Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012

Prof. Erik Sudderth

Lecture 18:

Clustering & K-Means Algorithm

EM for Gaussian Mixture Models

Many figures courtesy Kevin Murphy's textbook,

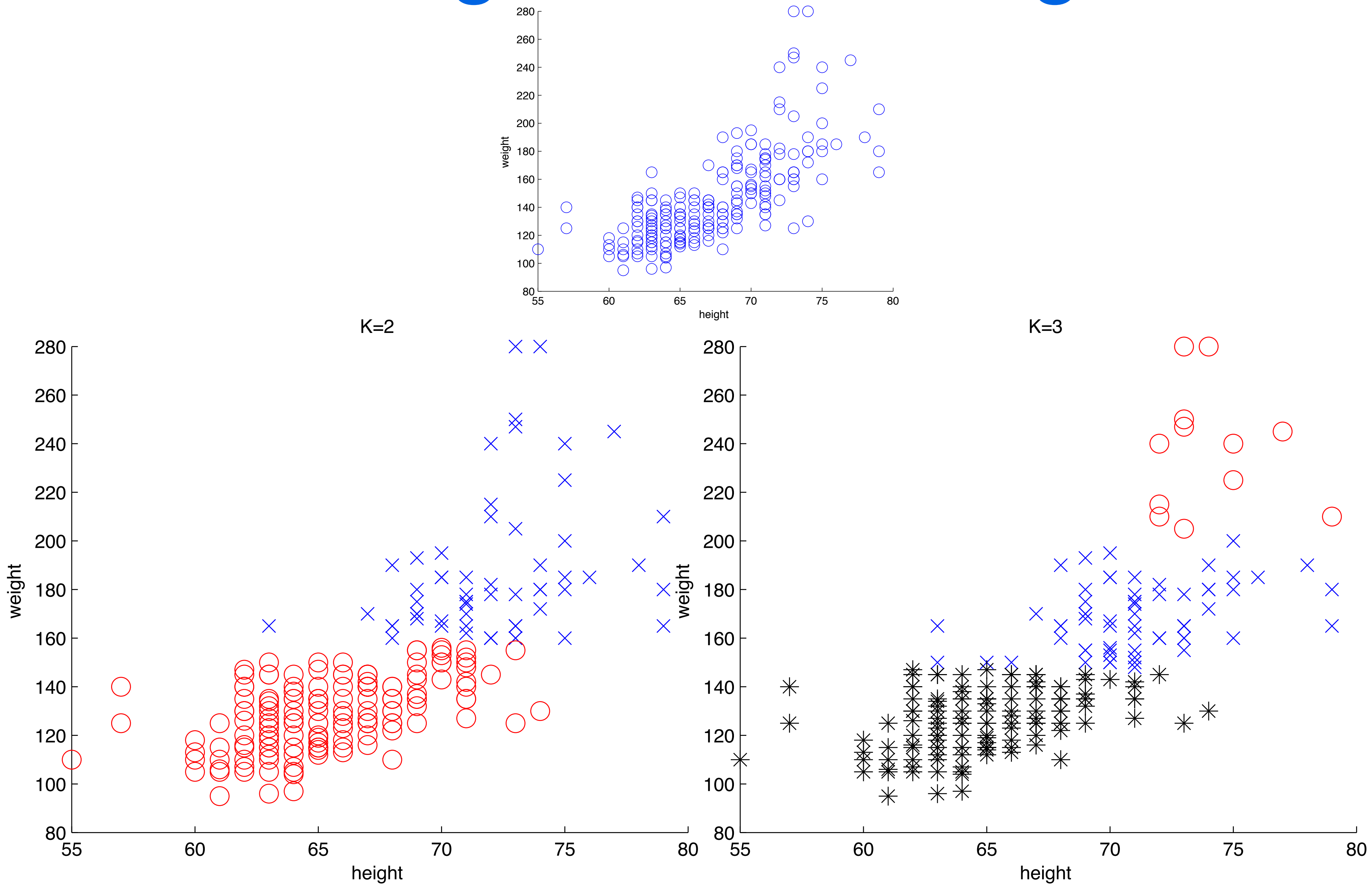
Machine Learning: A Probabilistic Perspective

On to Unsupervised Learning

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

- **Goal:** Infer label/response y given only features x
- **Classical:** Find latent variables y good for *compression* of x
- **Probabilistic learning:** Estimate parameters of joint distribution p(x,y) which *maximize marginal probability* p(x)

Clustering can be Ambiguous



K-Means Objective: Compression

- Observed feature vectors: $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels: $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden cluster centers: $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$

$$J(z, \mu \mid x, K) = \sum_{k=1}^{K} \sum_{i \mid z_i = k} ||x_i - \mu_k||^2 = \sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2$$

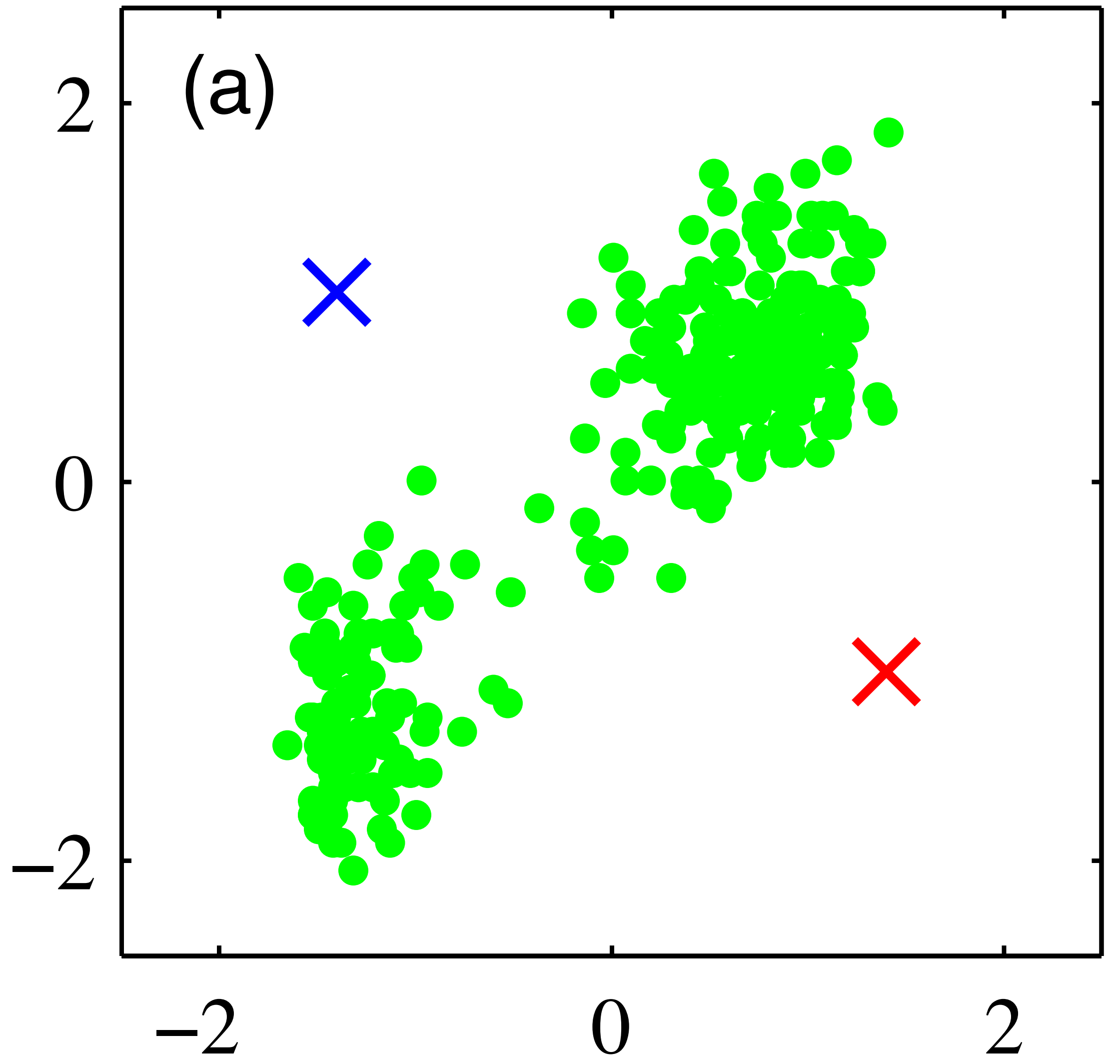
$$J(z, \mu \mid x, K) = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} ||x_i - \mu_k||^2 \qquad r_{ik} = \mathbb{I}(z_i = k)$$

K-Means alternates between

$$z^{(t)} = \arg\min_z J(z, \mu^{(t-1)} \mid x, K)$$

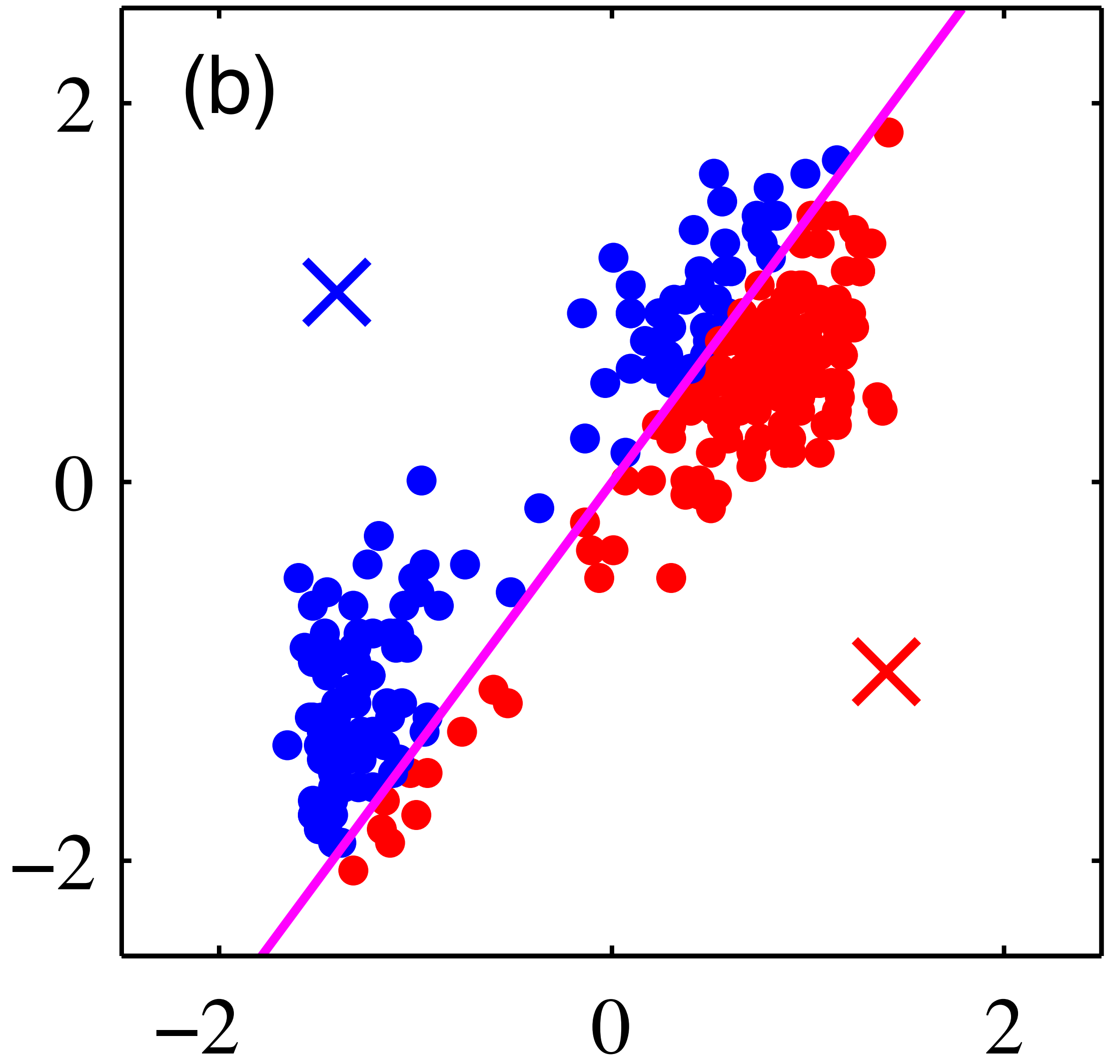
$$\mu^{(t)} = \arg\min_\mu J(z^{(t)}, \mu \mid x, K)$$

K-Means Algorithm



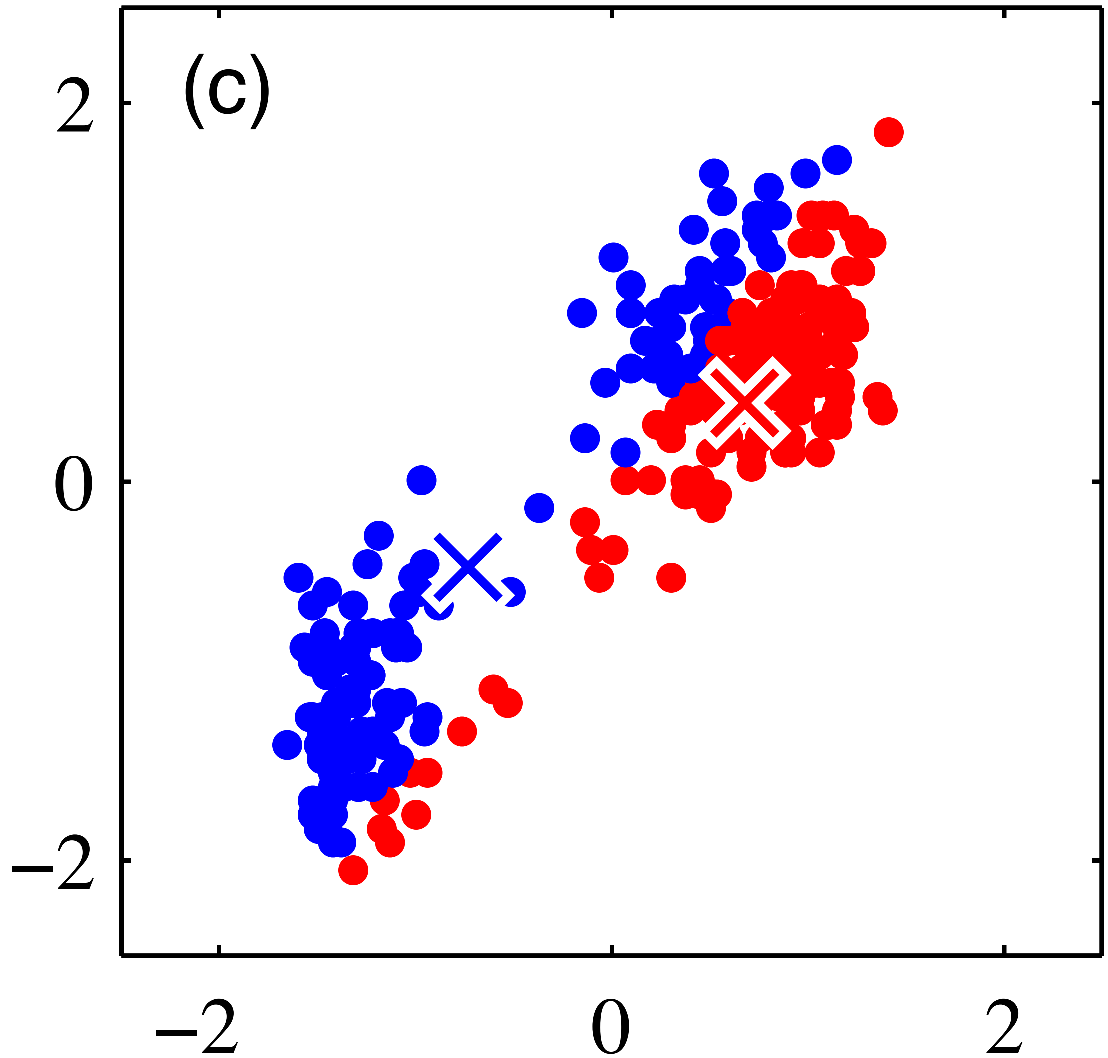
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



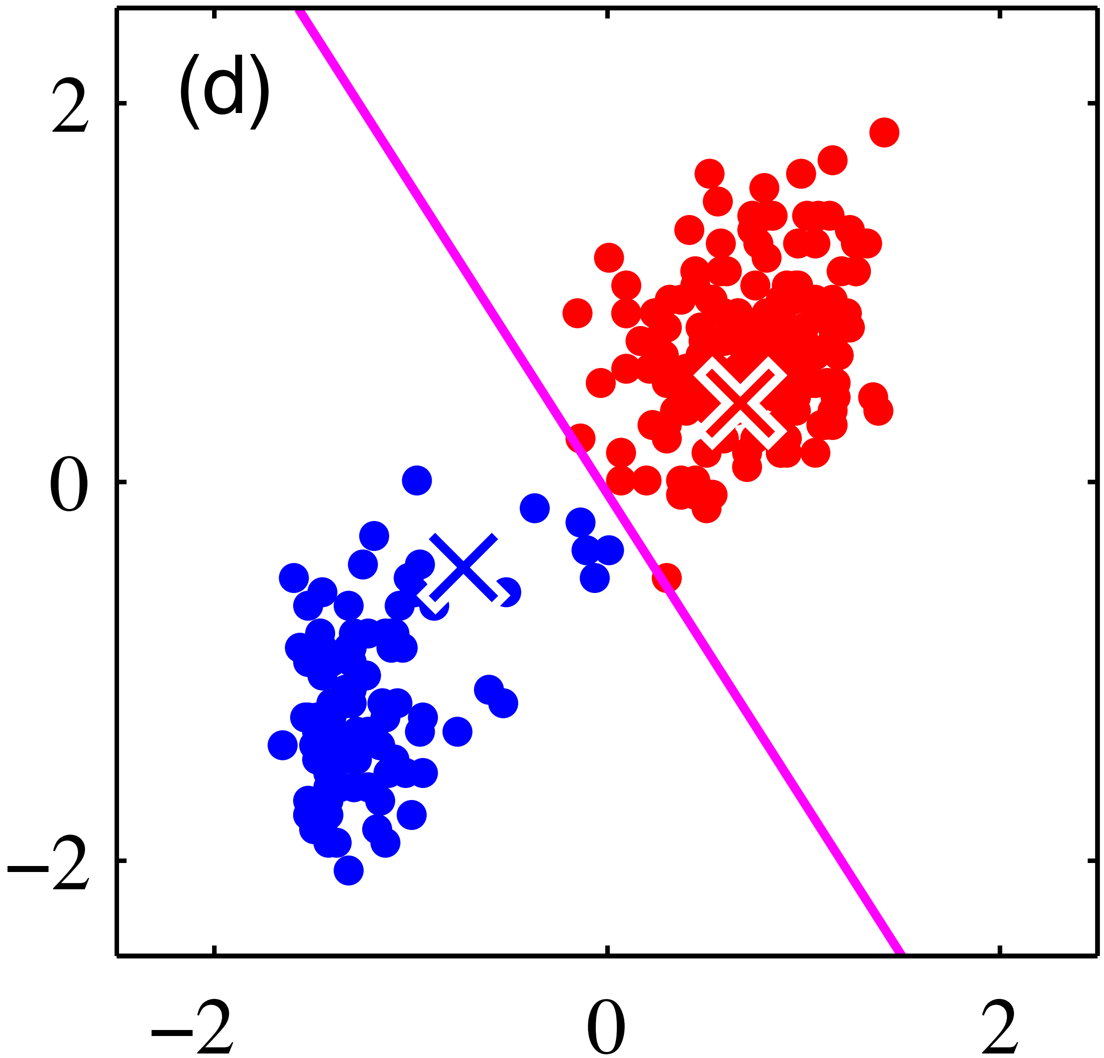
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



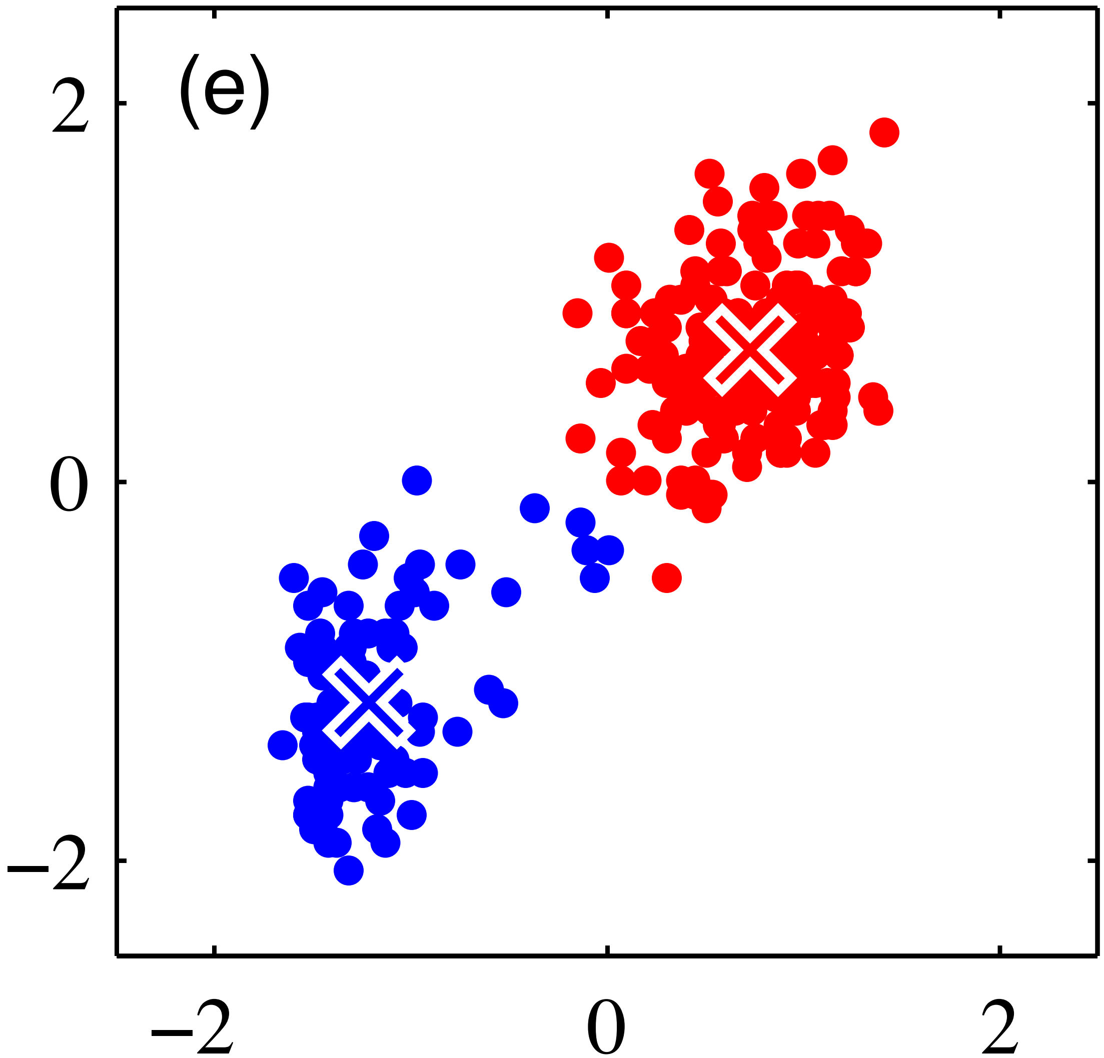
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



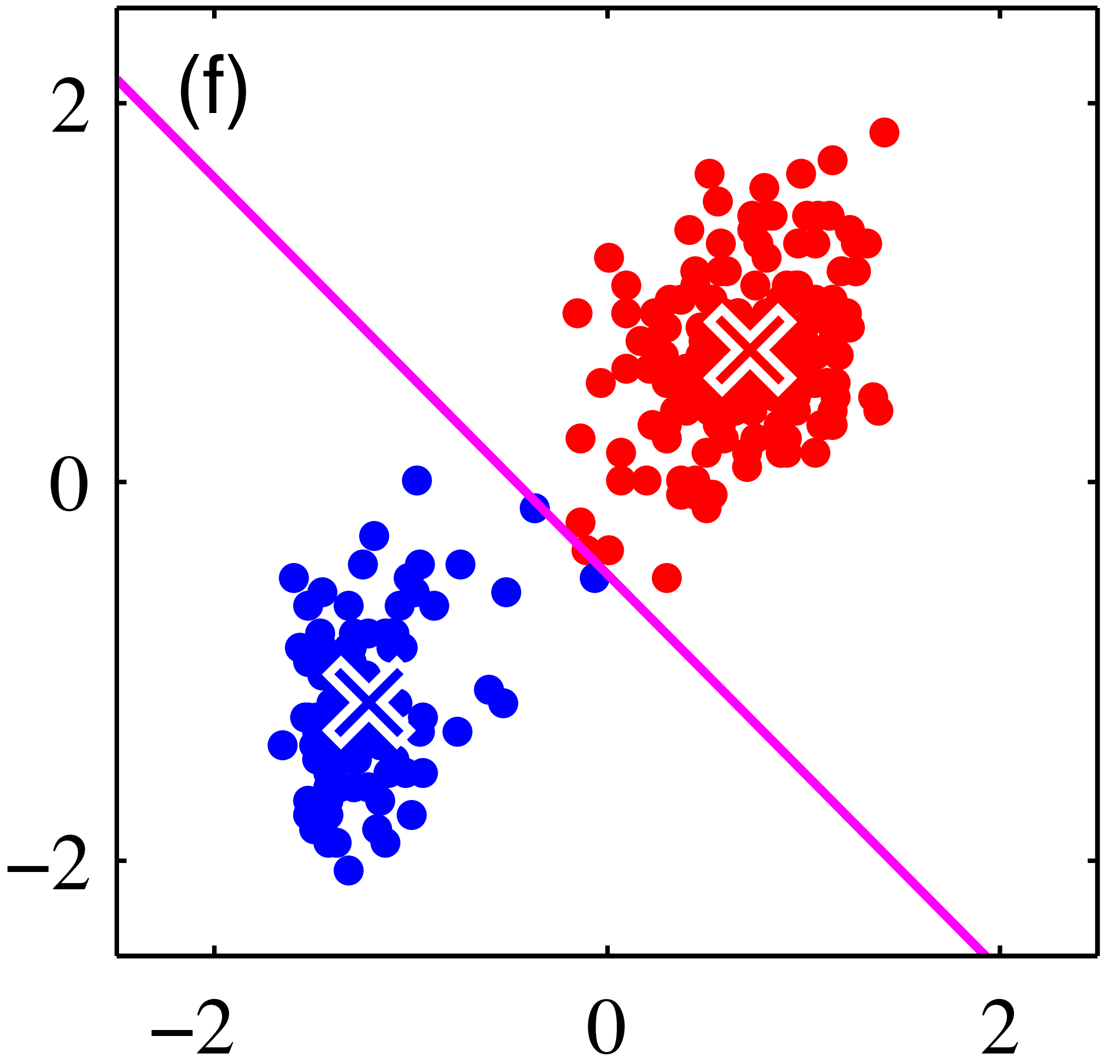
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



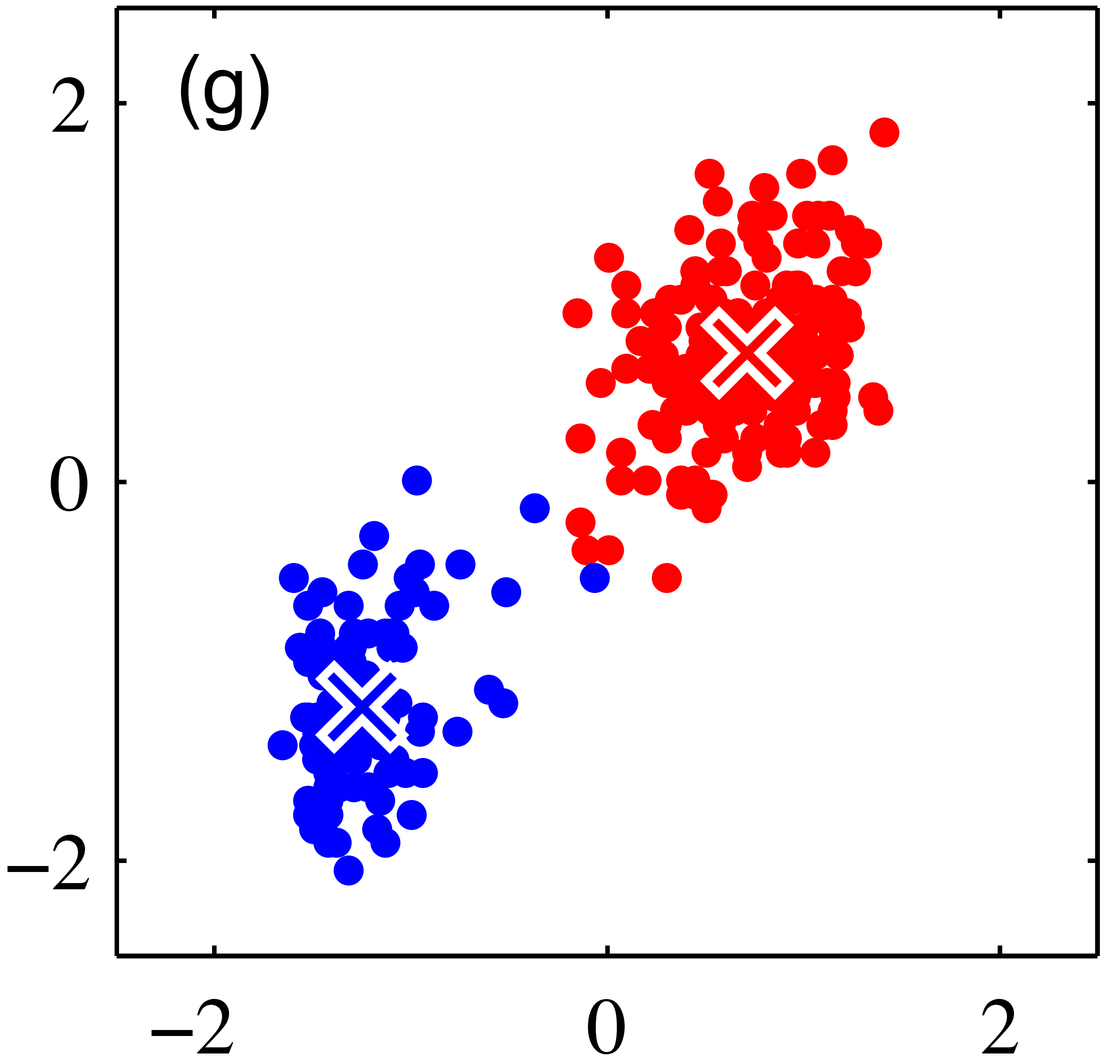
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



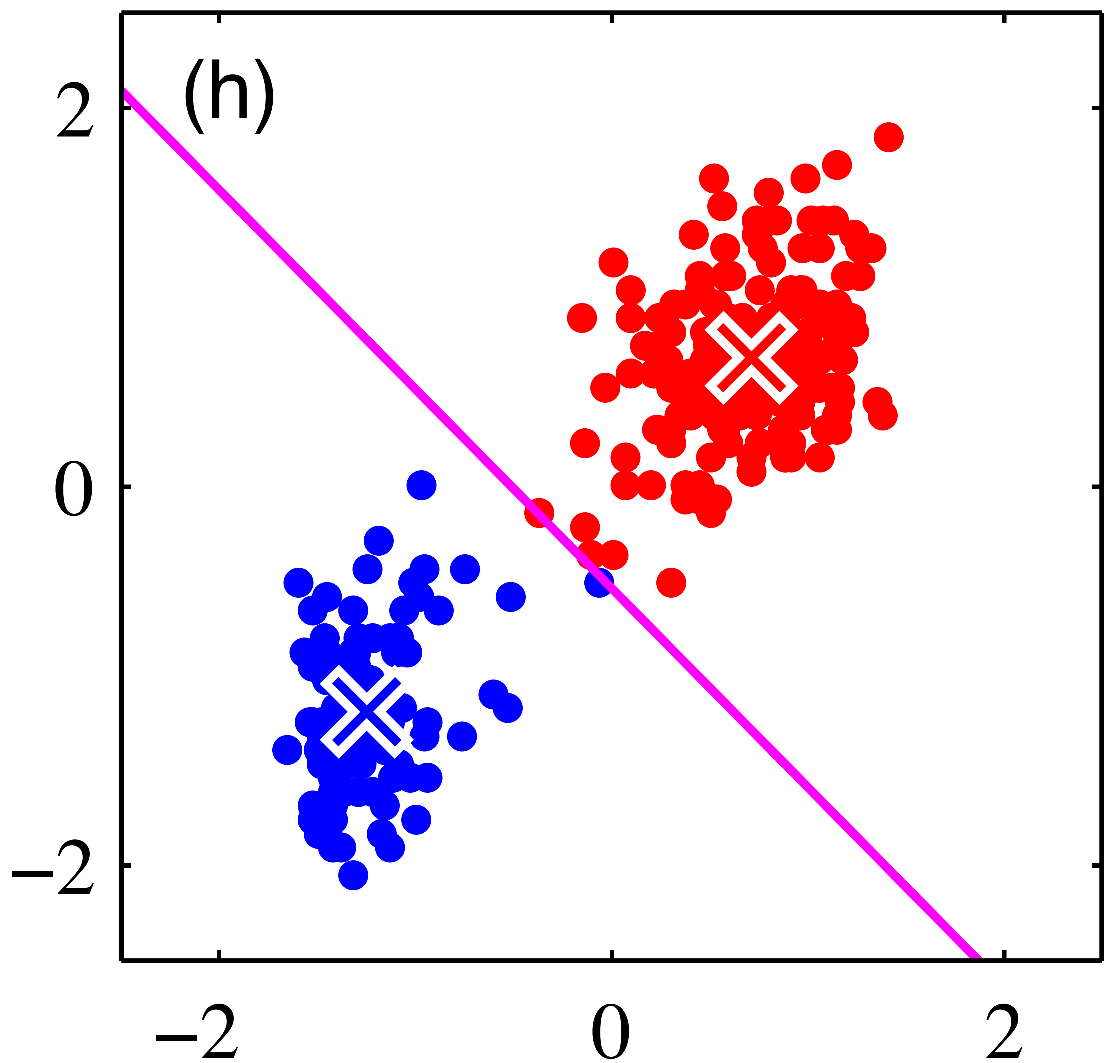
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



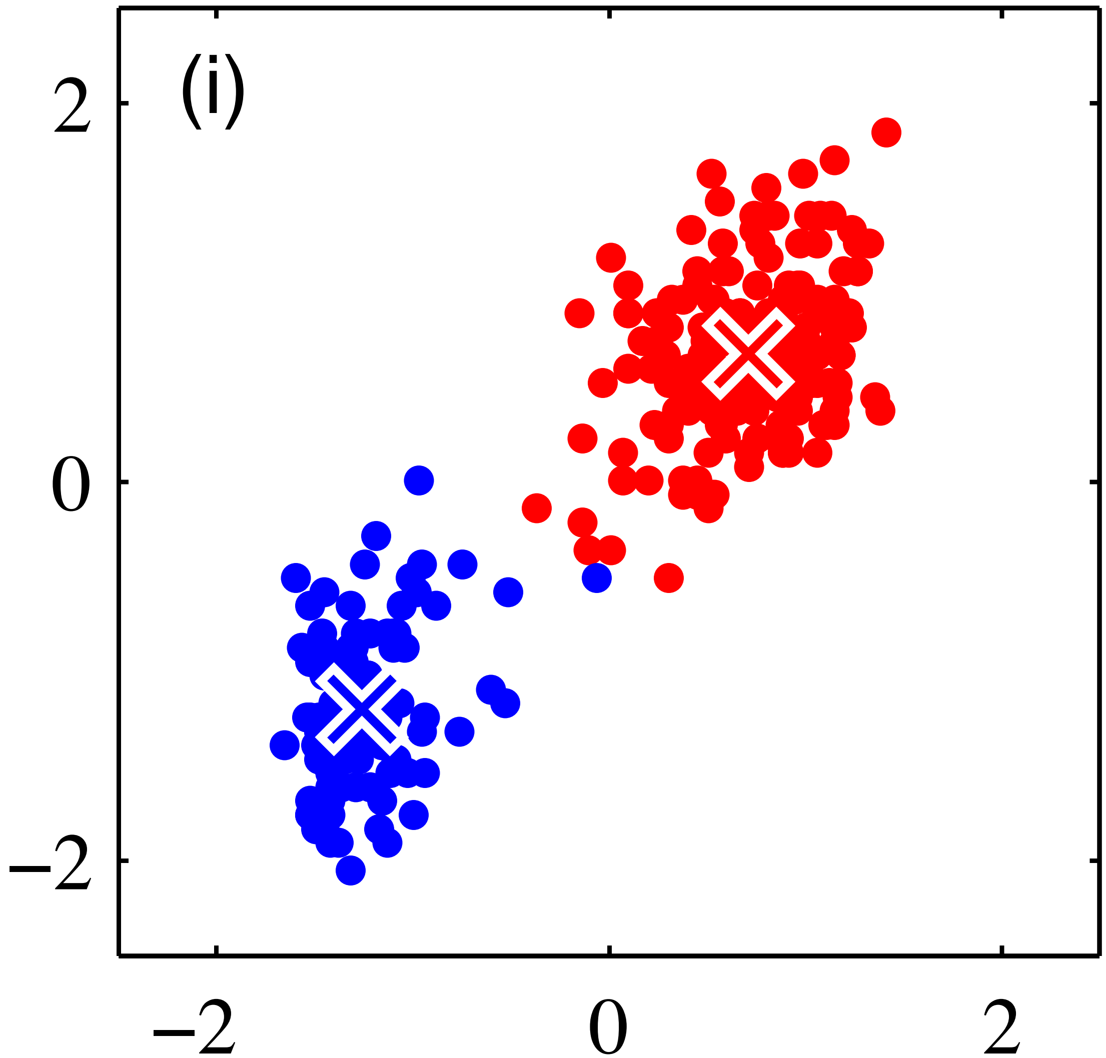
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



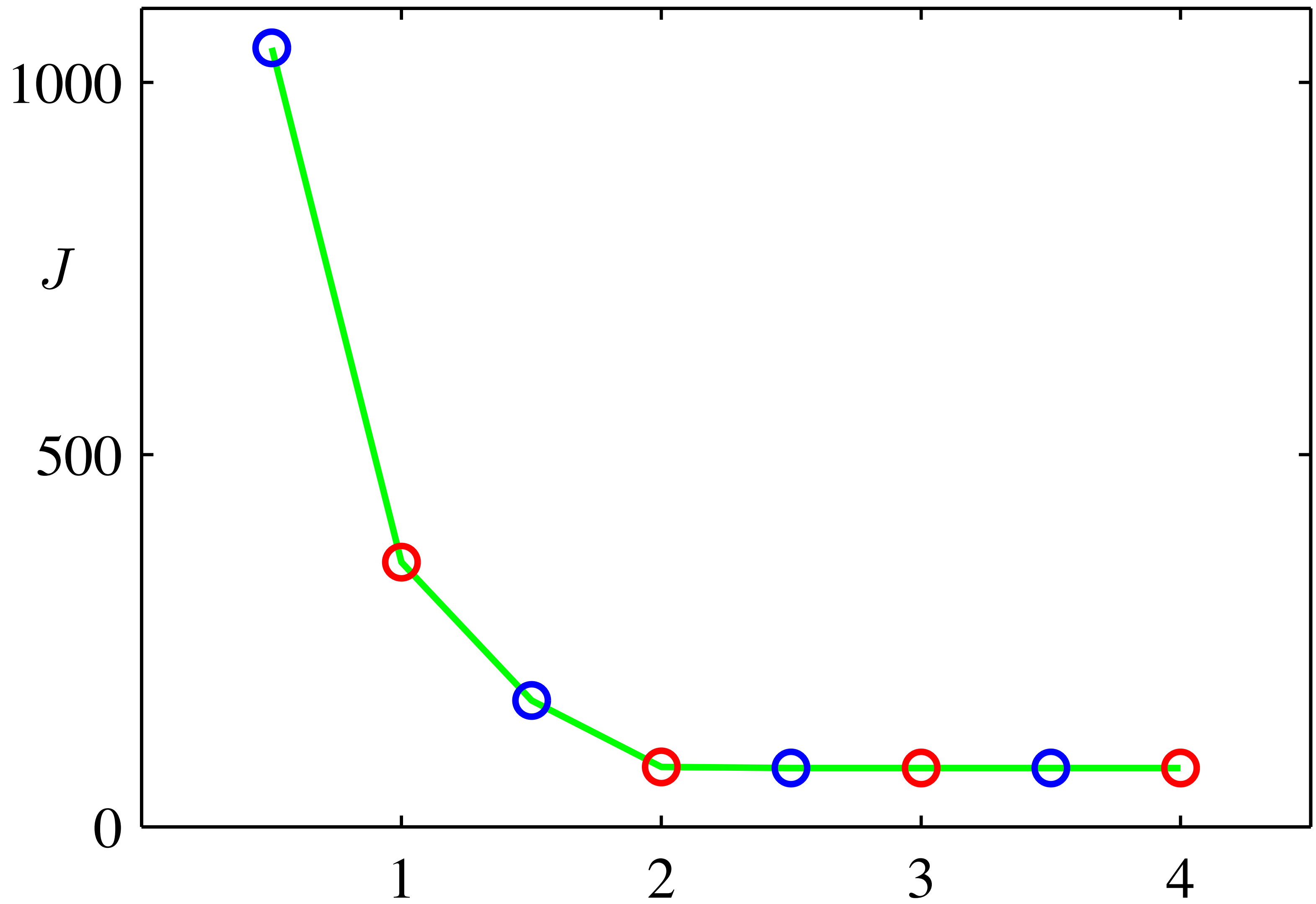
C. Bishop, Pattern Recognition & Machine Learning

K-Means Algorithm



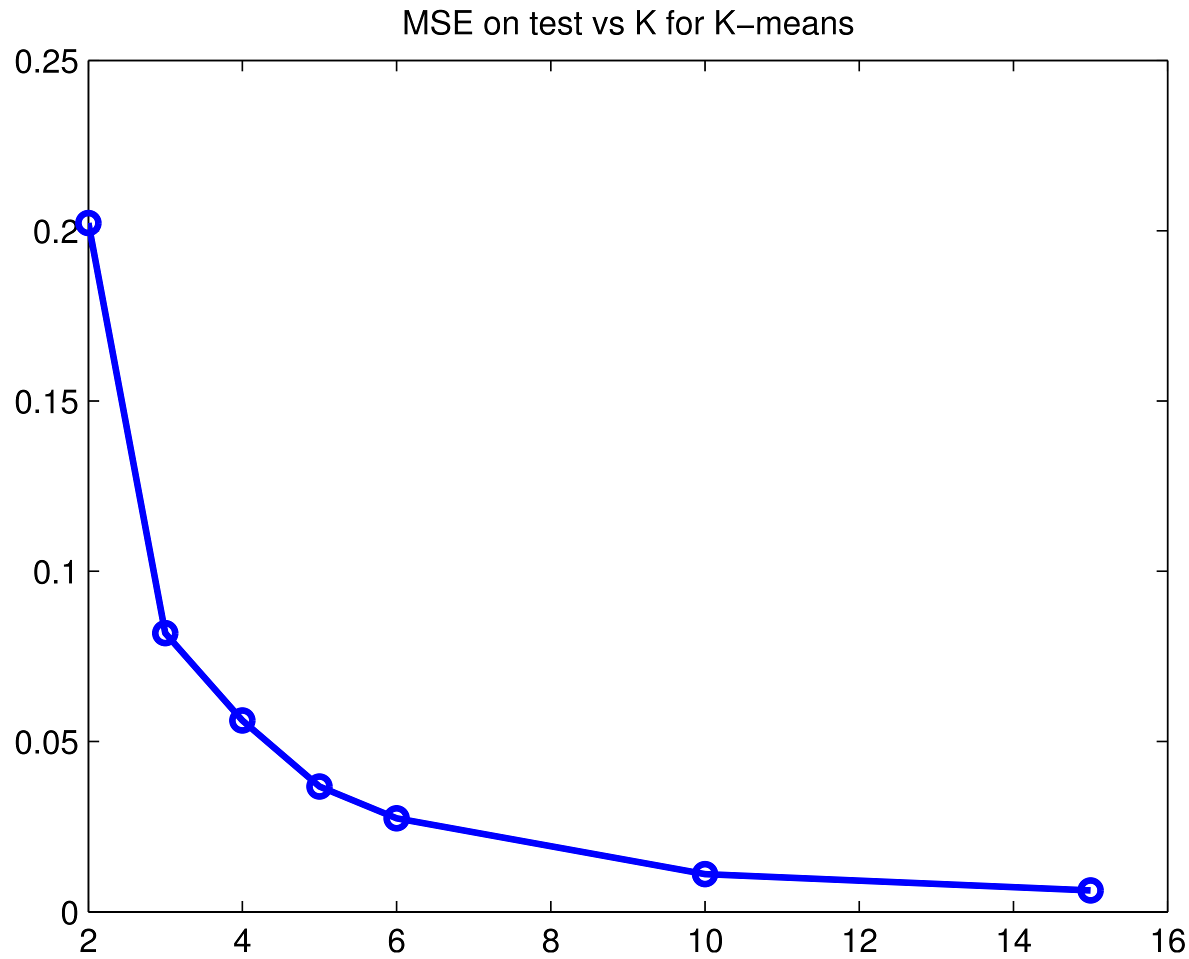
C. Bishop, Pattern Recognition & Machine Learning

Reconstruction Error versus Iteration



C. Bishop, Pattern Recognition & Machine Learning

Test Error versus K

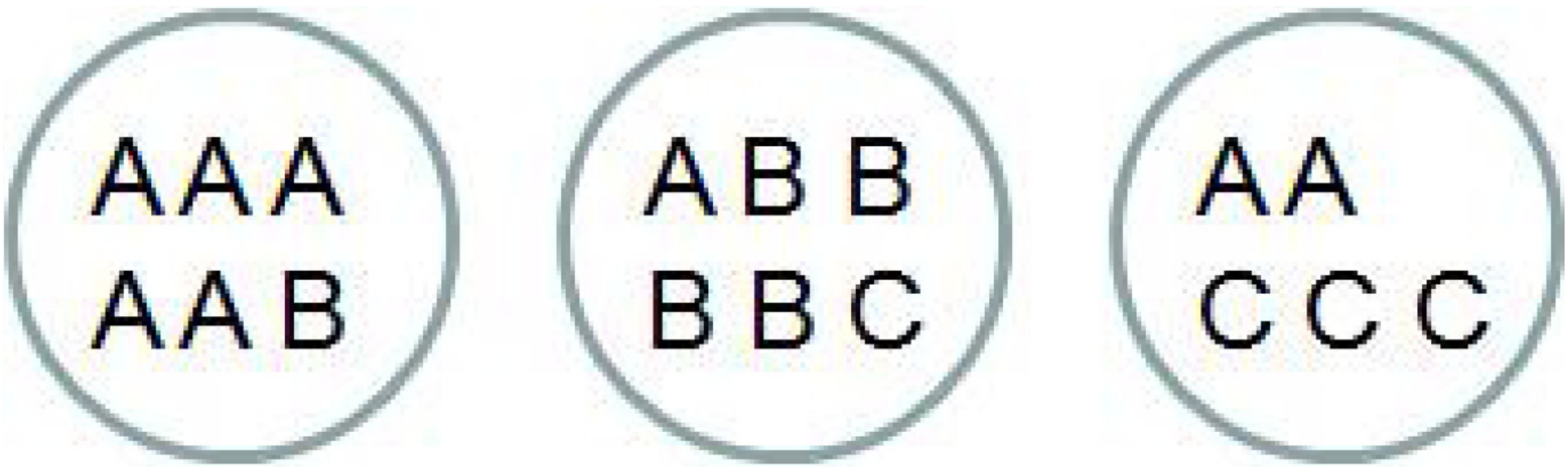


Data from a mixture of 3 1D Gaussian distributions

For compressing new data, more codewords is always better.

Cross-validation fails for unsupervised learning!

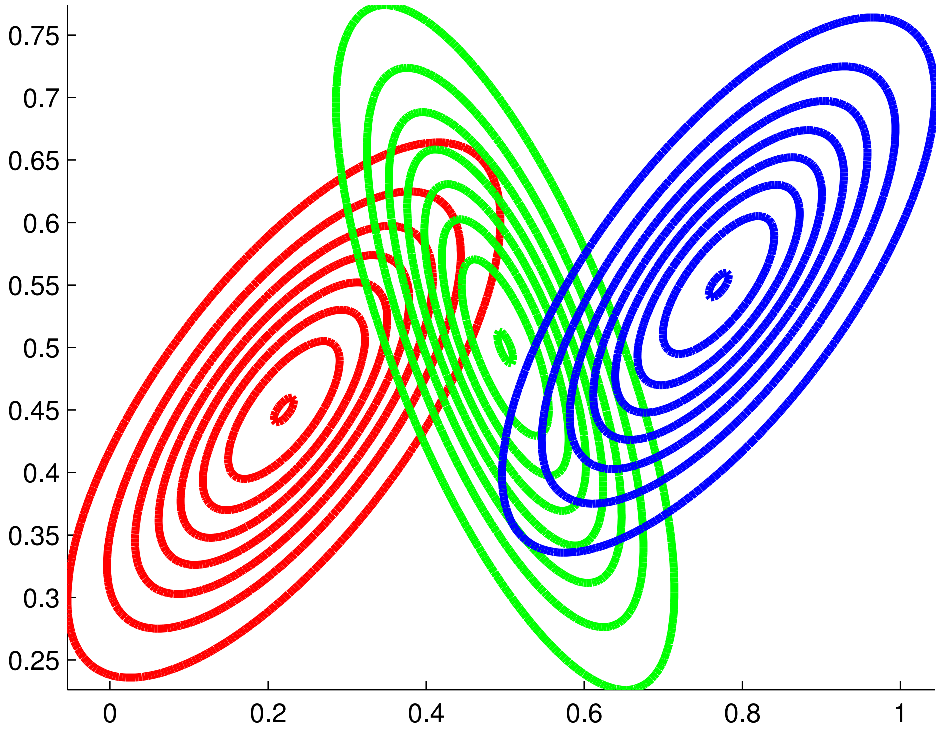
Clustering Evaluation: Rand Index



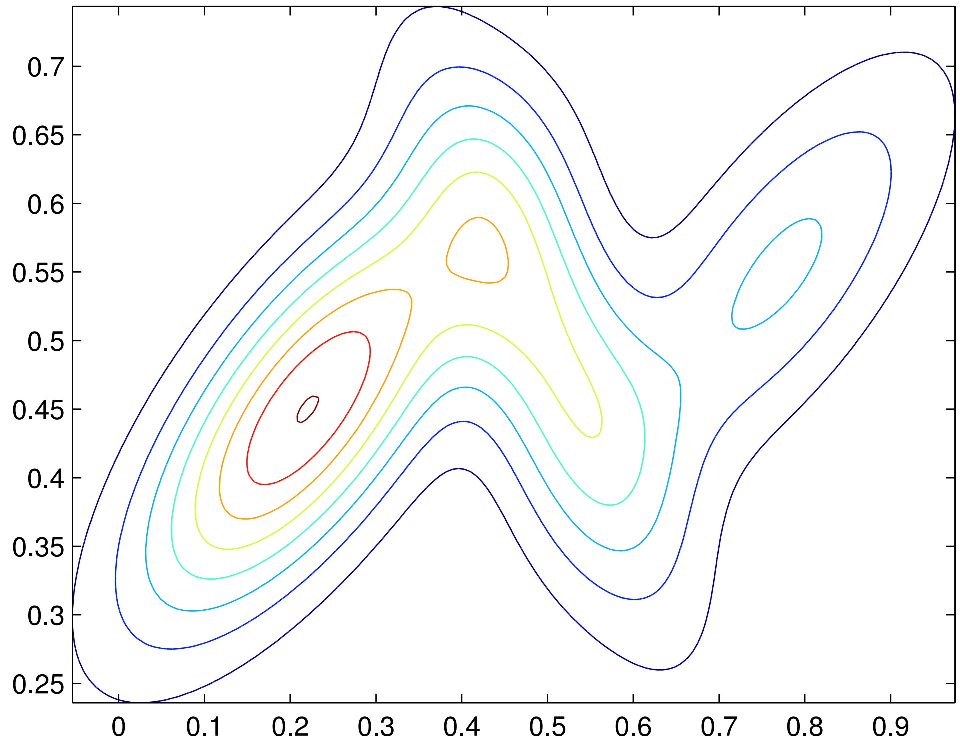
$$R := \frac{TP + TN}{TP + FP + FN + TN}$$

Consider all pairs of data points, and count fraction where hypothesized and target clusterings agree

Gaussian Mixture Models

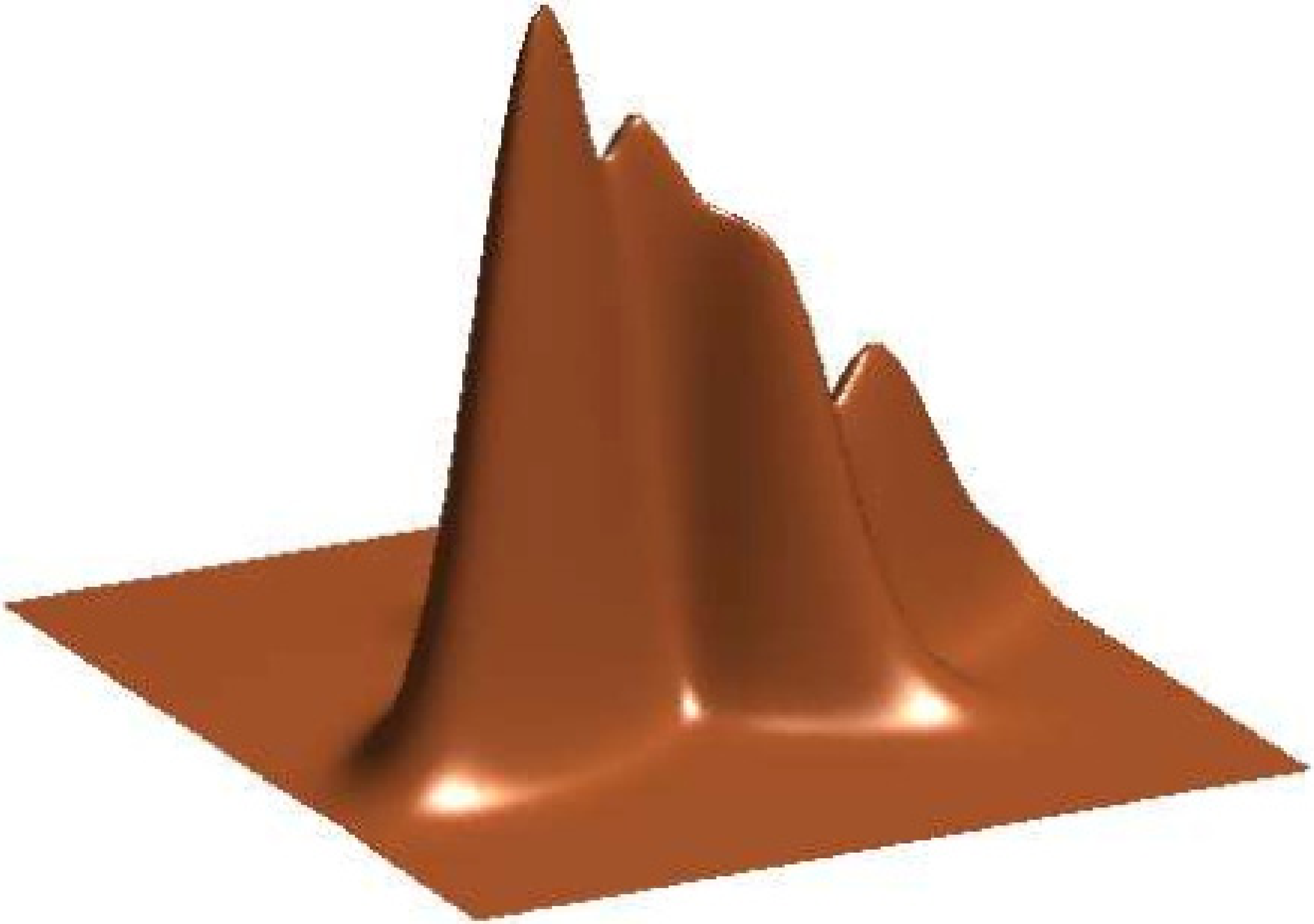


Mixture of 3 Gaussian Distributions in 2D



Contour Plot of Joint Density, Marginalizing Cluster Assignments

Gaussian Mixture Models



Surface Plot of Joint Density,
Marginalizing Cluster Assignments

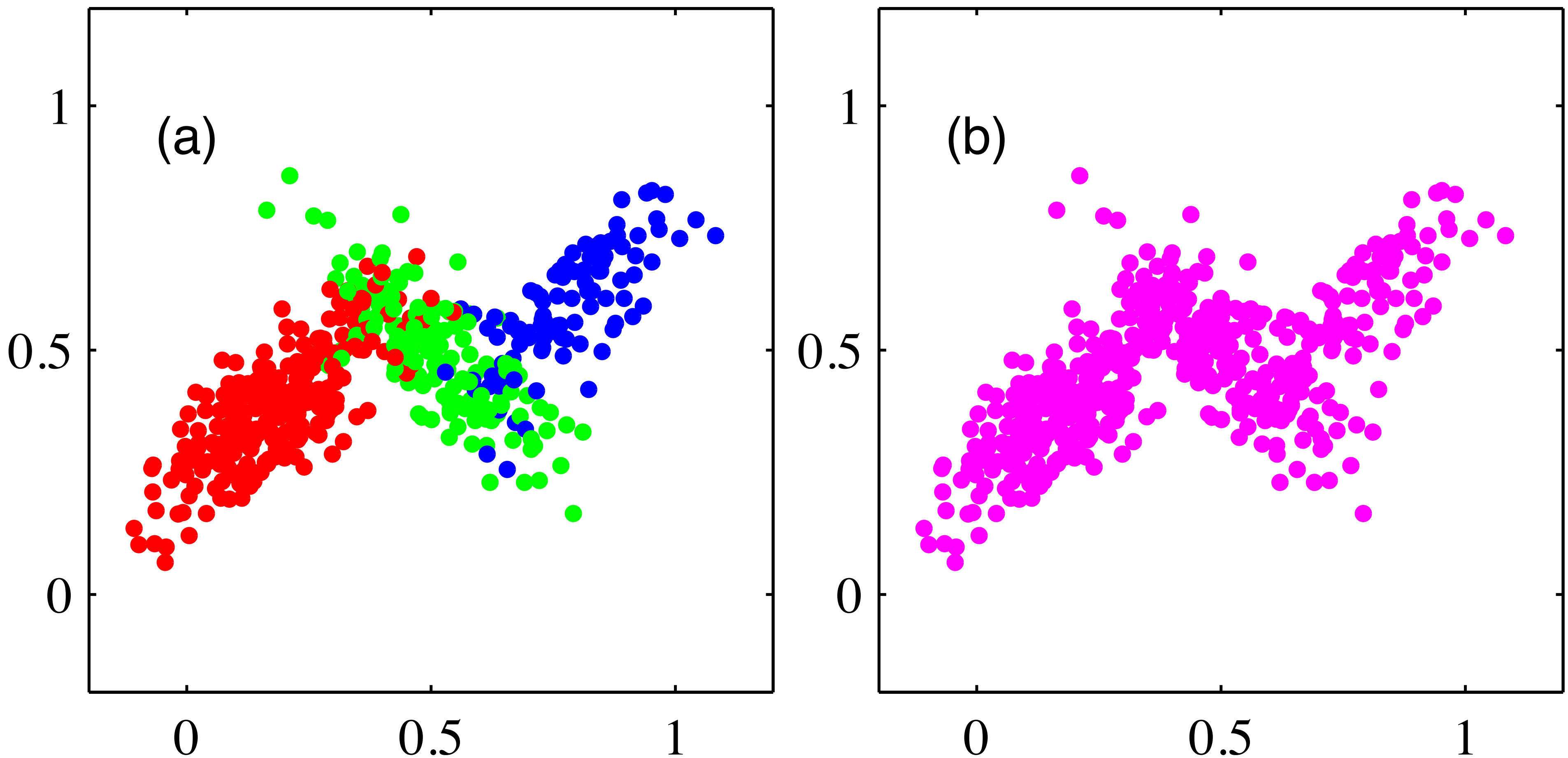
Gaussian Mixture Models

- Observed feature vectors: $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels: $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden mixture means: $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$
- Hidden mixture covariances: $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K$
- Hidden mixture probabilities: $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$
- Gaussian mixture marginal likelihood:

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(0 \mid \mu_{z_i}, \Sigma_{z_i})$$

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(0 \mid \mu_{z_i}, \Sigma_{z_i})$$

Fitting Gaussian Mixtures

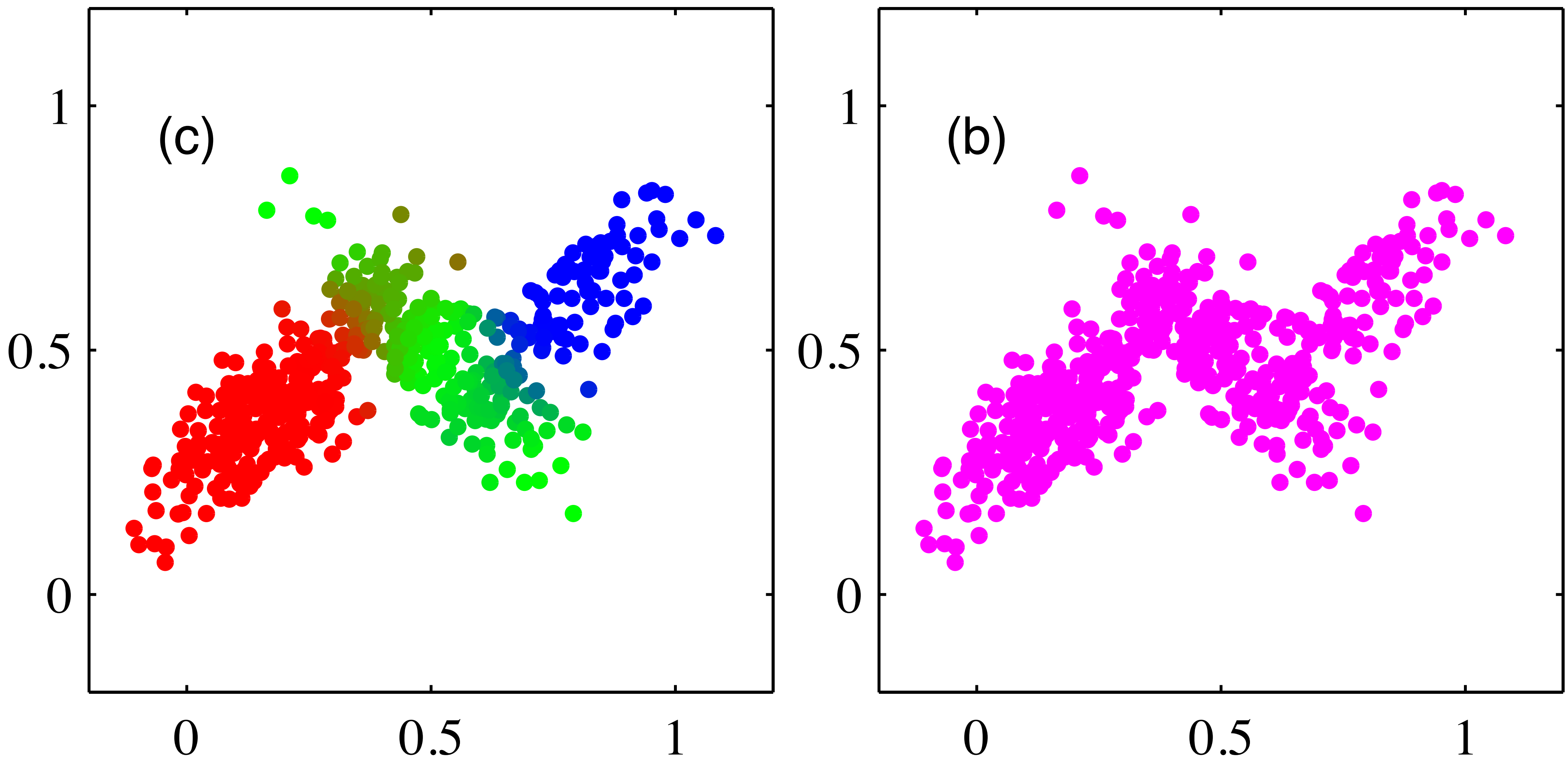


Complete Data Labeled
by True Cluster Assignments

Incomplete Data:
Points to be Clustered

C. Bishop, Pattern Recognition & Machine Learning

Posterior Assignment Probabilities

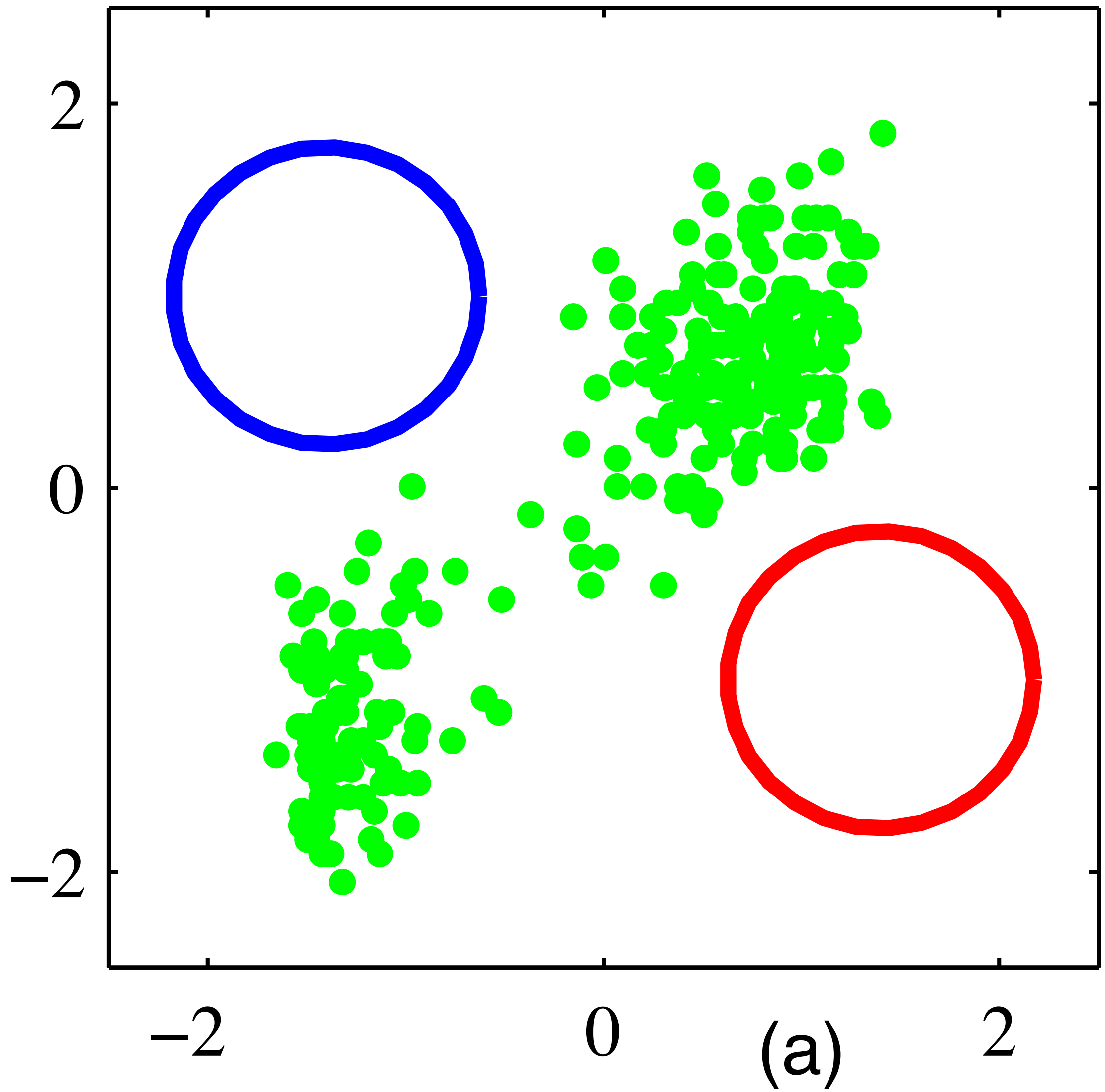


Posterior Probabilities of Assignment to Each Cluster

Incomplete Data: Points to be Clustered

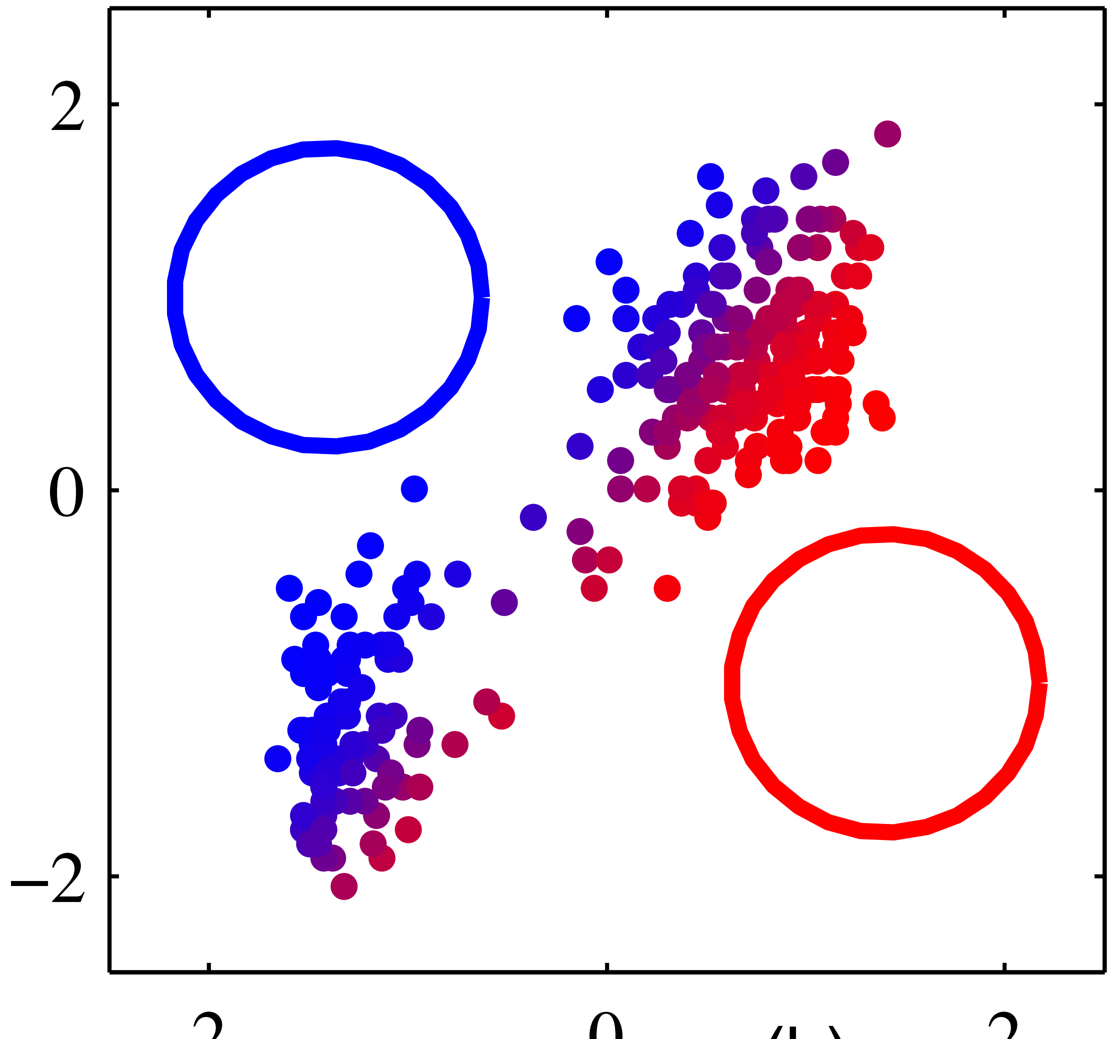
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

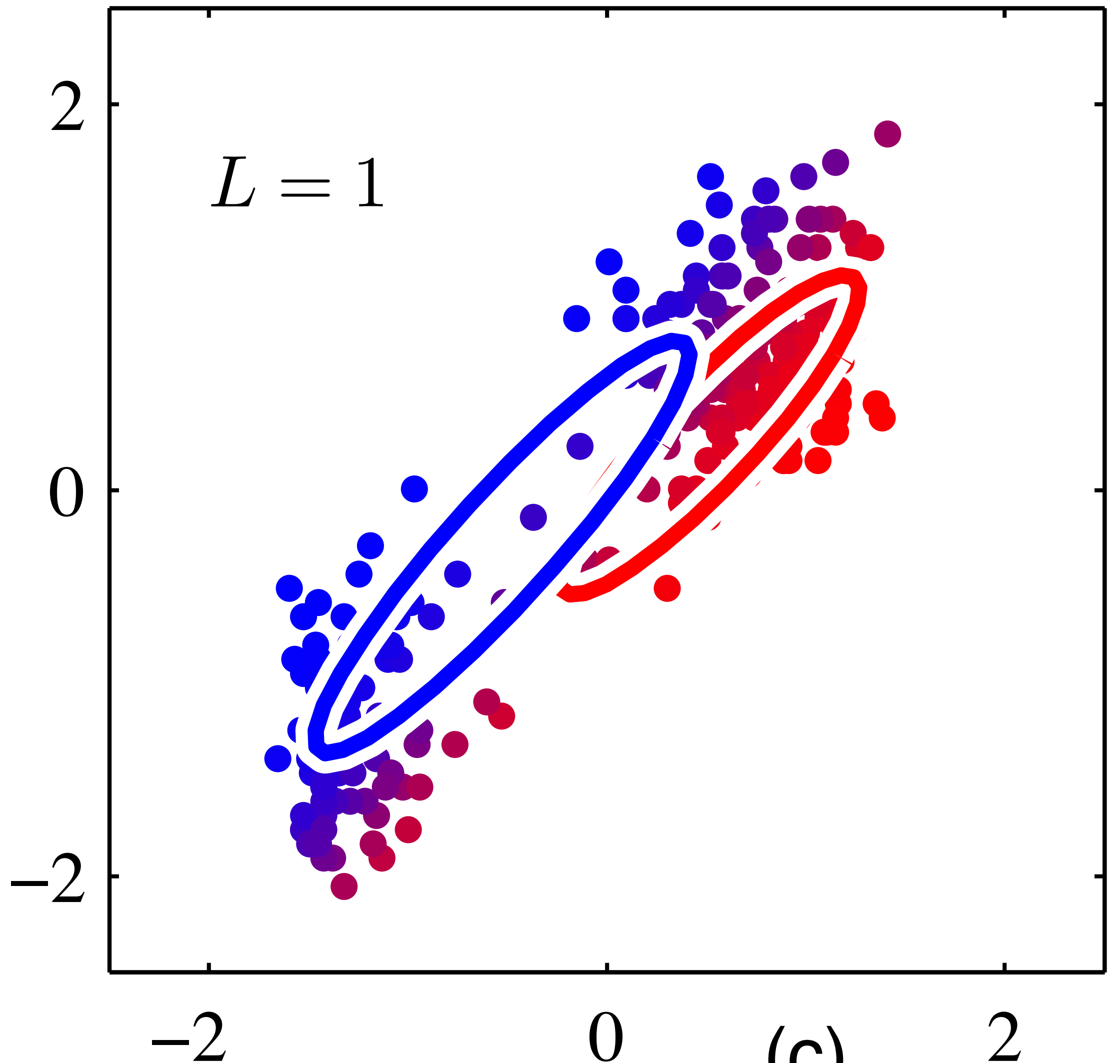
EM Algorithm



(b)

C. Bishop, Pattern Recognition & Machine Learning

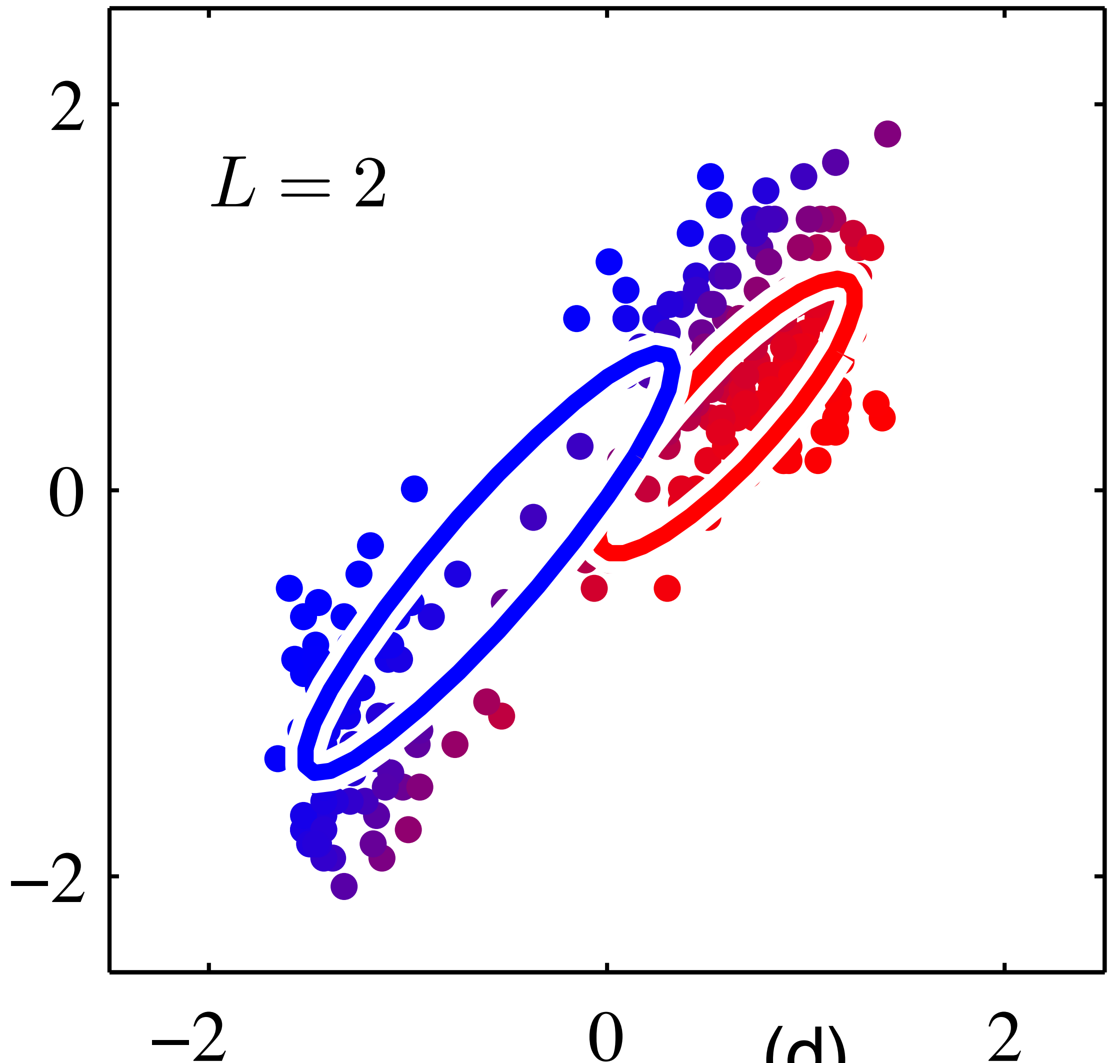
EM Algorithm



(c)

C. Bishop, Pattern Recognition & Machine Learning

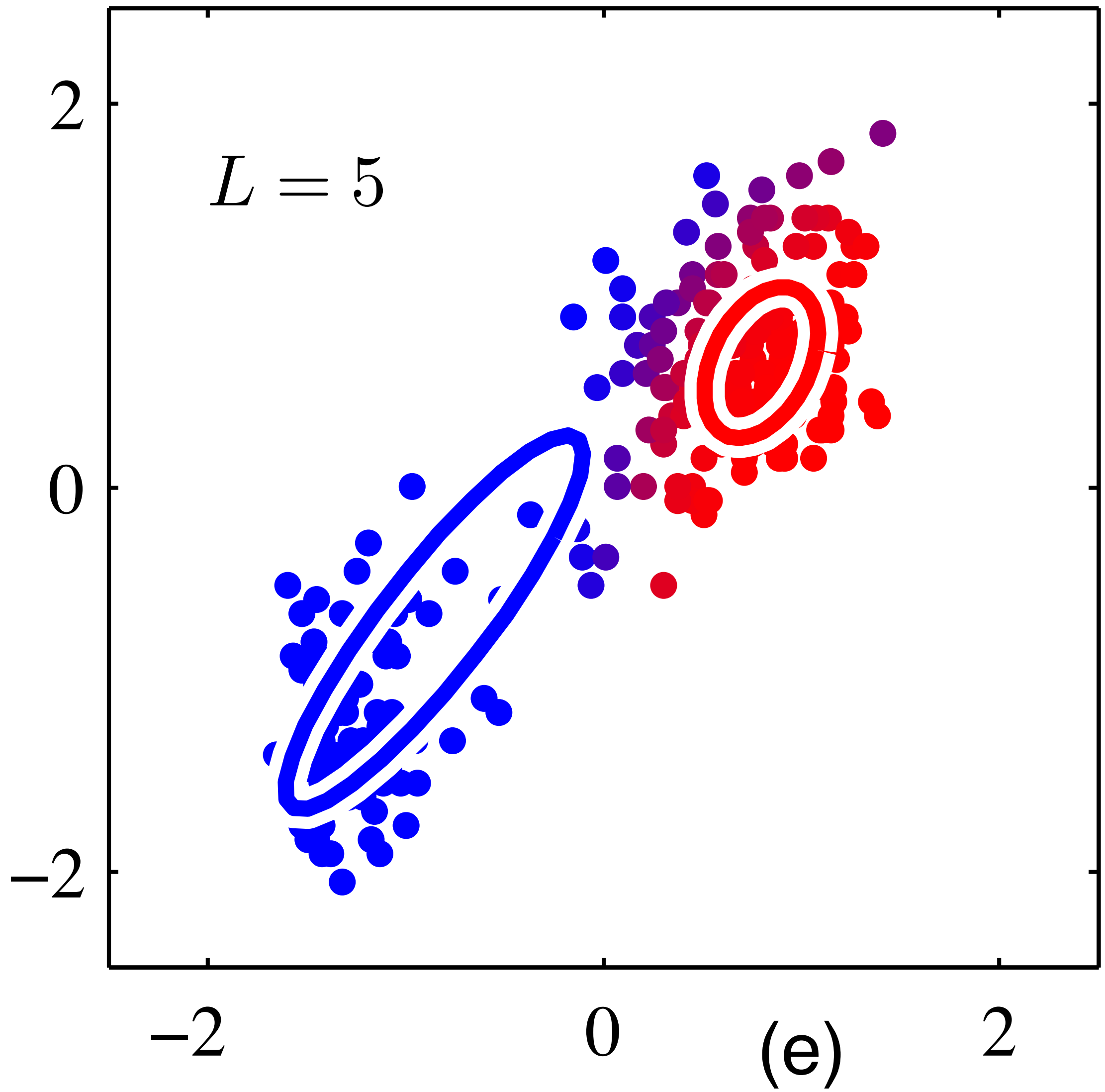
EM Algorithm



(d)

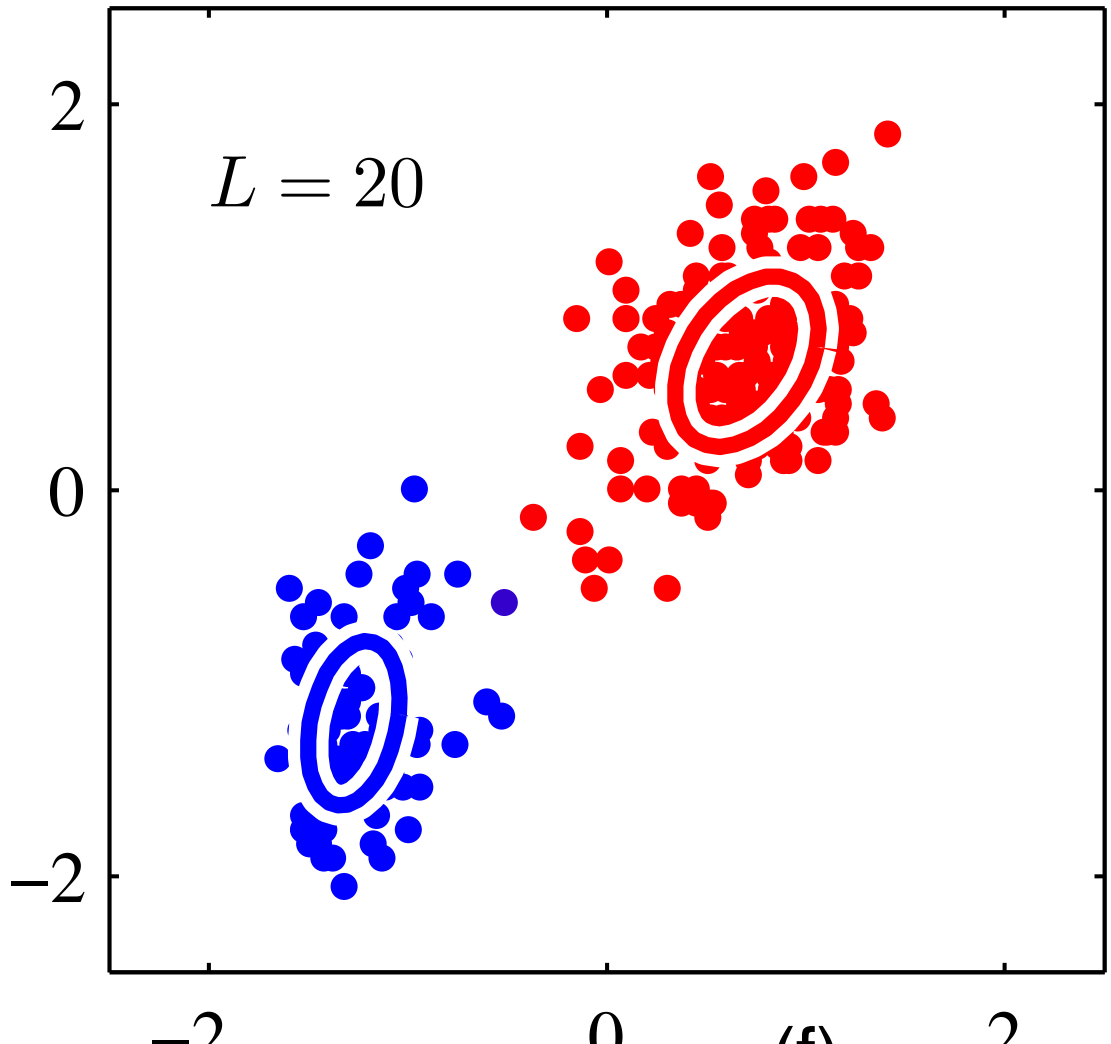
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



(f)

C. Bishop, Pattern Recognition & Machine Learning

EM for Gaussian Mixtures

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(0 \mid \mu_{z_i}, \Sigma_{z_i})$$

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(0 \mid \mu_{z_i}, \Sigma_{z_i})$$

- Can easily estimate parameters if know cluster labels. How?
- What if we don't know cluster labels?