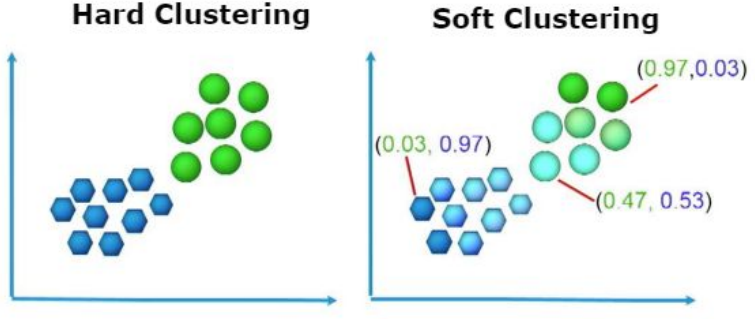
Hard and Soft Clusters

- K-Means uses ***hard* clustering assignment**.
  - A point belongs to exactly one cluster.

- Mixture of Gaussians uses ***soft* clustering**.
  - A point could be explained **by more than one cluster.**
  - Different clusters takedifferent levels of "responsibility" (posterior probability) for that point.

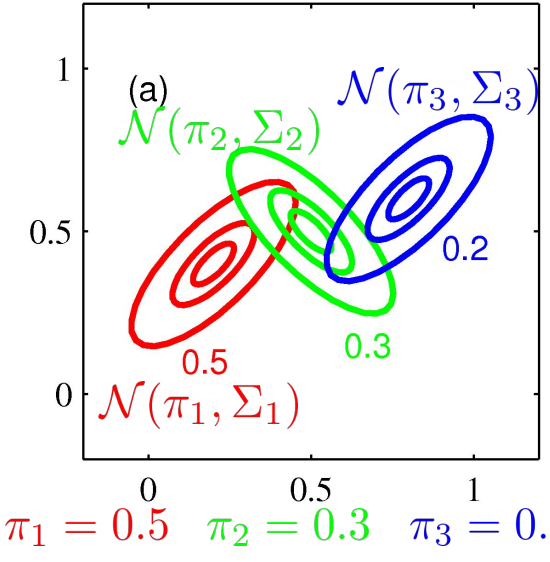


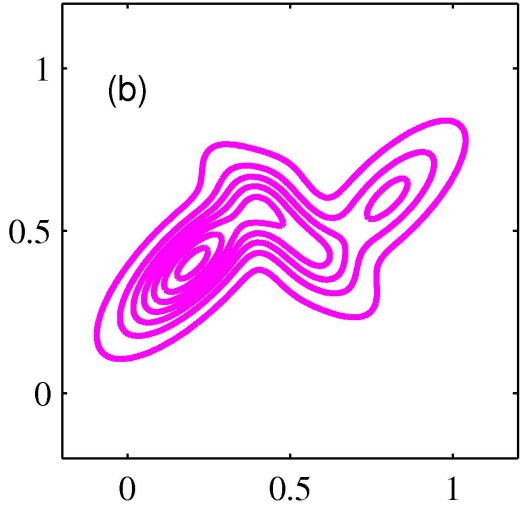
  - (It was actually generated by only one cluster, but we don't know which one; we assign a probability)

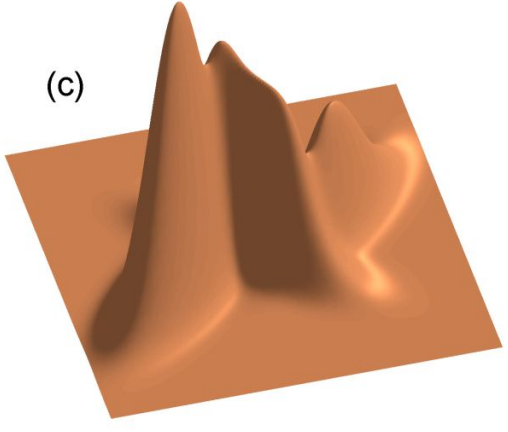
Mixtures of Gaussians

- Mixtures of Gaussians make it possible to describe much richer distributions.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

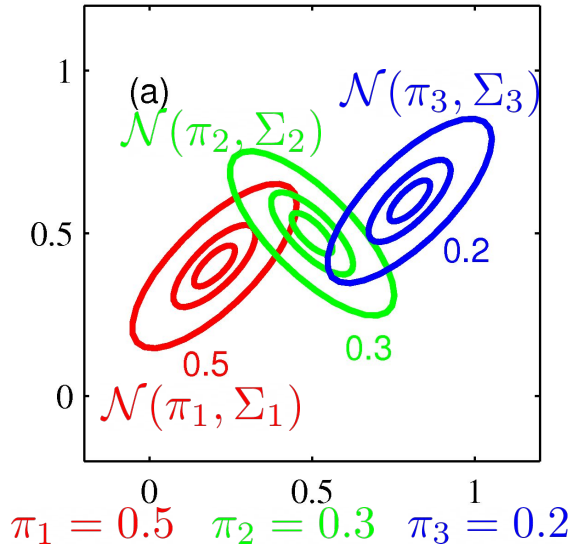






$\pi_1 = 0.5$ $\pi_2 = 0.3$ $\pi_3 = 0.2$

Mixtures of Gaussians



- Note the mixing coefficients in

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k) \qquad \sum_{k=1}^{K} \pi_k = 1$$

- Let **z** in $\{0,1\}^K$ be a 1-of-*K* random variable;

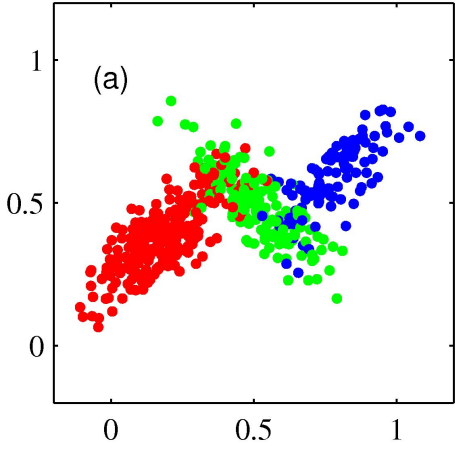
$$p(z_k = 1) = \pi_k \qquad p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

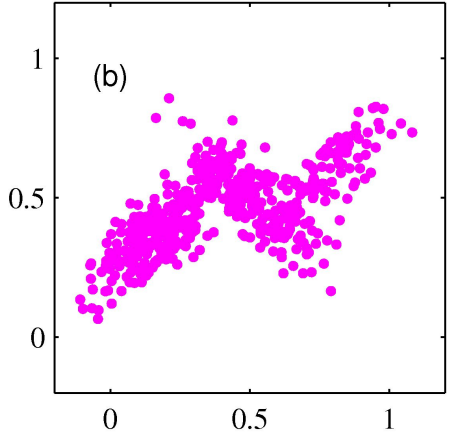
$$p(\mathbf{x} \mid z_k = 1) = \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

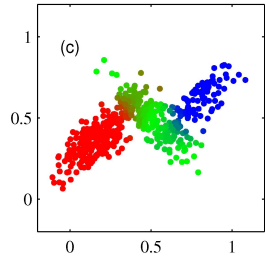
Mixtures of Gaussians

- To generate samples from a Gaussian mixture distribution $p(\mathbf{x})$, use $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})\, p(\mathbf{x} \mid \mathbf{z})$:
  - Select a value $\mathbf{z}$ from the marginal $p(\mathbf{z})$;
  - Then select a value $\mathbf{x}$ from $p(\mathbf{x} \mid \mathbf{z})$ for that $\mathbf{z}$.





EM for Gaussian Mixtures
(1-page summary)



- Initialize parameters randomly $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$
- Repeat until convergence (alternating optimization)
  - E Step: Given fixed parameters θ, optimize $q(\mathbf{Z})$: find posterior probabilities of Z given X, $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \theta)$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} = P(z_k = 1 | \mathbf{x}^{(n)})$$

  - M Step: Given fixed $q(\mathbf{Z})$ (or $\gamma(z_{nk})$), optimize θ (estimate mean, covariance, etc.): equivalent to MLE for Gaussian Discriminant Analysis using $q(\mathbf{Z})$ as pseudo-counts!

Relation to K-means

- In Guassian mixture, we fix the covariance matrix for each cluster as $\sigma^2 I$
- We take $\sigma^2 \to 0$ then
- the update equations coverge to doing K-means clustering

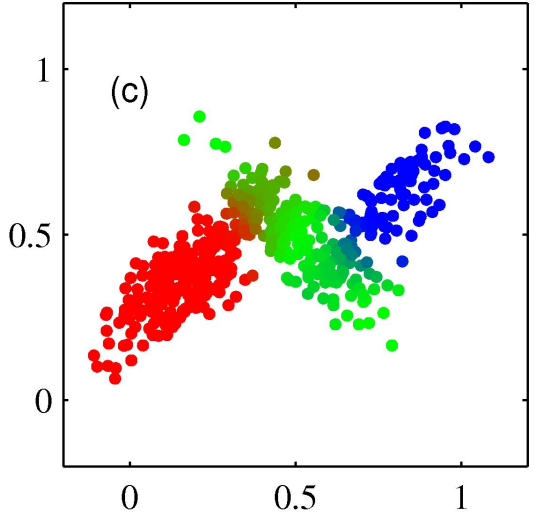
Derivation: in the next lecture!

EM for Gaussian Mixtures: E-step

- Initialize means, covariances, and mixing coefficients for the K Gaussians.

- E Step: Given the parameters, evaluate the responsibilities.

$$q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \theta)$$



$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} = P(z_k = 1 | \mathbf{x}^{(n)})$$

(Recap) E-step: compute posterior → optimal q(Z)!

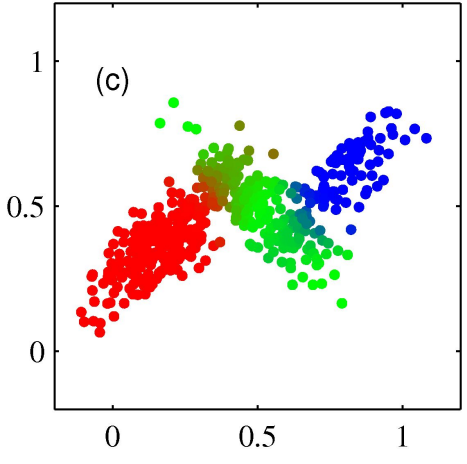
(Hint: Use Bayes Rule)

Mixtures of Gaussians: E-step

- Need to calculate $p(\mathbf{Z} \mid \mathbf{X}, \theta)$, i.e., *soft assignments*
- Responsibility is the degree (posterior prob.) to which each Gaussian explains an observation **X**.

Q. Verify this! (Hint: Use Bayes Rule)

$$q^{(n)}(\mathbf{z}_k) = p(\mathbf{z}_k|\mathbf{x}^{(n)}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} =: \gamma(\mathbf{z}_{nk})$$



Mixtures of Gaussians: M-step

General formula for M-step:

$$\arg\max_\theta \ \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)} \mid \theta)$$

Plug in for GMM: $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid k \in \{1 \dots K\}\}$

$$\operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \quad \underbrace{\sum_{n=1}^{N} \sum_{k=1}^{K} q^{(n)}(\mathbf{z}_k) \log p(\mathbf{z}_k, \mathbf{x}^{(n)} | \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{= J(\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

$$\text{s.t.} \sum_{k=1}^{K} \pi_k = 1$$

Mixtures of Gaussians: M-step

Let's first simplify the expression $J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$
\begin{aligned}
J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^{N} \sum_{k=1}^{K} q^{(n)}(\mathbf{z}_k) \log p(\mathbf{z}_k, \mathbf{x}^{(n)} \mid \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \left(\log \pi_k + \log \frac{1}{(2\pi)^{m/2} (\det \boldsymbol{\Sigma}_k)^{1/2}} - \frac{1}{2} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) \right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \left((2\pi)^{m/2} (\det \boldsymbol{\Sigma}_k)^{1/2} \right) \\
&\qquad - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k \\
&\qquad - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + \text{ const}
\end{aligned}
$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k$$
$$- \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + \text{ const}$$

- Maximize *J* w.r.t. $\boldsymbol{\mu}_k$ by differentiating w.r.t. $\boldsymbol{\mu}_k$ and setting the gradient to 0:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) = 0$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \mathbf{x}^{(n)}}{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k$$
$$- \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + \text{ const}$$

- To find $\boldsymbol{\Sigma}_k$, we use change of variables: $\mathbf{M}_k = \boldsymbol{\Sigma}_k^{-1}$

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{M}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \mathbf{M}_k$$
$$- \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \mathbf{M}_k(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + \text{ const}$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{M}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log\det \mathbf{M}_k$$
$$- \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top \mathbf{M}_k(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + \text{ const}$$

- Maximize *J* w.r.t. $\mathbf{M}_k$ by differentiating w.r.t. $\mathbf{M}_k$ and setting the gradient to 0:

*Note: $\frac{\partial \log|\det \mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^\top$ $\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^\top$

$$\frac{\partial J}{\partial \mathbf{M}_k} = \frac{1}{2}\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})\mathbf{M}_k^{-1} - \frac{1}{2}\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top = 0$$

$$\boldsymbol{\Sigma}_k = \mathbf{M}_k^{-1} = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\mathbf{z}_{nk})\log\pi_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\mathbf{z}_{nk})\log\det\boldsymbol{\Sigma}_k$$
$$- \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + \text{ const}$$

- Finally we need: $\max_{\boldsymbol{\pi}} \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\mathbf{z}_{nk})\log\pi_k \quad \text{s.t.} \sum_{k=1}^{K}\pi_k = 1$
- Use Lagrange multipliers

$$L(\boldsymbol{\pi}, \alpha) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\mathbf{z}_{nk})\log\pi_k - \alpha\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

Mixtures of Gaussians: M-step

- Finally we need: $\max_{\boldsymbol{\pi}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k \quad \text{s.t.} \sum_{k=1}^{K} \pi_k = 1$
- Use Lagrange multipliers

$$L(\boldsymbol{\pi}, \alpha) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

- Setting $\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \frac{1}{\pi_k} - \alpha = 0$ gives $\pi_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}{\alpha}$
- Using the constraint $\sum_{k=1}^{K} \pi_k = 1$, we get:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})} = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}{N}$$

Mixtures of Gaussians: M-step
(putting together)

- The mean of a cluster is the weighted mean, weighted by the responsibilities.

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \mathbf{x}^{(n)}}{N_k}$$

[N_k = sum of pseudo-counts γ_{nk} over n.]

  - N_k is the effective number of points in cluster *k*

$$N_k = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \qquad \pi_k = \frac{N_k}{N}$$

- Likewise for covariance:

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^\top}{N_k}$$

EM for Gaussian Mixtures: Summary

- Initialize means, covariances, and mixing coefficients for the *K* Gaussians.
- E Step: Given the parameters, evaluate the responsibilities or the posterior

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)}|\mu_j, \Sigma_j)} = P(z_k = 1|\mathbf{x}^{(n)})$$

EM for Gaussian Mixtures: Summary

- M Step: Given the responsibilities, re-evaluate the coefficients.

$$\pi_k^{\text{new}} = \frac{N_k}{N} = \frac{\sum_n \gamma(z_{nk})}{N}$$

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}^{(n)}$$

$$\sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}})(\mathbf{x}^{(n)} - \mu_k^{\text{new}})^\top$$

- Stop when either coefficients or log likelihood converges.

Summary: EM for Gaussian Mixtures

- Initialize parameters randomly $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
- Repeat until convergence (alternating optimization)
  - E Step: Given fixed parameters θ, set $q^{(n)}(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}^{(n)}, \theta)$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} = P(z_k = 1 | \mathbf{x}^{(n)})$$

  - M Step: Given fixed $q(\mathbf{z}^{(n)})$'s for $\mathbf{x}^{(n)}$'s (or $\gamma(z_{nk})$), update θ:

$$\pi_k^{\text{new}} = \frac{N_k}{N} = \frac{\sum_n \gamma(z_{nk})}{N}$$

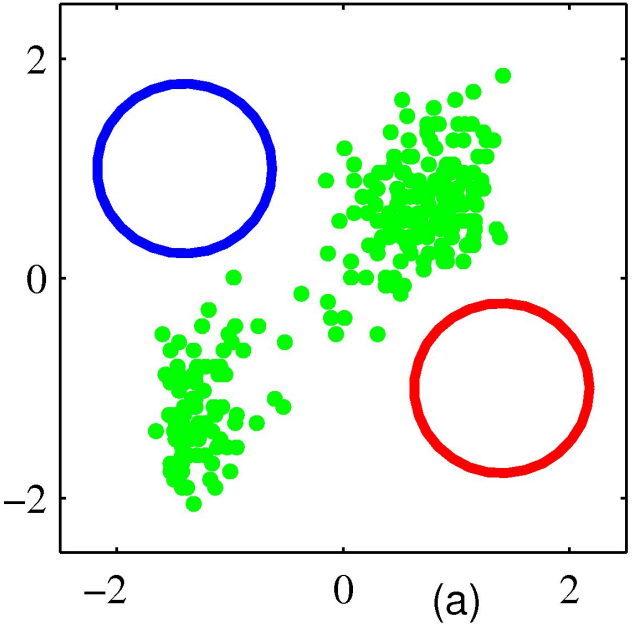
$$\arg\max_\theta \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)} \mid \theta)$$

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}^{(n)}$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}})(\mathbf{x}^{(n)} - \mu_k^{\text{new}})^\top$$

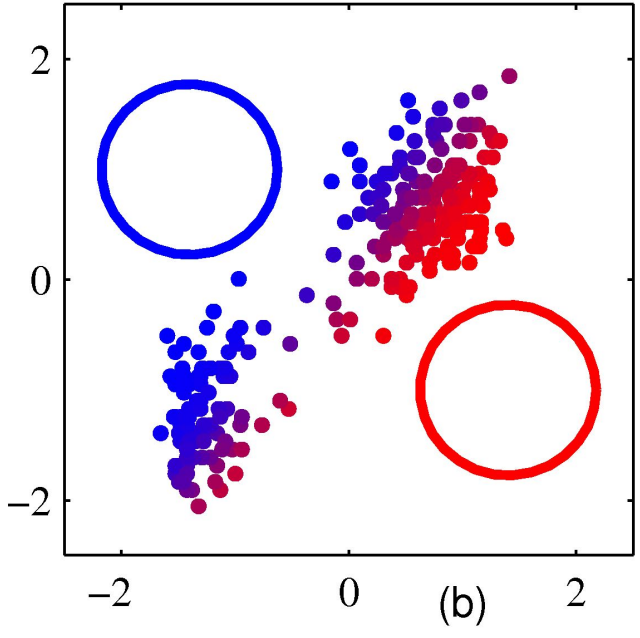
EM Example

- Initialize parameters: means, covariances, and mixing coefficients.



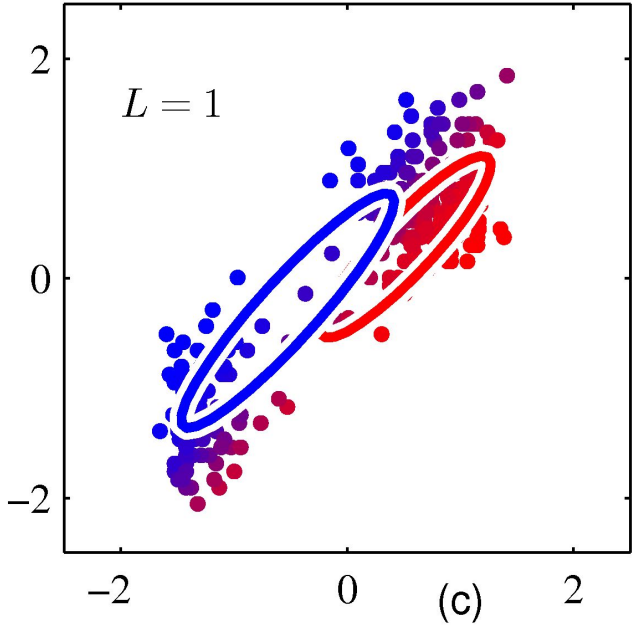
EM Example

- First E Step



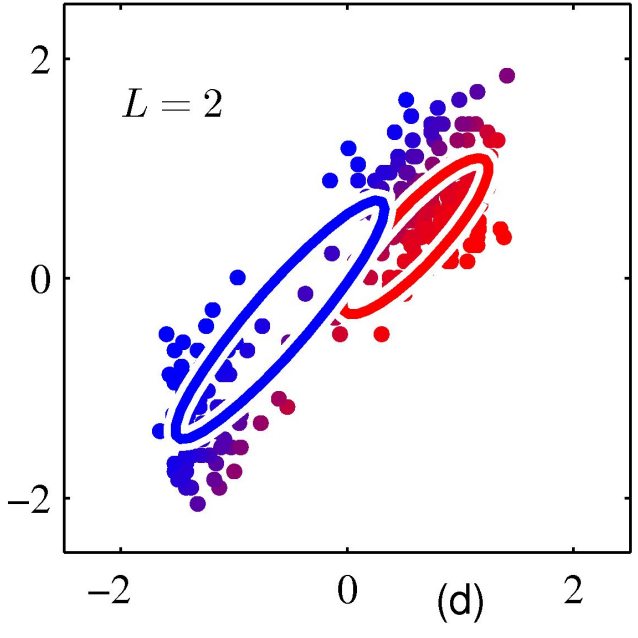
EM Example

- First M Step



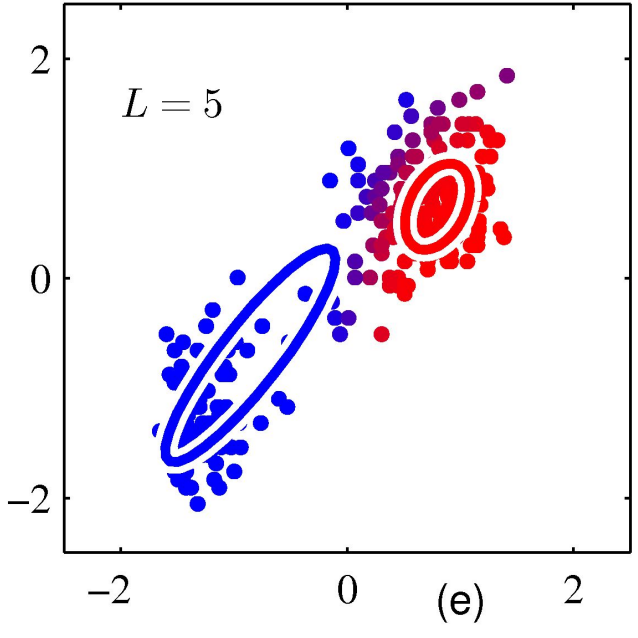
EM Example

- Second E and M Steps



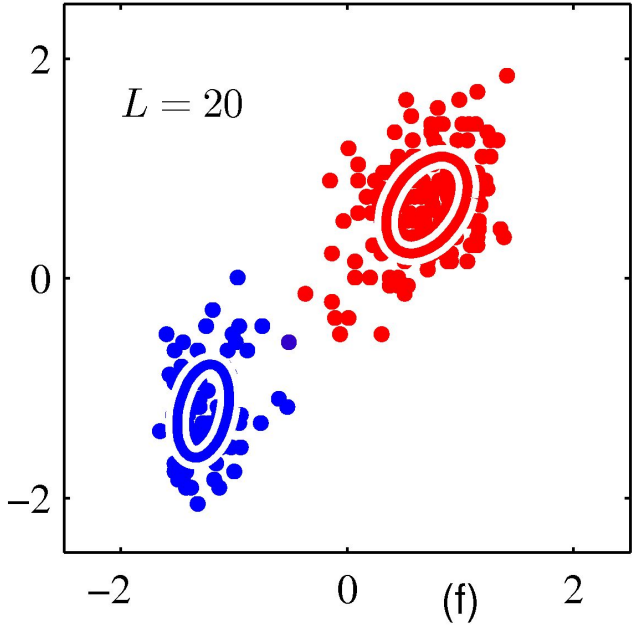
EM Example

- Three more E-M cycles



EM Example

- Fifteen E-M cycles later



Summary: Expectation Maximization

Core Idea:

- Iterative method for parameter estimation with hidden data (latent variables).
- Finds maximum likelihood estimates (local optimum) via EM.

Steps: (alternating optimization until convergence)

- **E-Step:** Estimate the posterior probability (expected value) of latent variables.
- **M-Step:** Maximize the “expected” likelihood to update parameters.

Key Points:

- Converges to a local optimum (initialization matters).
- Commonly used in clustering (e.g., Gaussian Mixture Models) and other models with missing data.

Summary: K-Means vs Gaussian Mixtures

K-means:

- Clusters data by minimizing within-cluster variance.
- Hard assignments; best for spherical, similar-sized clusters.
- Fast, but sensitive to initialization.

Gaussian Mixture Models (GMM):

- Models data as a mix of Gaussian distributions via EM.
- Soft assignments; handles elliptical, overlapping clusters.
- More flexible yet computationally intensive.
- K-means can be shown to be a special case of GMM

Key Considerations:

- Both need pre-specified # clusters and careful initialization.
- Use K-means for speed; choose GMM for probabilistic clustering or probabilistic modeling of data.