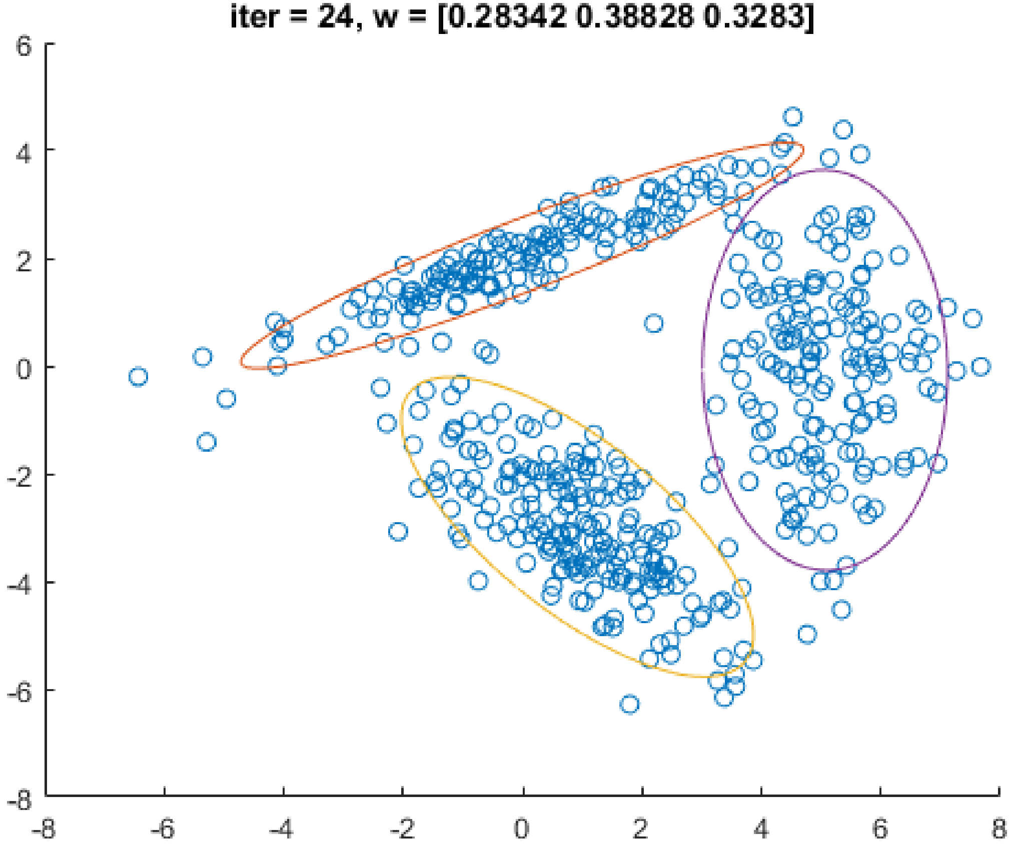
Latent Variable Models; The Expectation-Maximization Algorithm

Outline

- Review of Gaussian Mixture Models
- Latent Variable Models
- The Expectation-Maximization (EM) algorithm

Gaussian Mixture Models



Gaussian Mixture Models

- Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} > 0$.
- Recall the multivariate Gaussian density

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

- A random variable $\boldsymbol{X}$ follows a *Gaussian mixture model* (GMM) *with K components* if its probability density function f has the form

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

where $w_k \geq 0$, $\sum_k w_k = 1$, and for all k, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}, \boldsymbol{\Sigma}_k > 0$.

K known

Maximum Likelihood Estimation

- Observed data: $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
- The likelihood is

$$
\begin{aligned}
L(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) &:= \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \left(\sum_{k} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)
\end{aligned}
$$

  and the log-likelihood is

$$
\ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) := \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right).
$$

- When $K > 1$, there is no closed form solution.

Algorithm for Learning GMMs

Initialize $\boldsymbol{\theta}^{(0)}$, $j = 0$

Repeat

E-step:

$$\gamma_{i,k}^{(j)} = \frac{w_k^{(j)} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(j)}, \boldsymbol{\Sigma}_k^{(j)})}{\sum_{\ell=1}^{k} w_\ell^{(j)} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\ell^{(j)}, \boldsymbol{\Sigma}_\ell^{(j)})}$$

M-step:

$$\boldsymbol{\mu}_k^{(j+1)} = \frac{\sum_i \gamma_{i,k}^{(j)} \boldsymbol{x}_i}{\sum_i \gamma_{i,k}^{(j)}}$$

$$\boldsymbol{\Sigma}_k^{(j+1)} = \frac{\sum_i \gamma_{i,k}^{(j)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)})^T}{\sum_i \gamma_{i,k}^{(j)}}$$

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

$j = j + 1$

Until convergence criterion satisfied

Today's Main Message

- The previous algorithm is an instance of a more general algorithm called the Expectation-Maximization (EM) algorithm
- The EM algorithm is an iterative algorithm for maximum likelihood estimation for latent variable models
- A latent variable models is a probabilistic model for data, where each observation is explained by one or more latent, or unobserved, variables
- More concretely, in a LVM for an unsupervised learning problem, each observation $\boldsymbol{x}_i$ is explained by a latent variable $\boldsymbol{z}_i$

GMMs are LVMs

- A key to understanding GMMs is to know how to simulate a realization from a known GMM.
- Suppose
$$\boldsymbol{\theta} = (w_1, \ldots, w_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$$
is known.
- The following two-step procedure generates a realization of the GMM with parameter vector $\boldsymbol{\theta}$.
  - First, select $k \in \{1, \ldots, K\}$ at random, according to the pmf $w_1, \ldots, w_K$.
  - Then draw a realization of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Z_i = component from which x_i is drawn

$\in \{1, \ldots, K\}$

Bishop, Murphy

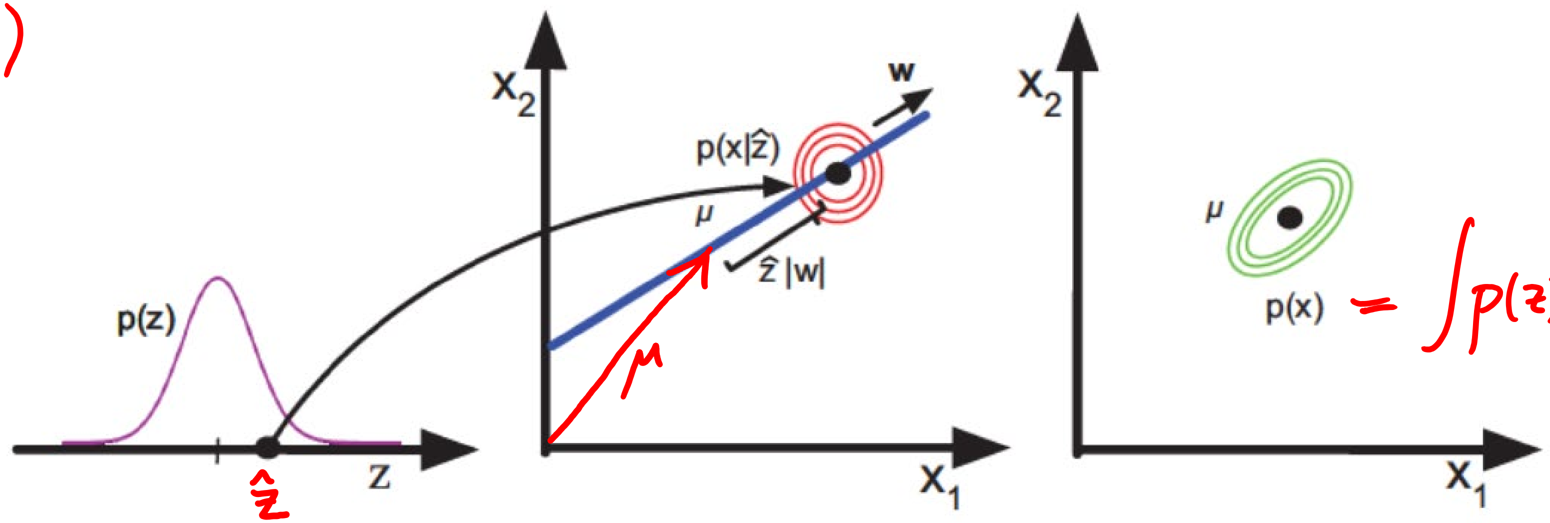
Probabilistic PCA

- Generative model for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$
- Assume each $\boldsymbol{x}_i$ is generated as follows:
  - $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $1 \le i \le n$
  - $\boldsymbol{x}_i \mid \boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, $1 \le i \le n$

where $\boldsymbol{z}_i \in \mathbb{R}^k$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{W} \in \mathbb{R}^{d\times k}$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\sigma^2 > 0$

$W = \begin{bmatrix} w \end{bmatrix} \in \mathbb{R}^{2\times 1}$

$p(z|x)$



$= \int p(z)p(x|z)\,dz$

Murphy **Figure 12.1** Illustration of the PPCA generative process, where we have ~~$L = 1$~~ $k=1$ latent dimension generating $d=2$ ~~$D = 2$~~ observed dimensions. Based on Figure 12.9 of (Bishop 2006b).

Latent Variable Models

- We will assume that every observation $\boldsymbol{X}_i$ is associated to an unobserved (or *hidden* or *latent*) variable $\boldsymbol{Z}_i$
- Let $\underline{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ denote all the latent variables.
- Note that the random variables $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are jointly distributed

Goal: $\max_{\theta} \underbrace{L(\theta\,;x_1,\ldots,x_n)}_{f(x_1,\ldots,x_n\,;\,\theta)}$

$\underline{z} = (z_1, \ldots, z_n)$

Complete Data

- We refer to $(\underline{\boldsymbol{x}}, \underline{\boldsymbol{z}})$ as the *complete data*
- The *complete-data likelihood* is

$$L(\theta; \underline{x}, \underline{z}) = f(\underline{x}, \underline{z}; \theta) = f(\underline{z}; \theta) \cdot f(\underline{x} \mid \underline{z}; \theta)$$

and the *complete-data log-likelihood* is

$$\ell(\theta; \underline{x}, \underline{z}) = \log L(\theta; \underline{x}, \underline{z})$$

want to maximize

- The basic idea behind the EM algorithm is to replace $\boxed{\ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}})}$ with

$$\mathbb{E}\left[\ell(\theta; \underline{X}, \underline{Z}) \mid \underline{X} = \underline{x}; \theta\right]$$

randomness due to $\underline{Z} \mid \underline{X} = \underline{x}$

- Since $\boldsymbol{\theta}$ is unknown, we use an estimate $\boldsymbol{\theta}^{(j)}$ to compute the expectation, and proceed iteratively

EM Algorithm

Initialize $\boldsymbol{\theta}^{(0)}$

$j \leftarrow 0$

Repeat

E-step: Compute

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j)}) = \mathbb{E}[\ell(\boldsymbol{\theta}; \underline{\boldsymbol{X}}, \underline{\boldsymbol{Z}}) \mid \underline{\boldsymbol{X}} = \underline{\boldsymbol{x}}; \boldsymbol{\theta}^{(j)}]$$

M-step: Solve

$$\boldsymbol{\theta}^{(j+1)} \leftarrow \mathrm{argmax}_{\boldsymbol{\theta}} \; Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j)})$$

$j \leftarrow j + 1$

Until convergence criterion satisfied

Poll

True or False: A reasonable termination criterion for the EM algorithm is to stop iterating when

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

for some small ϵ.

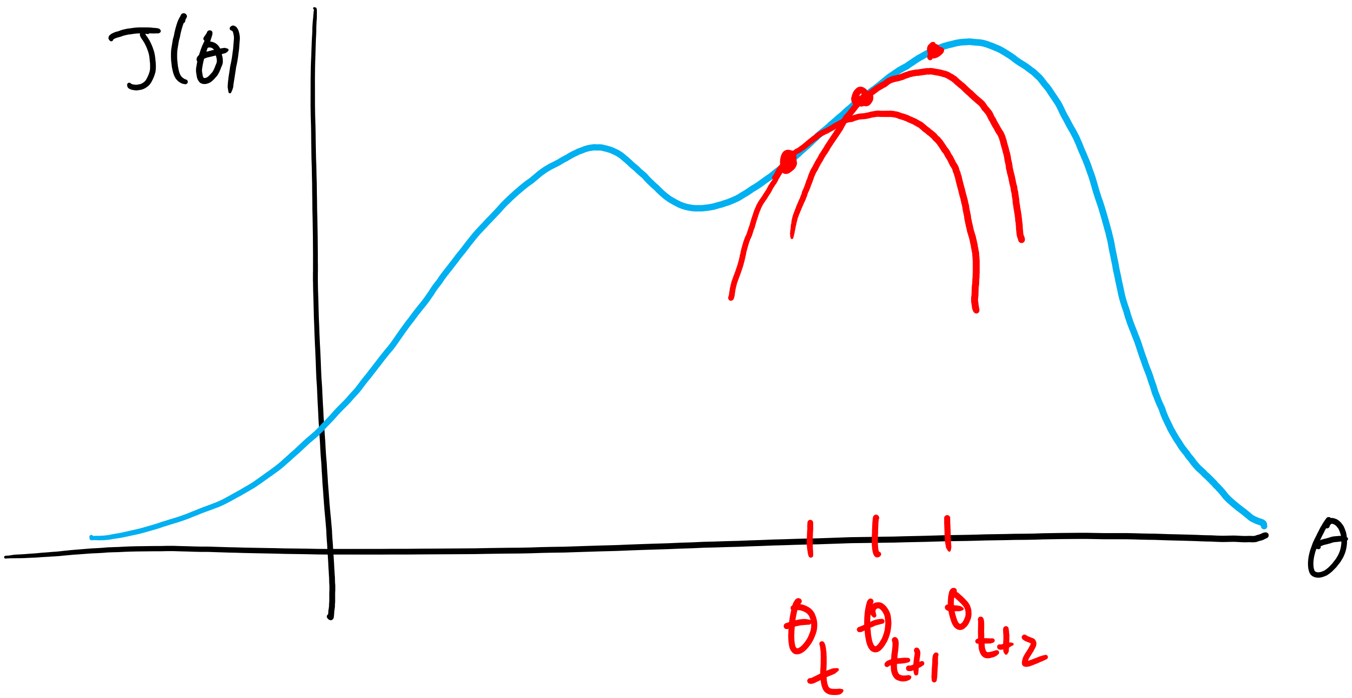
(A) True

(B) False

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

Minorize-Maximize Algorithms

$\max_{\theta} J(\theta)$



Minorize-Maximize Algorithms

$\theta_t \leftrightarrow \theta^{(j)}$

Suppose we wish to maximize the objective function $J(\boldsymbol{\theta})$

Initialize $\boldsymbol{\theta}_0$
$t \leftarrow 0$
Repeat
Minorize: Find a function $J_t(\boldsymbol{\theta})$ such that

$$\begin{aligned} J(\boldsymbol{\theta}_t) &= J_t(\boldsymbol{\theta}_t) \\ J(\boldsymbol{\theta}) &\geq J_t(\boldsymbol{\theta}) \qquad \forall \boldsymbol{\theta} \end{aligned}$$

Maximize: Solve

$$\boldsymbol{\theta}_{t+1} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} J_t(\boldsymbol{\theta})$$

$t \leftarrow t + 1$
Until convergence

Ascent Property of MM

- MM algorithms never decrease the objective: for all $t \geq 1$

$$J(\boldsymbol{\theta}_{t+1}) \geq J(\boldsymbol{\theta}_t)$$

$$J(\theta_t) = J_t(\theta_t) \leq J_t(\theta_{t+1}) \leq J(\theta_{t+1})$$

def of minorizing function

θ_{t+1} maximizes J_t

EM as Minorize-Maximize

- It can be shown that

$$J_t(\theta) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + \ell(\boldsymbol{\theta}_t; \underline{\boldsymbol{x}}) - Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t)$$

  minorizes the log-likelihood $\ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}})$. $= J(\theta)$

- Hence EM is an MM algorithm

- ~~Ascent property:~~ $\cancel{\ell(\boldsymbol{\theta}_0; \underline{\boldsymbol{x}}) \leq \ell(\boldsymbol{\theta}_1; \underline{\boldsymbol{x}}) \leq \ell(\boldsymbol{\theta}_2; \underline{\boldsymbol{x}}) \leq \cdots}$

$$J_t(\theta_t) = \cancel{Q(\theta_t, \theta_t)} + \ell(\theta_t; \underline{x}) - \cancel{Q(\theta_t, \theta_t)}$$

$$= \ell(\theta; \underline{x}) = J(\theta_t)$$

EM as Minorize-Maximize

Let p + q be two probability distributions

$$D_{KL}(p \| q) = \mathbb{E}_{Y \sim p}\left[\log\left(\frac{p(Y)}{q(Y)}\right)\right] \geq 0$$

Jensen's inequality

$$\begin{aligned} \ell(\theta; \underline{x}) - \ell(\theta_t; \underline{x}) &= Q(\theta; \theta_t) - Q(\theta_t, \theta_t) \\ &\quad + D_{KL}\left(p(z \mid \underline{x}; \theta_t) \| p(z \mid \underline{x}; \theta)\right) \\ &\geq Q(\theta, \theta_t) - Q(\theta_t, \theta_t) \end{aligned}$$

Poll

- True or False: The EM algorithm produces a sequence $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots$ satisfying

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

  for all $j \geq 0$.

  (A) True

  (B) False

EM for GMMs

- Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} > 0$.
- Recall the multivariate Gaussian density

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

- A random variable $\boldsymbol{X}$ follows a *Gaussian mixture model* (GMM) *with* K *components* if its probability density function f has the form

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

where $w_k \geq 0$, $\sum_k w_k = 1$, and for all k, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$, $\boldsymbol{\Sigma}_k > 0$.

K known

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

Complete-Data Log-Likelihood

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

$$\begin{aligned}
\ell(\theta; \underline{x}, \underline{z}) &= \log\left(f(\underline{z};\theta) \cdot f(\underline{x} \mid \underline{z};\theta)\right) \\
&= \log\left(\prod_{i=1}^{n} f(z_i;\theta)\, f(x_i \mid z_i;\theta)\right) \\
&= \log\left(\prod_{i=1}^{n} \Pr(Z_i = z_i;\theta)\, f(x_i \mid z_i;\theta)\right) \\
&= \log\left(\prod_{i=1}^{n} w_{z_i}\, \phi(x_i; \mu_{z_i}, \Sigma_{z_i})\right) \\
&= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \Delta_{i,k}\, w_k\, \phi(x_i; \mu_k, \Sigma_k)\right) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{i,k} \log\left(w_k\, \phi(x_i, \mu_k, \Sigma_k)\right)
\end{aligned}$$

E-Step for GMMs

- Denote the current iterate $\boldsymbol{\theta}^{(j)} = (w_1^{(j)}, \ldots, w_K^{(j)}, \boldsymbol{\mu}_1^{(j)}, \ldots, \boldsymbol{\mu}_K^{(j)}, \boldsymbol{\Sigma}_1^{(j)}, \ldots, \boldsymbol{\Sigma}_K^{(j)})$
- The E-step amounts to calculating, for all i, k,

$$\gamma_{i,k}^{(j)} = \mathbb{E}\left[\Delta_{i,k} \mid X_i = x_i \; ; \theta^{(j)}\right]$$

$$= \Pr\left\{\Delta_{i,k} = 1 \mid X_i = x_i \; ; \theta^{(j)}\right]$$

$$= \Pr\left\{Z_i = k \mid X_i = x_i \; ; \theta^{(j)}\right]$$

$$= \frac{\Pr\{Z_i = k \, ; \theta^{(j)}\} \cdot f(x_i \mid Z_i = k \, ; \theta^{(j)})}{f(x_i \, ; \theta^{(j)})}$$

$$= \frac{w_k^{(j)} \, \phi(x_i \, ; \mu_k^{(j)}, \Sigma_k^{(j)})}{\sum_{\ell=1}^{K} w_\ell^{(j)} \, \phi(x_i \, ; \mu_k^{(\ell)}, \Sigma_k^{(\ell)})}$$

M-Step for GMMs

- We need to compute

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

where

$\mathbb{E}[\Delta_{i,k} \mid X_i = x_i ; \theta^{(j)}]$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{i,k}^{(j)} \left[\log w_k - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right]$$

- The solution is

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

$$\boldsymbol{\mu}_k^{(j+1)} = \frac{\sum_i \gamma_{i,k}^{(j)} \boldsymbol{x}_i}{\sum_i \gamma_{i,k}^{(j)}}$$

$$\boldsymbol{\Sigma}_k^{(j+1)} = \frac{\sum_i \gamma_{i,k}^{(j)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)})^T}{\sum_i \gamma_{i,k}^{(j)}}$$

Summary: EM Algorithm for GMMs

Initialize $\boldsymbol{\theta}^{(0)}$, $j = 0$

Repeat

E-step:

$$\gamma_{i,k}^{(j)} = \frac{w_k^{(j)} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(j)}, \boldsymbol{\Sigma}_k^{(j)})}{\sum_{\ell=1}^{k} w_\ell^{(j)} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\ell^{(j)}, \boldsymbol{\Sigma}_\ell^{(j)})}$$

M-step:

$$\boldsymbol{\mu}_k^{(j+1)} = \frac{\sum_i \gamma_{i,k}^{(j)} \boldsymbol{x}_i}{\sum_i \gamma_{i,k}^{(j)}}$$

$$\boldsymbol{\Sigma}_k^{(j+1)} = \frac{\sum_i \gamma_{i,k}^{(j)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)})^T}{\sum_i \gamma_{i,k}^{(j)}}$$

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

$j = j + 1$

Until convergence criterion satisfied