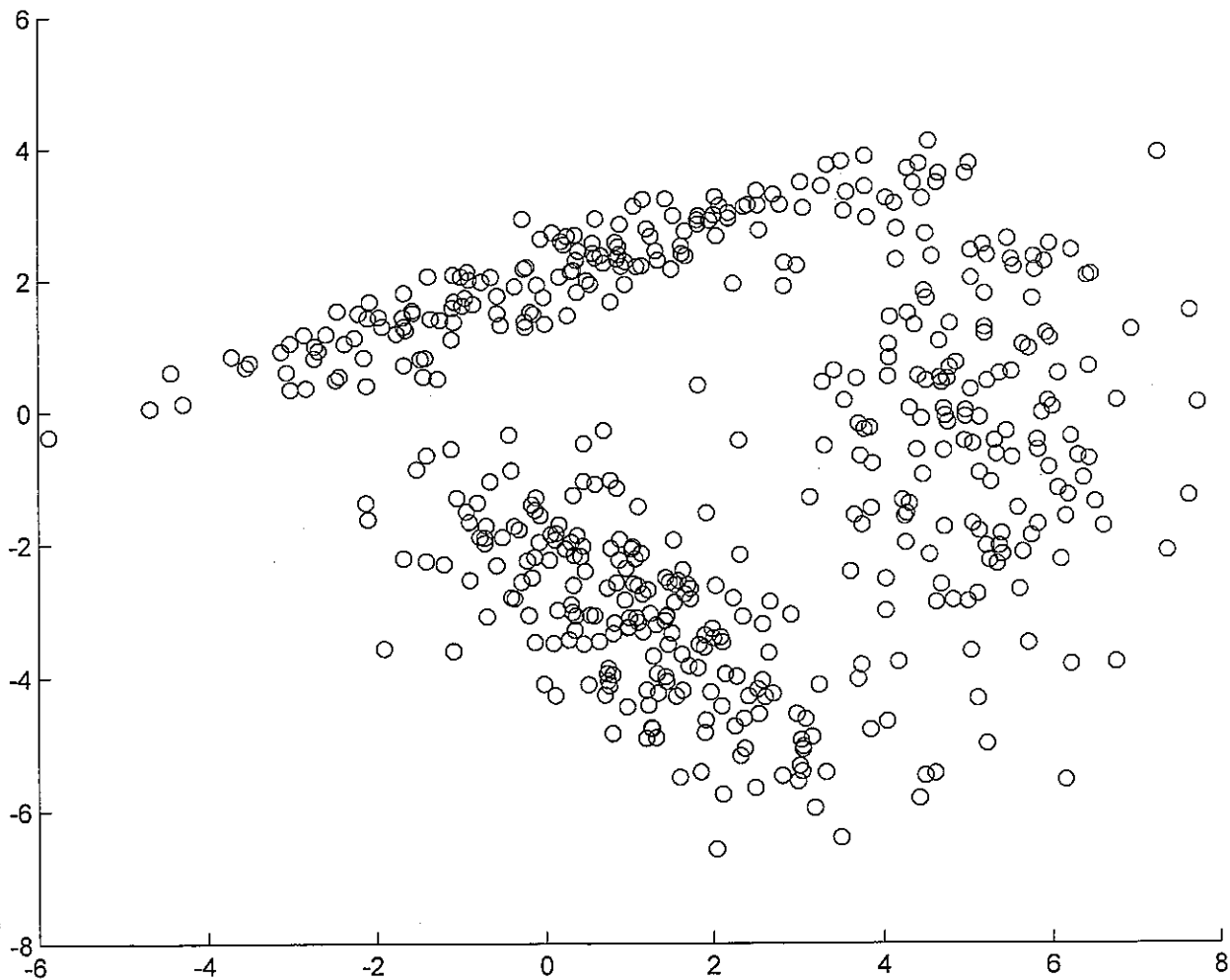
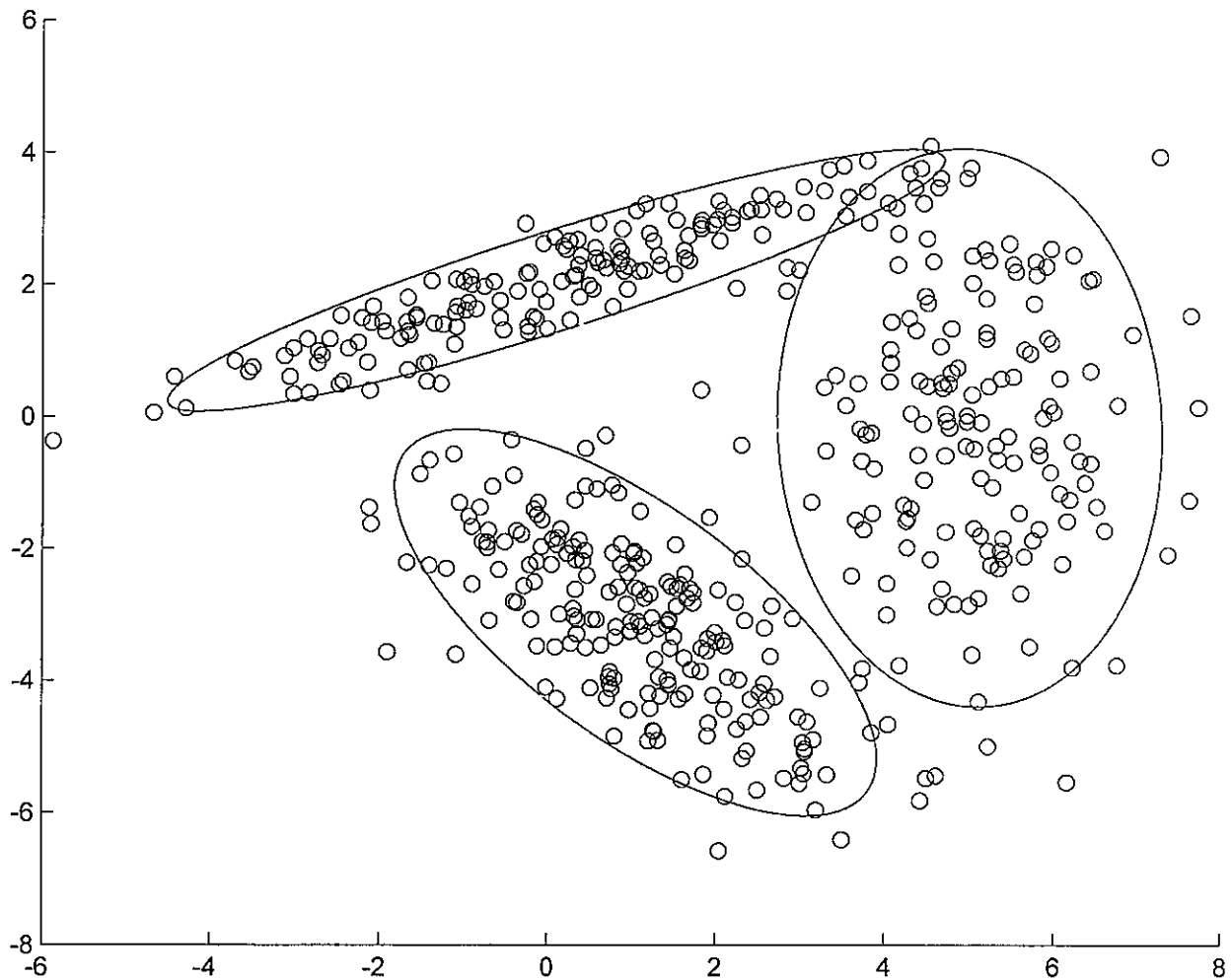


THE EM ALGORITHM FOR GAUSSIAN MIXTURE MODELS

Suppose we wish to cluster this dataset:



The data points cluster naturally into 3 groups. Each cluster is well modeled by a bivariate Gaussian density. The ellipses are "90% contours" based on a fitted Gaussian mixture model (GMM)



In these notes we'll develop a general method for clustering data when clusters are elliptical, based on maximum likelihood estimation of GMMs.

Gaussian Mixture Models

Recall the multivariate Gaussian density

$$\phi(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where

$$x \in \mathbb{R}^d$$

$$\mu \in \mathbb{R}^d$$

$$\Sigma \in \mathbb{R}^{d \times d}, \Sigma > 0$$

A random variable X follows a Gaussian mixture model if its density function has the form

$$f(x) = \sum_{k=1}^K w_k \phi(x; \mu_k, \Sigma_k)$$

where

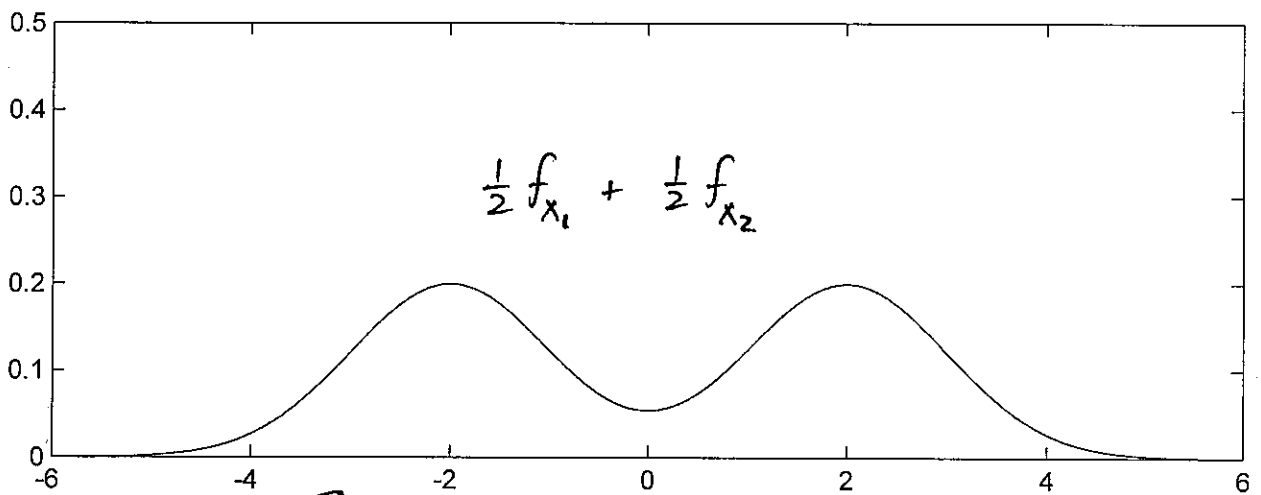
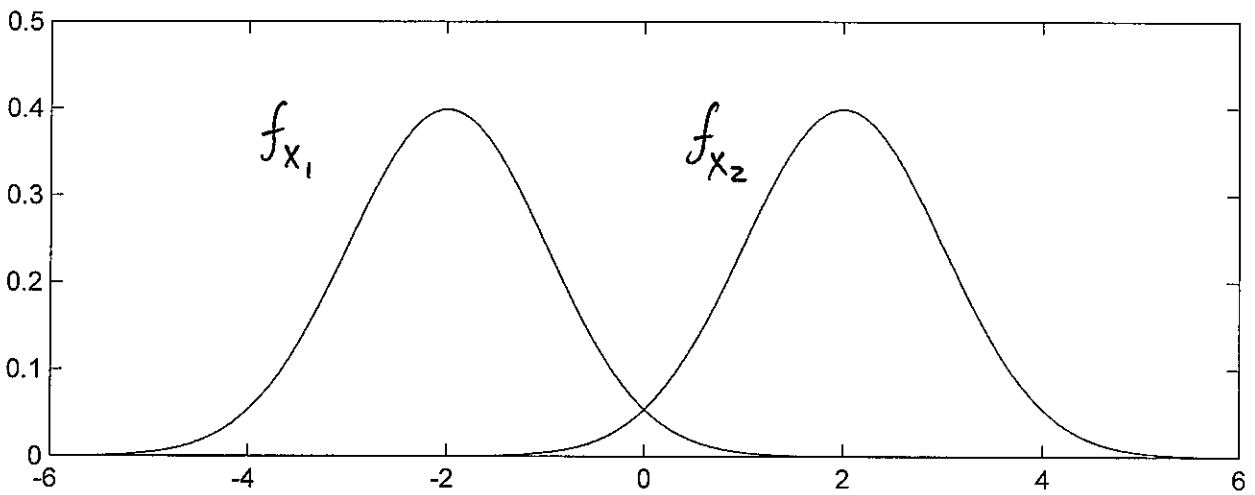
$$0 < w_k < 1, \sum_{k=1}^K w_k = 1$$

$$\mu_k \in \mathbb{R}^d$$

$$\Sigma_k \in \mathbb{R}^{d \times d}, \Sigma_k > 0$$

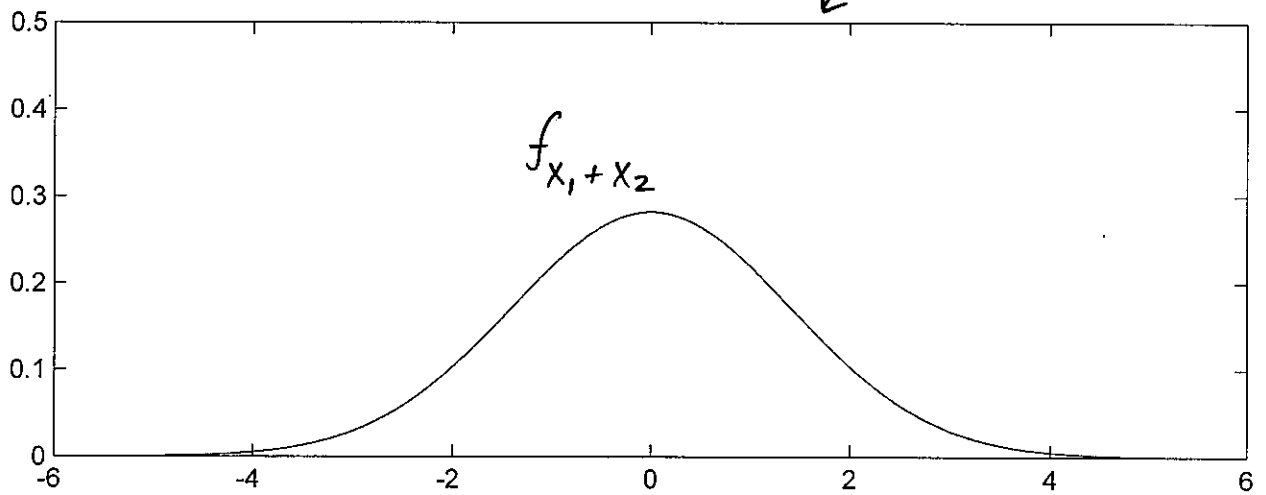
$$X_1 \sim \mathcal{N}(-2, 1)$$

$$X_2 \sim \mathcal{N}(2, 1)$$



mixture \nearrow

\nwarrow not a mixture



(B)

$$X_1 + X_2 \sim$$

Simulating a GMM

Suppose

$$\theta = (w_1, \dots, w_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$$

is known. How can we simulate a realization of the GMM?

Basic idea:

- 1: Choose a "component" at random, weighted according to w_k
- 2: Draw a realization from $\mathcal{N}(\mu_k, \Sigma_k)$

Why does this work?

Let $S \in \{1, 2, \dots, K\}$ be a discrete RV such that

$$\Pr \{S = k\} = w_k.$$

Generate X as follows:

1. Generate a realization s of S .
2. Generate $X \sim \mathcal{N}(\mu_s, \Sigma_s)$.

Then the density of X generated in this way is

$$\begin{aligned} f(x) &= \sum_{k=1}^K f(x | S=k) \cdot \Pr \{S=k\} \\ &= \sum_{k=1}^K w_k \phi(x; \mu_k, \Sigma_k) \end{aligned}$$

as desired.

The variable S is called a (hidden) state variable. We will imagine that every realization from a GMM is associated with a specific realization of a state variable.

In clustering, our objective is to estimate $\theta = (\{\omega_k\}, \{\mu_k\}, \{\Sigma_k\})$, and to define clusters in terms of the estimated GMM.

That is, we assume

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x; \theta)$$

and reduce clustering to parameter estimation

Maximum Likelihood Estimation

Suppose $f(x; \theta)$ is an arbitrary density parametrized by θ , and consider

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x; \theta).$$

Denote a fixed realization of X_1, \dots, X_n by

$$\underline{x} = (x_1, \dots, x_n)$$

The likelihood function of θ is

$$l(\theta; \underline{x}) := f(\underline{x}; \theta)$$

©

=

The maximum likelihood estimator (MLE) is

$$\hat{\theta} = \hat{\theta}(\underline{x}) := \arg \max_{\theta} l(\theta; \underline{x})$$

Example | Suppose

$$X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \Sigma)$$

so that $\theta = (\mu, \Sigma)$. Then

$$l(\theta; \underline{x}) = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left\{ -\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right\}$$

Computationally, it is convenient to maximize the log-likelihood

$$\log l(\theta; \underline{x}) = \sum_{i=1}^n \left[-\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right]$$

The MLE is

①

$$\hat{\mu} =$$

$$\hat{\Sigma} =$$

Now consider a GMM. Assume K is known.

The likelihood function is

$$l(\theta; \underline{x}) = \prod_{i=1}^n f(x_i; \theta)$$

⑤ =

and the log-likelihood is

$$\log l(\theta; \underline{x}) =$$

There is unfortunately no known closed-form maximizer.

Suppose, just for kicks, that we also had access to the realization of state variables

$$\underline{s} = (s_1, \dots, s_n)$$

associated with \underline{x} .

Notation

$$I_k = \{i : s_i = k\}$$

$$n_k = |I_k|$$

Then the likelihood is

$$\begin{aligned}l(\theta; \underline{x}, \underline{s}) &= \prod_{i=1}^n f(x_i, s_i; \theta) \\ &= \prod_{i=1}^n f(x_i | s_i; \theta) \cdot \Pr\{S_i = s_i; \theta\}\end{aligned}$$

(E)

=

=

and the log-likelihood is

$$\log l(\theta; \underline{x}, \underline{s}) =$$

\Rightarrow can maximize w.r.t (μ_k, Σ_k) independently
from other components

$$\Rightarrow \hat{\mu}_k =$$

$$\hat{\Sigma}_k =$$

To solve

$$\max \sum_{k=1}^K n_k \log w_k$$

$$\text{s.t. } \sum w_k = 1$$

we can use Lagrange multipliers

The Lagrangian is

$$L(w_1, \dots, w_K, \lambda) = \sum_{k=1}^K n_k \log w_k + \lambda \left(\sum_{k=1}^K w_k - 1 \right)$$

$$\frac{\partial L}{\partial w_k} = \frac{n_k}{w_k} + \lambda = 0$$

$$\Rightarrow w_k = -\frac{n_k}{\lambda}$$

$$\Rightarrow \sum \left(-\frac{n_k}{\lambda} \right) = 1$$

$$\Rightarrow \lambda = -\sum n_k = -n$$

$$\Rightarrow \boxed{\hat{w}_k = \frac{n_k}{n}}$$

The combined data

$$\underline{z} = (\underline{x}, \underline{s})$$

is called the complete data. More generally, the complete data is a combination of observed and unobserved data that makes the MLE tractable.

The Expectation - Maximization Algorithm

Define the "indicator" variable

$$\Delta_{i,k} = \begin{cases} 1 & \text{if } s_i = k \\ 0 & \text{if } s_i \neq k \end{cases}$$

The complete data log-likelihood can be written

$$\log l(\theta; \underline{x}, \underline{s})$$

$$= \sum_{i=1}^n \log \left(\sum_{k=1}^K \Delta_{i,k} \cdot w_k \phi(x_i; \mu_k, \Sigma_k) \right)$$

$$= \sum_{i=1}^n \sum_{k=1}^K \Delta_{i,k} \left[\log w_k + \log \phi(x_i; \mu_k, \Sigma_k) \right]$$

The EM algorithm is an iterative algorithm that produces a sequence $\theta^{(1)}, \theta^{(2)}, \dots$ of estimates

E-Step Given $\theta^{(j)}$, compute the expected complete-data
log likelihood

$$Q(\theta, \theta^{(j)}) = \mathbb{E} \left[\log l(\theta; \underline{x}, \underline{S}) \mid \underline{x}; \theta^{(j)} \right]$$

↑
w.r.t. $\underline{S} \mid \underline{x}$

In our case,

$$\mathbb{E} \left[\log l(\theta; \underline{x}, \underline{S}) \mid \underline{x}; \theta^{(j)} \right]$$

$$= \sum_{i=1}^n \sum_{k=1}^K \gamma_{i,k}(\theta^{(j)}) \cdot \left[\log w_k + \log \phi(x_i; \mu_k, \Sigma_k) \right]$$

where

$$\gamma_{i,k}(\theta^{(j)}) = \frac{w_k \phi(x_i; \mu_k, \Sigma_k)}{\sum_{k=1}^K w_k \phi(x_i; \mu_k, \Sigma_k)}$$

⑥

M-Step

$$\theta^{(j+1)} =$$

EM Algorithm for GMMs

Initialize $\theta^{(0)}$

Repeat

E-Step: Compute

$$\gamma_{i,k}^{(j)} := \gamma_{i,k}(\theta^{(j)}) = E[\Delta_{i,k} | \underline{x}; \theta^{(j)}]$$

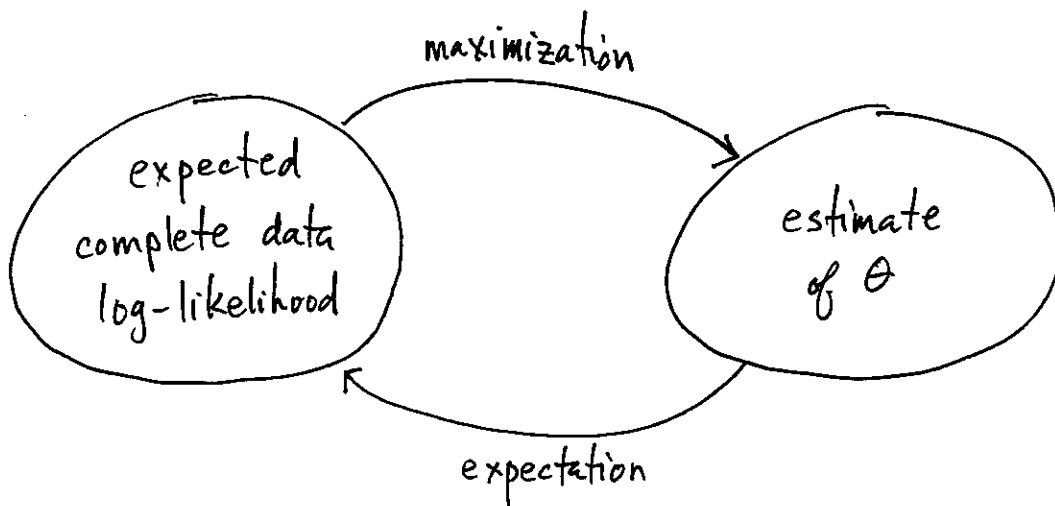
and form

$$Q(\theta, \theta^{(j)}) := E[\log l(\theta; \underline{x}, \underline{z}) | \underline{x}; \theta^{(j)}]$$

M-Step:

$$\theta^{(j+1)} = \arg \max_{\theta} Q(\theta, \theta^{(j)})$$

Until termination criterion satisfied



M-Step for GMM

Maximizing

$$Q(\theta, \theta^{(j)}) = \sum_{i=1}^n \sum_{k=1}^K \gamma_{i,k}^{(j)} \left[\log w_k - \frac{d}{2} \log 2\pi \right. \\ \left. - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right]$$

w.r.t. $\theta = (\{w_k\}, \{\mu_k\}, \{\Sigma_k\})$ yields

$$\mu_k^{(j+1)} = \frac{\sum_{i=1}^n \gamma_{i,k}^{(j)} x_i}{\sum_{i=1}^n \gamma_{i,k}^{(j)}}$$

$$\Sigma_k^{(j+1)} = \frac{\sum_{i=1}^n \gamma_{i,k}^{(j)} (x_i - \mu_k^{(j+1)}) (x_i - \mu_k^{(j+1)})^T}{\sum_{i=1}^n \gamma_{i,k}^{(j)}}$$

$$w_k^{(j+1)} = \frac{1}{n} \sum_{i=1}^n \gamma_{i,k}^{(j)}$$

E-Step for GMM

Exercise 1 Express $\gamma_{i,k}^{(j)}$ as a function of $x_i, \theta^{(j)}$.

Hint: Apply Bayes' rule.

Solution 1

$$\begin{aligned} \gamma_{i,k}^{(j)} &= E[\Delta_{i,k} \mid x_i; \theta^{(j)}] \\ &= \Pr\{\Delta_{i,k} = 1 \mid x_i; \theta^{(j)}\} \\ &= \Pr\{S_i = k \mid x_i; \theta^{(j)}\} \\ &= \frac{\Pr\{S_i = k; \theta^{(j)}\} \cdot f(x_i \mid S_i = k; \theta^{(j)})}{f(x_i; \theta^{(j)})} \\ &= \frac{w_k^{(j)} \phi(x_i; \mu_k^{(j)}, \Sigma_k^{(j)})}{\sum_{l=1}^K w_l^{(j)} \phi(x_i; \mu_l^{(j)}, \Sigma_l^{(j)})} \end{aligned}$$

Termination

One possible termination criterion is to stop iterating when

$$|\ell(\theta^{(j+1)}; \underline{x}) - \ell(\theta^{(j)}; \underline{x})| \leq \epsilon$$

or when

$$|Q(\theta^{(j+1)}, \theta^{(j)}) - Q(\theta^{(j)}, \theta^{(j)})| \leq \epsilon$$

for some pre-chosen tolerance ϵ .

Initialization

In general, the likelihood has several local maxima. Therefore initialization of the EM algorithm is critical.

In fact, a global maximum is obtained by putting $\mu_1 = x_i$ for some i , $\Sigma_1 = 0$, $w_1 = 1$, which is not a useful solution. Hence, we are actually seeking a local maximum.

A good initialization for the GMM is

$$\mu_k^{(0)} = \text{random } x_i \text{ (distinct)}$$

$$\Sigma_k^{(0)} = \text{sample covariance}$$

$$w_k^{(0)} = \frac{1}{K}$$

In practice, it may be beneficial to initialize the algorithm and run it several times, and select the final estimate with largest likelihood.

Defining Clusters

Recall

$$\gamma_{i,k}(\theta) = \Pr\{S_i = k \mid x_i; \theta\}$$

Therefore a reasonable "hard" assignment of points to clusters is given by

$$x_i \mapsto \arg \max_{k=1, \dots, K} \gamma_{i,k}(\hat{\theta})$$

Alternatively, for "soft" assignments, we may view $\gamma_{i,k}(\hat{\theta})$ as the affinity of x_i for cluster k .

The EM Algorithm in General

The EM is not specific to GMMs, but rather a general class of algorithms for computing ML/MAP estimators.

It is applicable when having knowledge of certain hidden variables renders the MLE tractable.

\underline{x} = observed data

\underline{z} = unobserved/hidden data

EM Algorithm

Initialize $\theta^{(0)}$

Repeat

E-Step: Form

$$Q(\theta, \theta^{(j)}) := E[\log \ell(\theta; \underline{x}, \underline{z}) \mid \underline{x}; \theta^{(j)}]$$

M-Step: Compute

$$\theta^{(j+1)} = \arg \max_{\theta} Q(\theta, \theta^{(j)})$$

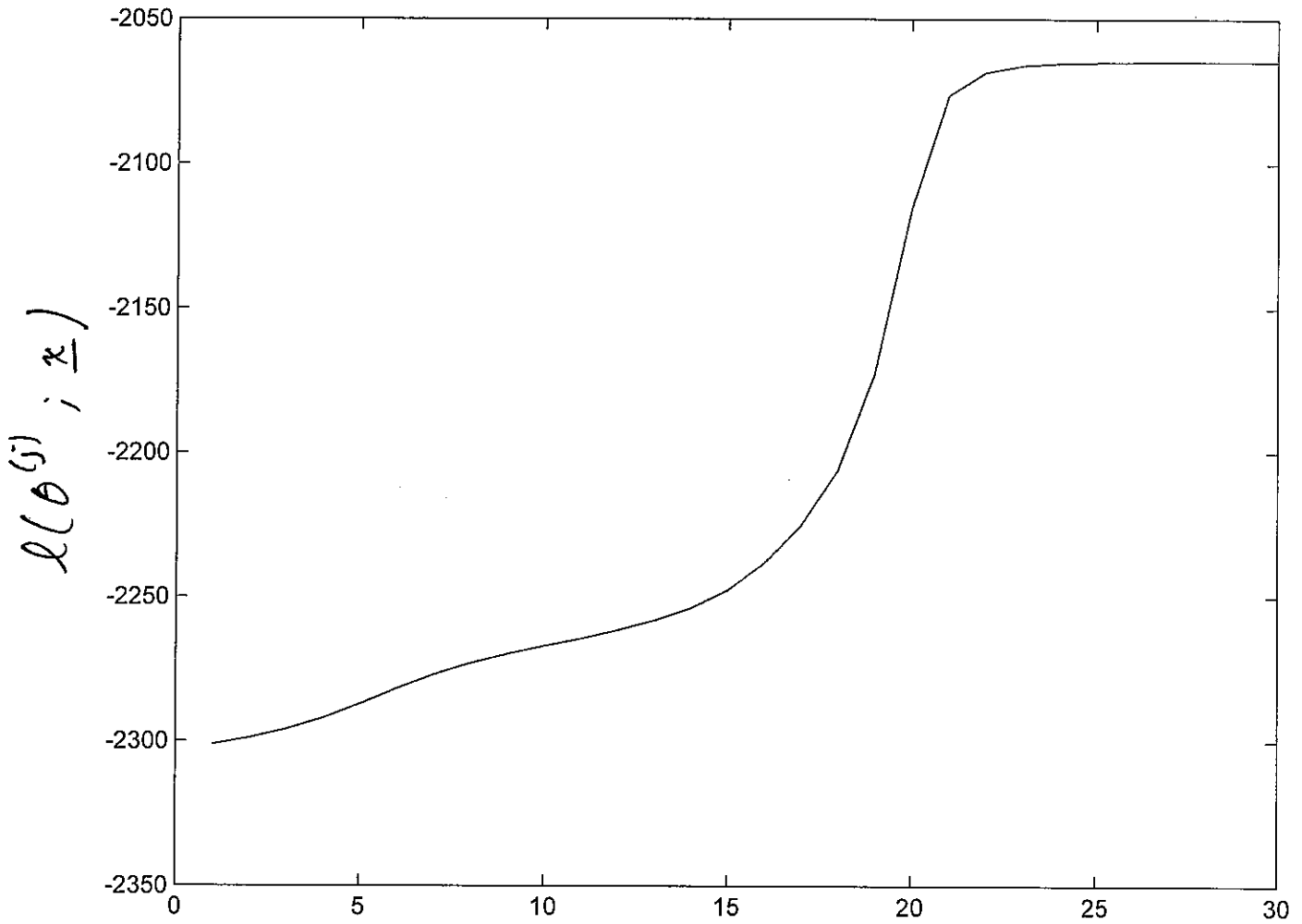
Until termination criterion satisfied

Theorem | For each $j = 1, 2, \dots$

$$\ell(\theta^{(j+1)}; \underline{x}) \geq \ell(\theta^{(j)}; \underline{x})$$

A proof based on Jensen's inequality may be found in Hastie, Tibshirani & Friedman

Example 1 3 component GMM shown at start
of notes



j

Connection to K-means

Consider a Gaussian mixture model (GMM)

$$f(x) = \sum_{k=1}^K w_k \phi(x; \mu_k, \sigma^2 \mathbf{I})$$

where σ^2 is fixed. The EM algorithm for ML estimation of $\{w_k\}_{k=1}^K$, $\{\mu_k\}_{k=1}^K$ is to iterate

$$\bullet \quad \mu_k = \frac{\sum_{i=1}^n \gamma_{i,k} x_i}{\sum_{i=1}^n \gamma_{i,k}}$$

$$w_k = \frac{1}{n} \sum_{i=1}^n \gamma_{i,k}$$

$$\bullet \quad \gamma_{i,k} = \frac{w_k \phi(x_i; \mu_k, \sigma^2 \mathbf{I})}{\sum_{l=1}^K w_l \phi(x_i; \mu_l, \sigma^2 \mathbf{I})}$$

when $\sigma^2 \rightarrow 0$, $\gamma_{i,k} \rightarrow \begin{cases} 1 & \text{if } k = \arg \min_l \|x_i - \mu_l\| \\ 0 & \text{otherwise} \end{cases}$
so the algorithm reduces to K-means.

Key A. $\phi(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$

B. $N(0, 2)$, assuming x_1 & x_2 are uncorrelated.

C. $\prod_{i=1}^n f(x_i; \theta)$

D. $\hat{\mu} = \frac{1}{n} \sum x_i$, $\hat{\Sigma} = \frac{1}{n} \sum (x_i - \hat{\mu})(x_i - \hat{\mu})^T$

E. $l(\theta; \mathbf{x}) = \prod_{i=1}^n \left(\sum_{k=1}^K w_k \phi(x_i; \mu_k, \Sigma_k) \right)$

$$\log l(\theta; \mathbf{x}) = \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k \phi(x_i; \mu_k, \Sigma_k) \right)$$

F. $l(\theta; \mathbf{x}, \underline{\varepsilon}) = \prod_{i=1}^n w_{s_i} \phi(x_i; \mu_{s_i}, \Sigma_{s_i})$

$$= \prod_{k=1}^K w_k^{n_k} \cdot \prod_{k=1}^K \prod_{i \in I_k} \phi(x_i; \mu_k, \Sigma_k)$$

$$\log l(\theta; \mathbf{x}, \underline{\varepsilon}) = \sum_{k=1}^K n_k \log w_k$$

$$+ \sum_{k=1}^K \sum_{i \in I_k} \log \phi(x_i; \mu_k, \Sigma_k)$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in I_k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i \in I_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$\begin{aligned} G. \quad \gamma_{i,k}(\theta^{(j)}) &:= E[\Delta_{i,k} \mid \underline{x}; \theta^{(j)}] \\ &= \Pr \{ S_i = k \mid \underline{x}; \theta^{(j)} \} \end{aligned}$$

$$\theta^{(j+1)} = \arg \max_{\theta} Q(\theta, \theta^{(j)})$$