Review: Gaussian-based classification

- **Conic decision boundaries in feature space**
- **Limited classification power (*Mathematica*)**
  - *Synthetic data examples*
  - *Object classification*

More expressive modeling approach: mixture modeling

Mixture modeling

Basic model:

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{k} p(\mathbf{x}|\phi_i)P(\omega_i)$$

where,

$p(\mathbf{x}|\phi_i) = N(\mathbf{x}, \mu_i, \Sigma_i) =$ Normal density

$P(\omega_i) =$ probability (weight) of ith Normal density

$\phi_i = \{\mu_i, \Sigma_i\}$

Given data, what needs to be estimated?

Mixture modeling: a look ahead

- **No closed form maximum-likelihood solution**
- **Iterative training (Expectation-Maximization)**
- **Random initialization of parameters.**

Some examples before the math (*Mathematica*)

- *Synthetic data examples*
- *Object classification*

Mixture modeling

Ok, but...

...how does the mixture-model estimation work?

(Why no closed-form solution?)

Simple two-parameter example

Example: 2-Gaussian mixture with known σ, and $P(\omega_i)$:

- $P(\omega_1) = 1/3$, $P(\omega_2) = 2/3$
- $\sigma_1 = \sigma_2 = 1$

Need to estimate $\{\mu_1, \mu_2\}$ for some data $\mathbf{X} = \{x_1, \ldots, x_n\}$

Simple two-parameter example

$$p(\mathbf{X}|\mu_1, \mu_2) = \prod_{j=1}^{n} [p(x_j|\mu_1)P(\omega_1) + p(x_j|\mu_2)P(\omega_2)]$$

$$\ln p(\mathbf{X}|\mu_1, \mu_2) = \sum_{j=1}^{n} \ln[p(x_j|\mu_1)P(\omega_1) + p(x_j|\mu_2)P(\omega_2)]$$

$$\ln p(\mathbf{X}|\mu_1, \mu_2) = \sum_{j=1}^{n} \ln\left\{\exp\left[-\frac{1}{2}(x_j - \mu_1)^2\right]P(\omega_1) + \exp\left[-\frac{1}{2}(x_j - \mu_2)^2\right]P(\omega_2)\right\} - n\ln\sqrt{2\pi}$$

Simple two-parameter example

Can this be simplified?

$$\ln p(\mathbf{X}|\mu_1, \mu_2) = \sum_{j=1}^{n} \ln\left\{\exp\left[-\frac{1}{2}(x_j - \mu_1)^2\right]P(\omega_1) + \exp\left[-\frac{1}{2}(x_j - \mu_2)^2\right]P(\omega_2)\right\} - n\ln\sqrt{2\pi}$$

Numeric example (*Mathematica*)

Maximum-likelihood estimation: mixture models

Conclusion: no closed-form solution to maximum-likelihood estimation for mixture models

Could use gradient-based algorithm; however, there is a better answer: Expectation-Maximization

From before: Maximum-likelihood parameter estimation for single Gaussian

For d-dimensional Gaussians:

$$\mu^* = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j$$

$$\Sigma^* = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j-\mu)(\mathbf{x}_j-\mu)^T$$

Familiar estimates of μ and Σ are maximum-likelihood estimates.

Understanding the EM algorithm

- Parameter estimation of mixture model would be no problem if...?
- Introduce hidden random variables y_{ij}:

$$y_{ij} = \begin{cases} 1 & \mathbf{x}_j \text{ came from class } \omega_i \\ 0 & \text{otherwise} \end{cases}$$

- Each data point has known and "hidden component."
- "Complete data" is given by:

$$\mathbf{z}_j = \{\mathbf{x}_j, y_{1j}, y_{2j}, \ldots, y_{kj}\}\ ,\ j \in \{1, 2, \ldots, n\}$$

Expressions for single Normal density

$$\mu_i^* = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_j^{(i)} = \left(\sum_{j=1}^{n} y_{ij}\mathbf{x}_j\right)\Big/\left(\sum_{j=1}^{n} y_{ij}\right)$$

$$\Sigma_i^* = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{x}_j^{(i)}-\mu_i)(\mathbf{x}_j^{(i)}-\mu_i)^T$$

$$= \left(\sum_{j=1}^{n} y_{ij}(\mathbf{x}_j-\mu_i)(\mathbf{x}_j-\mu_i)^T\right)\Big/\left(\sum_{j=1}^{n} y_{ij}\right)$$

$$P(\omega_i)^* = \frac{n_i}{n} = \left(\sum_{j=1}^{n} y_{ij}\right)\Big/ n$$

Very informal statement of EM

However, we don't know the unobserved y_{ij} variables ... EM to the rescue!

Expectation step: Calculate the expected value $E[y_{ij}|\Theta^{(t)}]$ for the hidden variables y_{ij}, given the current estimate of the parameters at step t ($\Theta^{(t)}$).

Maximization step: Calculate a new maximum-likelihood estimate for the parameters ($\mu_i^{(t+1)}$, $\Sigma_i^{(t+1)}$, $P(\omega_i)$) assuming that the value taken on by each hidden variable y_{ij} is its expected value $E[y_{ij}|\Theta^{(t)}]$ (calculated in the Expectation step). Then, iterate.

Expectation-Maximization algorithm

$$\mu_i^{(t+1)} = \left(\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]\mathbf{x}_j\right) \Big/ \left(\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]\right)$$

$$\Sigma_i^{(t+1)} = \frac{\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}](\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T}{\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]}$$

$$P(\omega_i)^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]$$

Let's see how this works...

(Mathematica)

What is $E[y_{ij}|\Theta^{(t)}]$?

By definition:

$$E[y_{ij}|\Theta^{(t)}] = \sum_{x=0}^{1} xP(y_{ij} = x|\Theta^{(t)})$$

So:

$$E[y_{ij}|\Theta^{(t)}] = 0 \times P(y_{ij} = 0|\Theta^{(t)}) + 1 \times P(y_{ij} = 1|\Theta^{(t)})$$

$$E[y_{ij}|\Theta^{(t)}] = P(y_{ij} = 1|\Theta^{(t)})$$

$$E[y_{ij}|\Theta^{(t)}] = P(\omega_i|\mathbf{x}_j, \Theta^{(t)})$$

So what is $P(\omega_i|\mathbf{x}_j, \Theta^{(t)})$?

$$P(\omega_i|\mathbf{x}_j, \Theta^{(t)}) = \frac{p(\mathbf{x}_j|\omega_i, \Theta^{(t)})P(\omega_i|\Theta^{(t)})}{p(\mathbf{x}_j|\Theta^{(t)})}$$ ***(Bayes Rule)***

where:

$$p(\mathbf{x}_j|\omega_i, \Theta^{(t)}) = N[\mathbf{x}_j, \mu_i^{(t)}, \Sigma_i^{(t)}], \quad P(\omega_i|\Theta^{(t)}) = P(\omega_i)^{(t)}$$

$$p(\mathbf{x}_j|\Theta^{(t)}) = \sum_{i=1}^{k} p(\mathbf{x}_j|\omega_i, \Theta^{(t)})P(\omega_i|\Theta^{(t)})$$

$$= \sum_{i=1}^{k} N[\mathbf{x}_j, \mu_i^{(t)}, \Sigma_i^{(t)}]P(\omega_i)^{(t)}$$

Expectation-Maximization algorithm

$$\mu_i^{(t+1)} = \left(\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]\mathbf{x}_j\right) \Big/ \left(\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]\right)$$

$$\Sigma_i^{(t+1)} = \frac{\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}](\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T}{\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]}$$

$$P(\omega_i)^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} E[y_{ij}|\Theta^{(t)}]$$

Expected value of hidden variables

$$E[y_{ij}|\Theta^{(t)}] = \frac{p(\mathbf{x}_j|\omega_i, \Theta^{(t)})P(\omega_i|\Theta^{(t)})}{p(\mathbf{x}_j|\Theta^{(t)})}$$

where:

$$p(\mathbf{x}_j|\omega_i, \Theta^{(t)}) = N[\mathbf{x}_j, \mu_i^{(t)}, \Sigma_i^{(t)}], \; P(\omega_i|\Theta^{(t)}) = P(\omega_i)^{(t)}$$

$$p(\mathbf{x}_j|\Theta^{(t)}) = \sum_{i=1}^{k} p(\mathbf{x}_j|\omega_i, \Theta^{(t)})P(\omega_i|\Theta^{(t)})$$

$$= \sum_{i=1}^{k} N[\mathbf{x}_j, \mu_i^{(t)}, \Sigma_i^{(t)}]P(\omega_i)^{(t)}$$

Informal definition of EM

1. **Choose an initial estimate for Θ.**
2. **(E)xpectation step: Calculate the expected value $E[y_{ij}|\Theta]$ for the hidden variables y_{ij}, given the current estimate of the parameters (Θ).**
3. **(M)aximization step: Calculate a new maximum-likelihood estimate for the parameters ($\bar{\Theta}$) assuming that the value taken on by each hidden variable y_{ij} is its expected value $E[y_{ij}|\Theta]$ (calculated in the Expectation step).**
4. **Iterate Expectation and Maximization steps.**

Formal problem statement for EM algorithm

Definitions:

- Let $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$ denote n *observed* (incomplete) data vectors.
- Let $\mathbf{Y} = \{\mathbf{y}_j\}$, $j \in \{1, 2, \ldots, n\}$ denote the corresponding n *unobserved (hidden)* data vectors.
- Let $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ denote the *full (or complete)* data where,

$$\mathbf{z}_j = \{\mathbf{x}_j, \mathbf{y}_j\}$$

Goal: Maximize $L(\mathbf{X}, \Theta)$ with respect to Θ such that,

$$L(\mathbf{X}, \Theta) \le L(\mathbf{X}, \Theta^*), \; \forall\Theta.$$

The EM algorithm (huh?)

1. **Choose an initial estimate for Θ.**
2. **(E)xpectation step: Compute $Q(\Theta, \bar{\Theta})$,**

$$Q(\Theta, \bar{\Theta}) = E[L(\mathbf{Z}|\bar{\Theta}) | \mathbf{X}, \Theta]\text{, where,}$$

$$L(\mathbf{Z}|\bar{\Theta}) = \sum_{j=1}^{n} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}).$$

3. **(M)aximization step: Replace the current estimate Θ with the new estimate $\bar{\Theta}$ where,**

$$\bar{\Theta} = \underset{\bar{\Theta}}{\operatorname{argmax}} \; Q(\Theta, \bar{\Theta})$$

4. **Iterate steps 2 and 3 until convergence.**

Some discussion on the EM algorithm

- **Huh?**
- **Important property: maximizing $Q(\Theta, \bar{\Theta})$ maximizes $L(\mathbf{X}, \bar{\Theta})$**
- **Proof of result for mixture modeling is non-trivial.**

Some key points

Question: How are $Q(\Theta, \bar{\Theta})$ and $L(\mathbf{X}, \bar{\Theta})$ related?

$$L(\mathbf{X}|\Theta) \equiv \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\Theta)$$

Let's express $p(\mathbf{x}|\Theta)$ in terms of $\mathbf{x}$ and $\mathbf{y}$:

$$p(\mathbf{y}|\mathbf{x}) \equiv \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} \quad \Rightarrow \quad p(\mathbf{y}|\mathbf{x}, \Theta) = \frac{p(\mathbf{x}, \mathbf{y}|\Theta)}{p(\mathbf{x}|\Theta)}$$

$$p(\mathbf{x}|\Theta) = \frac{p(\mathbf{x}, \mathbf{y}|\Theta)}{p(\mathbf{y}|\mathbf{x}, \Theta)}$$

$$\ln p(\mathbf{x}|\Theta) = \ln p(\mathbf{x}, \mathbf{y}|\Theta) - \ln p(\mathbf{y}|\mathbf{x}, \Theta)$$

How are $Q(\Theta, \bar{\Theta})$ and $L(\mathbf{X}, \bar{\Theta})$ related?

$$\ln p(\mathbf{x}|\Theta) = \ln p(\mathbf{x}, \mathbf{y}|\Theta) - \ln p(\mathbf{y}|\mathbf{x}, \Theta)$$

$$\ln p(\mathbf{x}|\Theta) = \ln p(\mathbf{z}|\Theta) - \ln p(\mathbf{y}|\mathbf{x}, \Theta)$$

$$\sum_{j=1}^{n} \ln p(\mathbf{x}_j|\Theta) = \sum_{j=1}^{n} \ln p(\mathbf{z}_j|\Theta) - \sum_{j=1}^{n} \ln p(\mathbf{y}_j|\mathbf{x}_j, \Theta)$$

$$L(\mathbf{X}|\Theta) = L(\mathbf{Z}|\Theta) - L(\mathbf{Y}|\mathbf{X}, \Theta)$$

Change Θ to $\bar{\Theta}$ and apply $E[\;\bullet\;|\mathbf{X}, \Theta]$:

$$L(\mathbf{X}|\bar{\Theta}) = L(\mathbf{Z}|\bar{\Theta}) - L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})$$

$$E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] = E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta] - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

How are $Q(\Theta, \bar{\Theta})$ and $L(X, \bar{\Theta})$ related?

$$E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] = E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta] - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

Note:

$$Q(\Theta, \bar{\Theta}) \equiv E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta]$$

so that:

$$E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] = Q(\Theta, \bar{\Theta}) - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

How are $Q(\Theta, \bar{\Theta})$ and $L(X, \bar{\Theta})$ related?

$$E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] = Q(\Theta, \bar{\Theta}) - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

What is $E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta]$?

- $L(\mathbf{X}|\bar{\Theta})$ is independent of Θ (constant for given $\mathbf{X}$, $\bar{\Theta}$)
- $E[L(\mathbf{X}|\bar{\Theta})|\mathbf{X}, \Theta] = L(\mathbf{X}|\bar{\Theta})$

So:

$$L(\mathbf{X}|\bar{\Theta}) = Q(\Theta, \bar{\Theta}) - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

How are $Q(\Theta, \bar{\Theta})$ and $L(X, \bar{\Theta})$ related?

$$L(\mathbf{X}|\bar{\Theta}) = Q(\Theta, \bar{\Theta}) - E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

Define:

$$H(\Theta, \bar{\Theta}) \equiv E[L(\mathbf{Y}|\mathbf{X}, \bar{\Theta})|\mathbf{X}, \Theta]$$

$$L(\mathbf{X}|\bar{\Theta}) = Q(\Theta, \bar{\Theta}) - H(\Theta, \bar{\Theta})$$

Can be shown that (see notes):

$$H(\Theta, \bar{\Theta}) \leq H(\Theta, \Theta)$$

Therefore:

$Q(\Theta, \bar{\Theta}) \geq Q(\Theta, \Theta)$ implies $L(\mathbf{X}|\bar{\Theta}) \geq L(\mathbf{X}|\Theta)$.

Expectation step: expression for $Q(\Theta, \bar{\Theta})$

By definition:

$$Q(\Theta, \bar{\Theta}) \equiv E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta]$$

What does $E[\ \bullet\ |\mathbf{X}, \Theta]$ mean?

Simpler example:

- $f(\mathbf{x}, \mathbf{y})$, $\mathbf{x}$ = known constant, $\mathbf{y}$ = random variable(s)
- $E[f(\mathbf{x}, \mathbf{y})] \equiv \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})p(\mathbf{y})$ (expected value)
- $E[f(\mathbf{x}, \mathbf{y})|\mathbf{x}, \Theta] \equiv \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})p(\mathbf{y}|\mathbf{x}, \Theta)$ (conditional expected value)

Expression for $Q(\Theta, \bar{\Theta})$

By definition:

$$Q(\Theta, \bar{\Theta}) \equiv E[L(\mathbf{Z}|\bar{\Theta})|\mathbf{X}, \Theta]$$

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \left(\sum_{\mathbf{y}_j} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}) p(\mathbf{y}_j|\mathbf{x}_j, \Theta) \right).$$

$$p(\mathbf{y}_j|\mathbf{x}_j, \Theta) = \frac{p(\mathbf{x}_j, \mathbf{y}_j|\Theta)}{p(\mathbf{x}_j|\Theta)}$$

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{\mathbf{y}_j} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}) \frac{p(\mathbf{x}_j, \mathbf{y}_j|\Theta)}{p(\mathbf{x}_j|\Theta)}$$

Expression for $Q(\Theta, \bar{\Theta})$

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{\mathbf{y}_j} \ln p(\mathbf{x}_j, \mathbf{y}_j|\bar{\Theta}) \frac{p(\mathbf{x}_j, \mathbf{y}_j|\Theta)}{p(\mathbf{x}_j|\Theta)}$$

For mixture model:

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{k} \ln[p(\mathbf{x}_j|\bar{\phi}_i)\overline{P(\omega_i)}] \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}$$

(why?)

Expression for $Q(\Theta, \bar{\Theta})$ (mixture model)

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{k} \ln[p(\mathbf{x}_j|\bar{\phi}_i)\overline{P(\omega_i)}] \frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)}$$

Note:

$$\frac{p(\mathbf{x}_j|\phi_i)P(\omega_i)}{p(\mathbf{x}_j|\Theta)} = P(\omega_i|\mathbf{x}_j, \Theta)$$

so that:

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i|\mathbf{x}_j, \Theta) \ln[p(\mathbf{x}_j|\bar{\phi}_i)\overline{P(\omega_i)}]$$

Expectation step (mixture model)

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i|\mathbf{x}_j, \Theta) \ln[p(\mathbf{x}_j|\bar{\phi}_i)\overline{P(\omega_i)}]$$

Final form of Q-function:

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i|\mathbf{x}_j, \Theta) \ln \overline{P(\omega_i)} + \sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i|\mathbf{x}_j, \Theta) \ln p(\mathbf{x}_j|\bar{\phi}_i)$$

Maximization step (mixture model)

To maximize w/respect to $\bar{\Theta}$:

$$\nabla_{\bar{\Theta}} Q(\Theta, \bar{\Theta}) = 0$$

Two terms:

$$\sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i | \mathbf{x}_j, \Theta) \ln \overline{P(\omega_i)}$$

$\Rightarrow$ constrained Lagrangian optimization:

$$\sum_{i=1}^{k} \overline{P(\omega_i)} = 1$$

Maximization step (mixture model)

Second term:

$$\sum_{j=1}^{n} \sum_{l=1}^{k} P(\omega_l | \mathbf{x}_j, \Theta) \ln p(\mathbf{x}_j | \bar{\phi}_l)$$

$\Rightarrow$ set gradient w/respect to $\bar{\phi}_i$ to zero and solve:

$$\nabla_{\bar{\phi}_i} \sum_{j=1}^{n} \sum_{l=1}^{k} P(\omega_l | \mathbf{x}_j, \Theta) \ln p(\mathbf{x}_j | \bar{\phi}_l) = 0$$

$$\sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta) \frac{\nabla_{\bar{\phi}_i} p(\mathbf{x}_j | \bar{\phi}_i)}{p(\mathbf{x}_j | \bar{\phi}_i)} = 0$$

Maximization step (mixture model)

$$\sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta) \frac{\nabla_{\bar{\phi}_i} p(\mathbf{x}_j | \bar{\phi}_i)}{p(\mathbf{x}_j | \bar{\phi}_i)} = 0$$

For exponential $p(\mathbf{x}_j | \bar{\phi}_i)$:

$$\nabla_{\bar{\phi}_i} p(\mathbf{x}_j | \bar{\phi}_i) \Rightarrow f(\mathbf{x}_j, \bar{\phi}_i) p(\mathbf{x}_j | \bar{\phi}_i)$$

so that:

$$\sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta) f(\mathbf{x}_j, \bar{\phi}_i) = 0$$

This property typically allows us to solve for $\bar{\phi}_i$ ...

Final solution of EM update equations for mixture models

many equations later...

$$\overline{P(\omega_i)} = \frac{1}{n} \sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta)$$

$$\bar{\mu}_i = \left(\sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta) \mathbf{x}_j \right) \Big/ \left(\sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta) \right)$$

$$\bar{\Sigma}_i = \sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta)(\mathbf{x}_j - \bar{\mu}_i)(\mathbf{x}_j - \bar{\mu}_i)^T \Big/ \sum_{j=1}^{n} P(\omega_i | \mathbf{x}_j, \Theta)$$

EM Summary

- **General convergence proof**

$$Q(\Theta, \bar{\Theta}) \geq Q(\Theta, \Theta) \text{ implies } L(\mathbf{X}|\bar{\Theta}) \geq L(\mathbf{X}|\Theta) .$$

- **Expressions for $Q(\Theta, \bar{\Theta})$ (mixture models):**

$$Q(\Theta, \bar{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i|\mathbf{x}_j, \Theta) \ln \overline{P(\omega_i)} +$$
$$\sum_{j=1}^{n} \sum_{i=1}^{k} P(\omega_i|\mathbf{x}_j, \Theta) \ln p(\mathbf{x}_j|\bar{\phi}_i)$$

EM concluding thoughts...

Powerful algorithm for training complex statistical models:

- Mixture models (Gaussian, non-Gaussian example in new notes)
- Vector quantization (next)
- Hidden Markov models (temporal mixture models)
- Bayesian belief networks, Markov tree models, etc...

Applications:

- Computer vision (object detection/recognition)
- Speech recognition
- Gesture recognition, human performance analysis, etc....