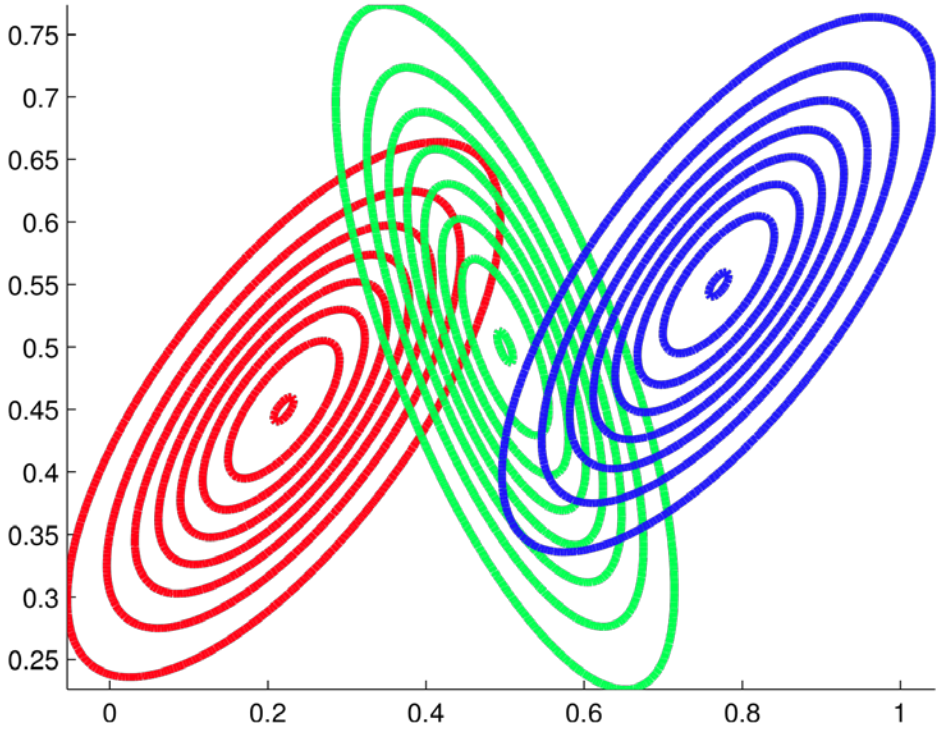
Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011
Prof. Erik Sudderth

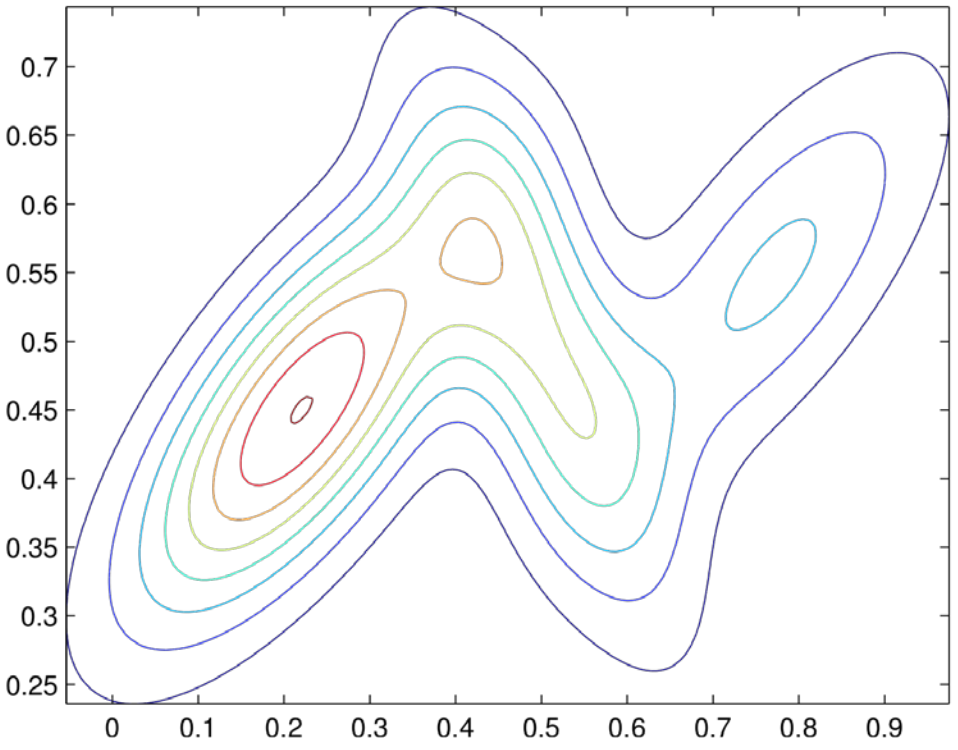
Lecture 19: EM Algorithm

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Gaussian Mixture Models

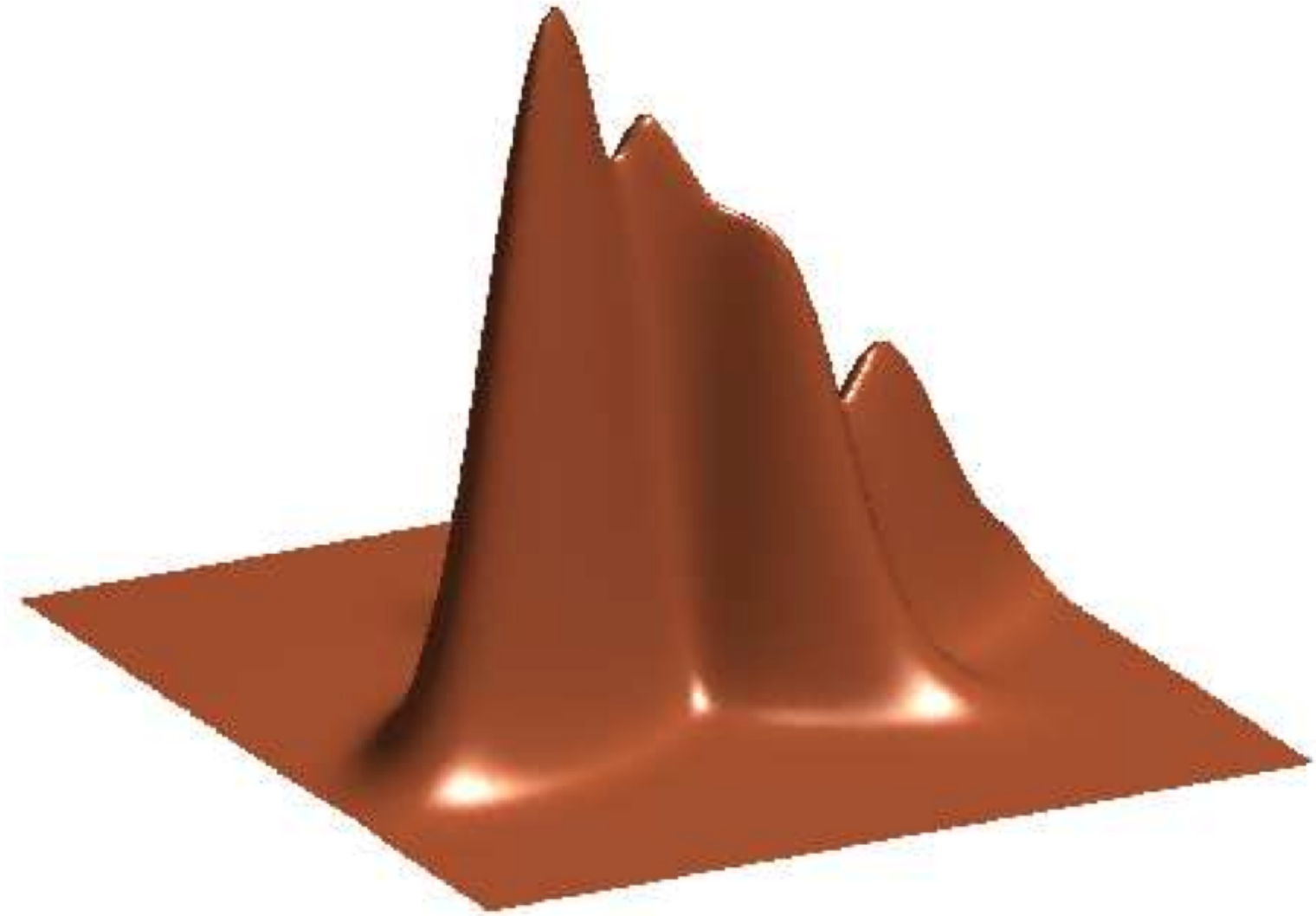


Mixture of 3 Gaussian Distributions in 2D



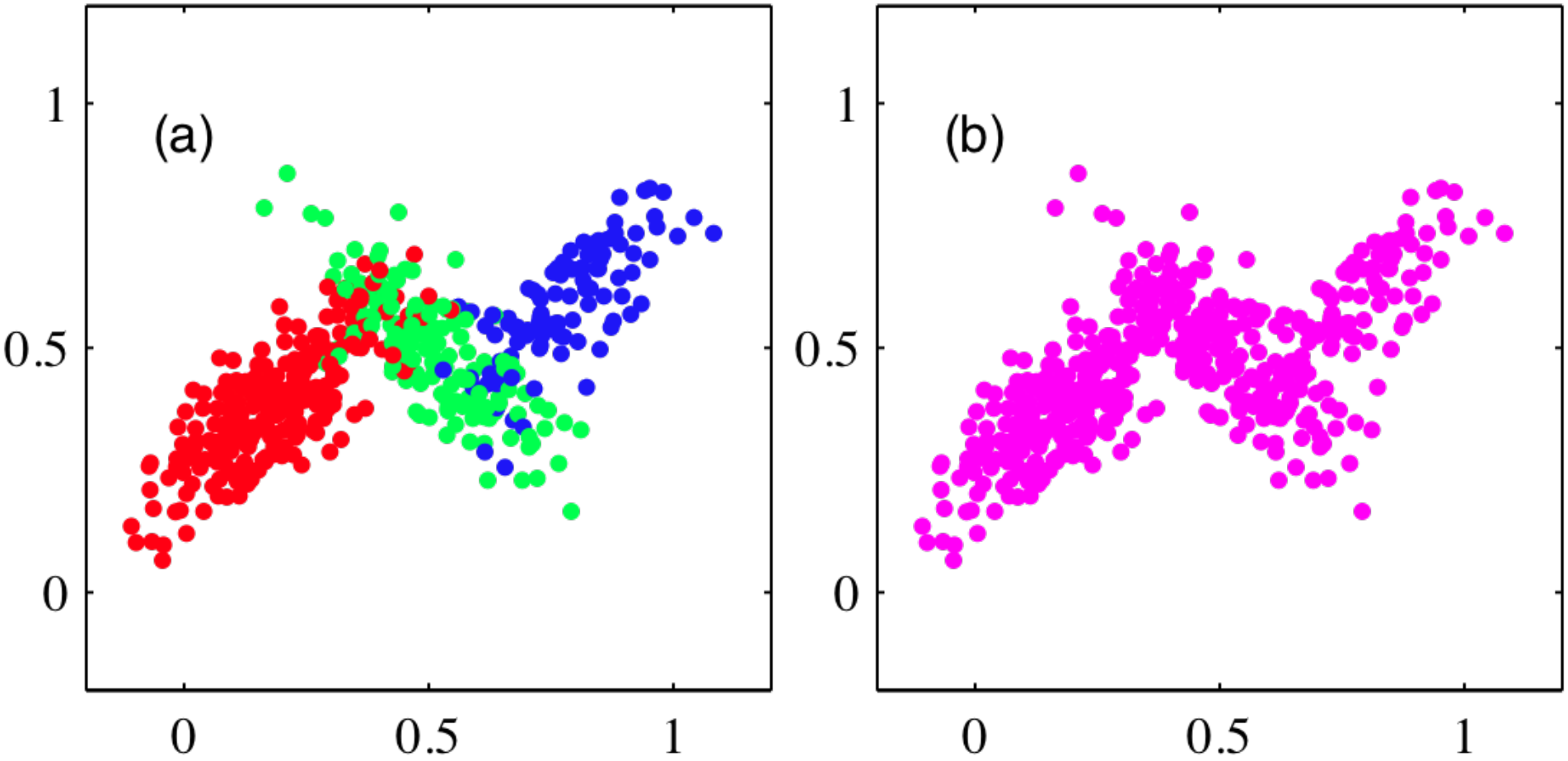
Contour Plot of Joint Density, Marginalizing Cluster Assignments

Gaussian Mixture Models



Surface Plot of Joint Density,
Marginalizing Cluster Assignments

Fitting Gaussian Mixtures

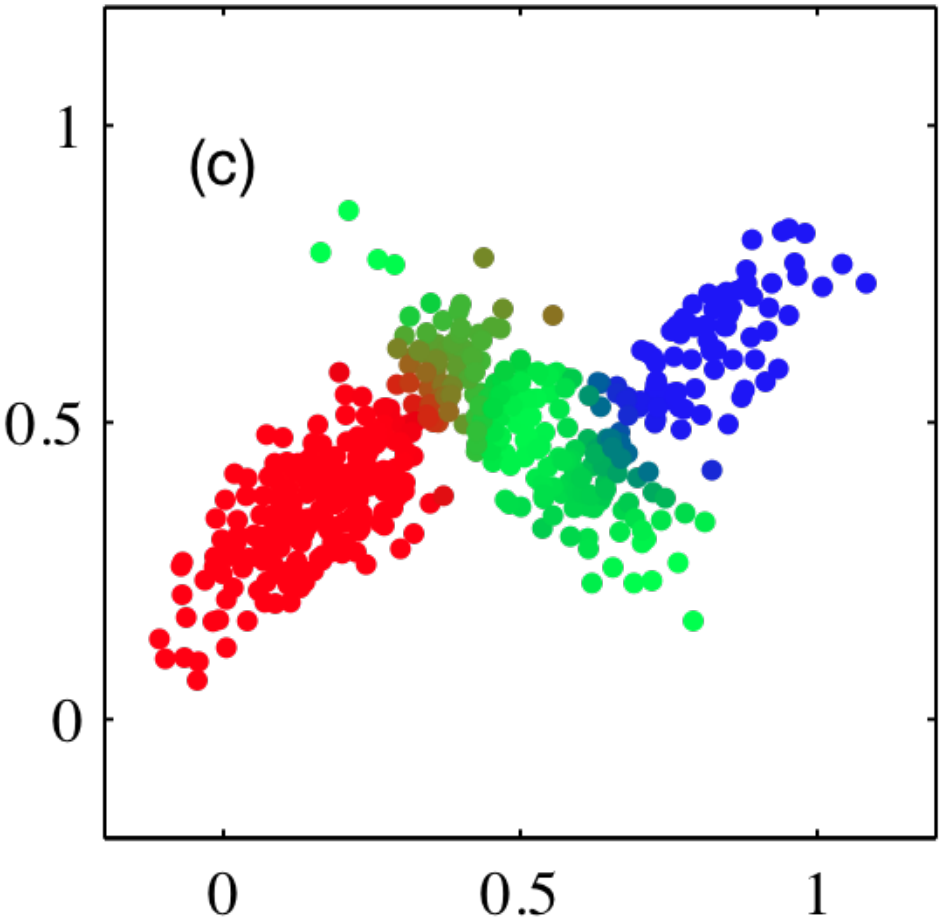


Complete Data Labeled by True Cluster Assignments

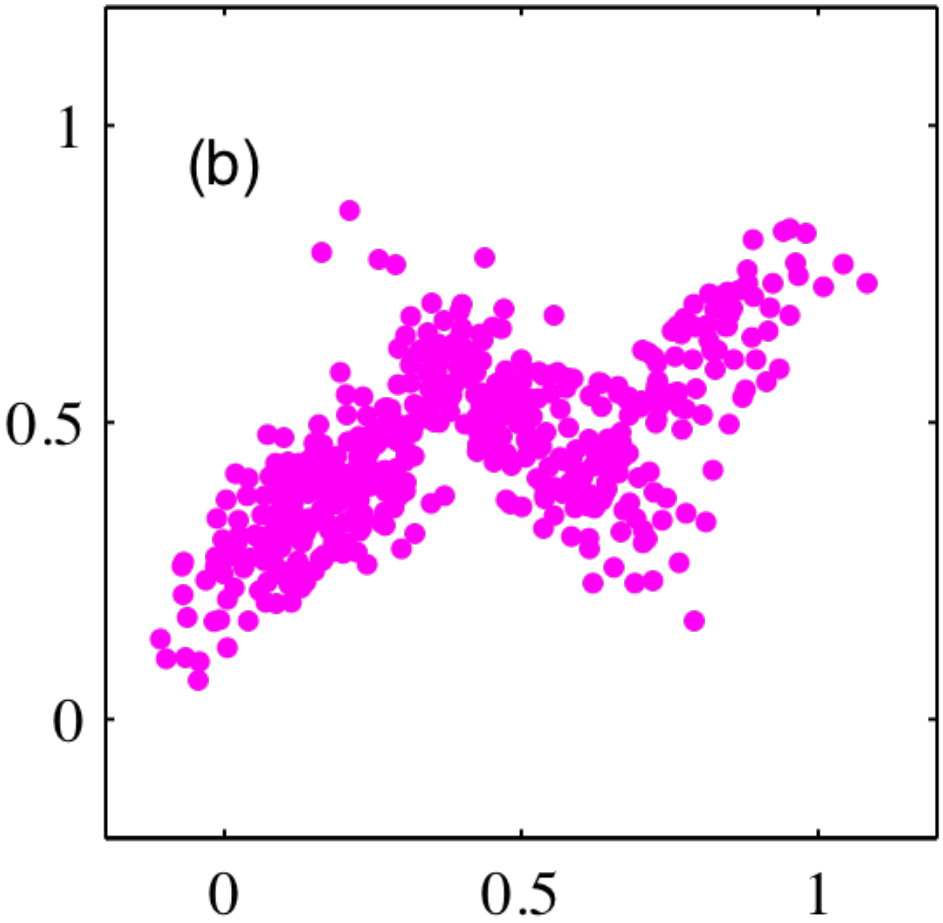
Incomplete Data: Points to be Clustered

C. Bishop, Pattern Recognition & Machine Learning

Posterior Assignment Probabilities



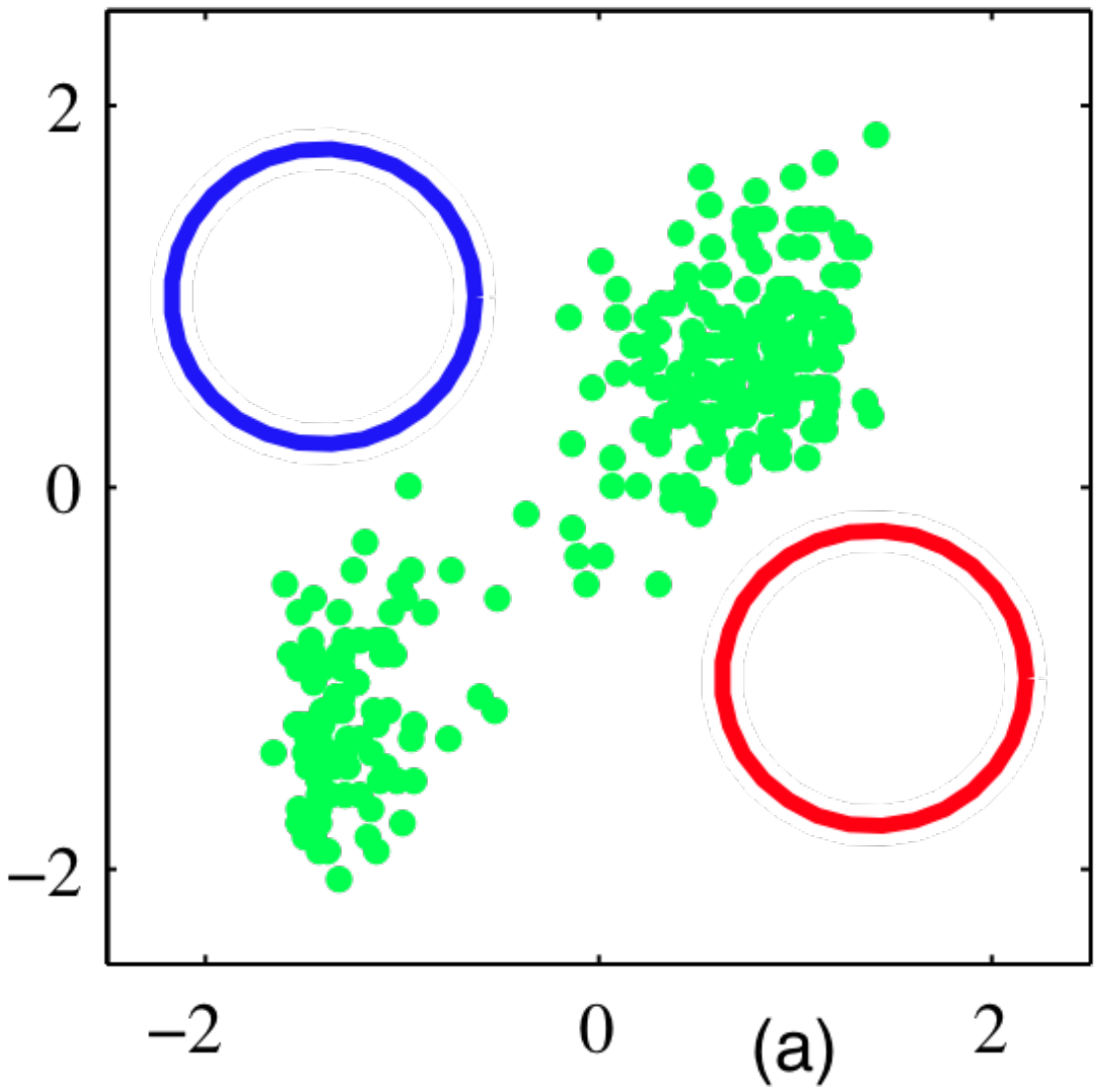
Posterior Probabilities of Assignment to Each Cluster



Incomplete Data: Points to be Clustered

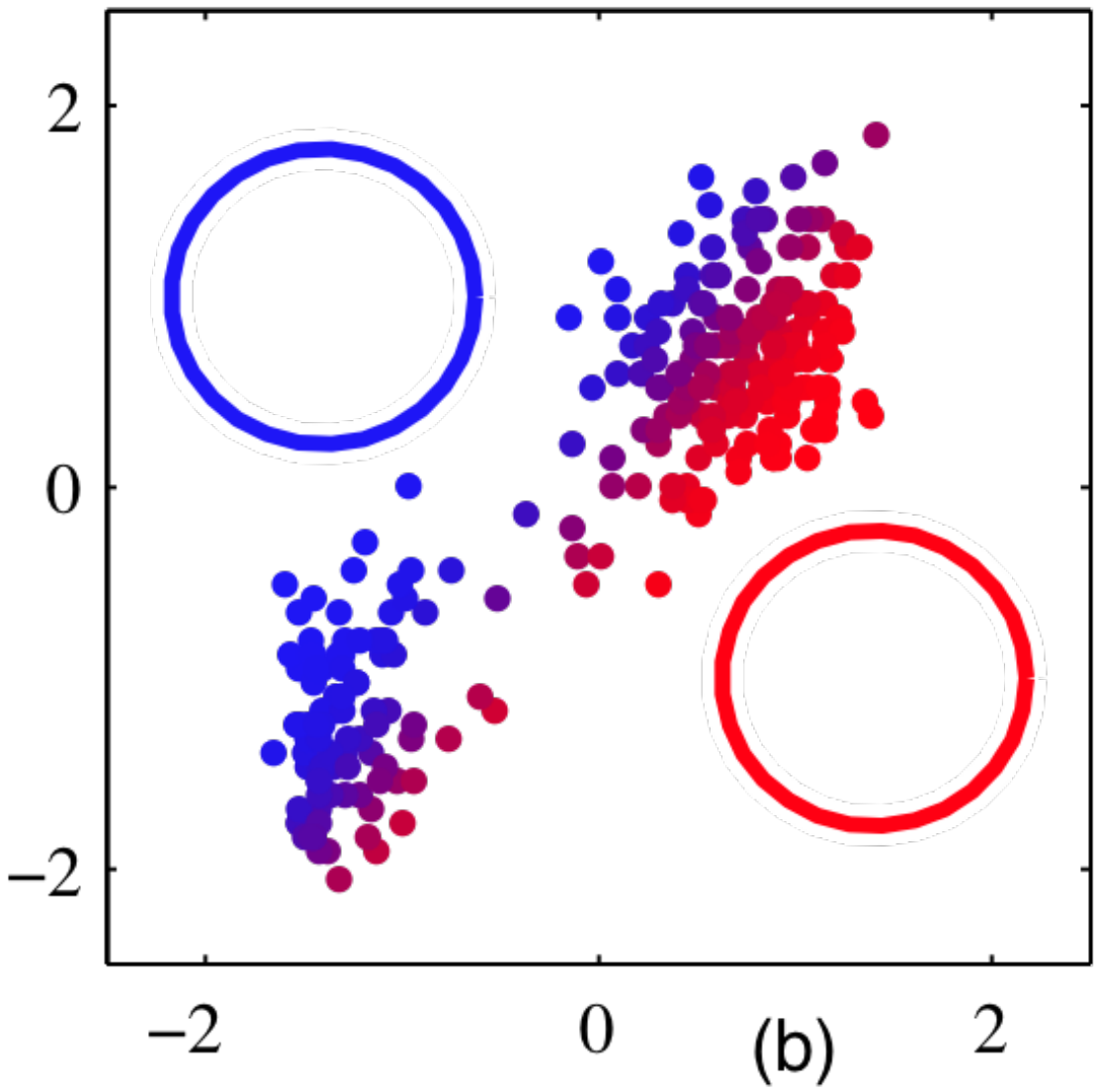
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



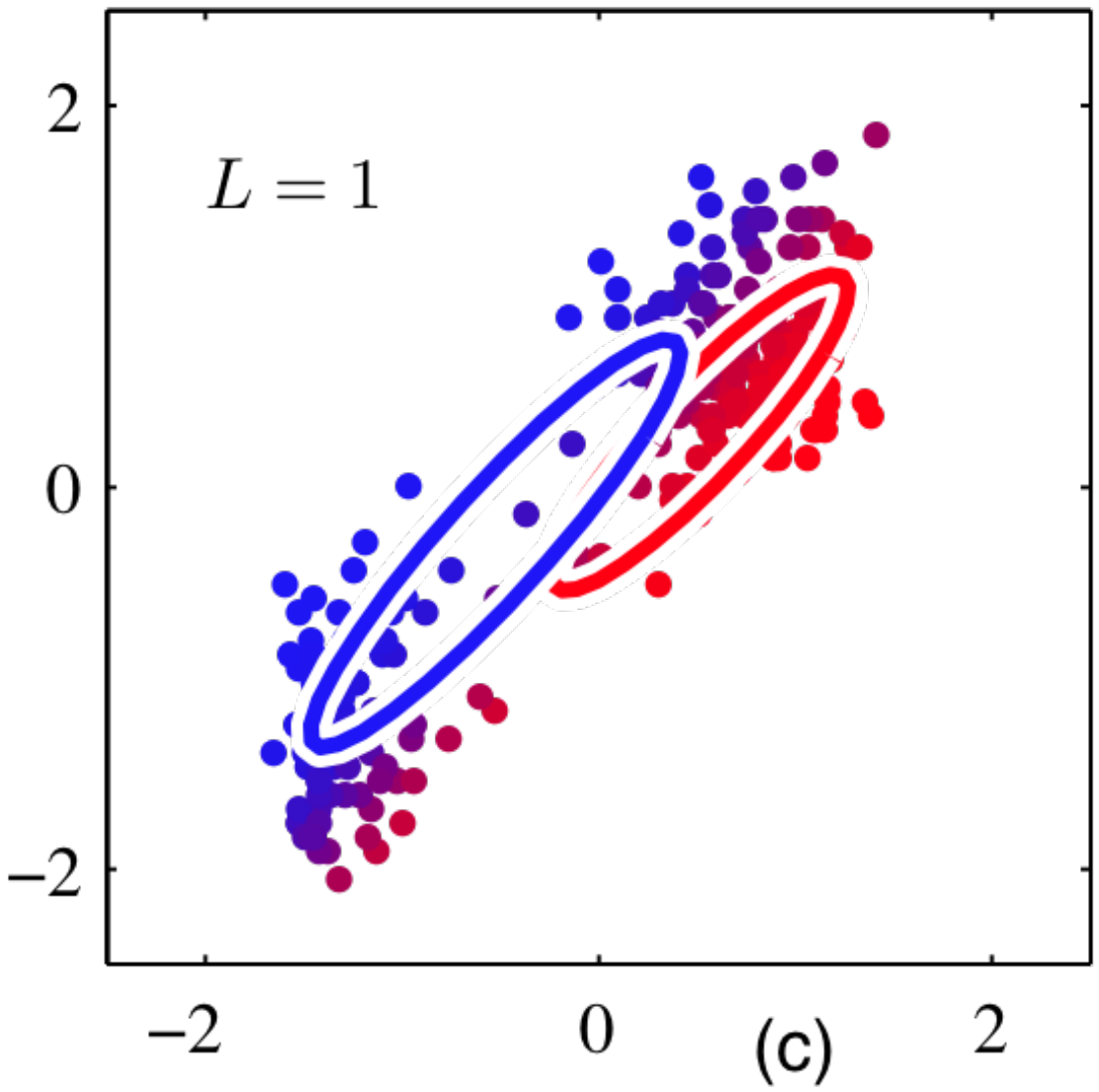
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



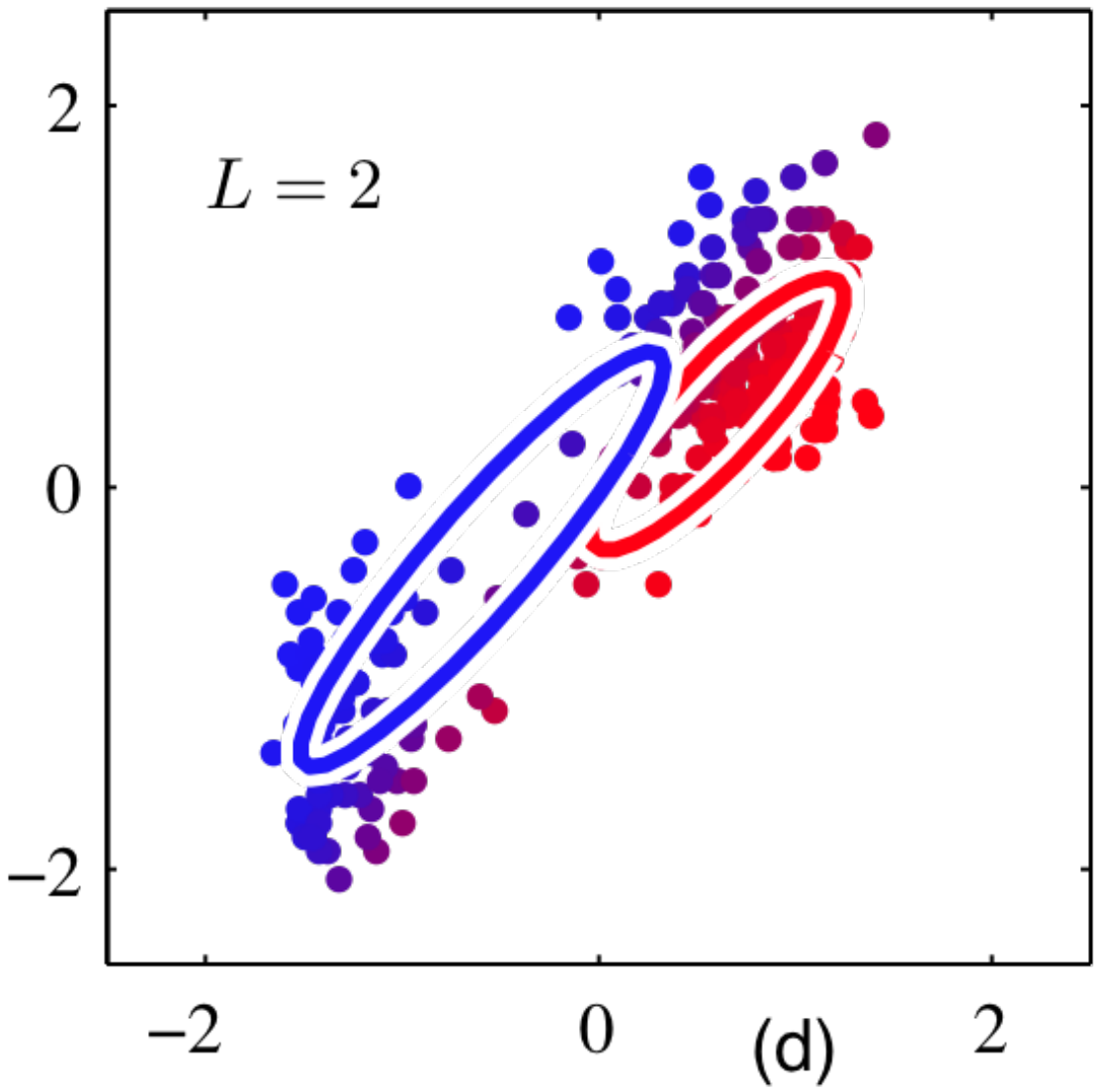
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



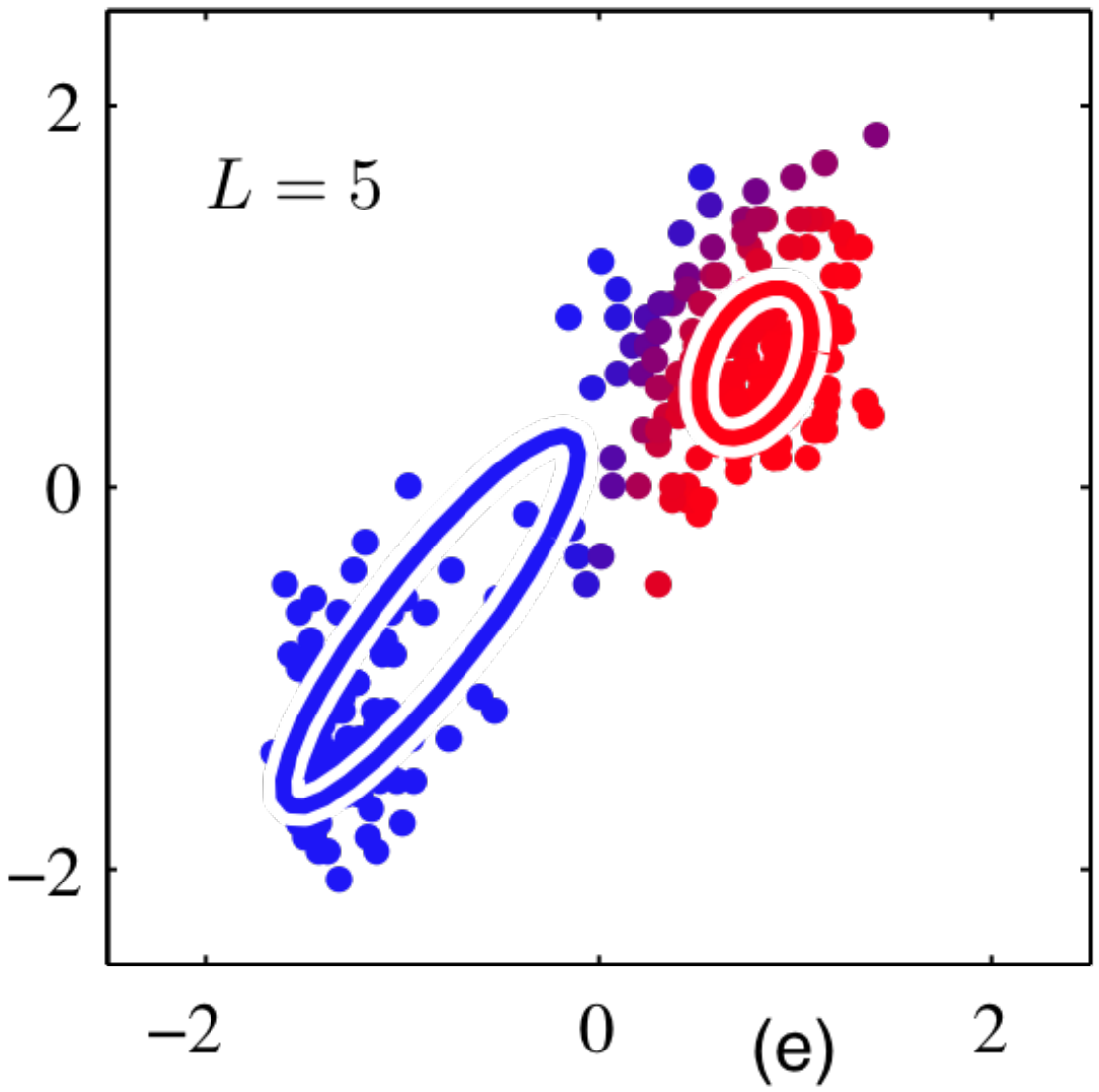
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



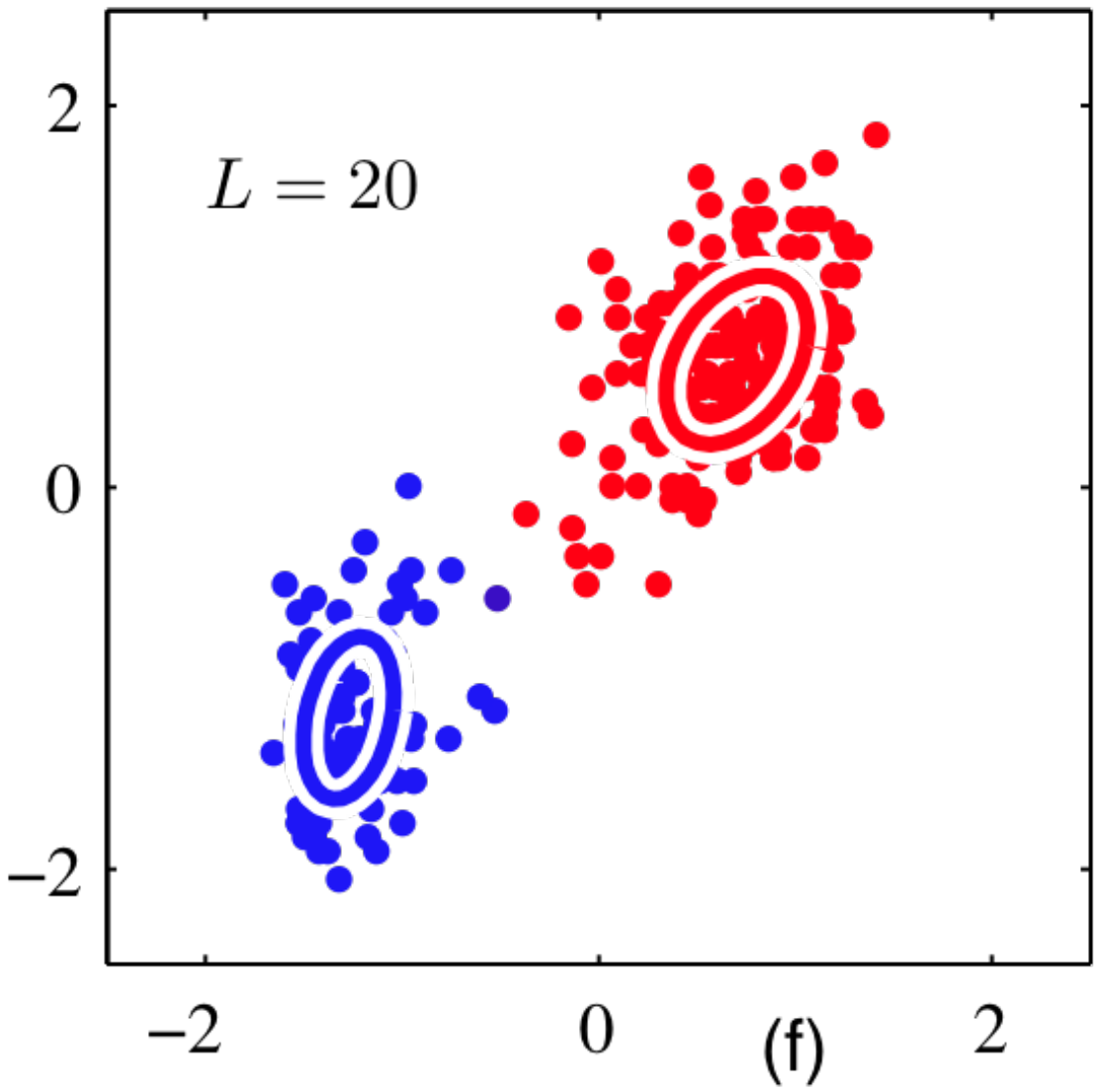
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



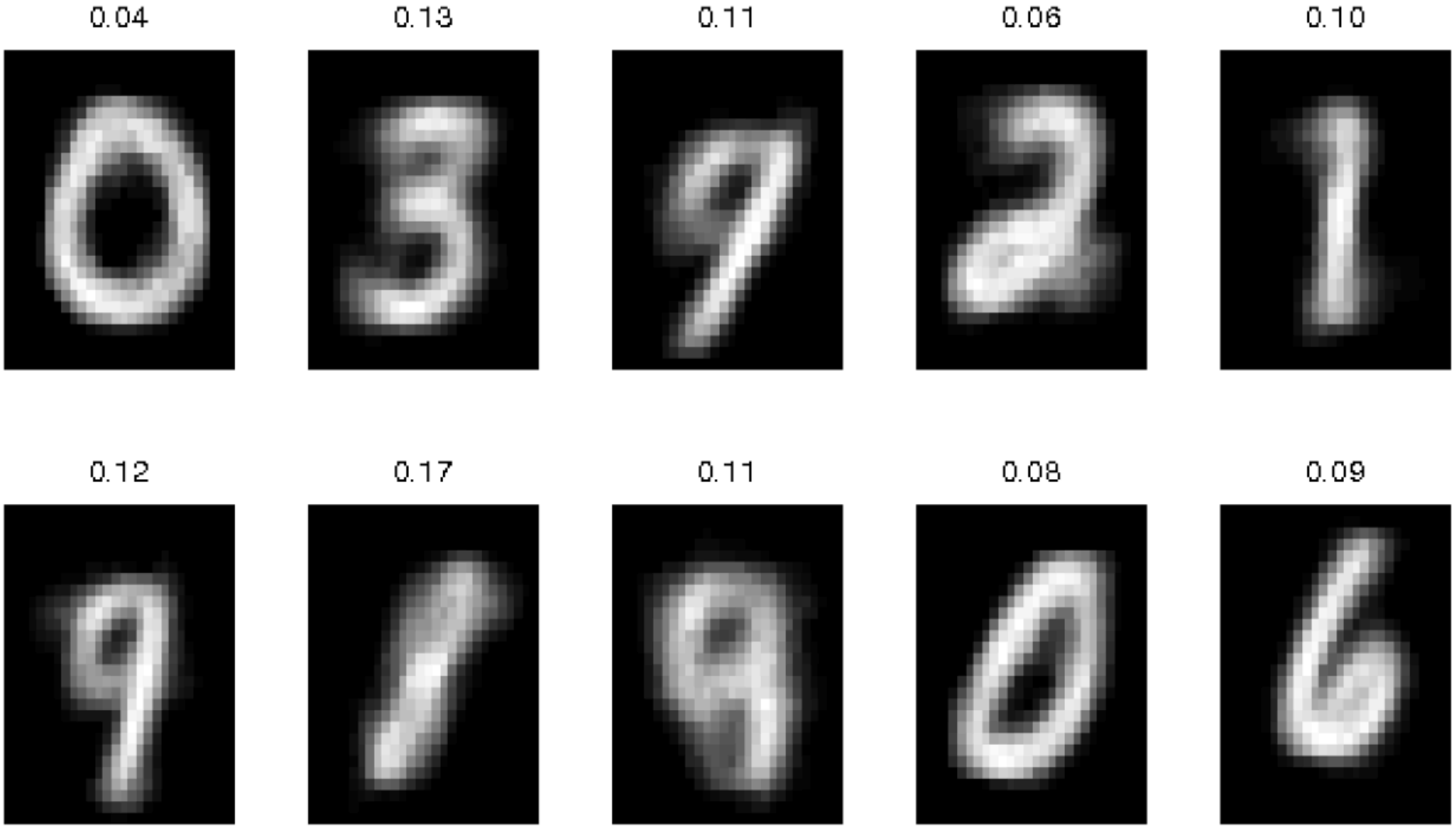
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



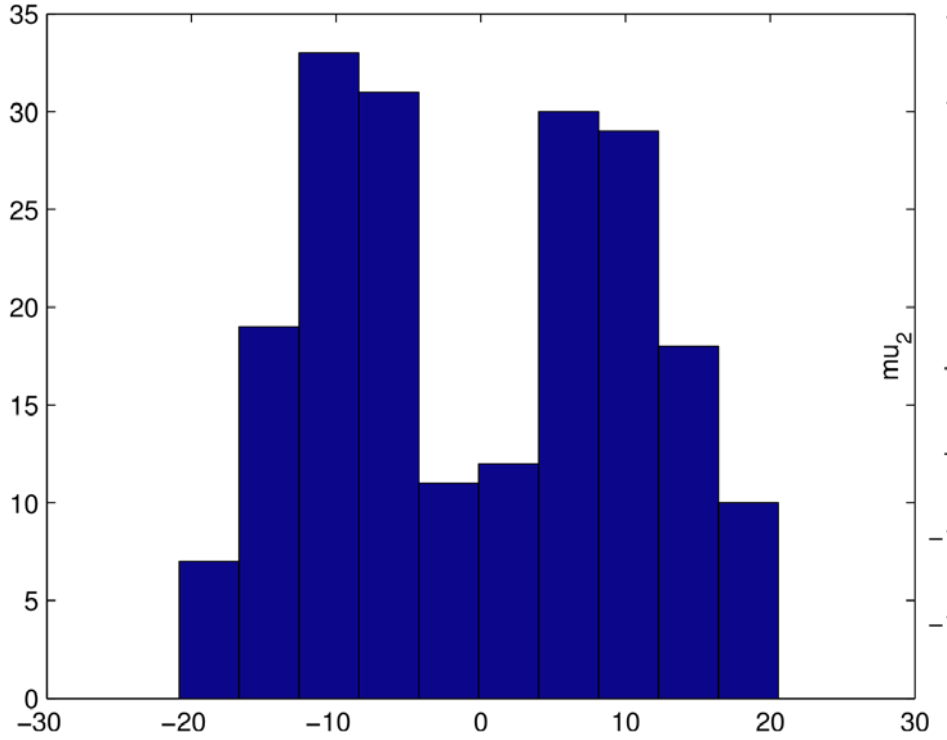
C. Bishop, Pattern Recognition & Machine Learning

Binary Features: Mixtures of Bernoullis

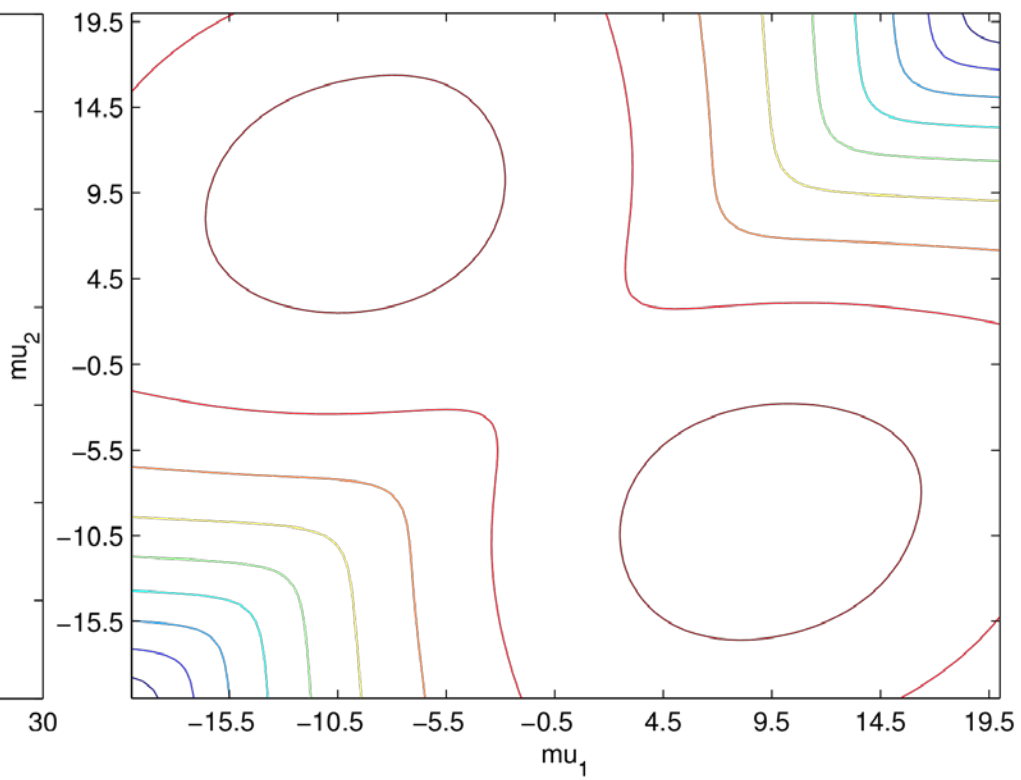


10 Clusters Identified via EM Algorithm from Binarized MNIST Digits

Label Switching in Mixture Models

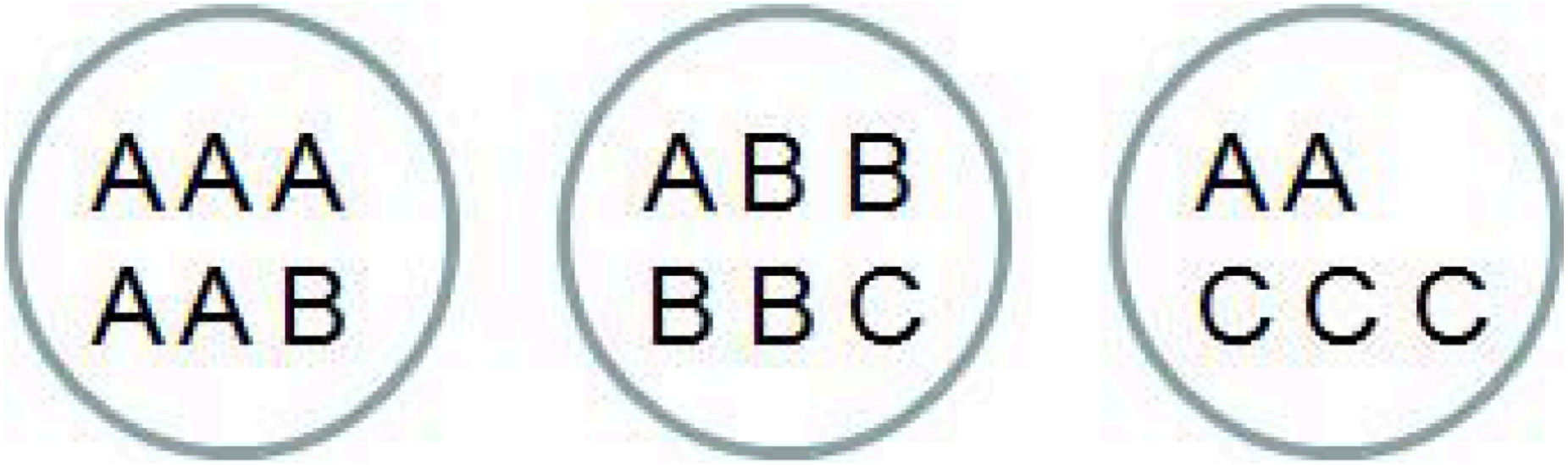


Histogram of 200 samples from a mixture of two 1D Gaussians



Two-component Gaussian mixture likelihood surface as function of means, for fixed variances

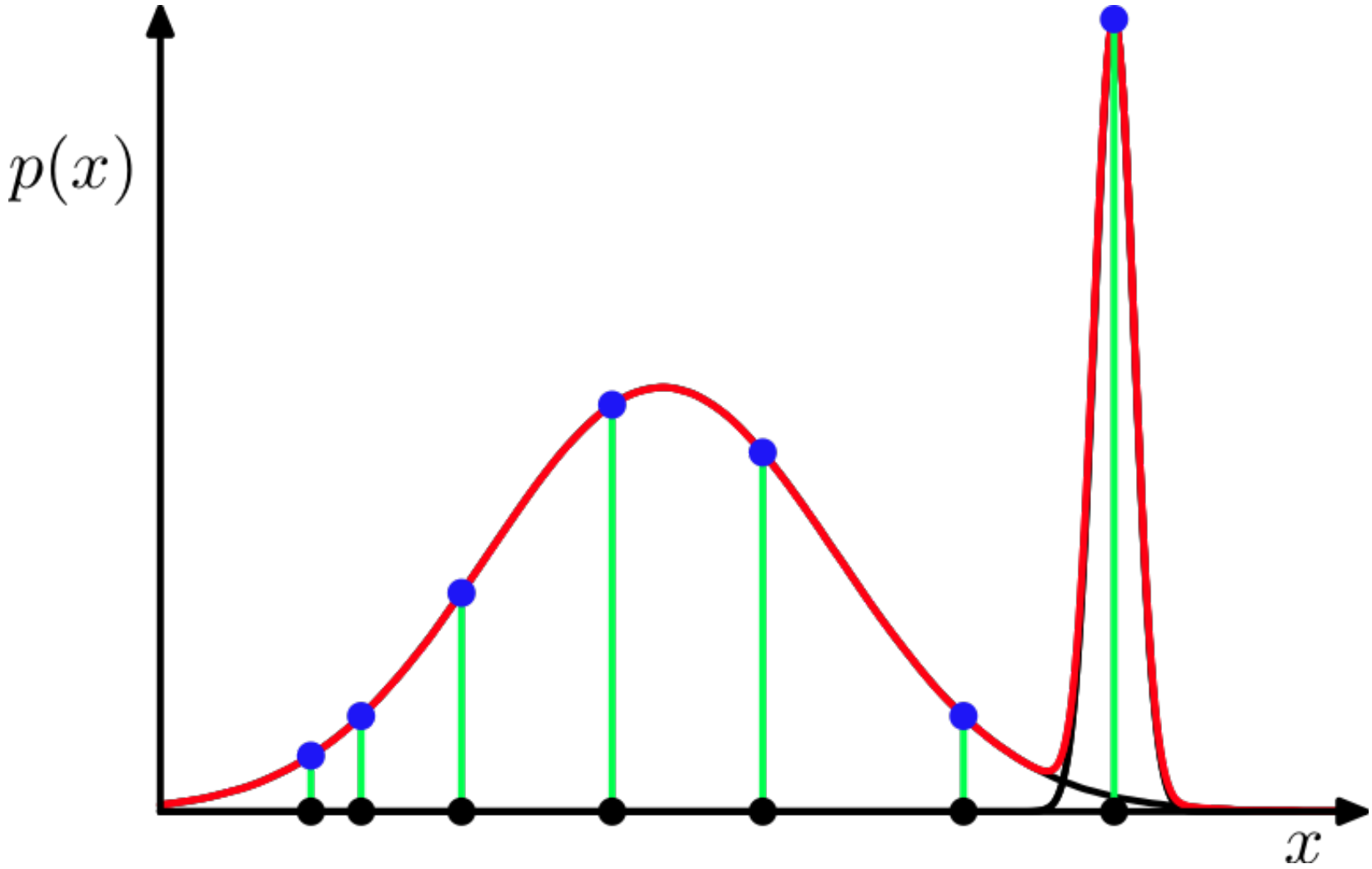
Clustering Evaluation: Rand Index



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

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

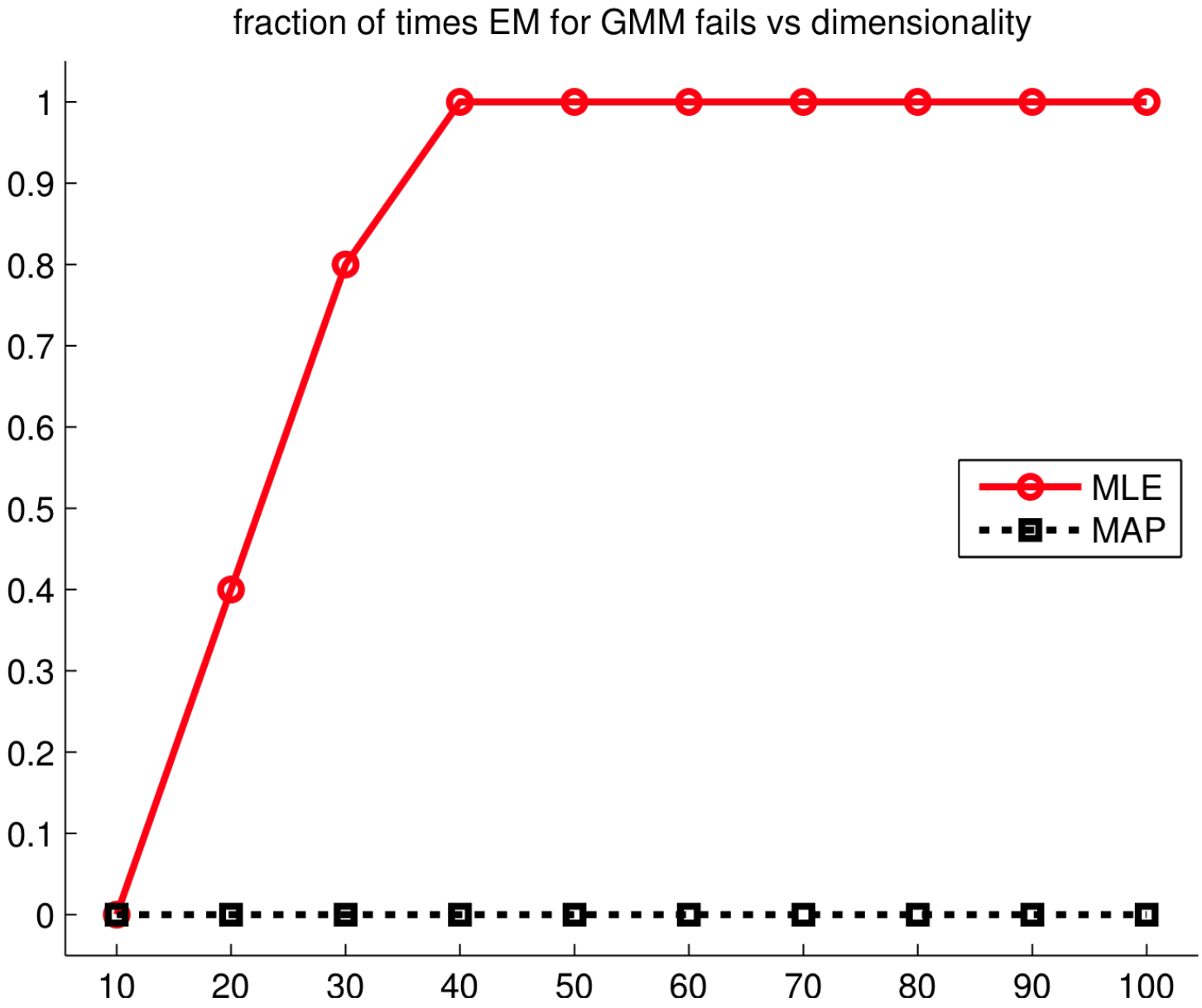
Singularities: ML for Gaussian Mixtures



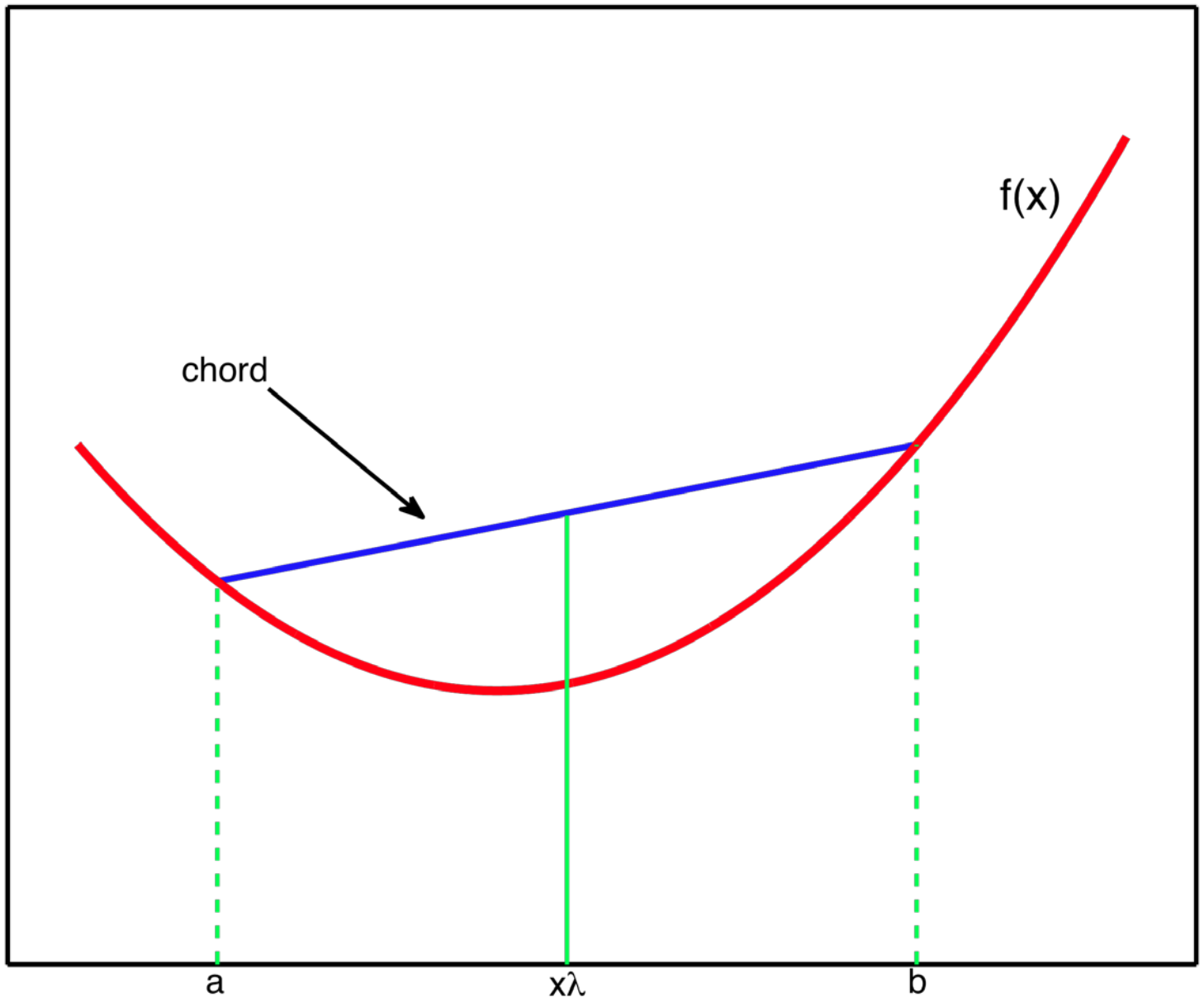
Avoid via placing prior on mixture variances. EM still applies.

C. Bishop, Pattern Recognition & Machine Learning

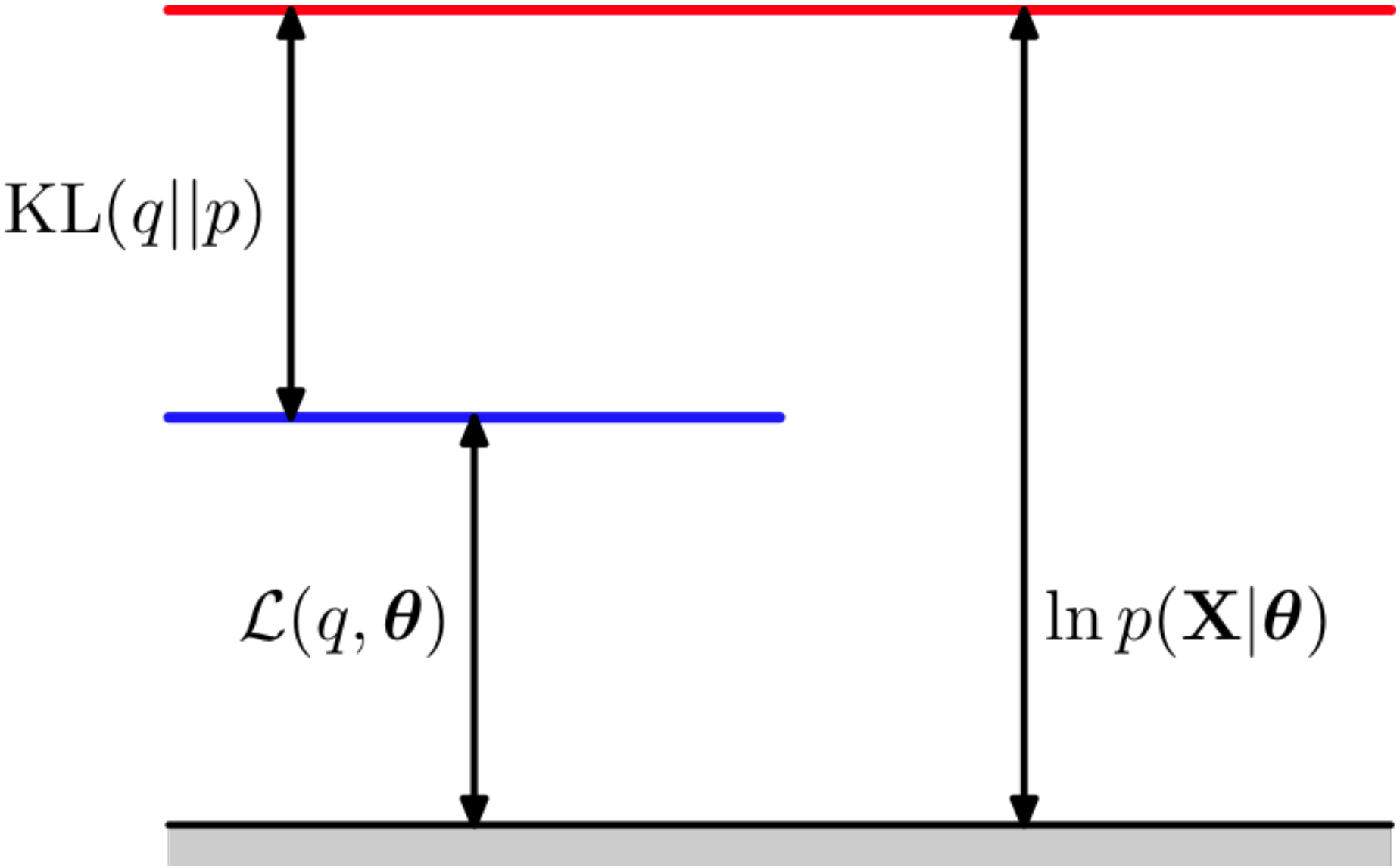
Numerical Instability: Gaussian Mixtures



Convexity & Jensen's Inequality

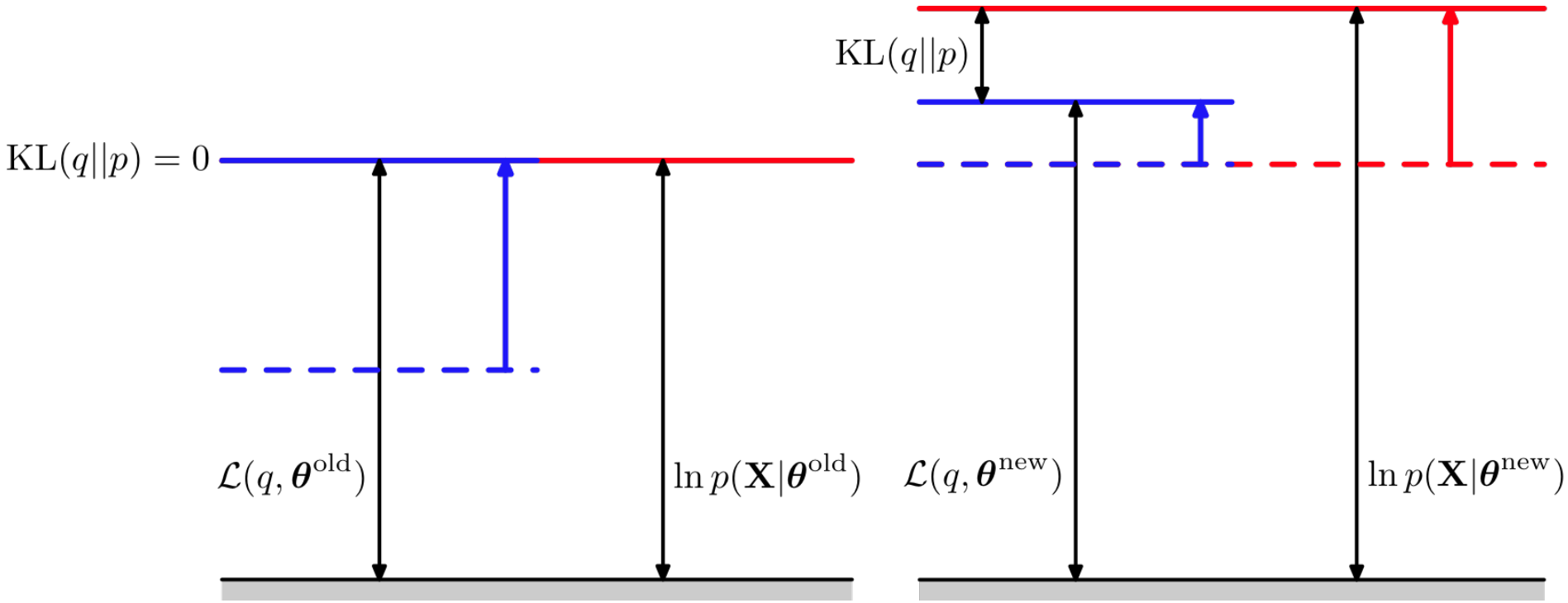


Lower Bounds on Marginal Likelihood



C. Bishop, Pattern Recognition & Machine Learning

Expectation Maximization Algorithm

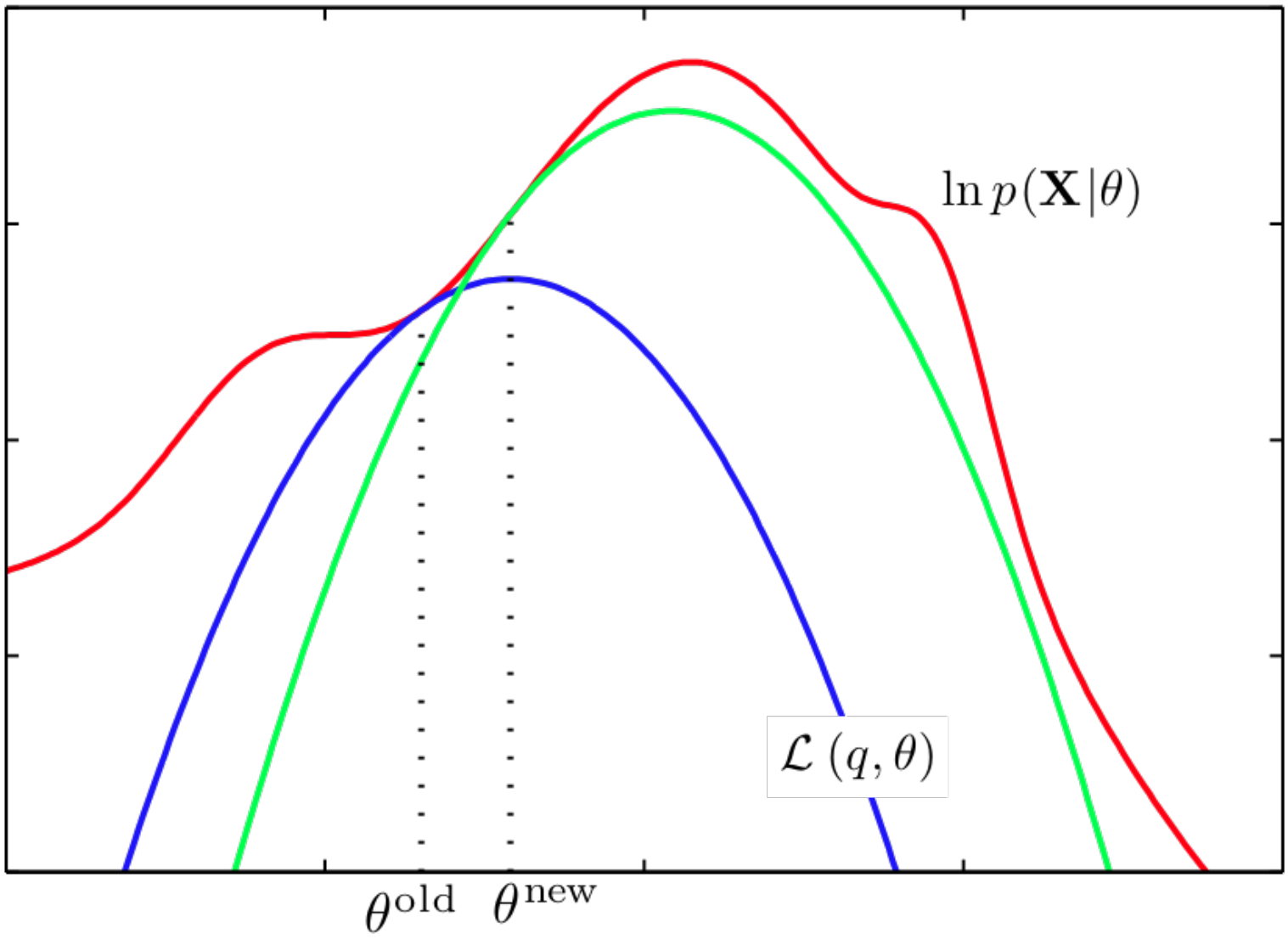


E Step: *Optimize distribution on hidden variables given parameters*

M Step: *Optimize parameters given distribution on hidden variables*

C. Bishop, Pattern Recognition & Machine Learning

EM: A Sequence of Lower Bounds



C. Bishop, Pattern Recognition & Machine Learning

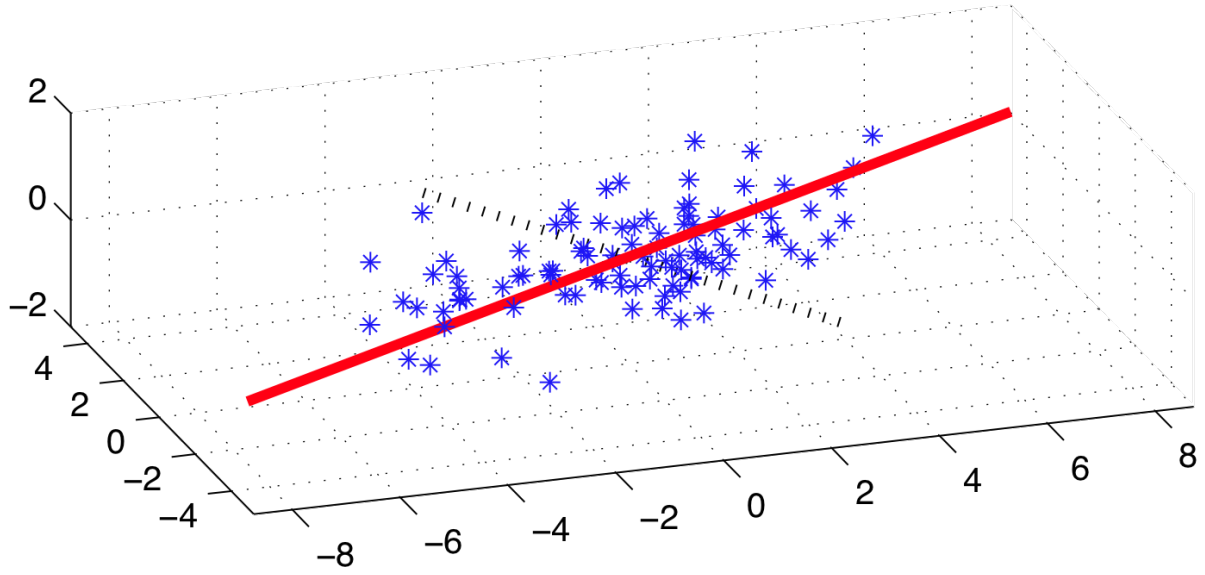
Dimensionality Reduction

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

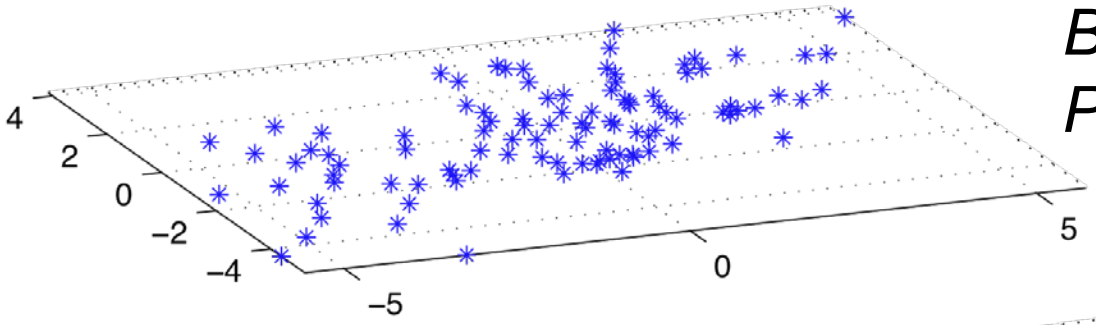
- **Goal:** Infer label/response y given only features x
- **Classical:** Find latent variables y good for *compression* of x
- **Probabilistic learning:** Estimate parameters of joint distribution p(x,y) which *maximize marginal probability* p(x)

Principal Components Analysis (PCA)

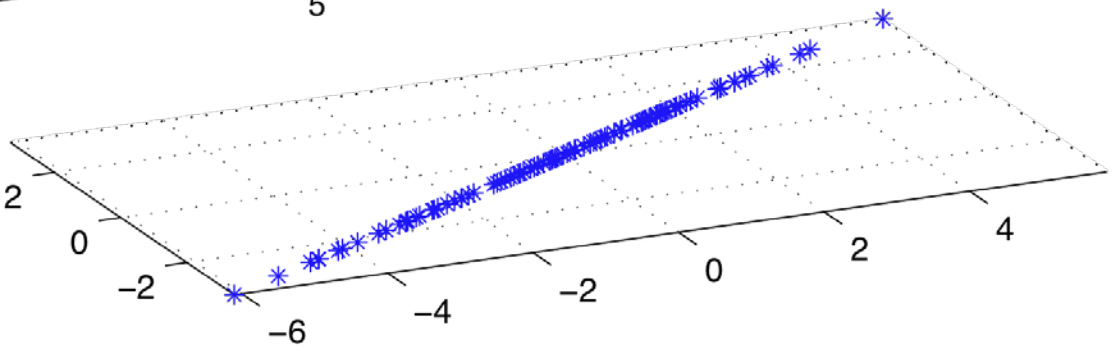
3D Data



Best 2D Projection



Best 1D Projection



Probabilistic PCA