Expectation Maximization
Chapter 5
Gaussian Mixture Models

T.L. Grobler

Supervised versus Unsupervised

- Supervised Learning:
  - Chapter 4
  - Class label comes with observations
- Unsupervised Learning:
  - Chapter 5
  - Class label inferred from observations

Expectation Maximization

- Expectation:
  - Chapter 5
  - Current estimate of parameters
  - Estimate class labels
- Maximization (MLE):
  - Chapter 4
  - Current estimate of class labels
  - Estimate parameters
- Alternate E and M step till convergence

Introduction

- Single Gaussian pdf
  - Useful analytic properties
  - Not always practical
  - Multi-modal data set
  - Mixture of k Gaussians

A Gaussian Mixture

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{k} \pi_j \mathcal{N}(\mathbf{x}|\mathbf{u}_j, \mathbf{\Sigma}_j)$$

- $\mathcal{N}$ – Gaussian distribution
- **u**, $\boldsymbol{\Sigma}$ – Mean and Covariance
- $\boldsymbol{\pi}$ – Mixture coefficients (class probabilities)
- $\Sigma\pi=1$
- $\boldsymbol{\theta}$ – Mixture Parameters (**u**, $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}$)

Problem

- Estimate $\boldsymbol{\theta}$ from $\mathbf{x}$
  - Contribution to mixture components unknown
- Similar to *k*-means
  - Cluster label unkown

A General Mixture Model

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$$

Marginalization

$$= \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z}) P(\mathbf{z})$$

Chain-Rule

Enumeration

$$= \sum_{j=1}^{k} p(\mathbf{x}|z_j = 1) P(z_j = 1)$$

$\mathcal{N}$

π_j

Gaussian Ass.

$$\mathbf{z} = [z_1, \ \cdots \ , z_n]^T \text{ where } z_j \in \{0, 1\} \text{ and } \sum_j z_j = 1$$

Latent-variable

Mixture Component Prob γ_j

Bayes-law

$$
\begin{aligned}
\gamma(z_j) &= P(z_j = 1|\mathbf{x}) \\
&= \frac{P(z_j = 1)p(\mathbf{x}|z_j = 1)}{\sum_{i=1}^{k} P(z_i = 1)p(\mathbf{x}|z_i = 1)} \\
&= \frac{\pi_j \mathcal{N}(\mathbf{x}|\mathbf{u}_j, \mathbf{\Sigma}_j)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}|\mathbf{u}_i, \mathbf{\Sigma}_i)}
\end{aligned}
$$

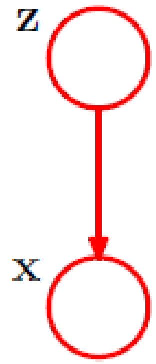
Gaussian Ass.

Interpreting γ_j

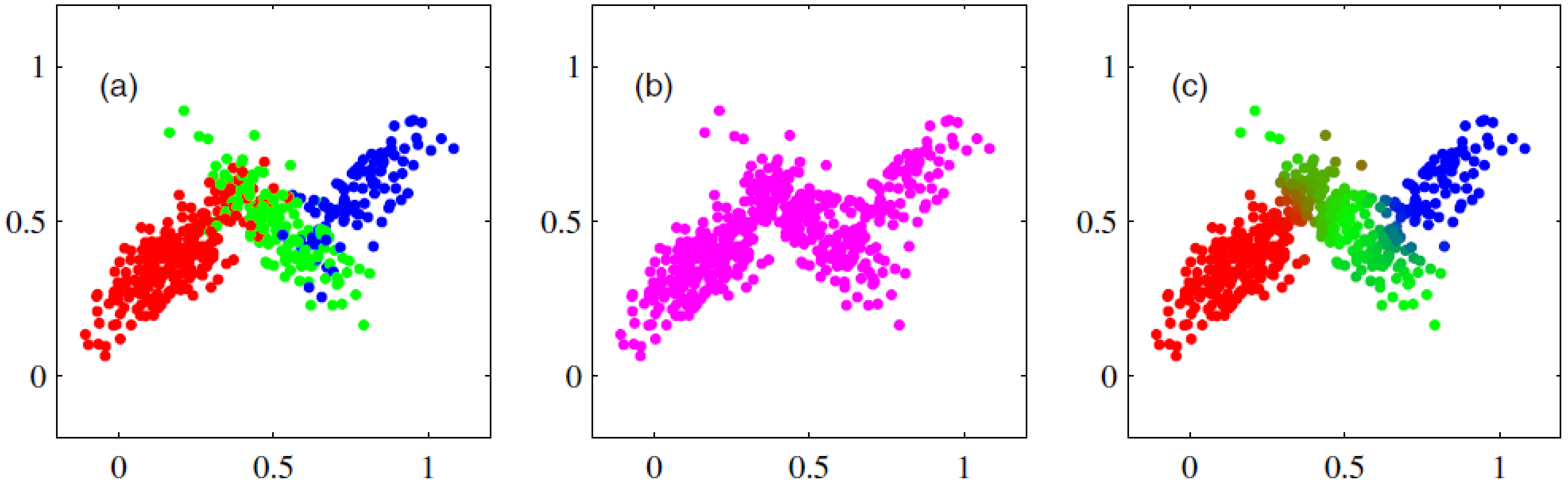
- π_j – prior probability of *j*-th mixture comp
  - Before any observations
- γ_j – posterior probability of *j*-th mixture comp
  - Once **x** has been observed
    - Prob of **x** belonging to *j*-th mixture comp
    - Responsibility *j*-th comp takes for expl **x**
    - Soft-assignment

Ancestral Sampling

- Draw a $\mathbf{z}$ from $p(\mathbf{z})$
- Draw an $\mathbf{x}$ from $p(\mathbf{x}|\mathbf{z})$
- Compute γ
- Proportion of R, B or G intensity used represents γ



Example



Matrix *Nxd*

Log-likelihood

Chain-Rule

$$p(\mathbf{X}, \pi, \mathbf{u}, \Sigma) = p(\mathbf{X}|\pi, \mathbf{u}, \Sigma)p(\boldsymbol{\pi}, \boldsymbol{u}, \Sigma)$$

Independence

$$= p(\boldsymbol{\pi}, \boldsymbol{u}, \Sigma) \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{u}, \Sigma)$$

Substitution

$$= p(\boldsymbol{\pi}, \boldsymbol{u}, \Sigma) \prod_{n=1}^{N} \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_n|\mathbf{u}_i, \boldsymbol{\Sigma}_i)$$

Discard

$$\boxed{\ln p(\mathbf{X}|\pi, \mathbf{u}, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_n|\mathbf{u}_i, \boldsymbol{\Sigma}_i) \right\}}$$

Singularity

- Comp covariances – $\mathbf{\Sigma}_k = [\sigma_k]^2\mathbf{I}$
- Data point equal to mean $\mathbf{u}_j = \mathbf{x}_j$
- Contribution to likelihood $\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n,\mathbf{\Sigma}_j)=1/[\sqrt{(2\pi)}\sigma_j]$
- $\mathcal{N}$ –> ∞ as σ_j –> 0 – Comp collapses onto a point
- Maximum likelihood estimation can fail (add prior)
- Does not occur for single Gaussian