Expectation Maximization

Daniel Sadoc Menasché

(Text book: C. Bishop)

Unsupervised Learning

- Clustering
  - Given a set of points, automatically group them
- Solution
  - K-means Algorithm
- Generalization
  - Expectation Maximization (EM) Algorithm

Clustering Problem

- Given set of n points, # of clusters K
- Goal: compute
  - For each point n, r_{nk} - binary indicators
    - r_{nk} = 1 if point n in clusters k
    - r_{nk} = 0 otherwise
  - For each cluster k, μ_k – centers of K cluster
- Metric to be minimized: distortion

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Solution: Divide Into Two Problems

- Given set of n points, # of clusters K
- Goal: compute
  - PROBLEM 1 (E step):

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

  - PROBLEM 2 (M step): $\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$
- Metric to be minimized: distortion

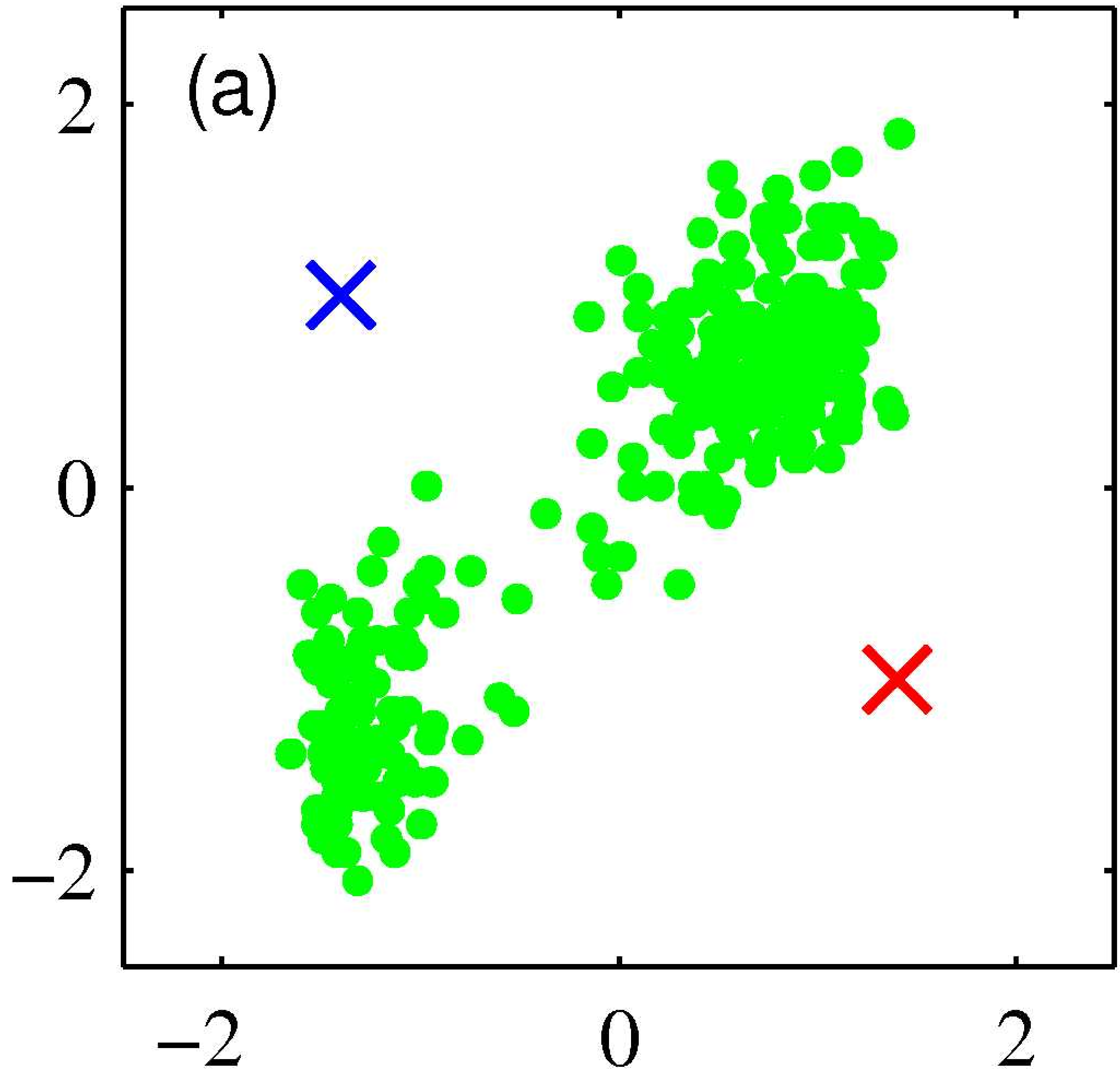
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

K Means Algorithm

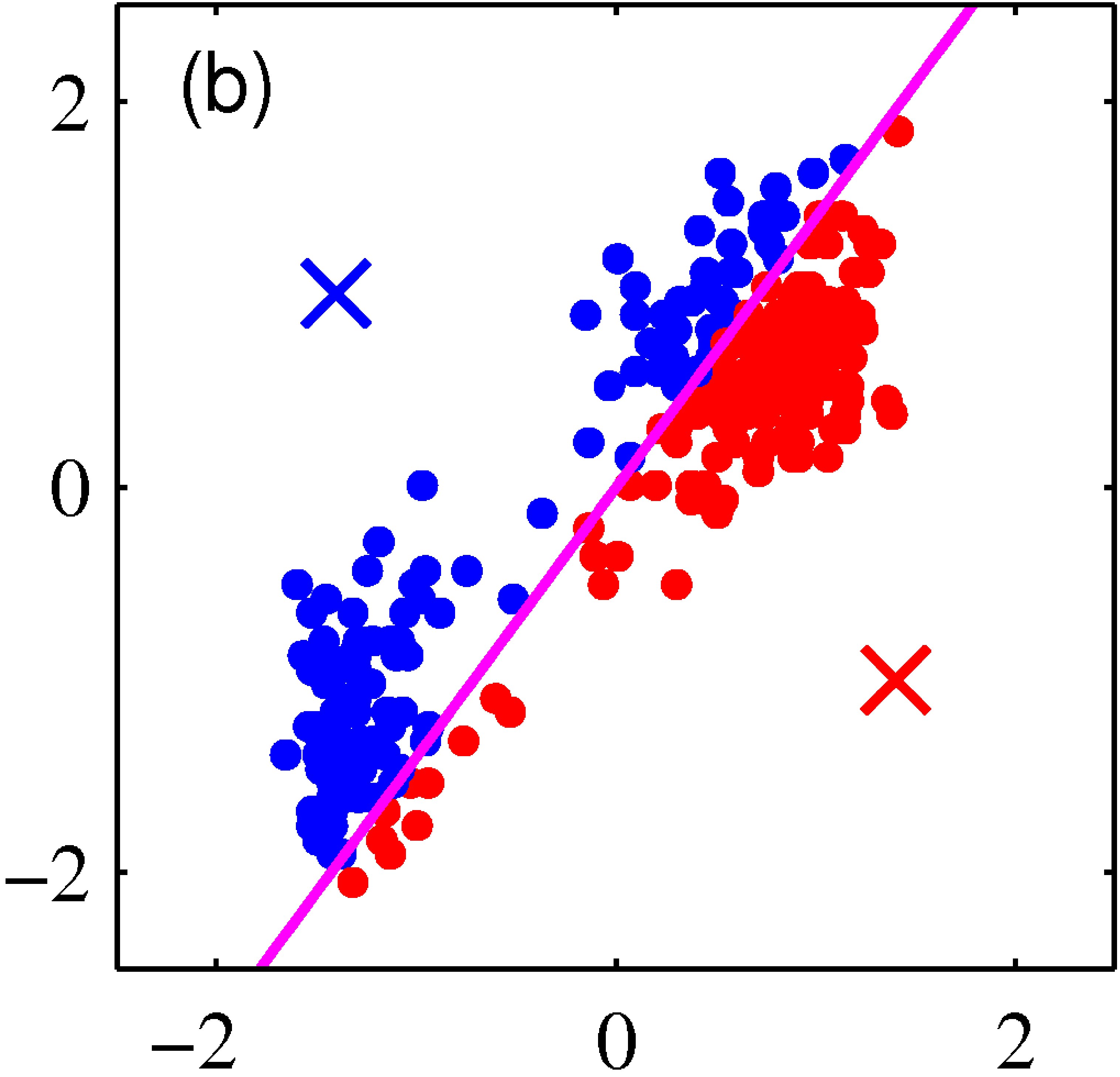
- Given set of n points, # of clusters K
- Iterate between the following two steps
  - E step:

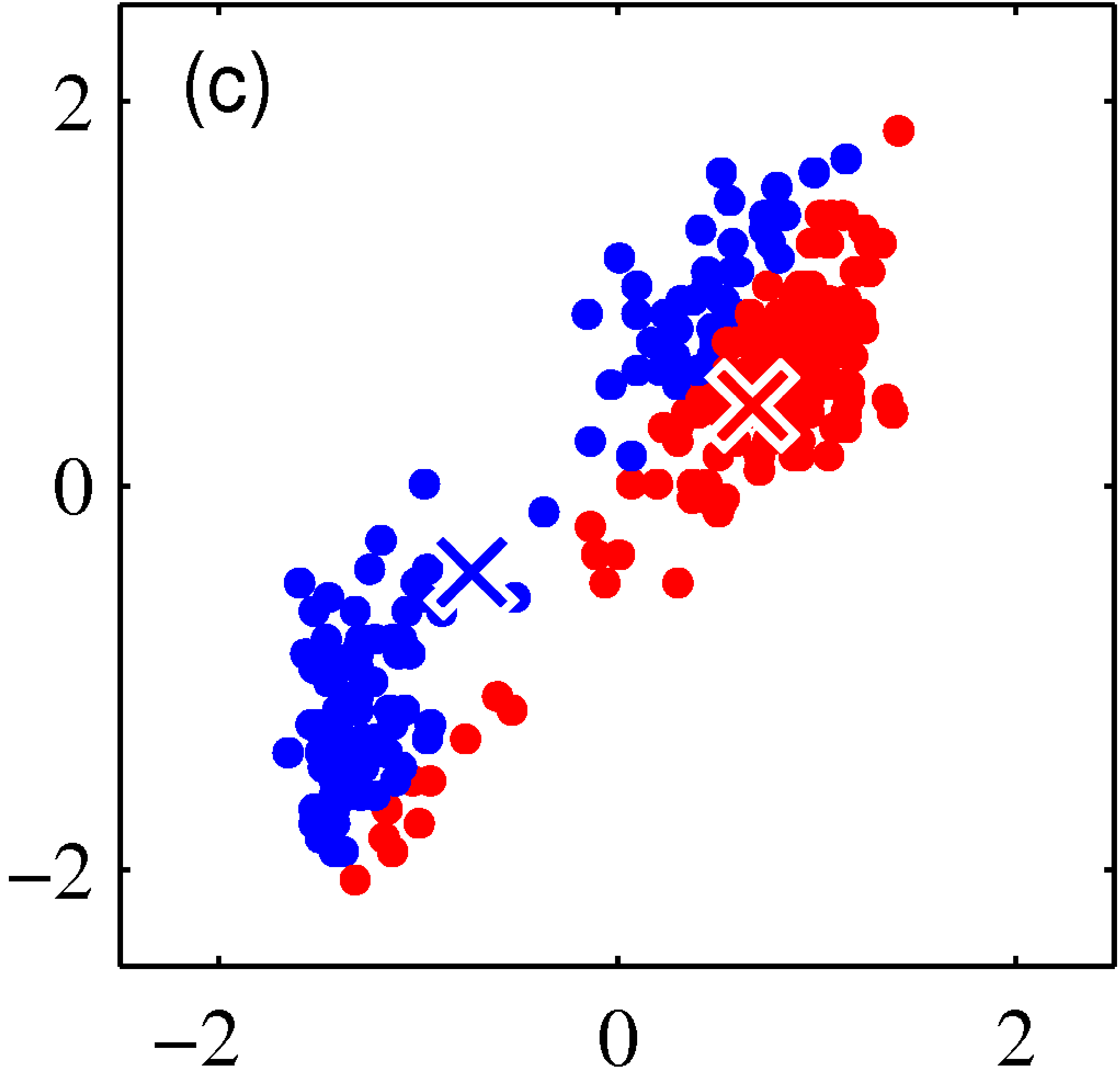
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

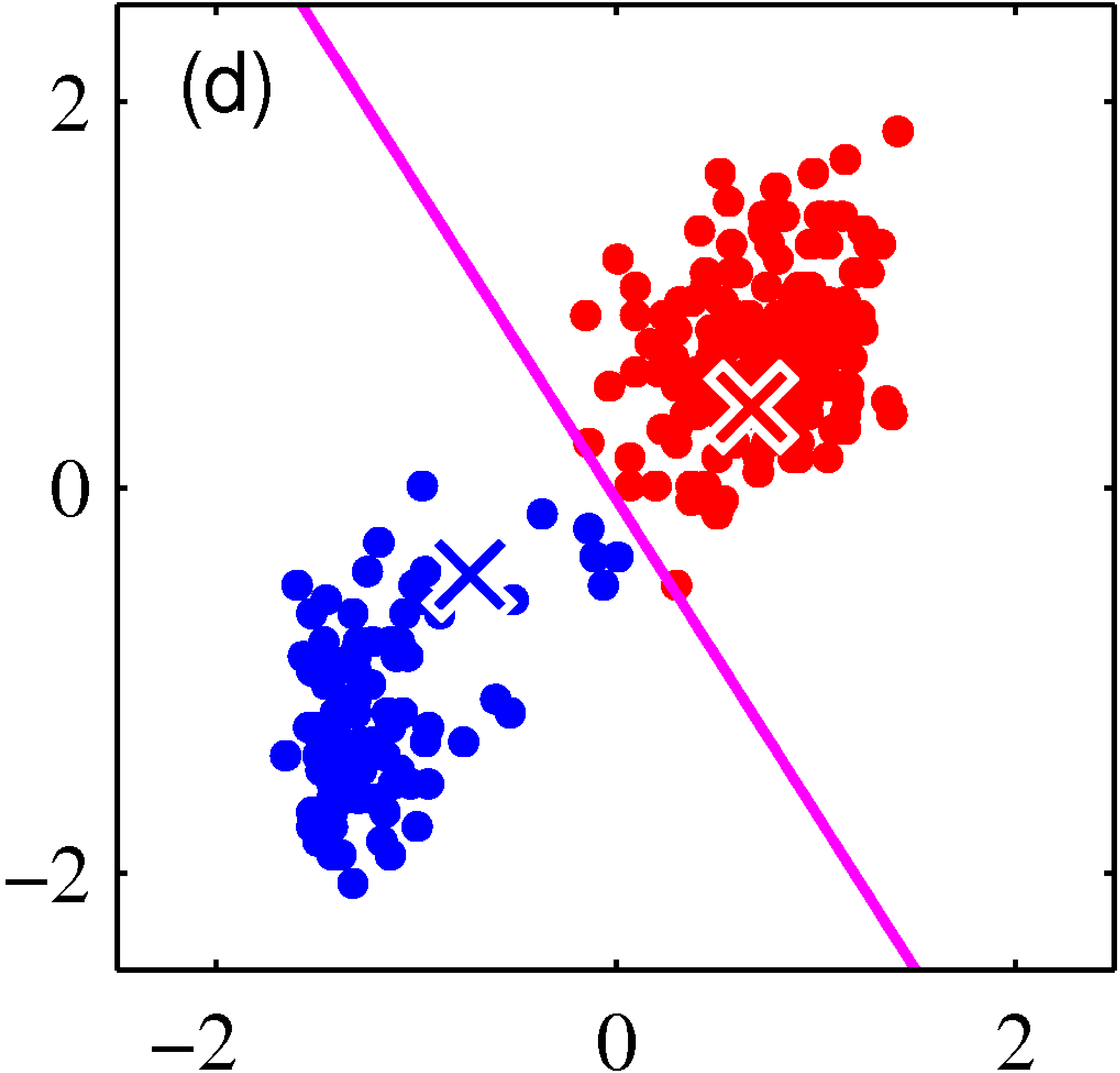
  - M step: $$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$$

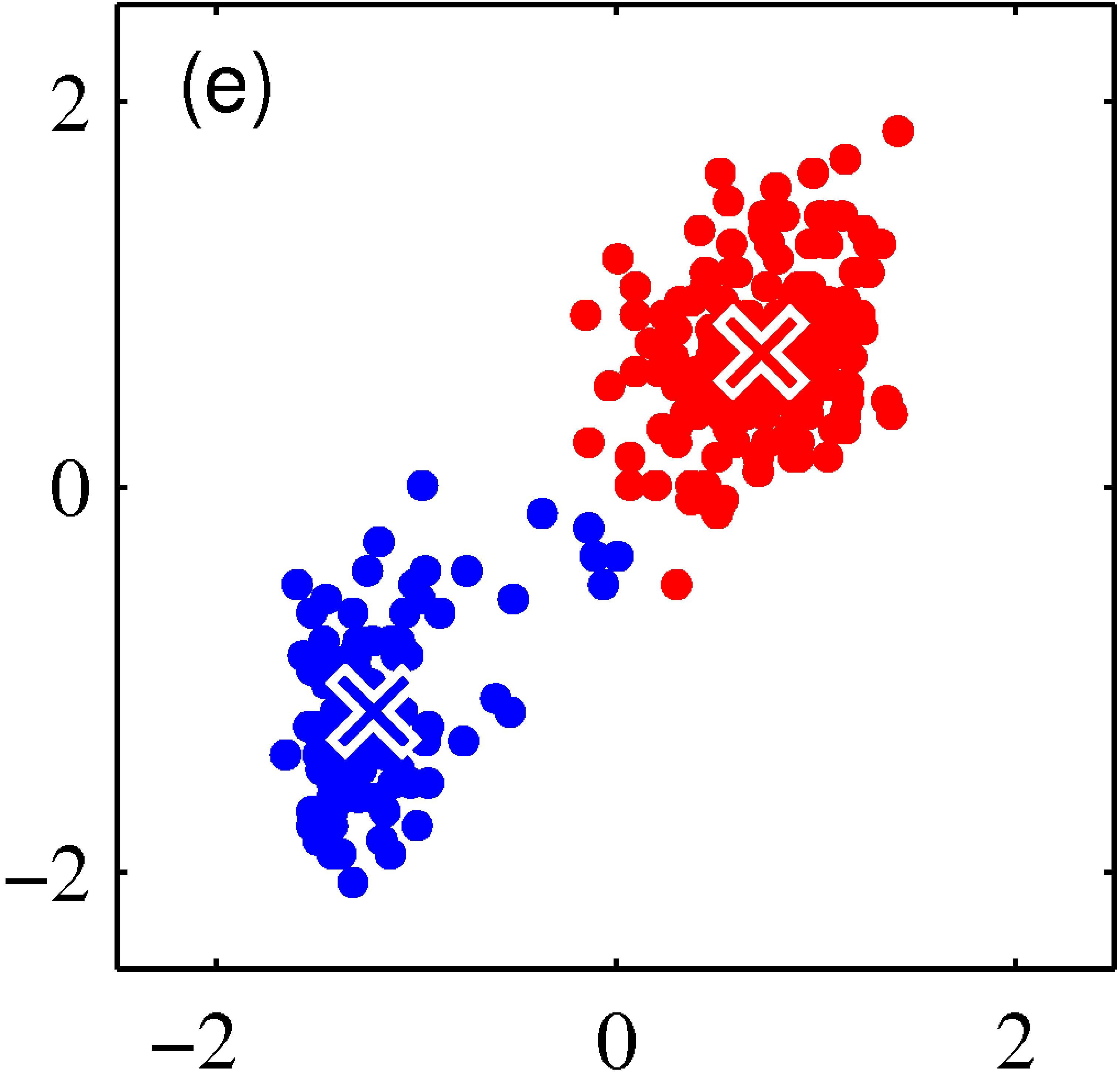


(a)
2
0
−2
−2
0
2

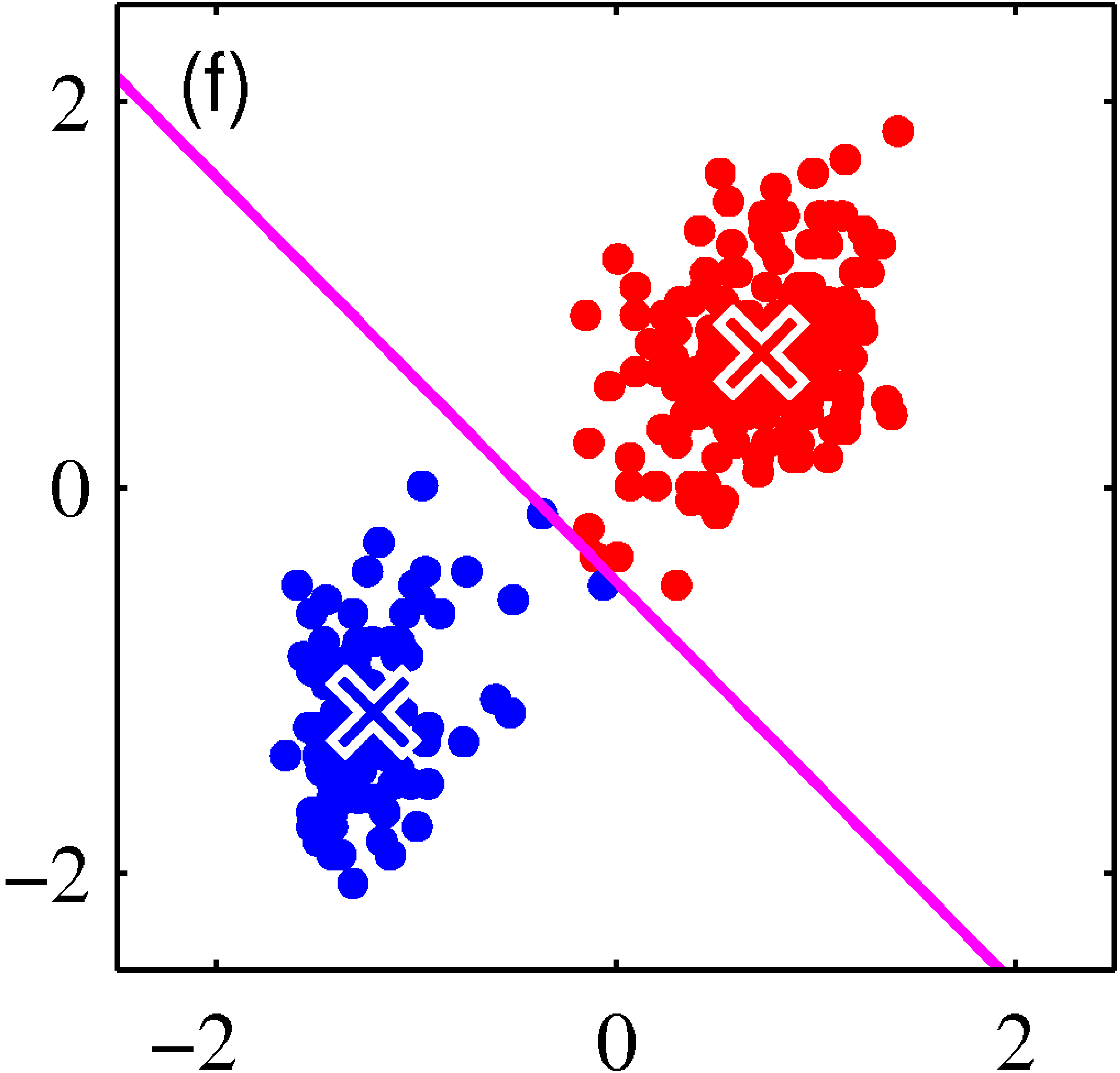

(b)
2
0
−2
−2
0
2


(c)
2
0
−2
−2
0
2

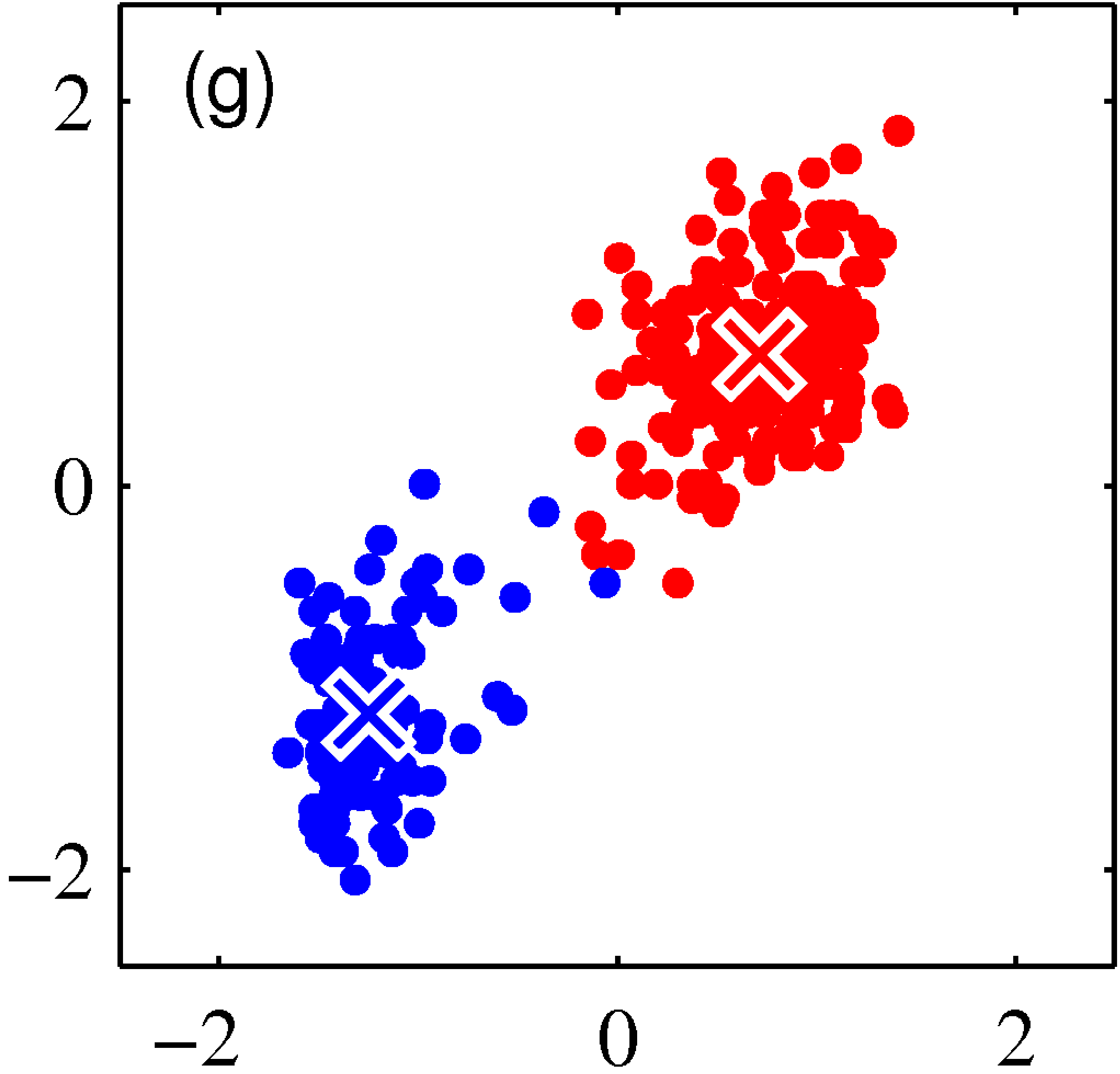

(d)
2
0
−2
−2
0
2

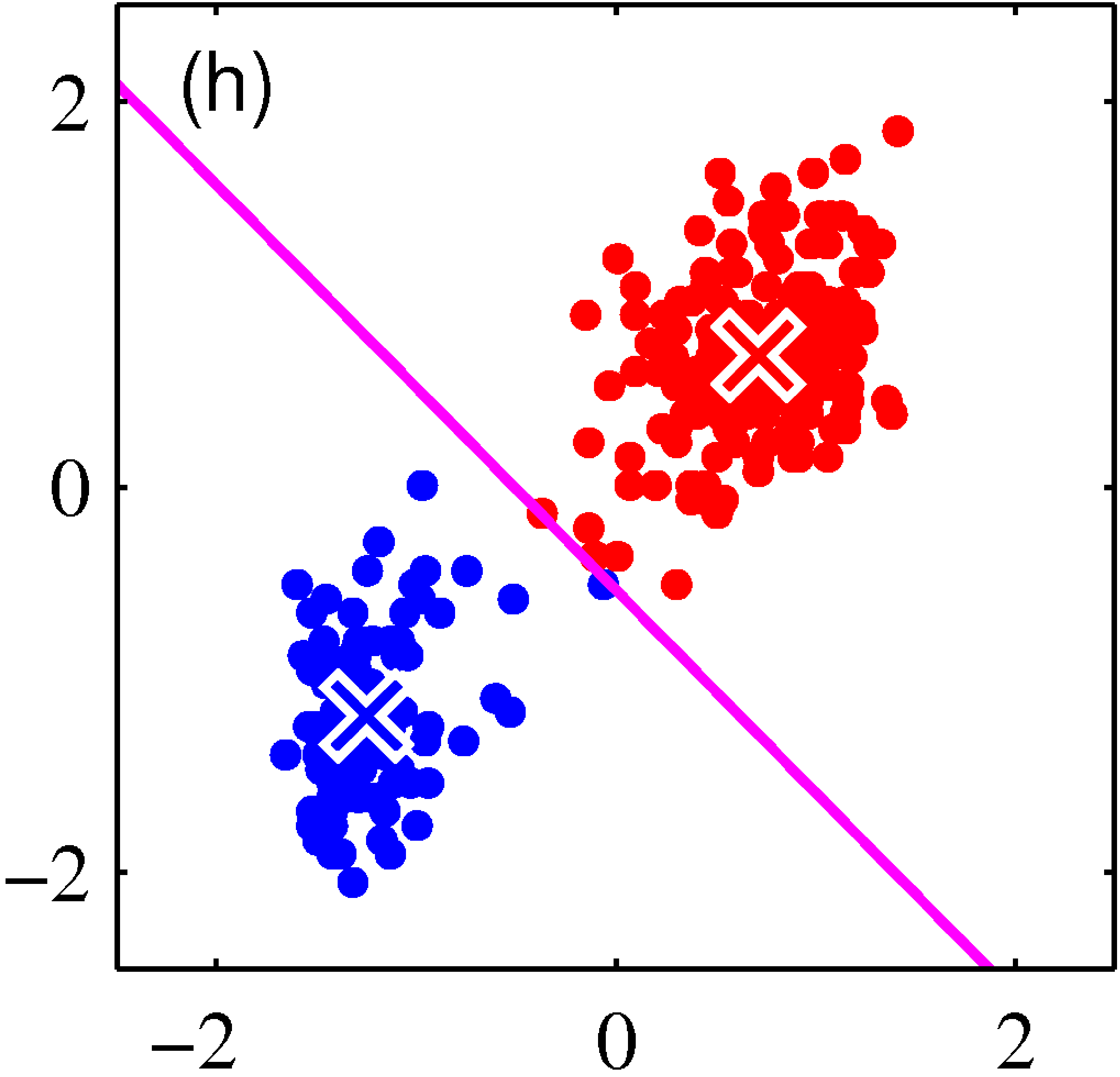


(e)
2
0
−2
−2
0
2

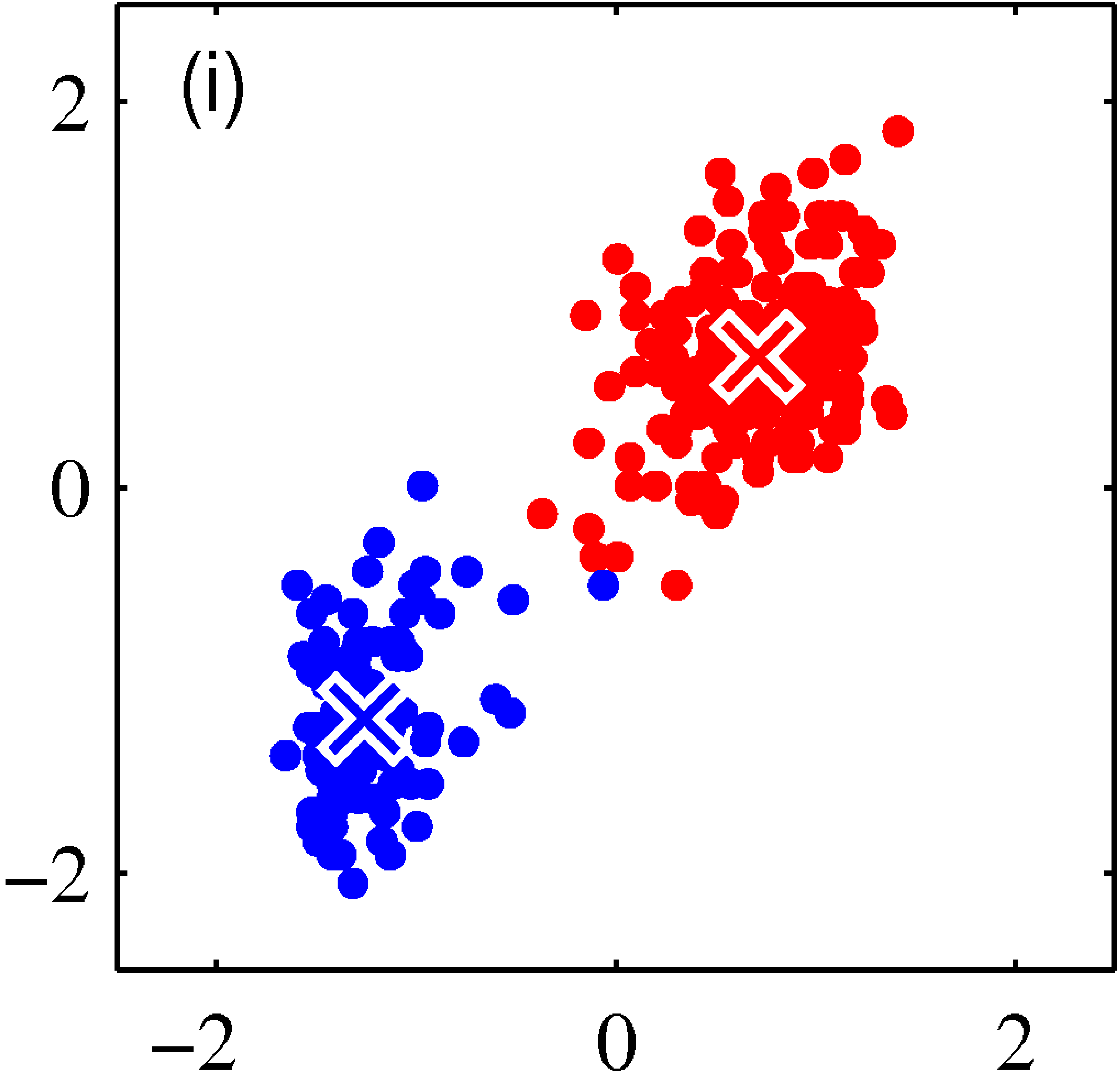


(f)
2
0
−2
−2
0
2

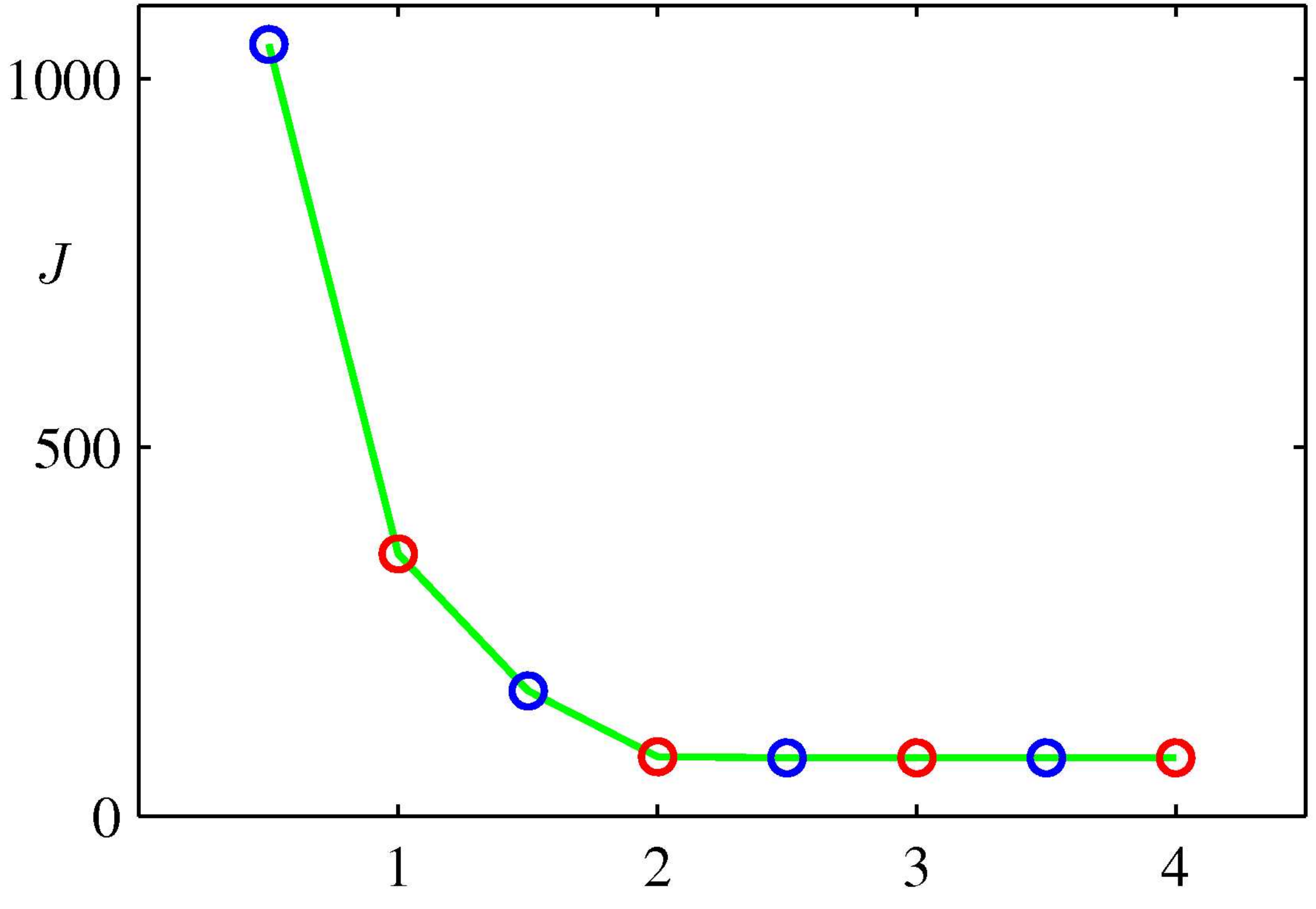

(g)
2
0
−2
−2
0
2



(h)
2
0
−2
−2
0
2


(i)
2
0
−2
−2
0
2

Error Decreases w/ Iterations



K Means Applicability for Image Processing

- Segmentation
- Compression
- Key idea: (R,G,B) defines point in 3D

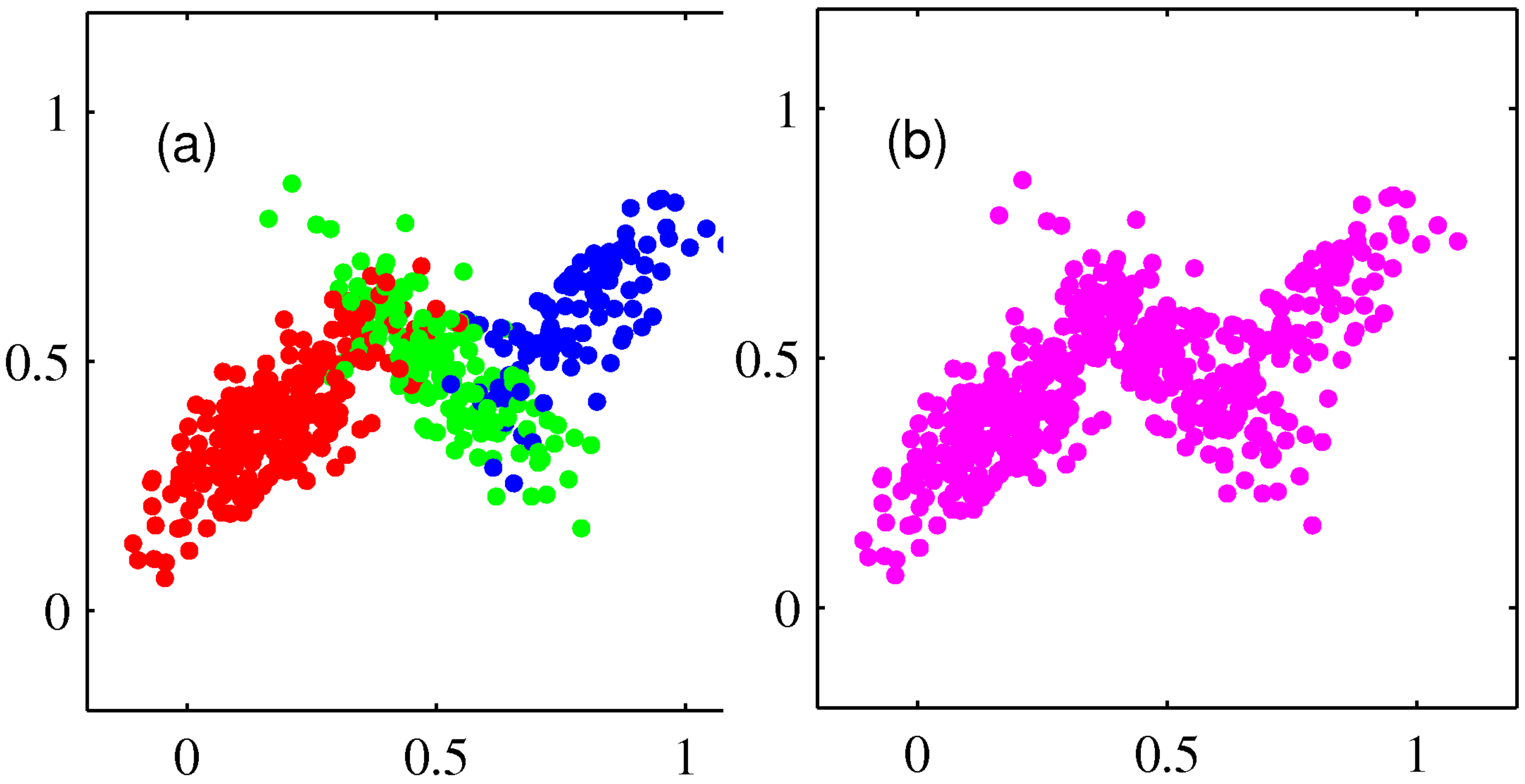
Original image

$K = 2$

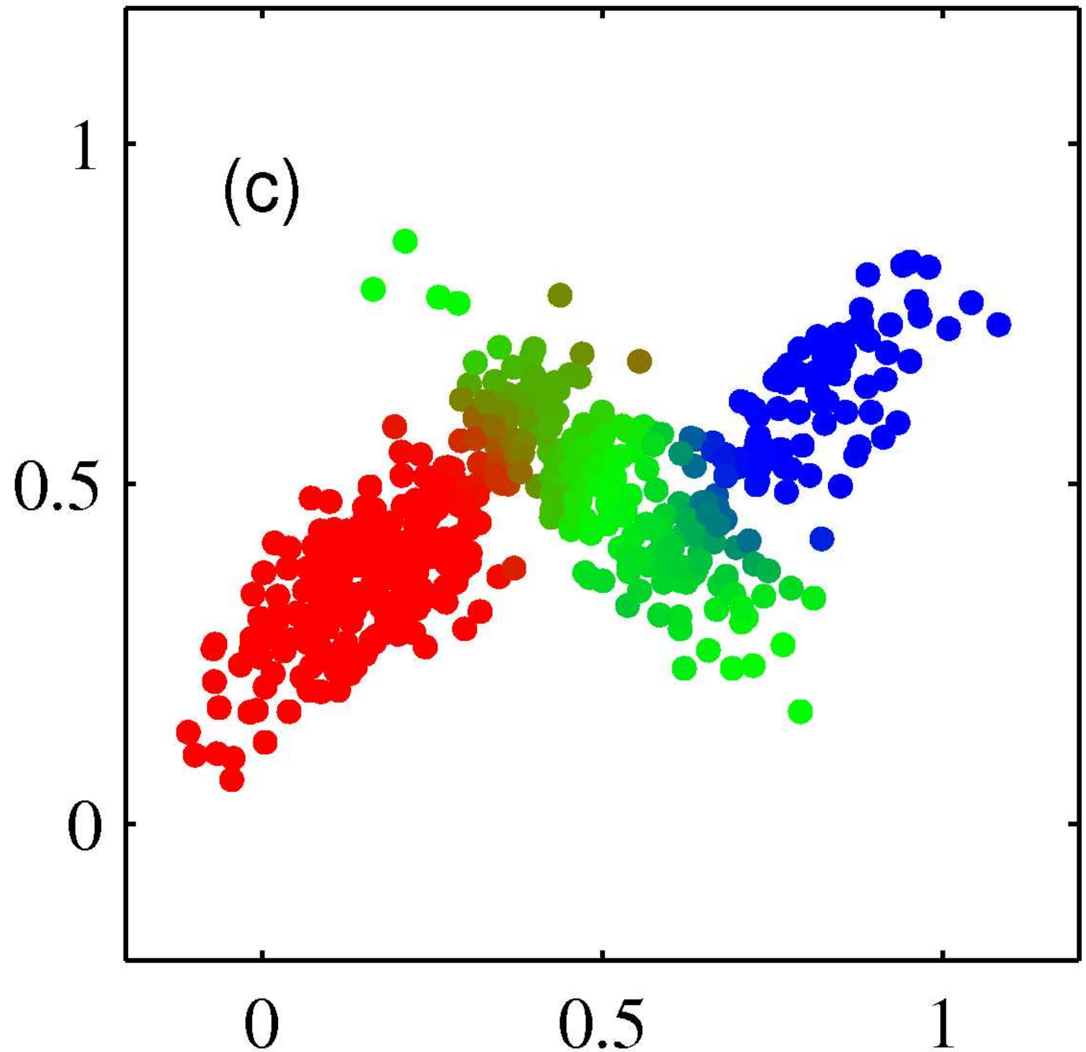
$$K = 3$$

$$K = 10$$

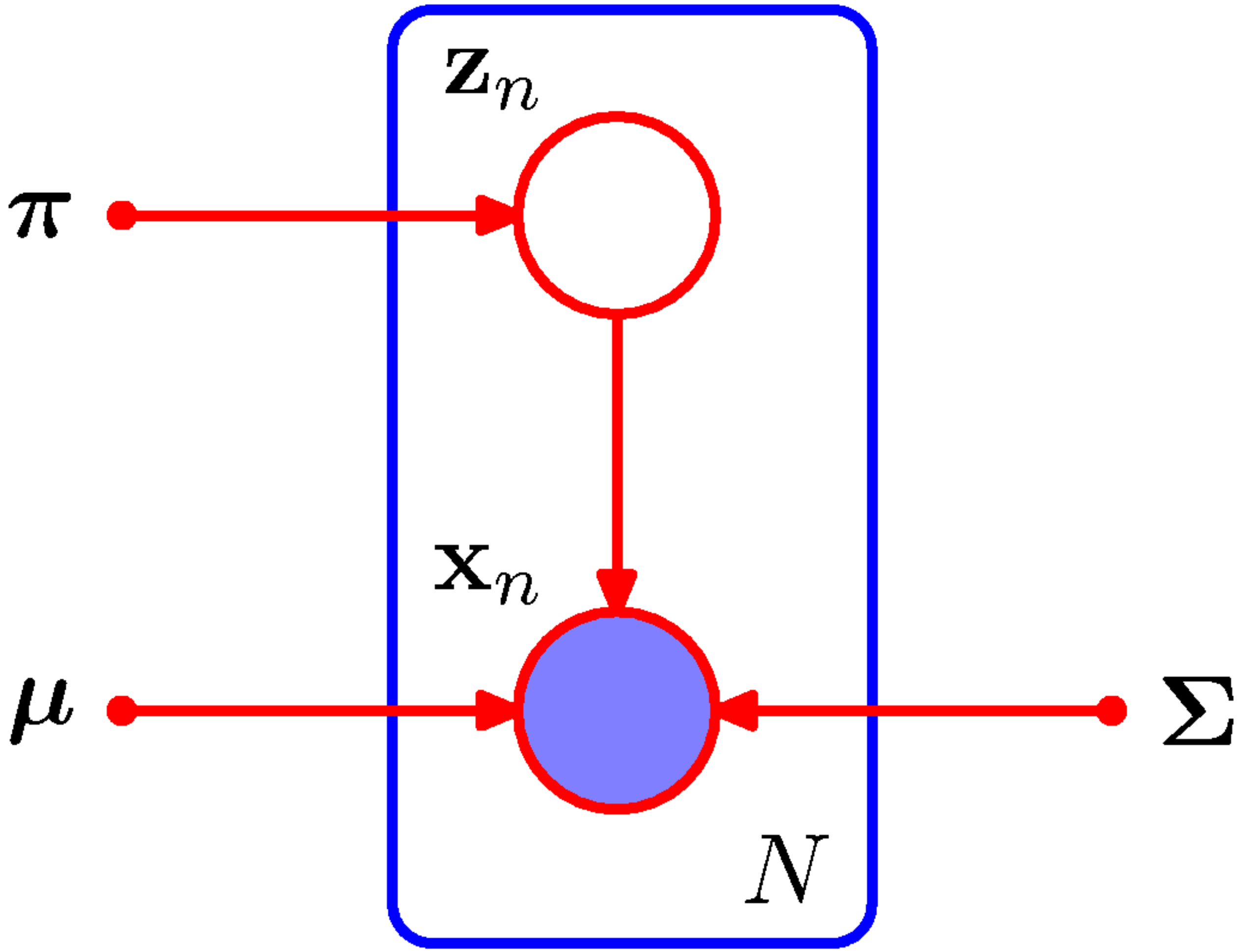
Mixture of Gaussians

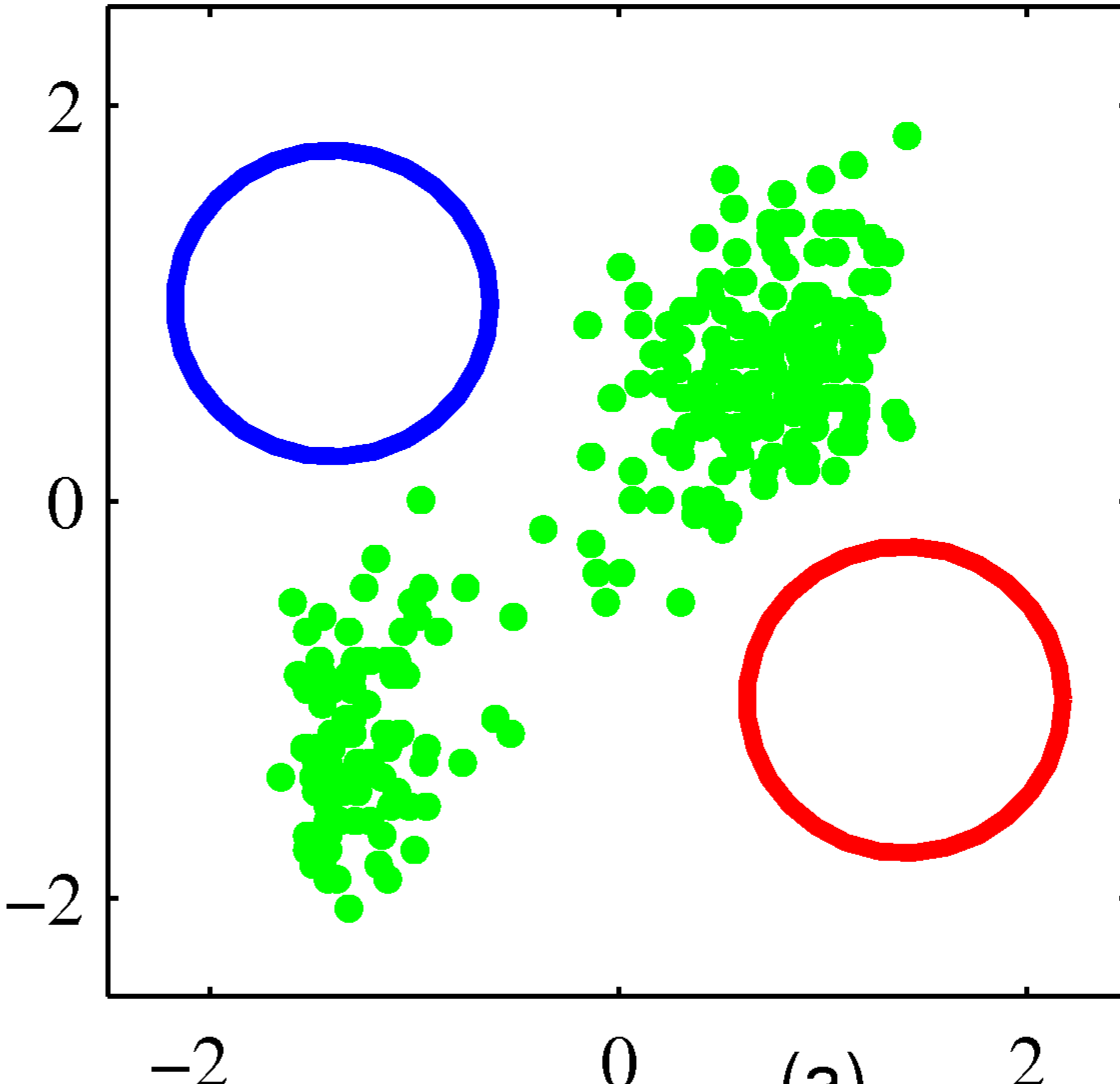


Mixture of Gaussians: Output

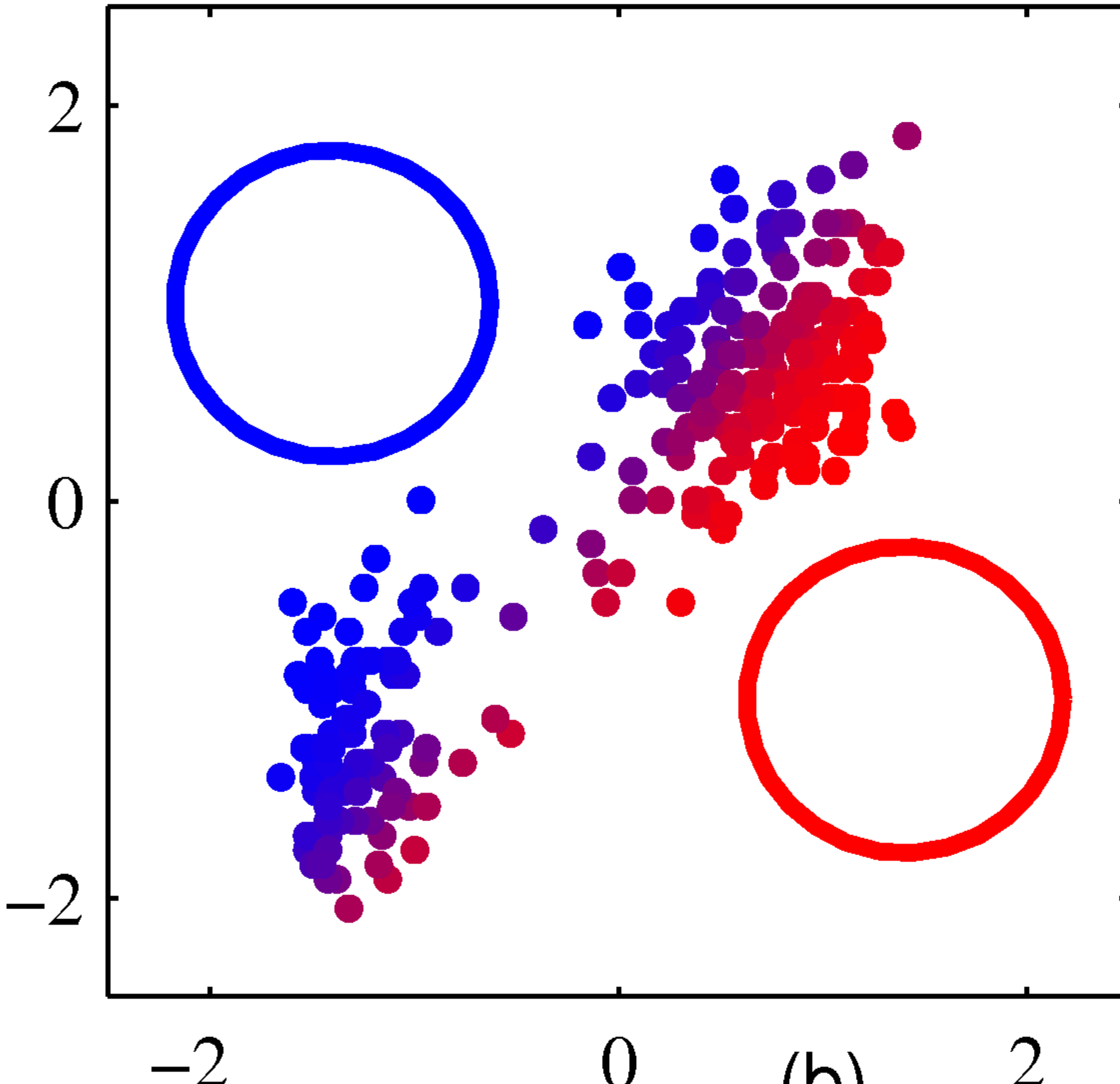


Graphical Model of Mixture of Gaussians

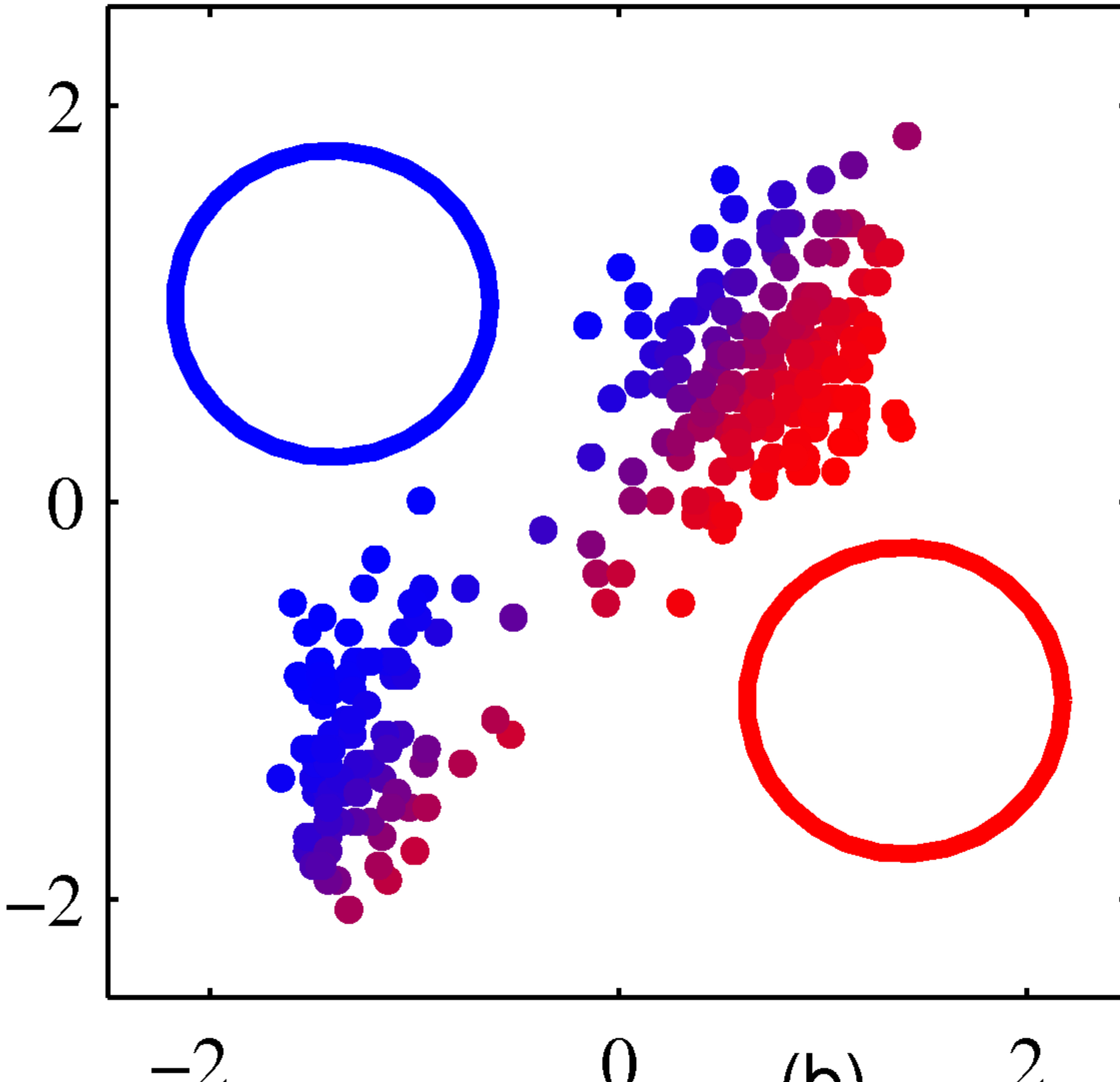




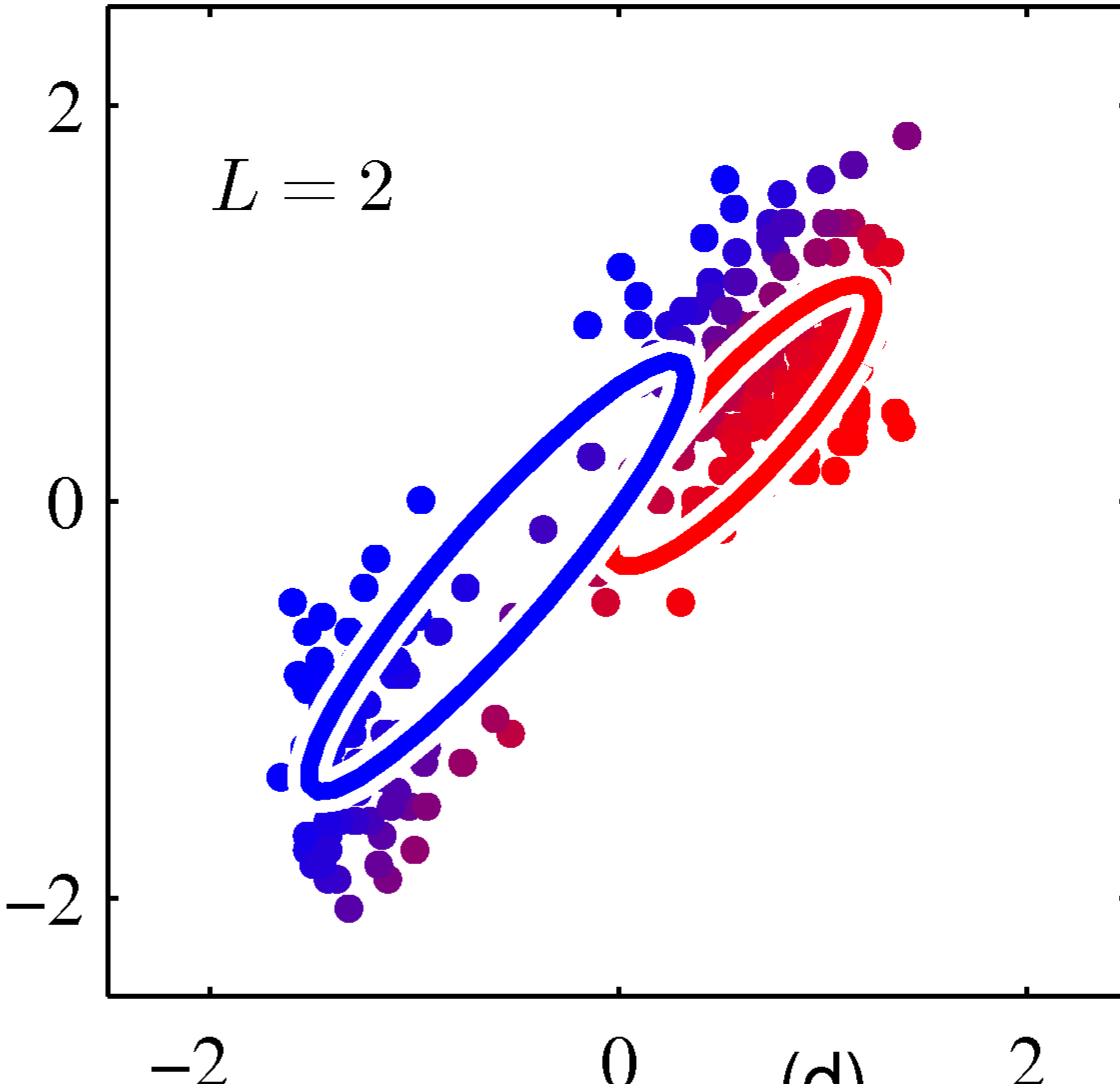
(a)



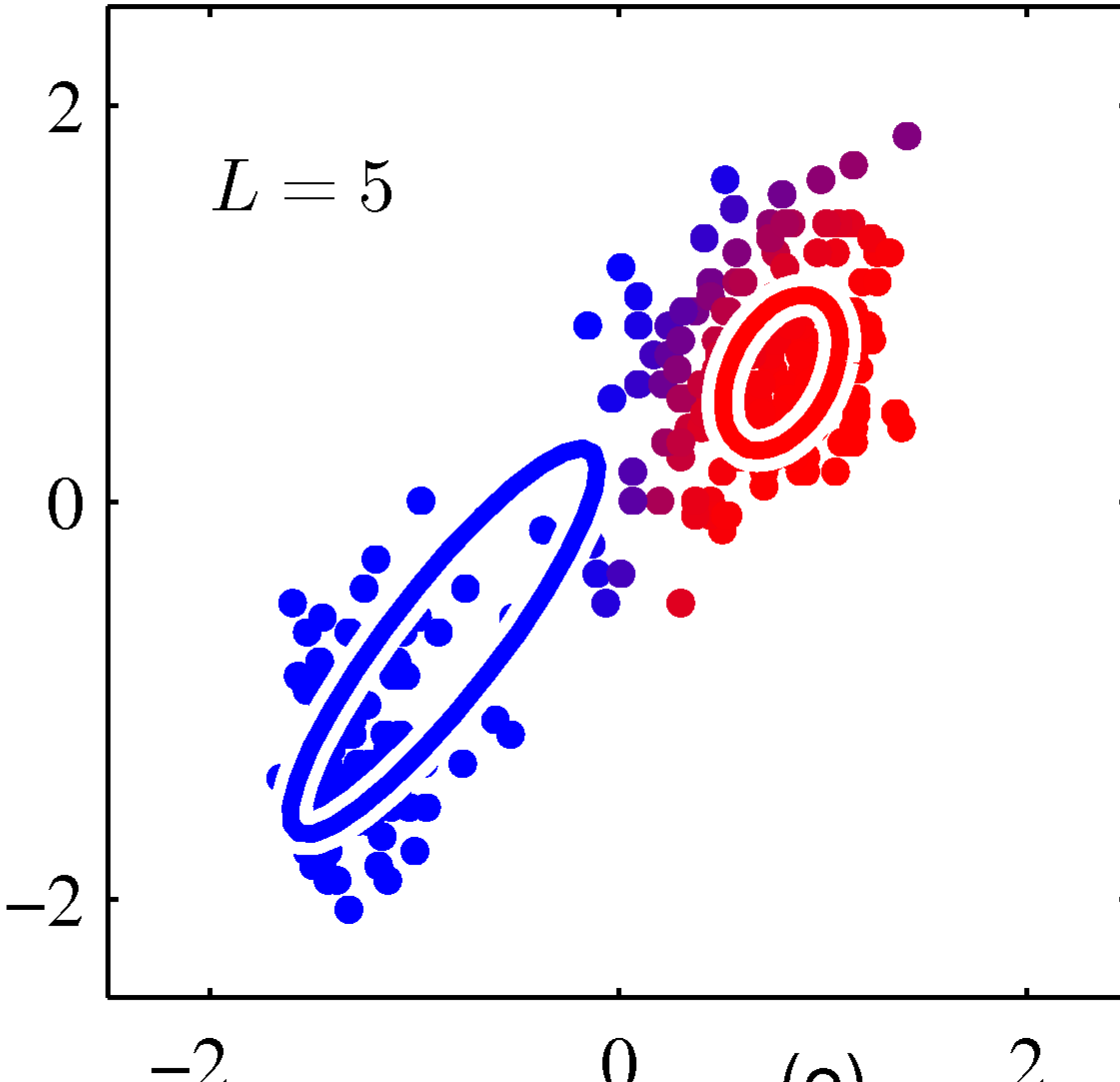
(b)



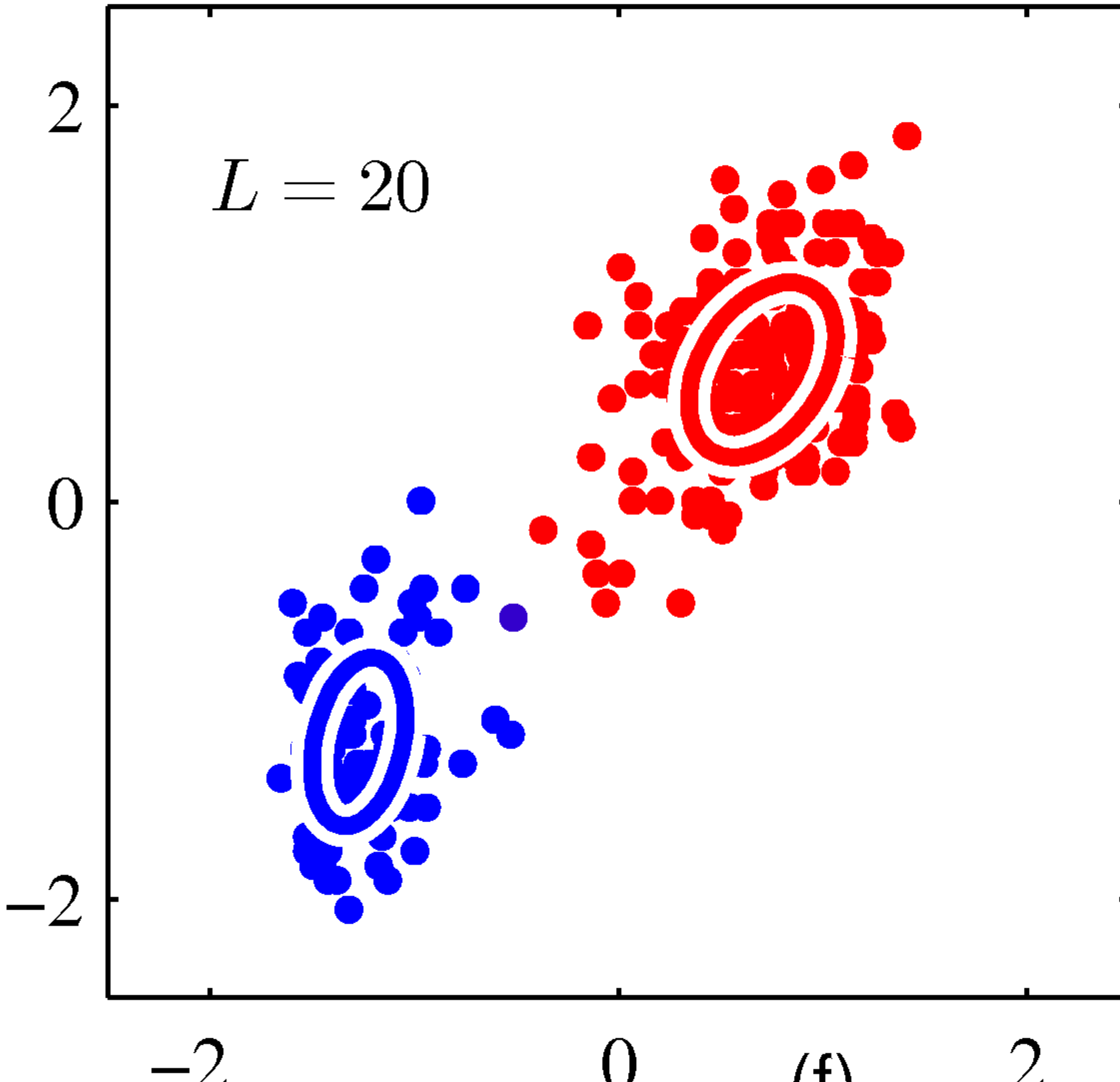
(b)



(d)



(e)



(f)

Mixture of Gaussians

- Problem

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1$$

$$p(z_k = 1) = \pi_k \qquad \sum_{k=1}^{K} \pi_k = 1$$

- Using 1-of-K represenation

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

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- $$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Mixture of Gaussians

- Responsibilities

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Using 1-of-K represenation

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

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Expectation Maximization For Mixture of Gaussians

- 1. Initialize
- 2. **E-Step**: compute responsibilities
- 3. **M-Step**: re-estimate parameters
    - means,
    - standard deviations,
    - latent probabilities

  using responsibilities
- 4. Compute likelikhood of data.
- 5. If likelihood > epsilon, return to 2

Generalizing Expectation-Maximization

- Goal (difficult): maximize

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}.$$

- Instead (easy)
  - Assume know classes of elements

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

  - Maximize function of

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Generalizing Expectation-Maximization

- Goal (difficult): maximize

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}.$$

- Instead (easy)
  - (E Step) Assume know classes of elements

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

  - (M Step) Maximize function of

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

General E-M Algorithm

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.

2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

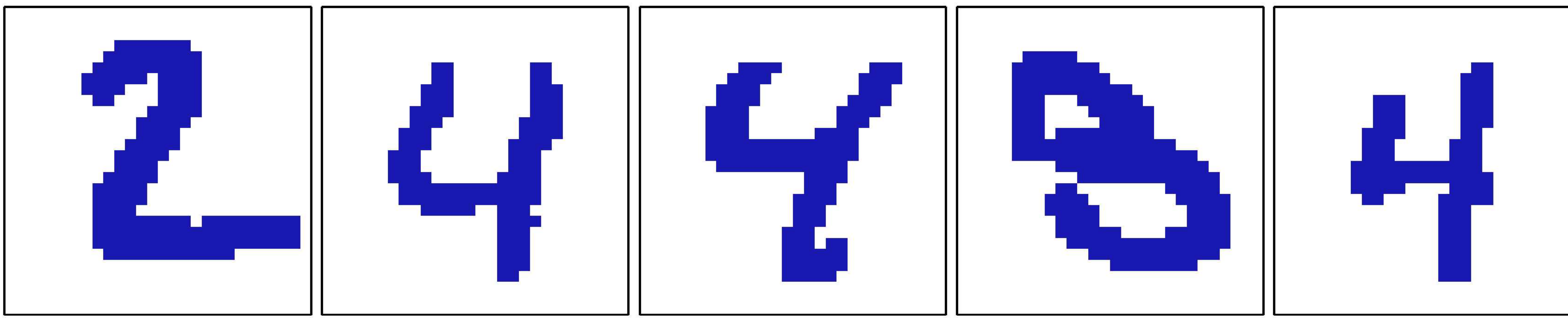
4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

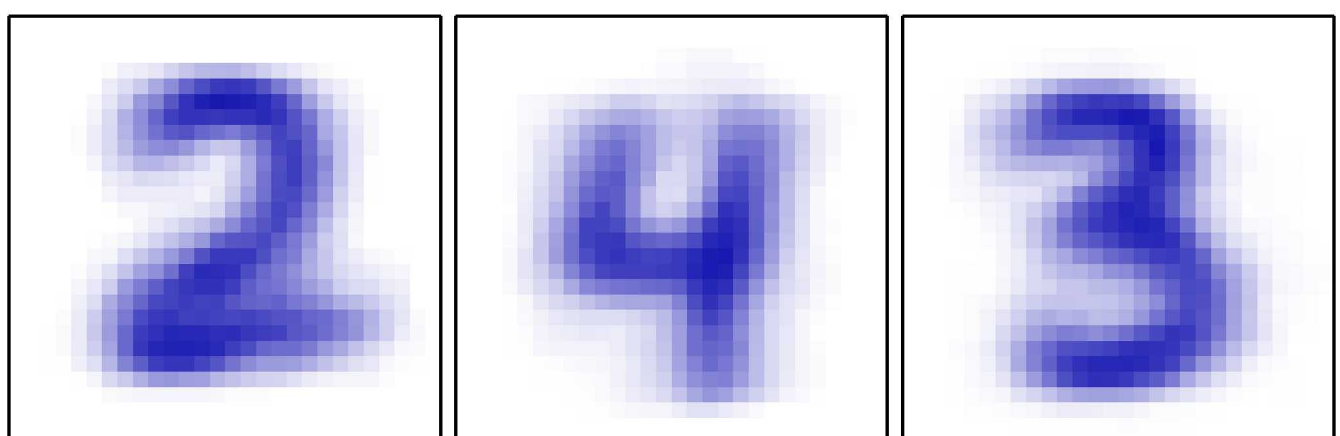
and return to step 2.

Unsupervised Digit Recognition
Mixture of Bernoullis

- Input



- Output



EM Algorithm In General

- Want to maximize

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

- Key observation

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \end{aligned}$$

EM Algorithm In General

- Want to maximize

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

Lower bound to be maximized

- Key observation

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

- Follows from

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta})$$

EM Algorithm In General

- Want to maximize

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

- Key observation

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \end{aligned}$$

- E-step: maximize lower bound with respect to q
- M-step: maximize lower bound with respect to θ

EM Algorithm In General

- Want to maximize

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

- Key observation

Lower bound to be maximized

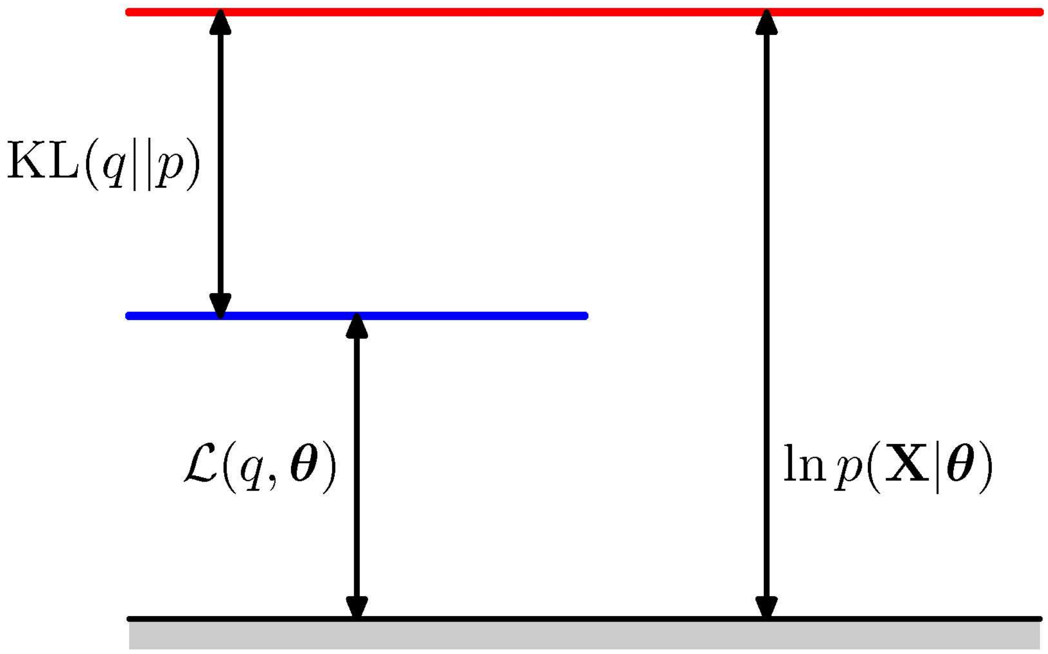
$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \end{aligned}$$

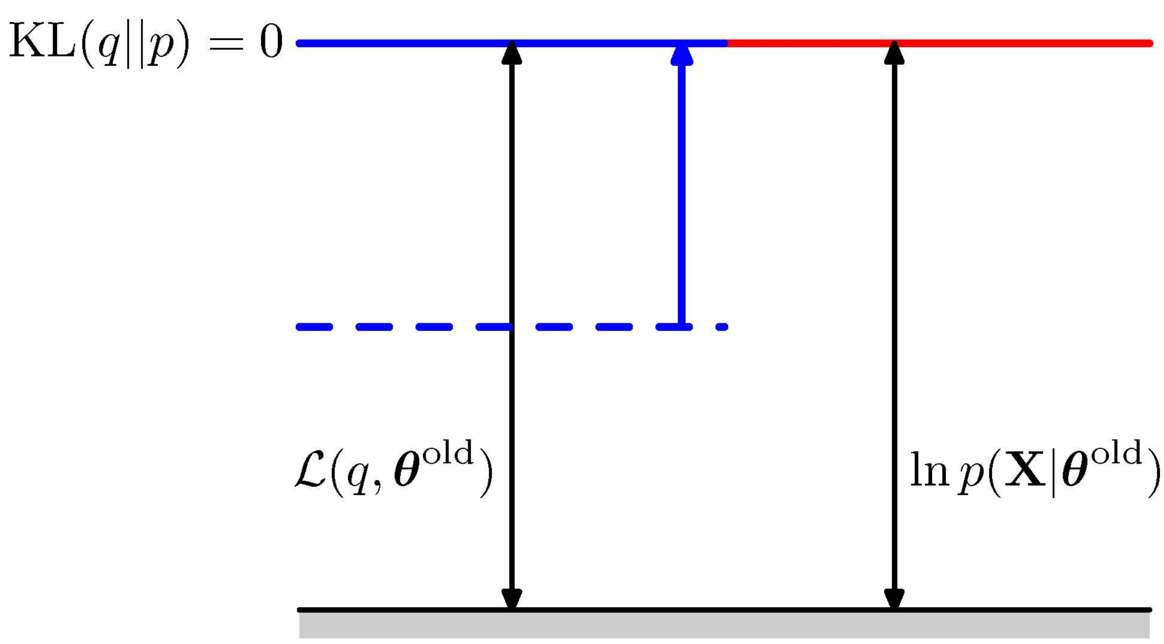
- E-step: q=p
- M-step: max

$$\begin{aligned} \mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \text{const} \end{aligned} \tag{9.74}$$

Expectation Maximization Convergence Proof



Expectation Maximization Convergence Proof



Expectation Maximization Convergence Proof