Machine Learning for Data Science (CS4786)
Lecture 6

Gaussian Mixture Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2017fa/

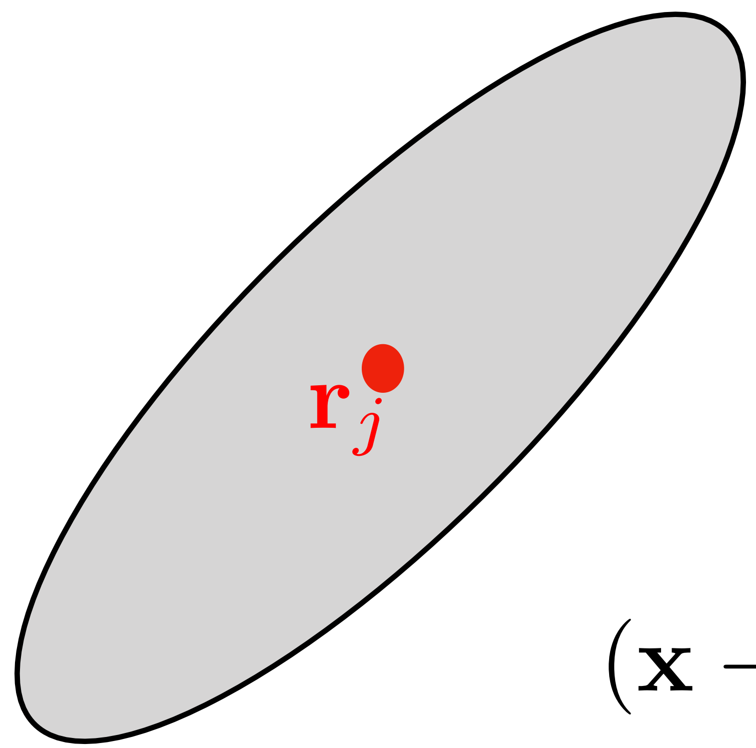
K-means: pitfalls

- Looks for spherical clusters
- Of same radius
- And with roughly equal number of points

K-means: pitfalls

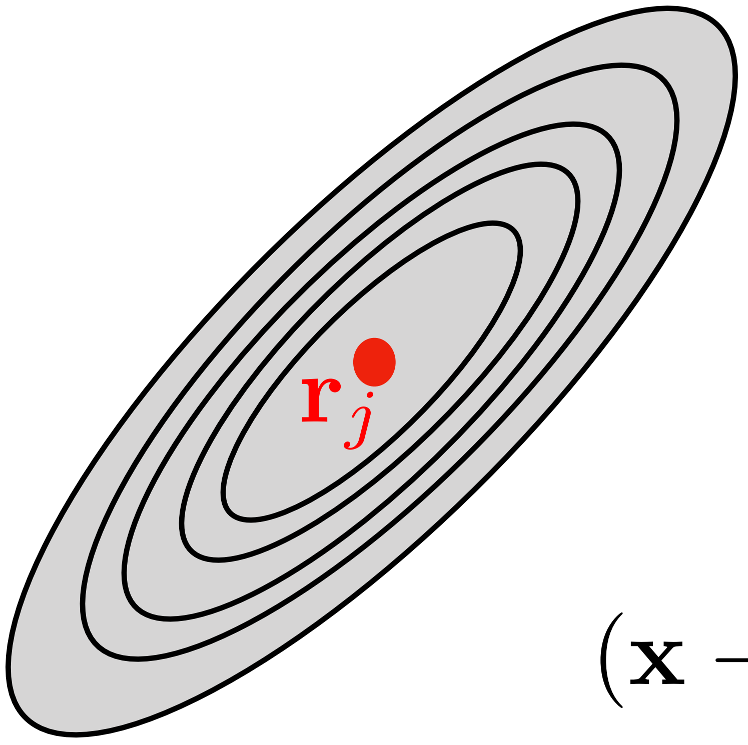
- Can we design algorithm that can address these shortcomings?

Ellipsoid



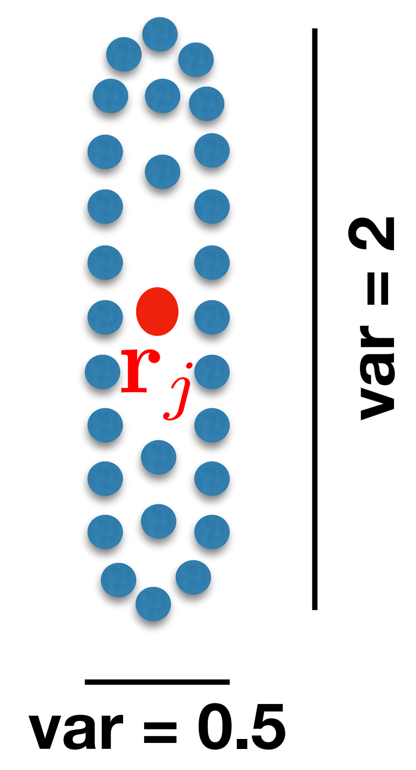
$$(\mathbf{x} - \mathbf{r}_j)^\top A (\mathbf{x} - \mathbf{r}_j) \leq 1$$

Ellipsoid

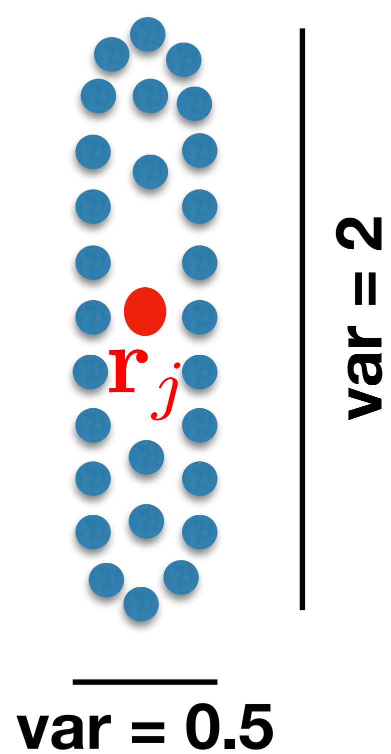


$$(\mathbf{x} - \mathbf{r}_j)^\top A(\mathbf{x} - \mathbf{r}_j) \leq 1$$

Axis Aligned Ellipsoid

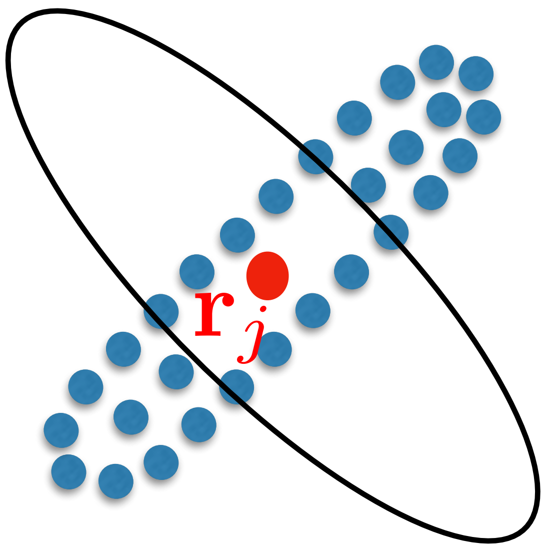


Axis Aligned Ellipsoid



$$(\mathbf{x} - \mathbf{r}_j)^\top \begin{bmatrix} 1/0.5 & 0 \\ 0 & 1/2 \end{bmatrix} (\mathbf{x} - \mathbf{r}_j)$$

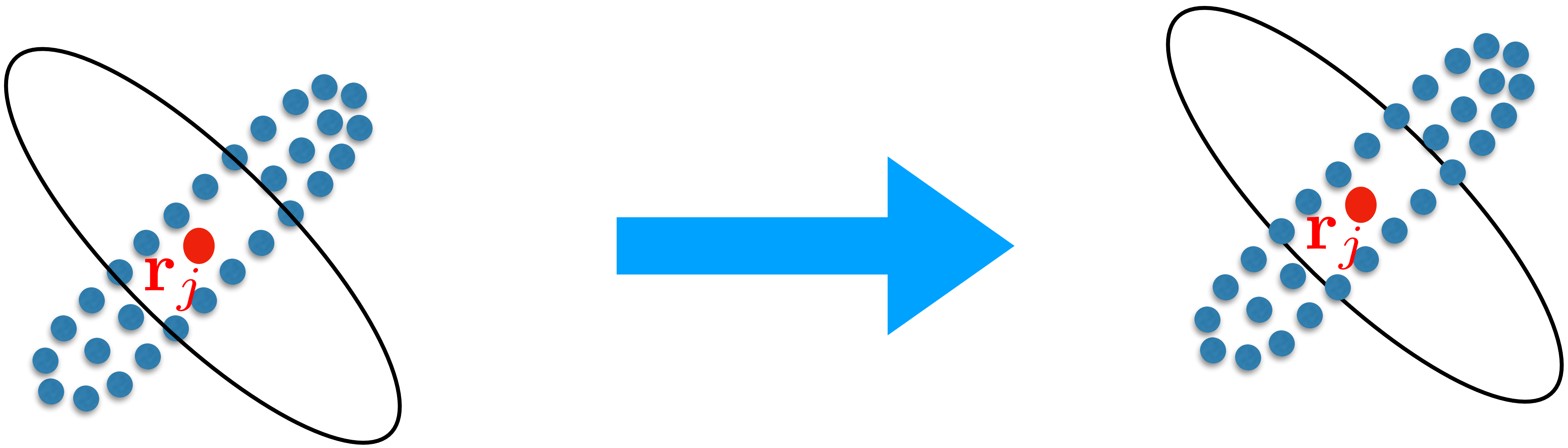
Error From Last Slide



$$(\mathbf{x} - \mathbf{r}_j)^\top \Sigma (\mathbf{x} - \mathbf{r}_j) \leq 1$$

Bug last lecture!

Error From Last Slide



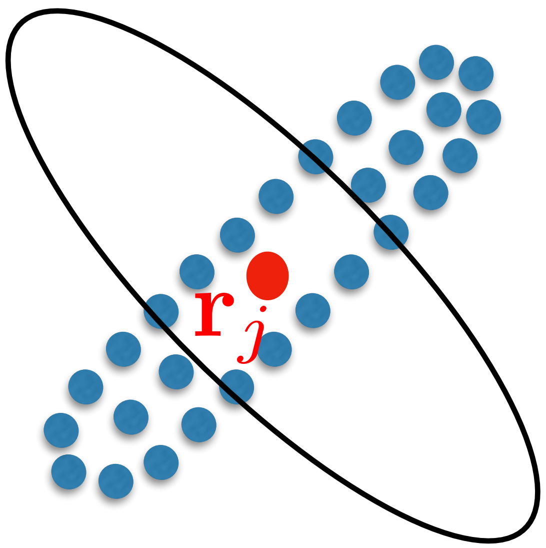
$$(\mathbf{x}-\mathbf{r}_j)^\top \Sigma (\mathbf{x}-\mathbf{r}_j) \leq 1$$

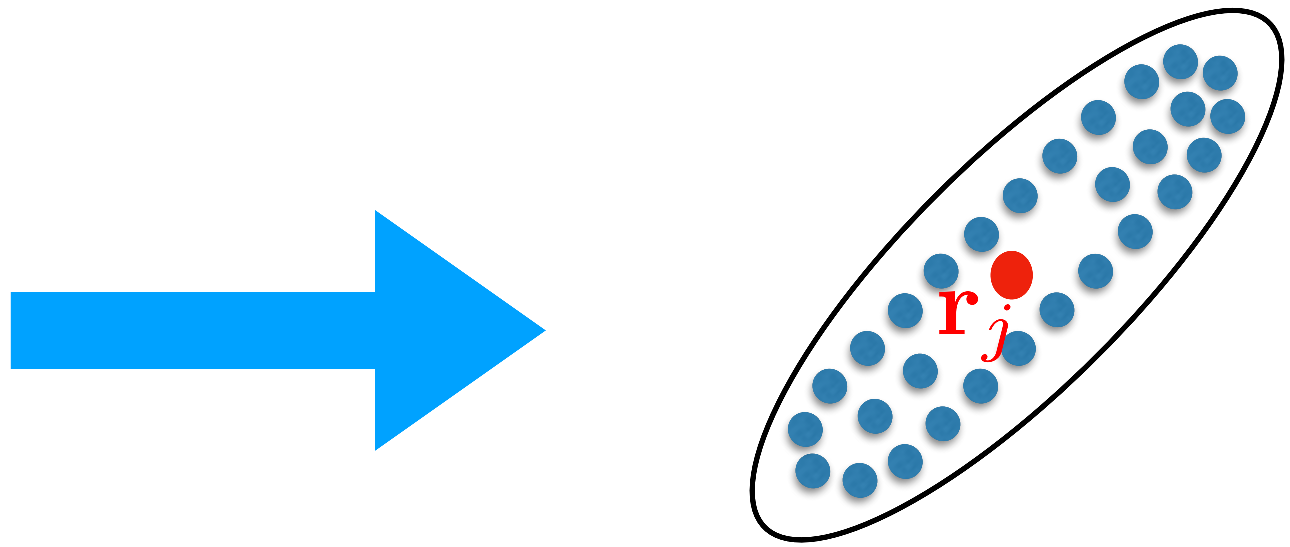
$$(\mathbf{x}-\mathbf{r}_j)^\top \Sigma^{-1} (\mathbf{x}-\mathbf{r}_j) \leq 1$$

Bug last lecture!

Correction!

Error From Last Slide





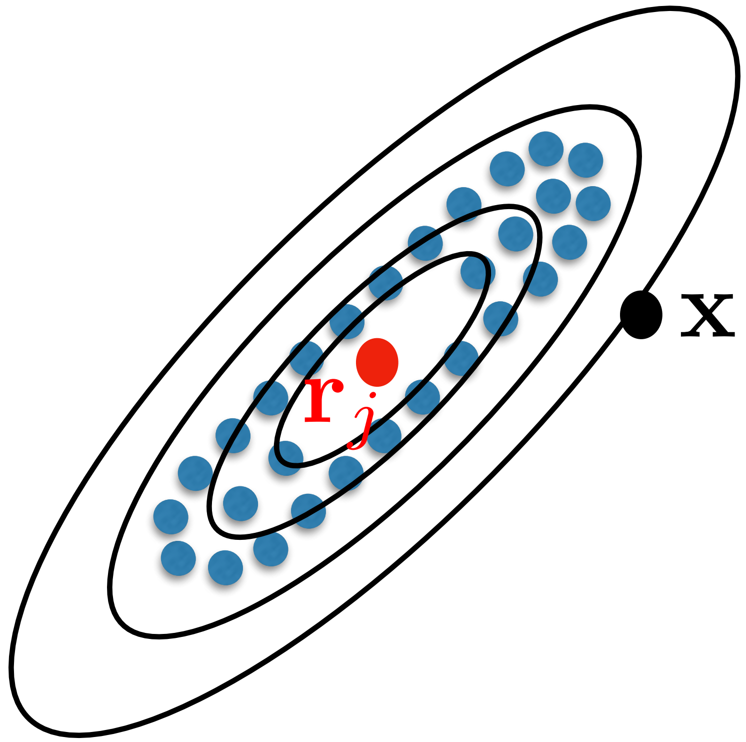
$$(\mathbf{x} - \mathbf{r}_j)^\top \Sigma (\mathbf{x} - \mathbf{r}_j) \leq 1$$

$$(\mathbf{x} - \mathbf{r}_j)^\top \Sigma^{-1} (\mathbf{x} - \mathbf{r}_j) \leq 1$$

Bug last lecture!

Correction!

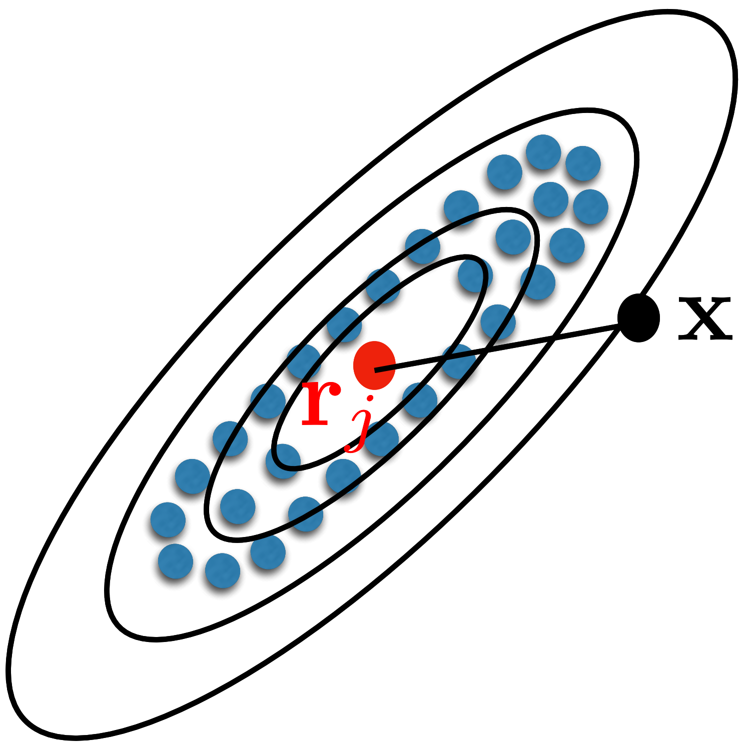
General Ellipsoid



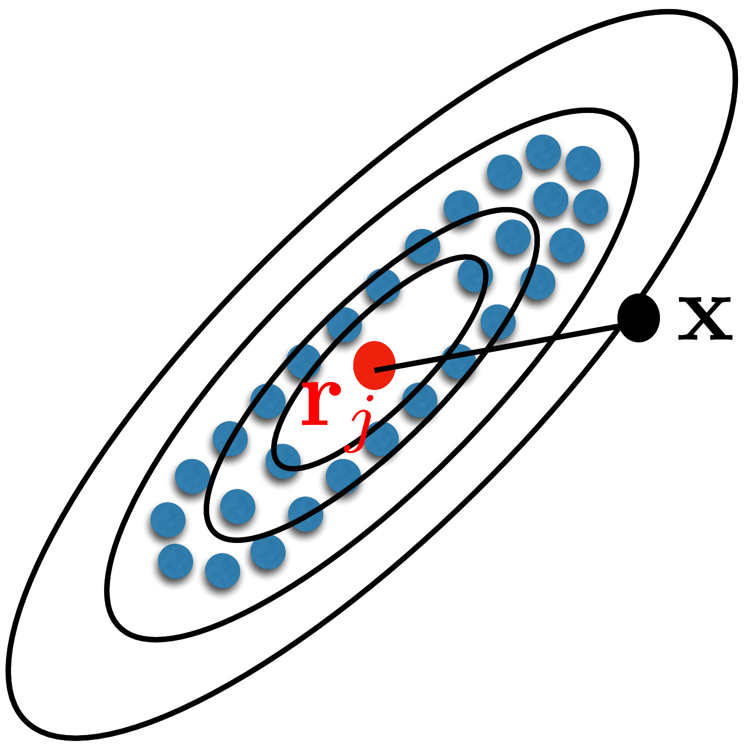
$$(\mathbf{x} - \mathbf{r}_j)^\top \Sigma^{-1} (\mathbf{x} - \mathbf{r}_j)$$

$$\Sigma = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \mathbf{r}_j)(\mathbf{x}_t - \mathbf{r}_j)^\top$$

General Ellipsoid

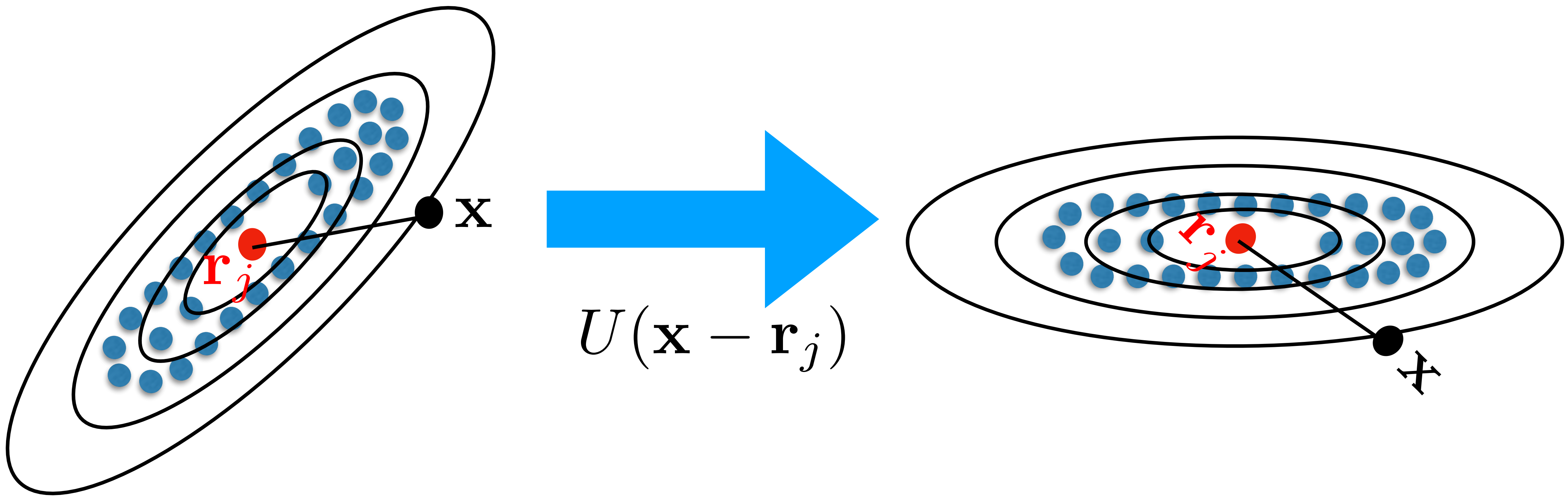


General Ellipsoid



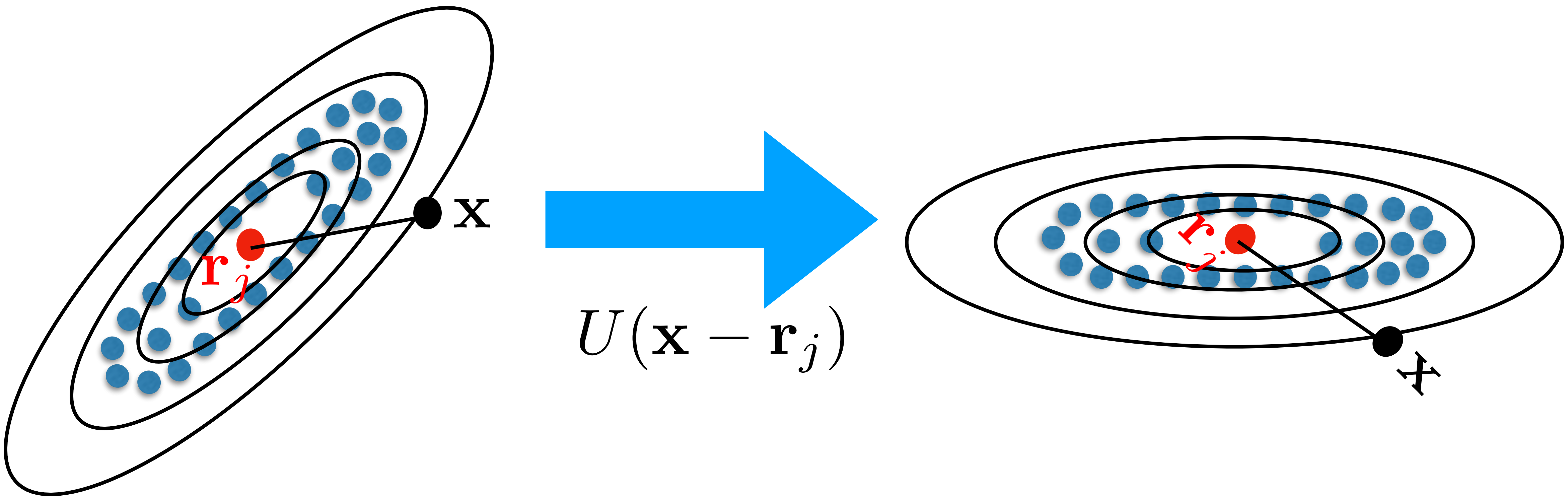
$$\Sigma = U^\top \Lambda U \qquad \Sigma^{-1} = U^\top \Lambda^{-1} U$$

General Ellipsoid



$$\Sigma = U^\top \Lambda U \qquad \Sigma^{-1} = U^\top \Lambda^{-1} U$$

General Ellipsoid



$$\Sigma = U^\top \Lambda U \qquad \Sigma^{-1} = U^\top \Lambda^{-1} U$$

$$(\mathbf{x} - \mathbf{r}_j)^\top \Sigma^{-1} (\mathbf{x} - \mathbf{r}_j) = (U(\mathbf{x} - \mathbf{r}_j))^\top \Lambda \, (U(\mathbf{x} - \mathbf{r}_j))$$

K-MEANS CLUSTERING

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \|\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1}\|$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t$$

  3. $m \leftarrow m + 1$

Ellipsoidal Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ and ellipsoids $\hat{\Sigma}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \; (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})^\top \left(\hat{\Sigma}^{m-1}\right)^{-1} (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top$$

  3. $m \leftarrow m + 1$

Ellipsoidal Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ and ellipsoids $\hat{\Sigma}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \operatorname*{argmin}_{j \in [K]} \; (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})^\top \left(\hat{\Sigma}^{m-1}\right)^{-1} (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})$$

$$d(\mathbf{x}_t, C_j)$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top$$

  3. $m \leftarrow m + 1$

K-means: pitfalls

- Looks for spherical clusters
- Of same radius
- And with roughly equal number of points

K-means: pitfalls

- Looks for spherical clusters ✓
- Of same radius
- And with roughly equal number of points

K-means: pitfalls

- Looks for spherical clusters ✓
- Of same radius ✓
- And with roughly equal number of points

K-means: pitfalls

- Looks for spherical clusters ✓
- Of same radius ✓
- And with roughly equal number of points ✗

Hard Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$, ellipsoids $\hat{\Sigma}_j^0$ and initial proportions π^0 randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \; (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})^\top \left(\hat{\Sigma}^{m-1}\right)^{-1} (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1}) - \log(\pi_j^{m-1})$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top \qquad \pi_j^m = \frac{|C_j^m|}{n}$$

  3. $m \leftarrow m + 1$

Hard Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$, ellipsoids $\hat{\Sigma}_j^0$ and initial proportions π^0 randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \; (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})^\top \left(\hat{\Sigma}^{m-1}\right)^{-1} (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1}) - \log(\pi_j^{m-1})$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top \qquad \pi_j^m = \frac{|C_j^m|}{n}$$

  3. $m \leftarrow m + 1$

Multivariate Gaussian

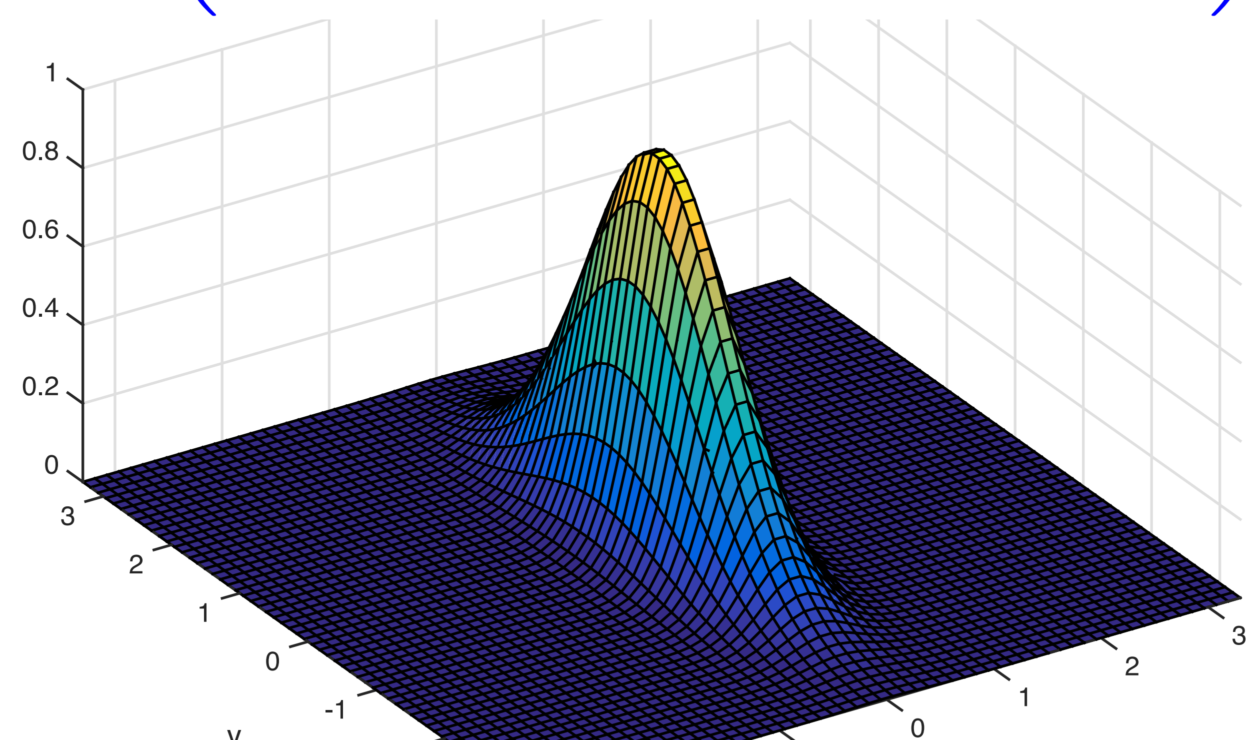
- Two parameters:
  - Mean $\mu \in \mathbb{R}^d$
  - Covariance matrix Σ of size dxd

$$p(x; \mu, \Sigma) = (2\pi)^{d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Multivariate Gaussian

- Two parameters:
  - Mean $\mu \in \mathbb{R}^d$
  - Covariance matrix Σ of size dxd

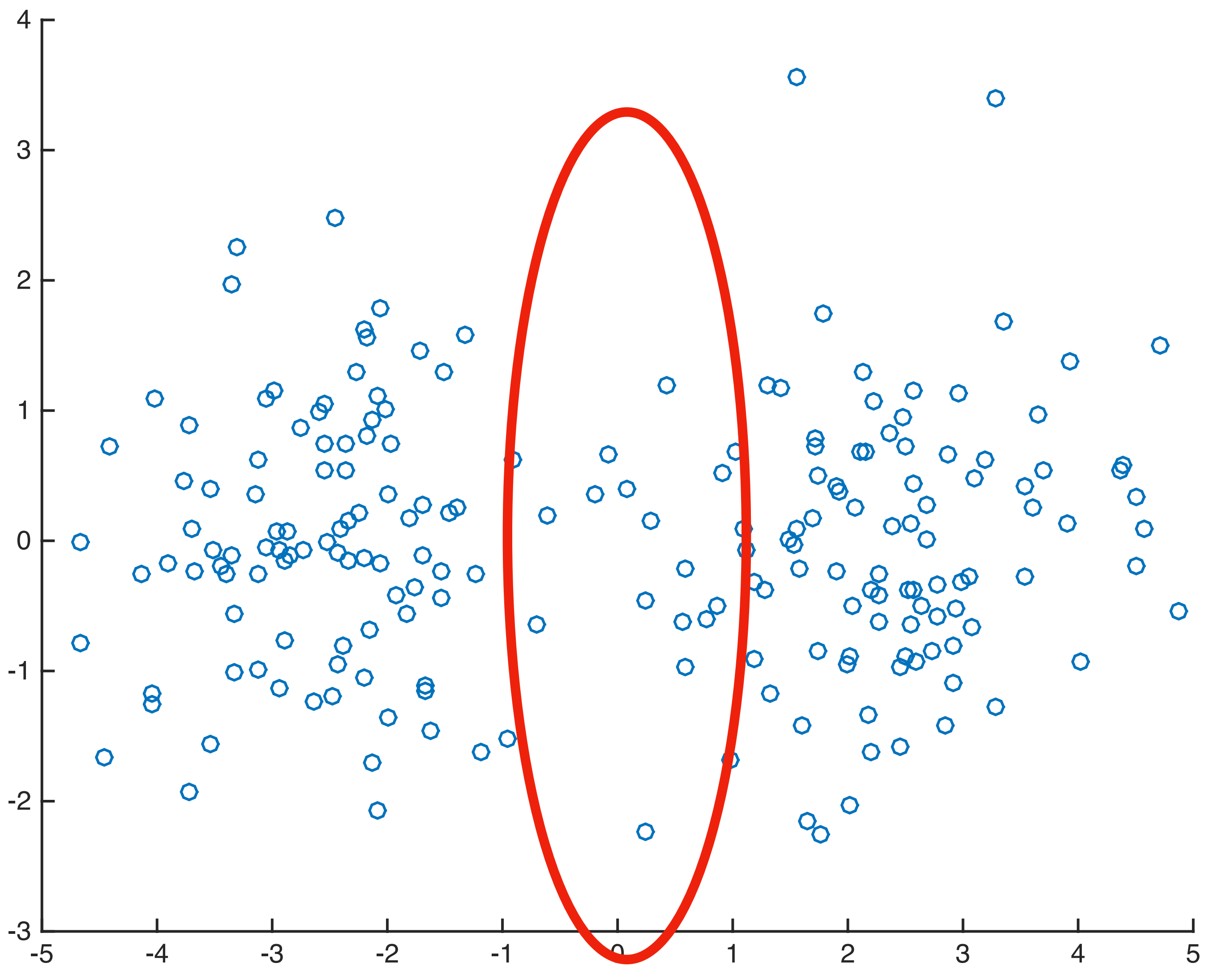
$$p(x; \mu, \Sigma) = (2\pi)^{d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right)$$



Hard Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$, ellipsoids $\hat{\Sigma}_j^0$ and initial proportions π^0 randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point
  $$\hat{c}^m(\mathbf{x}_t) = \arg\max_{j \in [K]} \; p(\mathbf{x}_t, \hat{\mathbf{r}}_j^{m-1}, \hat{\Sigma}^{m-1}) \times \pi^m(j)$$
  2. For each $j \in [K]$, set new representative as
  $$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top \qquad \pi_j^m = \frac{|C_j^m|}{n}$$
  3. $m \leftarrow m + 1$

Pitfall of Hard Assignment



(Soft) Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ and ellipsoids $\hat{\Sigma}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$Q_t^m(j) = p(\mathbf{x}_t, \hat{\mathbf{r}}_j^{m-1}, \hat{\Sigma}^{m-1}) \times \pi^m(j)$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{\sum_{t=1}^n Q_t(j)\mathbf{x}_t}{\sum_{t=1}^n Q_t(j)} \qquad \hat{\Sigma}^m = \frac{\sum_{t=1}^n Q_t(j)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top}{\sum_{t=1}^n Q_t(j)}$$

$$\pi_j^m = \frac{\sum_{t=1}^n Q_t(j)}{n}$$

  3. $m \leftarrow m + 1$