CSci5521: Machine Learning Fundamentals

- Mixture Models and EM II

Catherine Qi Zhao
Computer Science and Engineering
University of Minnesota

K-Means Clustering: Distortion Measure

- Dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Partition in K clusters
- Cluster prototype: μ_k
- Binary indicator variable, 1-of-K Coding scheme
  $r_{nk} \in \{0, 1\}$
  $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$.
  Hard assignment.
- Distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \tag{9.1}$$

K-Means Clustering: Expectation Maximization

- Find values for $\{r_{nk}\}$ and $\{\mu_k\}$ to minimize:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \tag{9.1}$$

- Iterative procedure:
  1. Minimize J w.r.t. r_{nk}, keep μ_k fixed (Expectation)

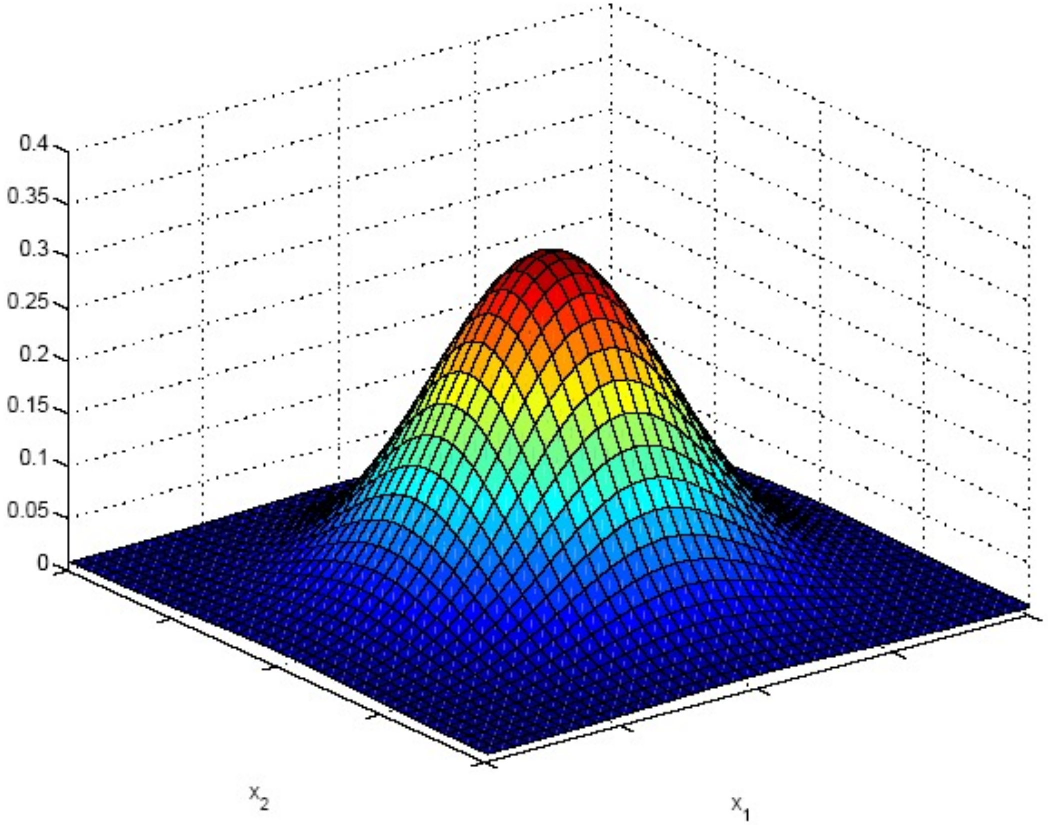
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{9.2}$$

  2. Minimize J w.r.t. μ_k, keep r_{nk} fixed (Maximization)

$$2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \mu_k) = 0 \tag{9.3}$$

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} \tag{9.4}$$

Multivariate Gaussian Distribution



$$\mathbf{x} \sim \mathcal{N}_d(\mu, \Sigma)$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)]$$

Mixture of Gaussians: Latent variables

- Gaussian Mixture Distribution:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \tag{9.7}$$

- Introduce latent variable $\mathbf{z}$
    - $\mathbf{z}$ is binary 1-of-K coding variable
    - $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$

Mixture of Gaussians: Latent variables (2)

- $p(z_k = 1) = \pi_k$
  constraints: $0 \leq \pi_k \leq 1$, and $\sum_k \pi_k = 1$
  $p(\mathbf{z}) = \prod_k \pi_k^{z_k}$
- $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$
  $p(\mathbf{x}|\mathbf{z}) = \prod_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$
- $p(\mathbf{x}) = \sum_z p(\mathbf{x}, \mathbf{z}) = \sum_z p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$
- The use of the joint probability $p(\mathbf{x}, \mathbf{z})$, leads to significant simplifications

Mixture of Gaussians: Latent variables (3)

- responsibility of component k to generate observation $\mathbf{x}$ (9.13):

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_k p(z_k = 1)p(\mathbf{x}|z_k = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}$$

  is the *posterior probability*

Mixture of Gaussians: Maximum Likelihood

- Log Likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right\} \tag{9.14}$$

Mixture of Gaussians: EM for Gaussian Mixtures

- Informal introduction of *expectation-maximization* algorithm (Dempster *et al.*, 1977).
- Maximum of log likelihood:
  derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t parameters to 0.

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right\} \qquad (9.14)$$

- For the μ_k[1]:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}}_{\gamma(z_{nk})} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k) \qquad (9.16)$$

$$\mu_k = \frac{1}{\sum_n \gamma(z_{nk})} \sum_n \gamma(z_{nk}) \mathbf{x}_n \qquad (9.17)$$

probability (think like "r" in hard case)

[1]Error in book, see erata file

Mixture of Gaussians: EM for Gaussian Mixtures

- For Σ_k:

weight

$$\Sigma_k = \frac{1}{\sum_n \gamma(z_{nk})} \sum_n \gamma(z_{nk}) \underbrace{(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}_{\text{variance}} \qquad (9.19)$$

- For the π_k:
  - Take into account constraint $\sum_k \pi_k = 1$
  - Lagrange multiplier

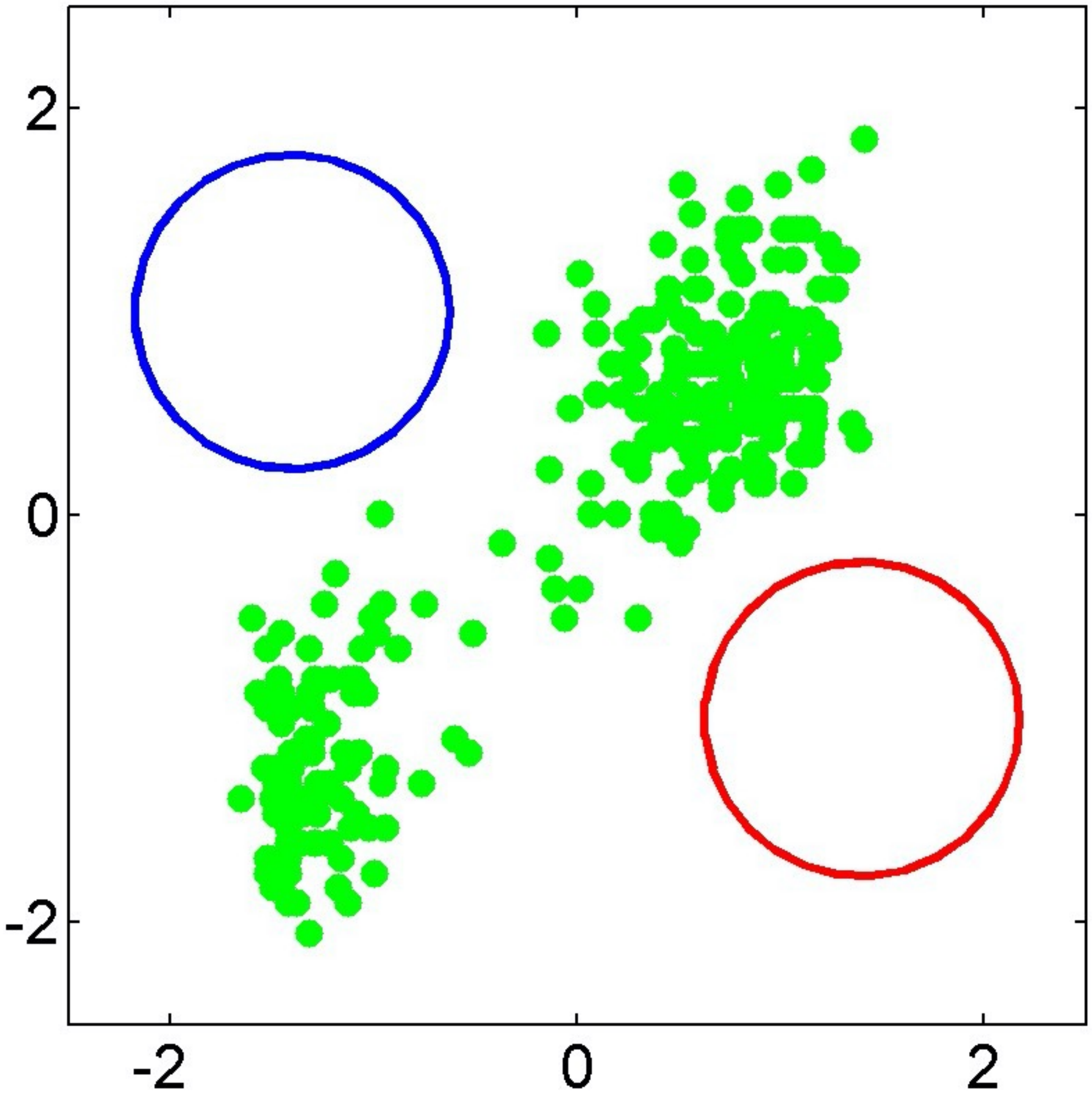
$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda(\sum_k \pi_k - 1) \qquad (9.20)$$

$$0 = \sum_n \frac{\mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)} + \lambda \qquad (9.21)$$

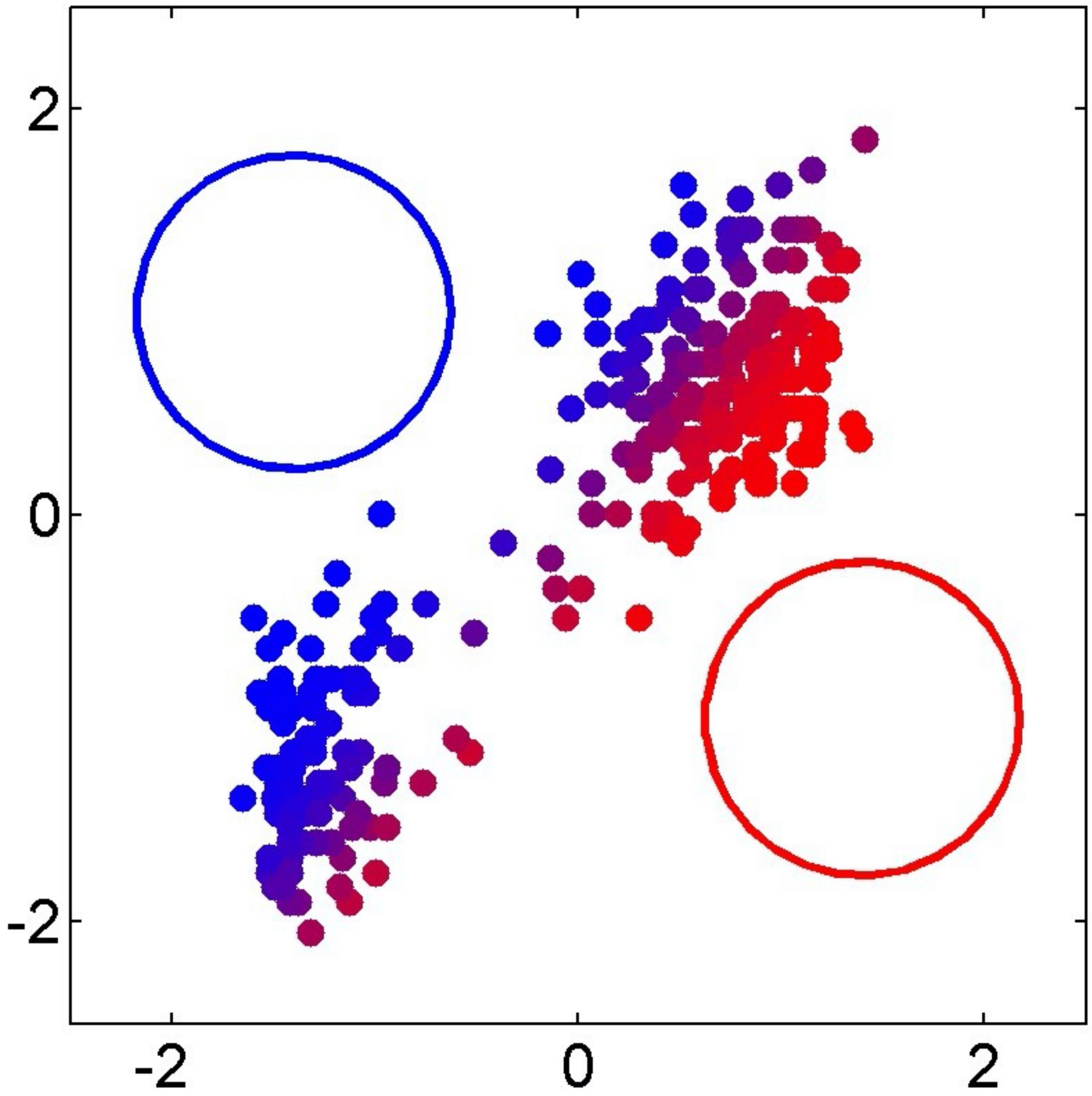
num of pts. @ k

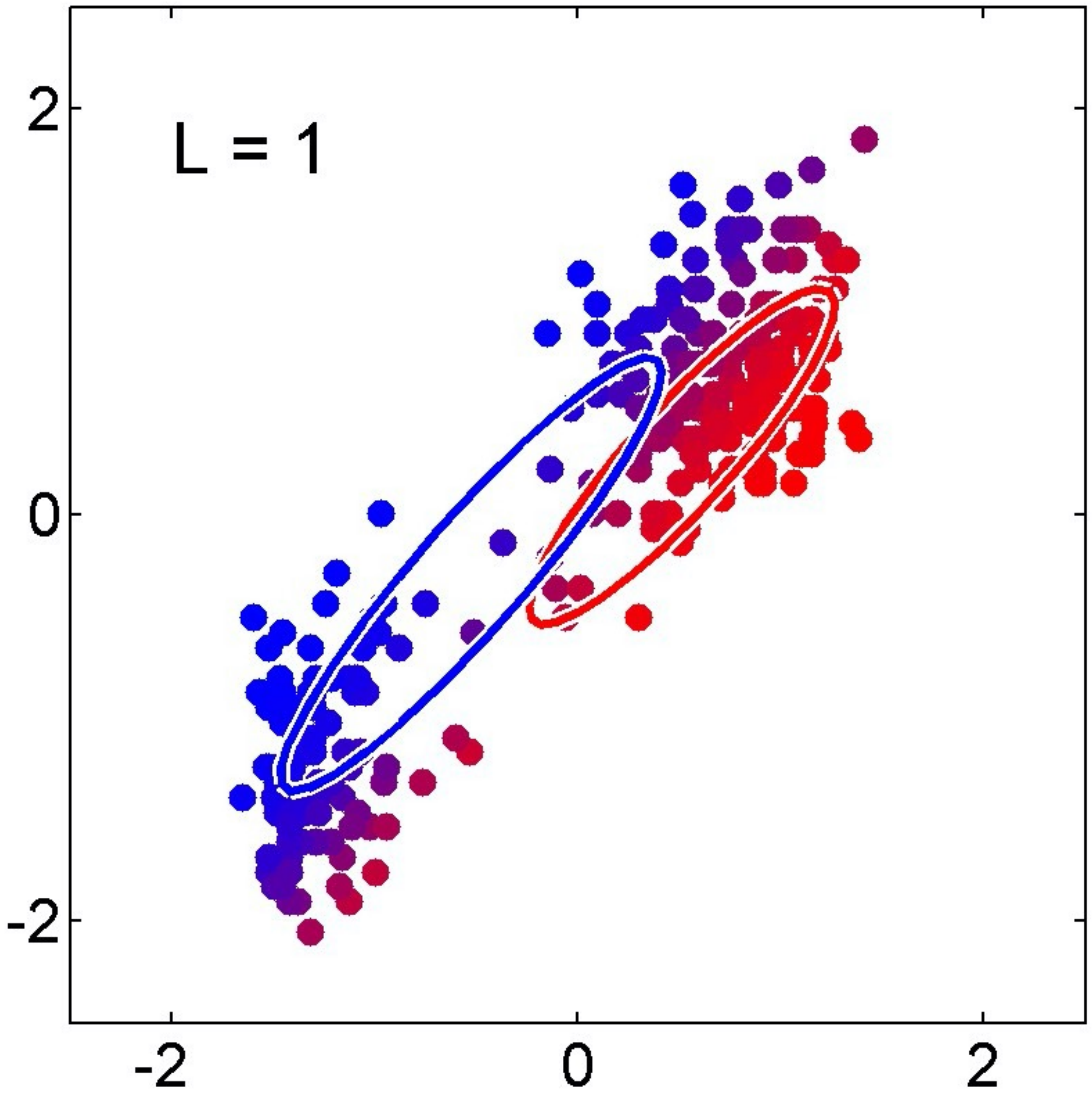
$$\pi_k = \frac{\sum_n \gamma(z_{nk})}{N} \qquad (9.22)$$

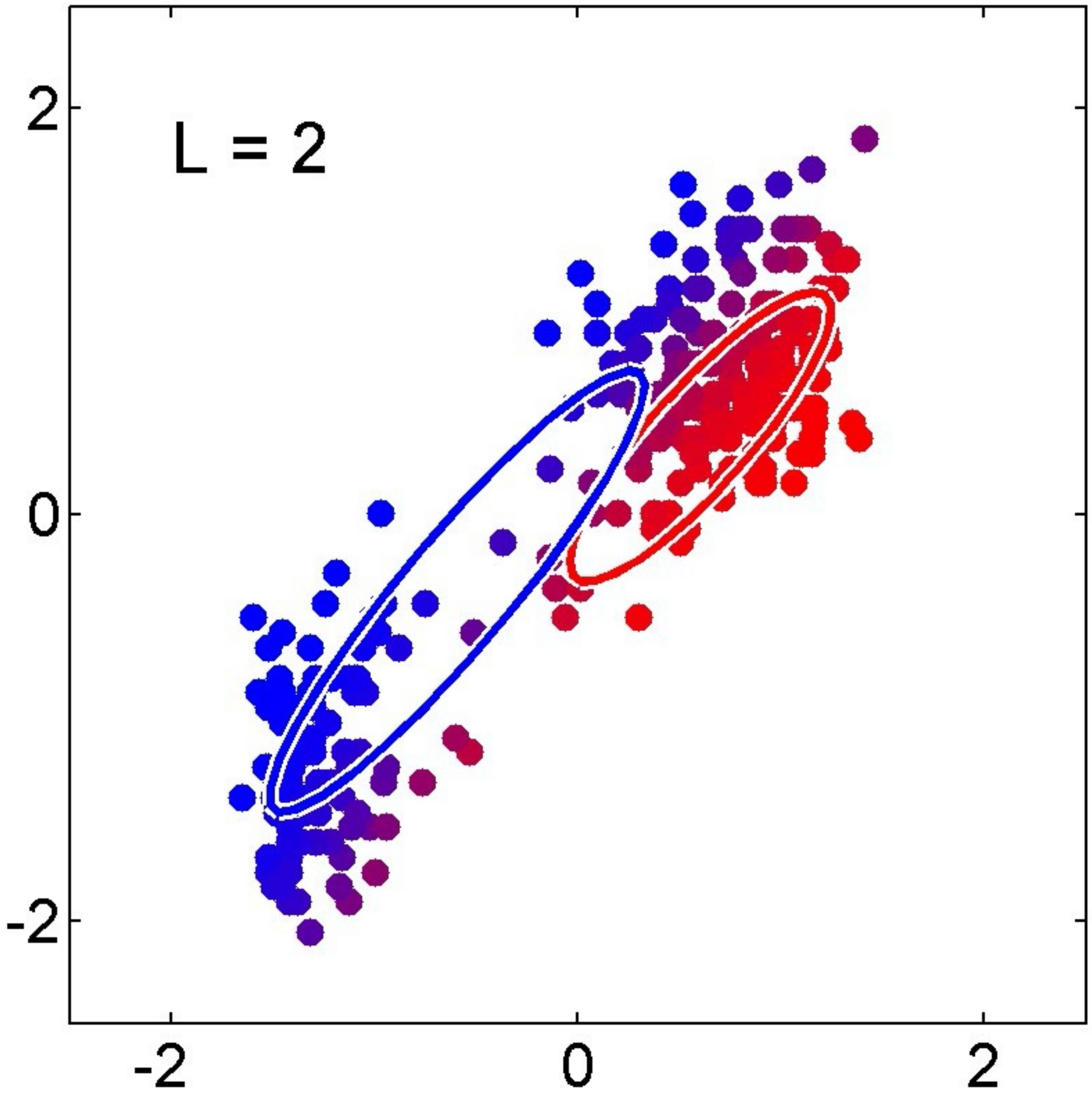
total num of pts.

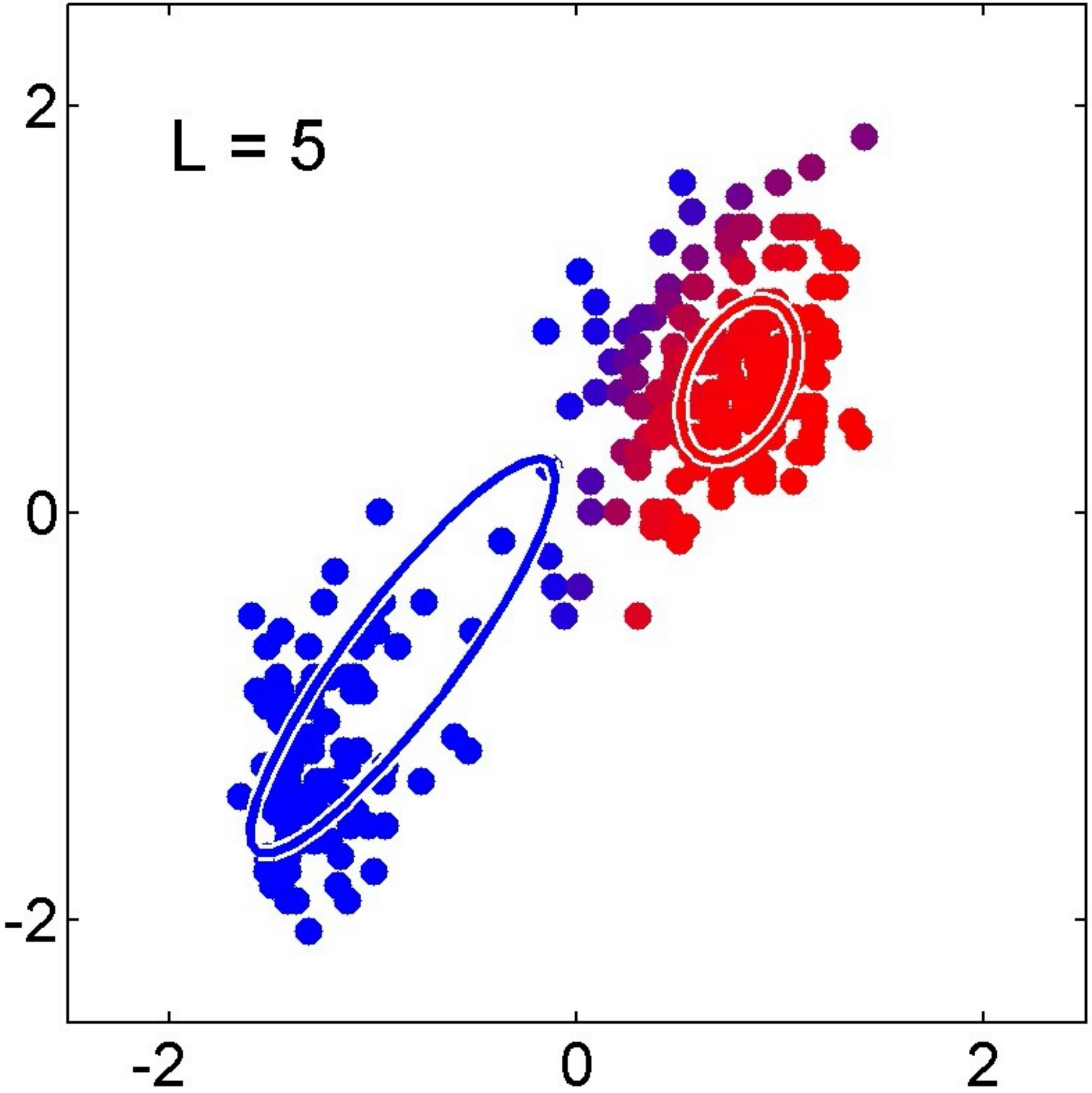

2
0
-2
-2
0
2

E



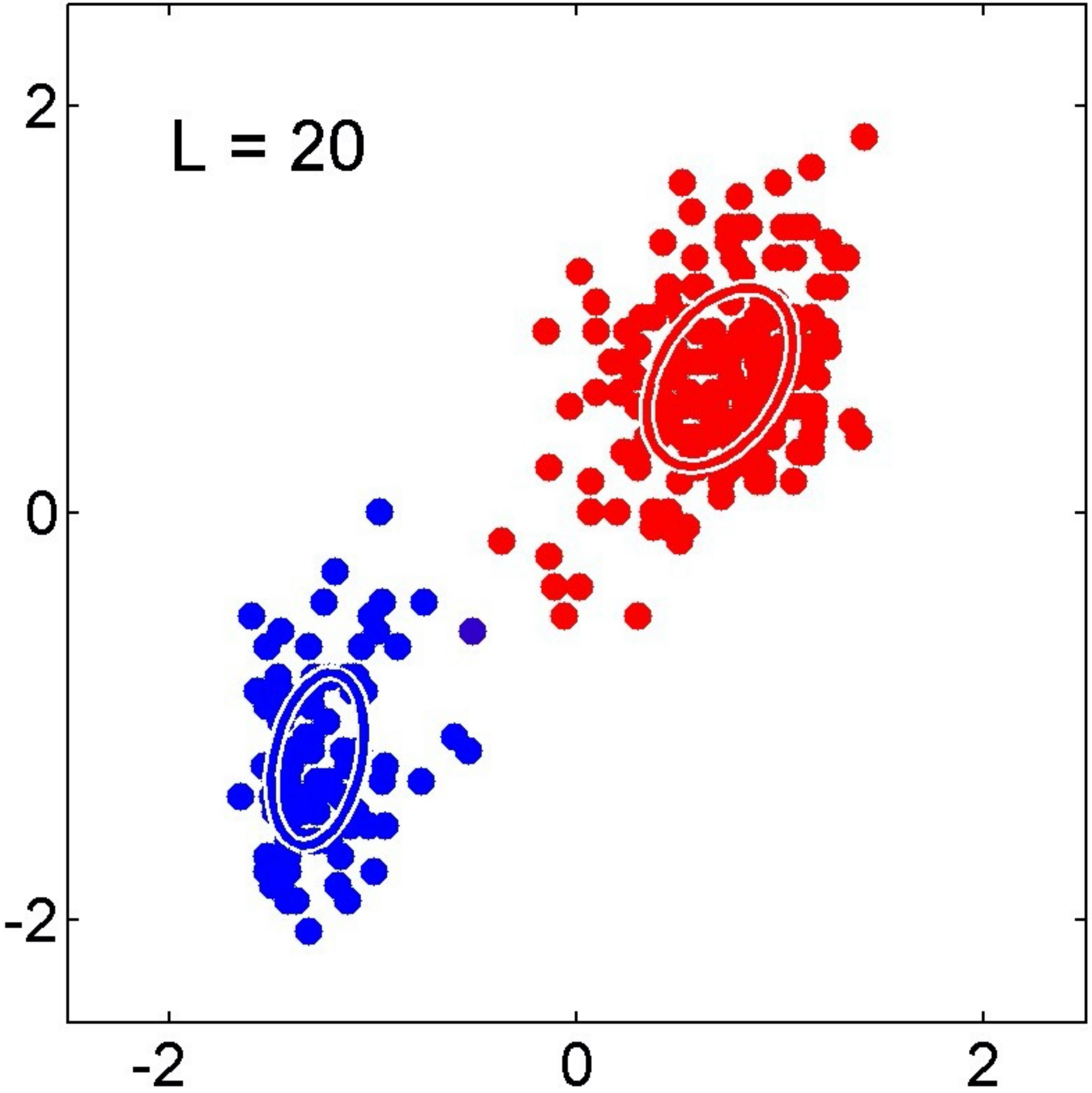

L = 1
2
0
-2
-2
0
2


L = 2
2
0
-2
-2
0
2



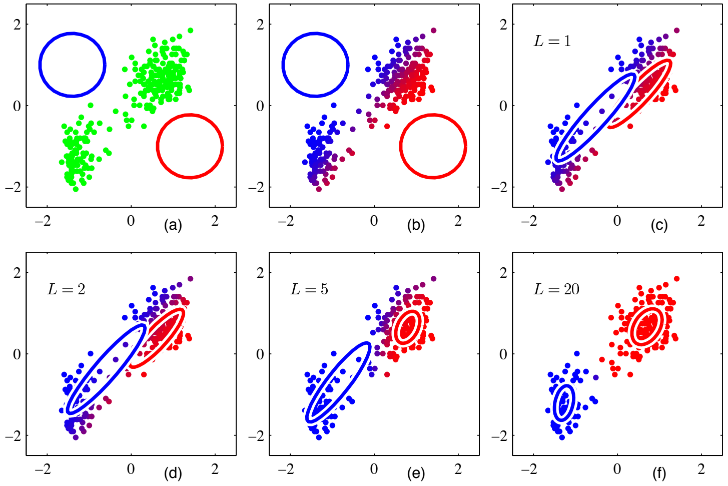
L = 5
2
0
-2
-2
0
2

E



Mixture of Gaussians: EM for Gaussian Mixtures Example

- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood:
  Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\mu_k, \Sigma_k, \pi_k\}$.



- More iterations needed to converge than K-means algorithm, and each cycle requires more computation
- Common, initialise parameters based K-means run.

Mixture of Gaussians: EM for Gaussian Mixtures Summary

1. Initialize $\{\mu_k, \Sigma_k, \pi_k\}$ and evaluate log-likelihood
2. E-Step Evaluate responsibilities $\gamma(z_{nk})$
3. M-Step Re-estimate paramters, using current responsibilities:

$$\mu_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_{nk})} \sum_n \gamma(z_{nk}) \mathbf{x}_n \quad (9.23)$$

$$\Sigma_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_{nk})} \sum_n \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \quad (9.24)$$

$$\pi_k^{\text{new}} = \frac{\sum_n \gamma(z_{nk})}{N} \quad (9.25)$$

4. Evaluate log-likelihood $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ and check for convergence (go to step 2).

Practice Question with White Board