Estimate Mixture of $K$ Gaussians

- Pick Gaussian $k$ with prob. $\pi_k$ and sample from that Gaussian.

- Generative Distribution, $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$

- A mixture distribution with mixture coefficients $\pi$

- For iid sample $X$ and parameters $\theta = \{\pi, \mu, \Sigma\}$, we have:

$$p(X \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} p(x_n \mid \pi, \mu, \Sigma)$$

$$= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$
log-likelihood: \[ \mathcal{L}(\pi, \mu, \Sigma) = \ln p(X | \pi, \mu, \Sigma) \]

\[ = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right) \]

- Which Gaussian picked for each \( x_i \in X \) is a latent variable \( z_i \) in \( \{0, 1\}^K \)
- Want to maximize \( \ln(p(X | \theta)) = \ln \left( \sum_{Z} p(X, Z | \theta) \right) \)
- Sum inside \( \ln() \) is a big problem – if “know” \( Z \), then sum disappears, loglikelihood becomes just \( \ln p(X, Z | \theta) \)
- **Complete data** is \( \{X, Z\}; \) **incomplete data** is just \( X \)
- Don’t know \( Z \), but from \( \theta^{old} \) can get \( p(Z | X, \theta^{old}) \)
- EM key Idea: Approximate $\ln(p(X, Z | \theta))$ for known $Z$ with its expectation over (the unknown) $Z$, estimate goodness of $\theta$ with

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z | X, \theta^{\text{old}}) \ln p(X, Z | \theta)$$

- View $Q$ as function of $\theta$ since $\theta^{\text{old}}, X$ are fixed (and $Z$ summed over)
- The E-step computes $p(Z | X, \theta^{\text{old}})$
- In the M-step,

$$\theta^{\text{new}} = \arg \max_\theta Q(\theta, \theta^{\text{old}})$$
Red: log-likelihood  
Blue: $Q(\theta, \theta^{old}) + \text{const}$

With more math (9.4), $Q(\theta, \theta^{old})$ curves tangent and below log-likelihood.
Back to Gaussian Mixtures:
Need $\ln p(X, Z \mid \mu, \Sigma, \pi)$ and $p(Z \mid X, \mu, \Sigma, \pi)$ for $Q()$ function

$$p(X, Z \mid \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left( \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)^{z_{n,k}}$$

$$\ln p(X, Z \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \left( \ln(\pi_k) + \ln \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)$$

Now logarithm is on the inside and sum on the outside!

Note that $z_{n,k}$ is either 0 or 1; add in term only when $z_{n,k} = 1$
Now for $p(\mathbf{Z} \mid \mathbf{X}, \mu, \Sigma, \pi)$:

$$p(\mathbf{Z} \mid \mathbf{X}, \mu, \Sigma, \pi) = \prod_{n=1}^{N} p(z_n \mid x_n, \mu, \Sigma, \pi)$$

$$p(z_n \mid x_n, \mu, \Sigma, \pi) = \frac{p(x_n \mid z_n, \mu, \Sigma, \pi)p(z_n \mid \mu, \Sigma, \pi)}{p(x_n \mid \mu, \Sigma, \pi)}$$

$$p(z_{n,k} = 1 \mid x_n, \mu, \Sigma, \pi) = \frac{p(x_n \mid \mu_k, \Sigma_k)\pi_k}{p(x_n \mid \mu, \Sigma, \pi)}$$

$$p(z_{n,k} = 1 \mid x_n, \mu, \Sigma, \pi) = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\text{normalization}}$$

$$p(z_{n,k} = 1 \mid x_n, \mu, \Sigma, \pi) = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} \equiv \gamma_{n,k}$$

- $\gamma_{n,k}$ is the relative responsibility of Gaussian $k$ for $x_n$. $\gamma(z_{n,k})$ in Bishop
Back to $Q$: (recall $\theta$ is $\{\mu, \Sigma, \pi\}$)

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z \mid X, \theta^{\text{old}}) \ln p(X, Z \mid \theta)$$

$$= \sum_Z p(Z \mid X, \theta^{\text{old}}) \sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \left( \ln(\pi_k) + \ln \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)$$

included when $z_{n,k} = 1$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{n,k} = 1 \mid X, \theta^{\text{old}}) \left( \ln(\pi_k) + \ln \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{n,k} \left( \ln(\pi_k) + \ln \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)$$

maximize this to get new parameters $\mu, \Sigma, \pi$

$\gamma_{n,k}$ responsibilities with respect to old parameters
For $\pi$ parameters, maximize:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{n,k} \ln(\pi_k) + \lambda \left(1 - \sum_{j=1}^{N} \pi_j\right)$$

Add Lagrange multiplier for $\sum_{k=1}^{K} \pi_k = 1$, then derivatives (wrt $\pi_k$’s) give:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,k},$$

The average relative responsibility of Gaussian $k$ for the examples.
- Find $\mu$, $\Sigma$ maximizing:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{n,k} \ln \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

- For each $k$, find $\mu_k$, $\Sigma_k$, maximizing

$$\sum_{n=1}^{N} \gamma_{n,k} \ln \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

- Fit a max-likelihood Gaussian to a $\gamma_{n,k}$-weighted data set,

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{n,k} x_n}{\sum_{n=1}^{N} \gamma_{n,k}}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{n,k} (x_n - \mu_k)^T (x_n - \mu_k)}{\sum_{n=1}^{N} \gamma_{n,k}}$$
EM for Gaussian Mixtures

Initialize means $\mu_k$, covariances $\Sigma_k$, mixture coefficients $\pi_k$ for each $1 \leq k \leq K$, and repeat the following until convergence:

1. **E-step**: Calculate responsibilities of each Gaussian for each data point (the expectation of the hidden variables):

   $$\gamma_{n,k} = E(z_{n,k} \mid X, \mu, \Sigma, \pi) = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

2. **M-step**: Re-estimate parameters using $\gamma_{n,k}$’s

   $$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,k}$$

   $$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{n,k} x_n}{\sum_{n=1}^{N} \gamma_{n,k}}$$

   $$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{n,k} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma_{n,k}}$$
EM for mixtures of Bernoulli distributions

- $D$ binary $\{0, 1\}$ features per example
- Basic Distribution:
  1. Features independent
  2. Each feature $i$ has prob $\mu_i$ of being 1
  3. $p(x | \mu) = \prod_{i=1}^{D} \mu_i^{x_i}(1 - \mu_i)^{1-x_i}$
  4. Max likelihood estimate $\hat{\mu}_i$ is fraction of data with $x_i = 1$
- Consider mixture of $K$ basic distributions:

  $$p(x | \mu, \pi) = \sum_{k=1}^{K} \pi_k p(x | \mu_k)$$

  $$p(x | \mu_k) = \prod_{i=1}^{D} \mu_{k,i}^{x_i}(1 - \mu_{k,i})^{1-x_i}$$

- Mixture can capture correlations between features
Given data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,

$$
\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n | \mu_k) \right)
$$

Again, sum inside $\ln$

Use latent $\mathbf{z}_n$ for each $\mathbf{x}_n$ indicating the basic distribution used for $\mathbf{x}_n$:

$$
p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}_n | \mu_k)^{z_{n,k}}
$$

$$
p(\mathbf{z}_n | \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{k}}
$$
\[
p(X, Z \mid \mu, \pi) = p(X \mid Z, \mu, \pi)p(Z \mid \mu, \pi)\\
= \left( \prod_{n=1}^{N} \prod_{k=1}^{K} p(x_n \mid \mu_k)^{z_{n,k}} \right) \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{k,n}}
\]

- Complete log-likelihood, \( p(X, Z \mid \mu, \pi) \) is:

\[
\sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \left( \ln \pi_k + \sum_{i=1}^{D} (x_{n,i} \ln \mu_{k,i} + (1 - x_{n,i}) \ln (1 - \mu_{k,i})) \right)
\]

(Recall \( p(x_n \mid \mu_k) = \prod_{i=1}^{D} \mu_{k,i}^{x_{n,i}} (1 - \mu_{k,i})^{1-x_{n,i}} \))

- Expectation wrt \( Z \) of complete log-likelihood is

\[
\sum_{n=1}^{N} \sum_{k=1}^{K} E(z_{n,k} \mid X, \mu, \pi) \left( \ln \pi_k + \sum_{i=1}^{D} (x_{n,i} \ln \mu_{k,i} + (1 - x_{n,i}) \ln (1 - \mu_{k,i})) \right)
\]
\[
p(\mathbf{Z} \mid \mathbf{X}, \mu, \pi) = \prod_{n=1}^{N} p(z_n \mid x_n, \mu, \pi)
\]

\[
p(z_n \mid x_n, \mu, \pi) = \frac{p(x_n \mid z_n, \mu, \pi)p(z_n \mid \mu, \pi)}{p(x_n \mid \mu, \pi)}
\]

\[
p(z_{n,k} = 1 \mid x_n, \mu, \pi) = \frac{p(x_n \mid \mu_k, \pi_k)}{p(x_n \mid \mu, \pi)}
\]

\[
p(z_{n,k} = 1 \mid x_n, \mu, \pi) = \frac{\pi_k p(x_n \mid \mu_k)}{\sum_{j=1}^{K} \pi_j p(x_n \mid \mu_j)} \equiv \gamma_{n,k}
\]

Again, \(\gamma_{n,k}\) is responsibility of \(k\)th distribution for \(x_n\).
Thus have $\gamma$’s and maximize expected log-likelihood

$$
\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{n,k} \left( \ln \pi_k + \sum_{i=1}^{D} (x_{n,i} \ln \mu_{k,i} + (1 - x_{n,i}) \ln (1 - \mu_{k,i})) \right)
$$

wrt $\pi$ and $\mu$.

Maximizing mixture $\pi$ is:

$$
\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,k}
$$

Maximizing means are:

$$
\mu_k = \frac{\sum_{n=1}^{N} \gamma_{n,k} x_n}{\sum_{n=1}^{N} \gamma_{n,k}}
$$