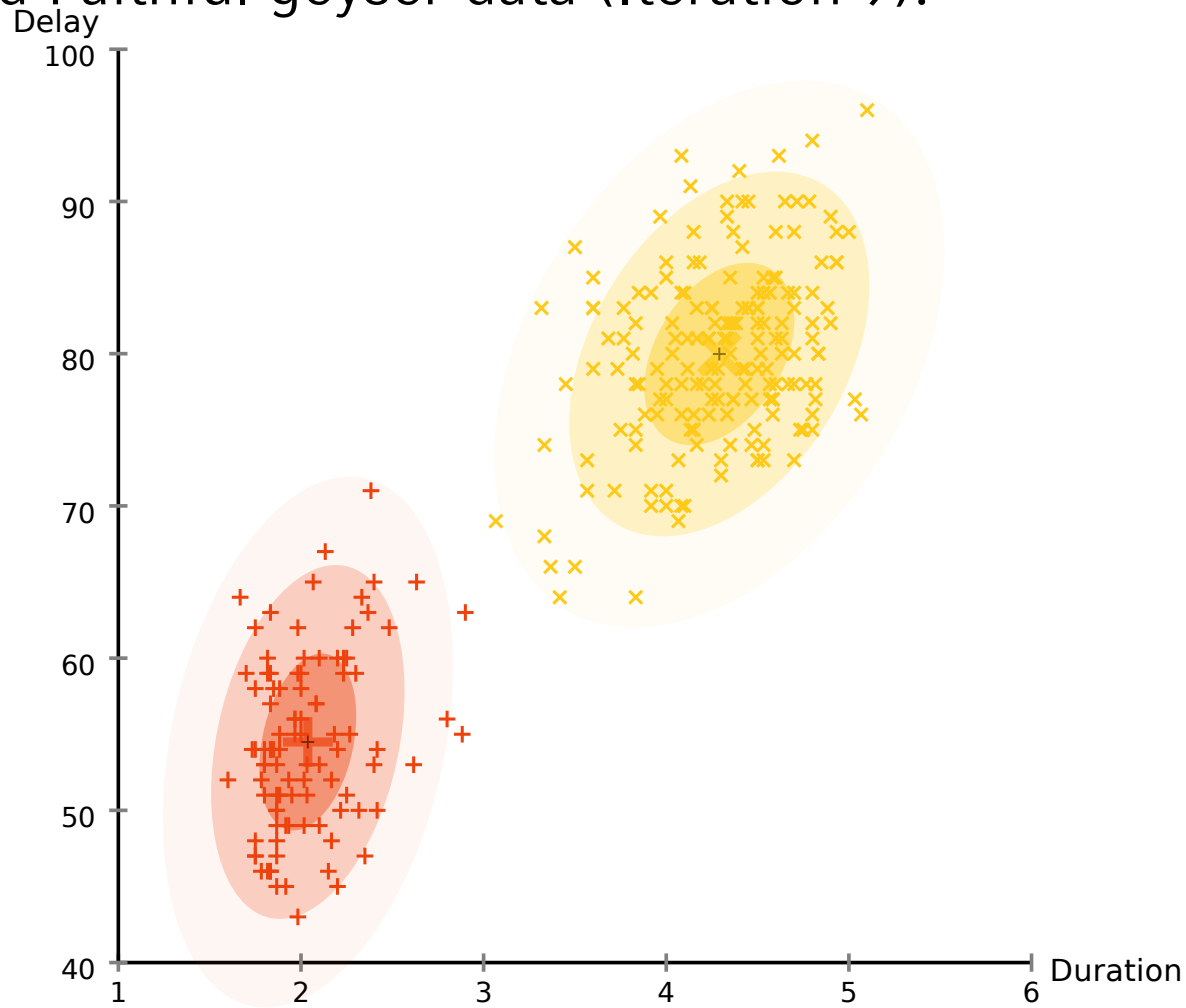
# EM demonstration

## Old Faithful geyser data (Iteration 8):



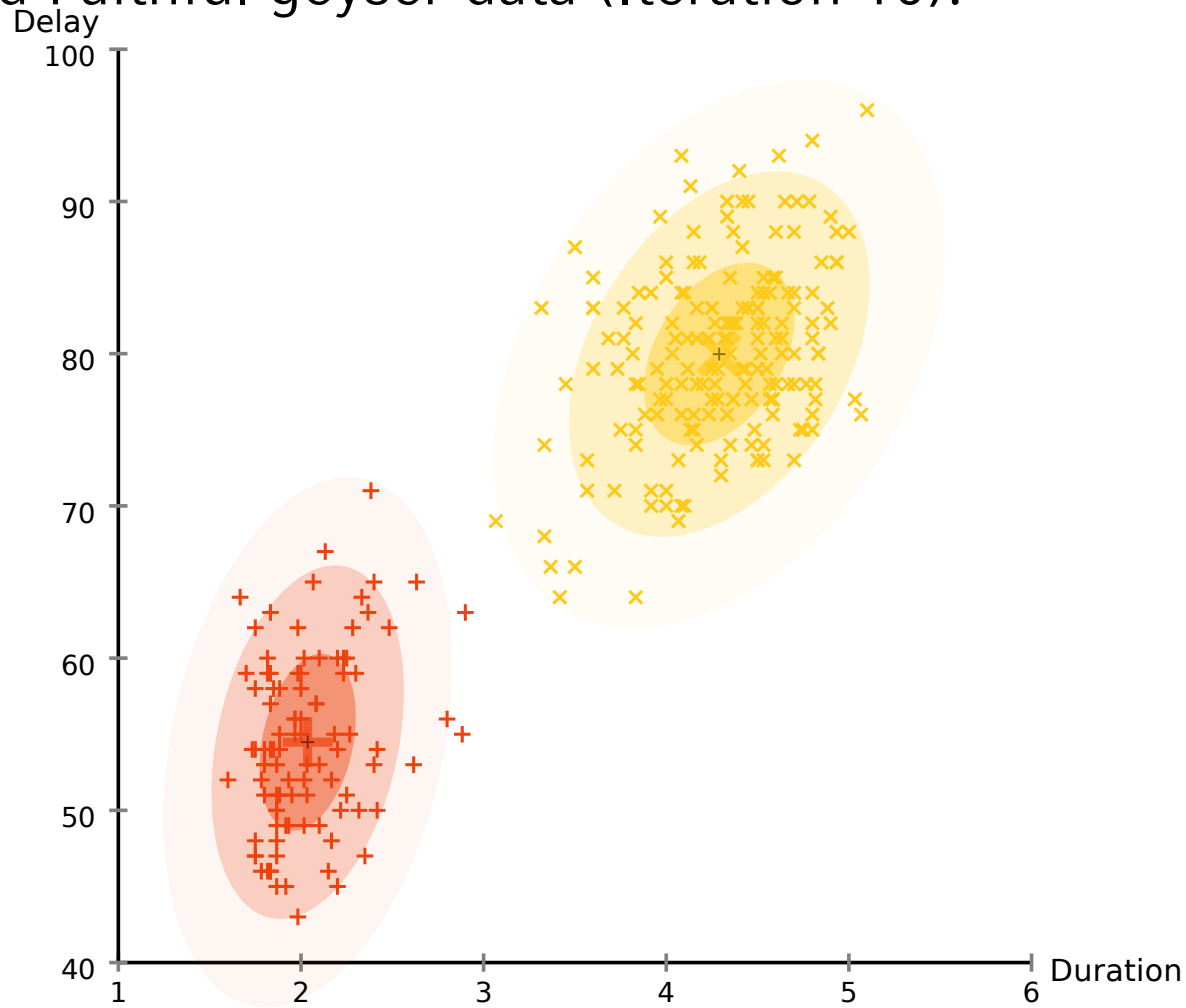
# EM demonstration

## Old Faithful geyser data (Iteration 9):



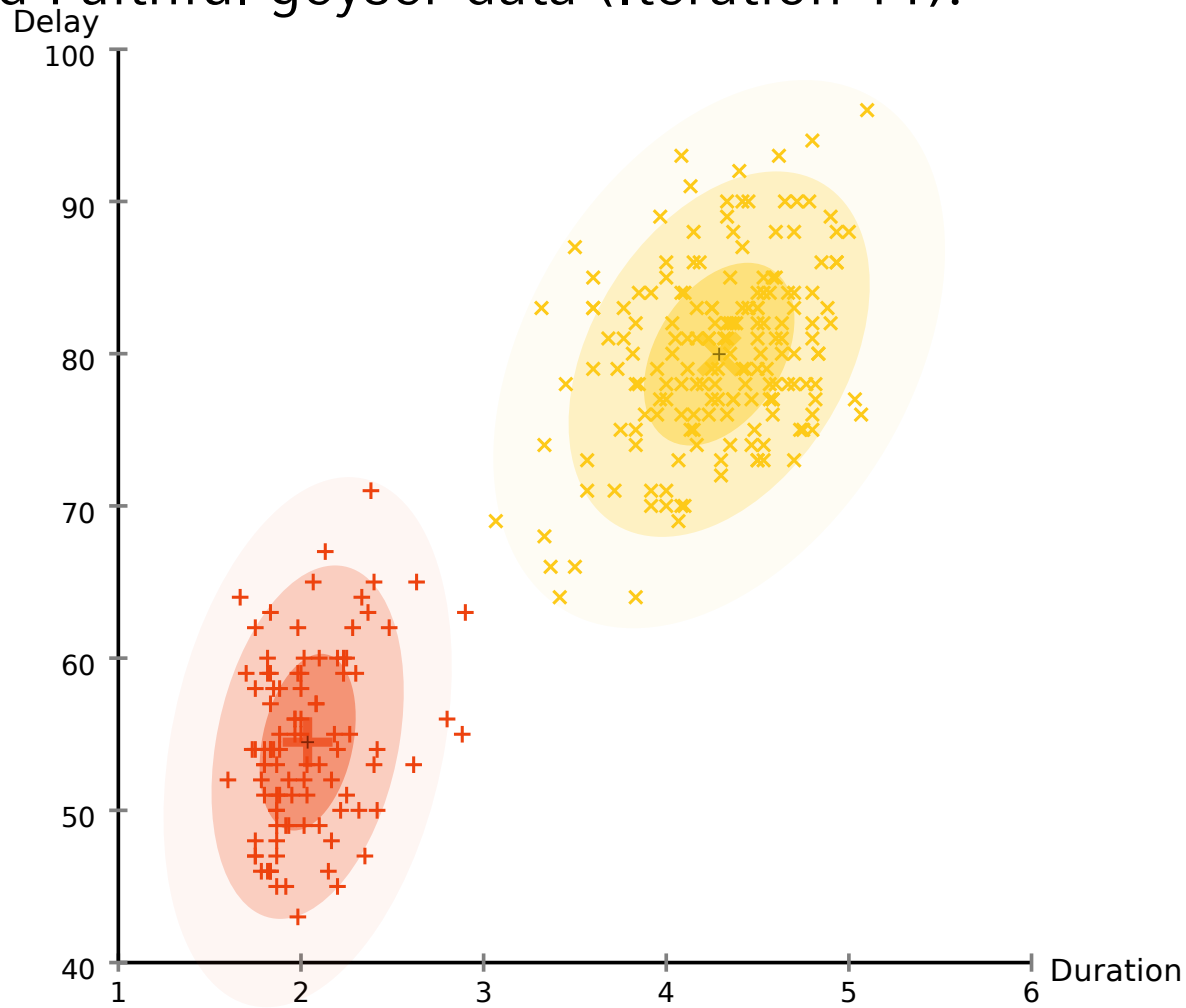
# EM demonstration

Old Faithful geyser data (Iteration 10):



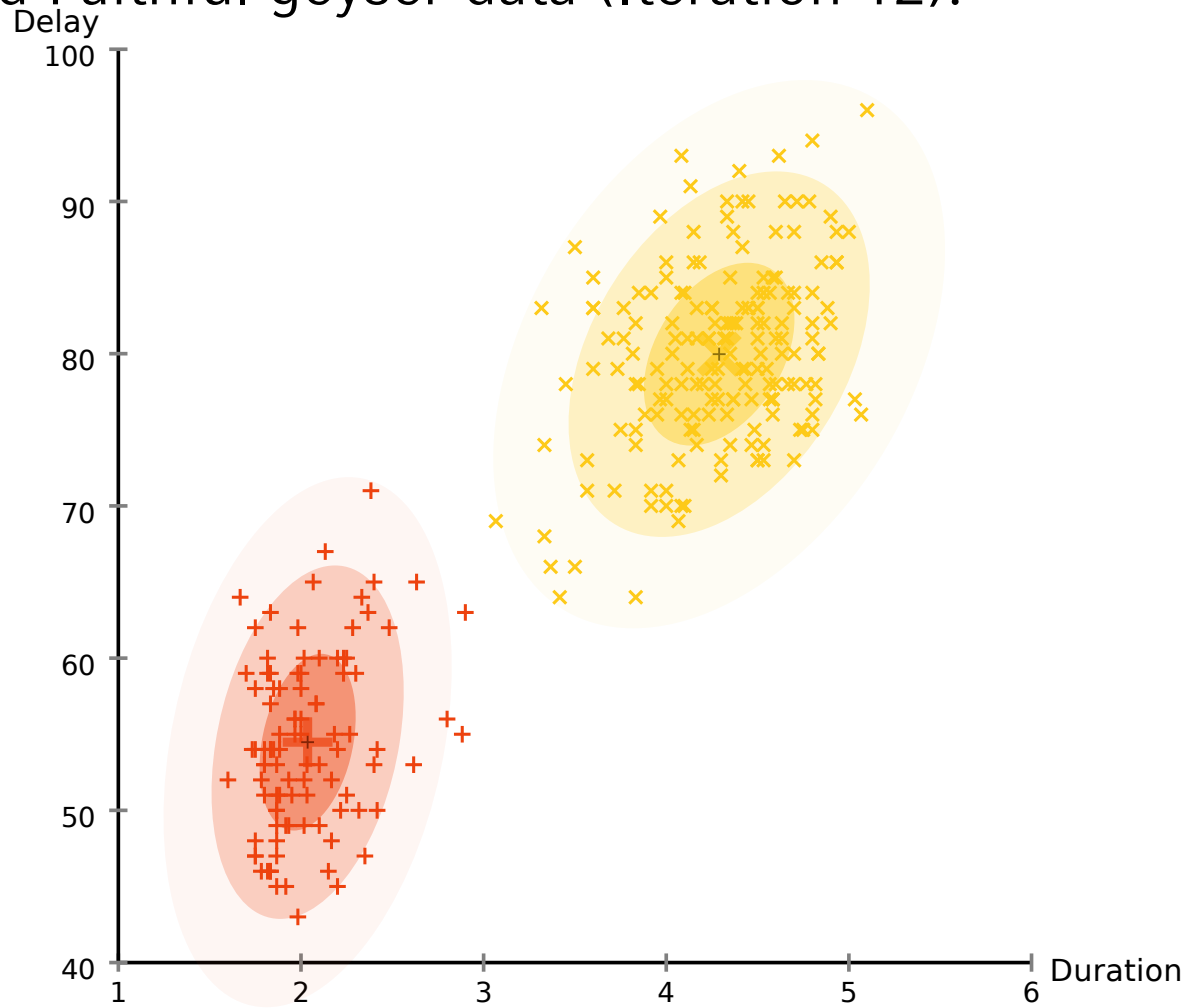
# EM demonstration

## Old Faithful geyser data (Iteration 11):



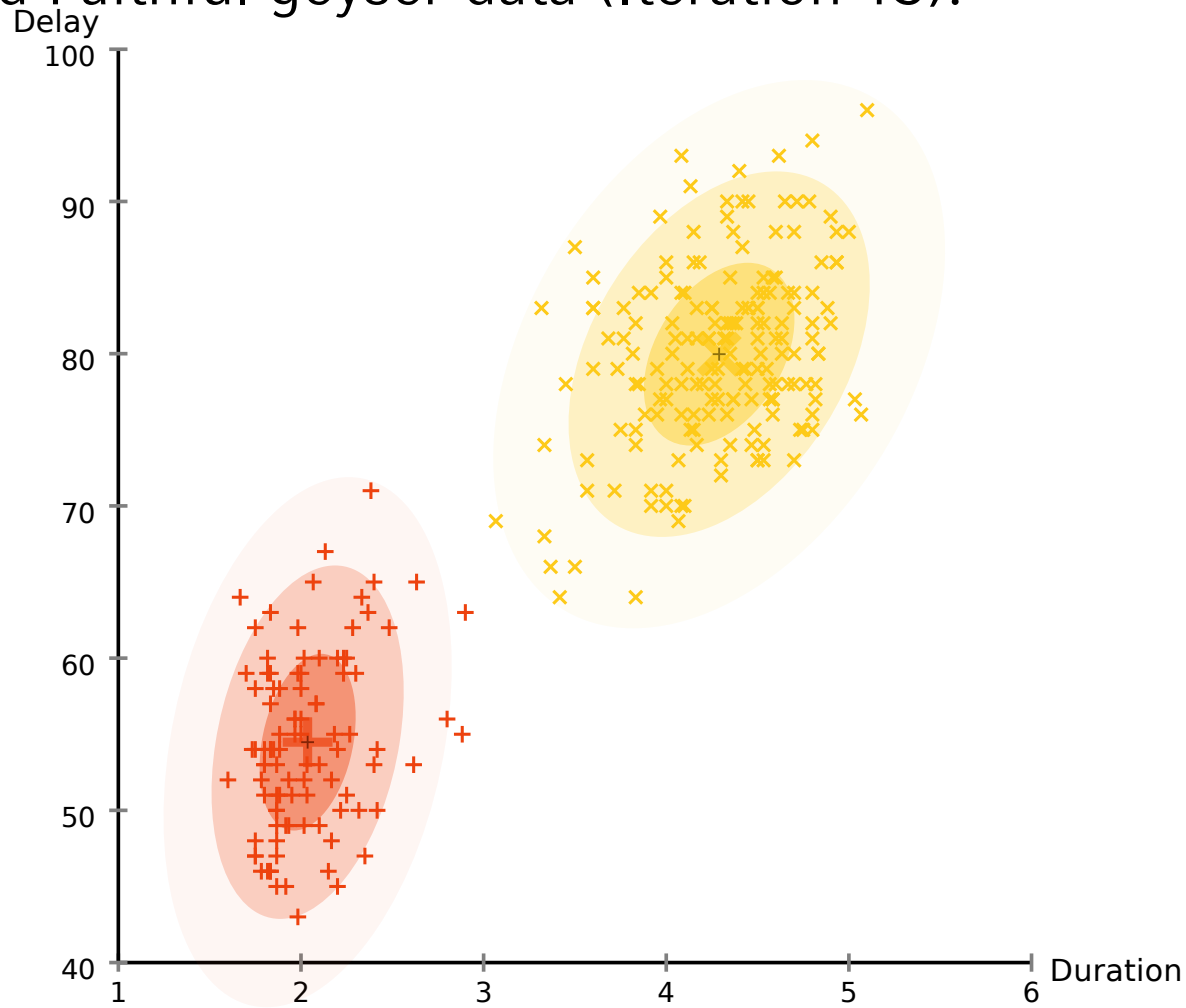
# EM demonstration

Old Faithful geyser data (Iteration 12):



# EM demonstration

Old Faithful geyser data (Iteration 13):



# EM demonstration

## Old Faithful geyser data (Iteration 14):