EM Algorithm:

$X = \{x_i\}$ $\quad$ $Y = \{y_i\}$. latent

Note $q(Y) = \prod_i q_i(y_i)$

$q_i(y_i) = P(y_i | x_i, \theta)$

Log-likelihood:

$$\ell(\theta) = \log P(X|\theta) = \log \int P(X, Y|\theta)\, dY$$

$$= \log \int \frac{P(X,Y|\theta)}{q(Y)} \cdot q(Y)\, dY$$

$$\geq \int \log\left(\frac{P(X,Y|\theta)}{q(Y)}\right) \cdot q(Y)\, dY \equiv \mathcal{F}(q, \theta)$$

Jensen's Inequality. (Appendix)

Definition: Free Energy

$$\mathcal{F}(q(Y), \theta) = \int \log\left(\frac{P(X,Y|\theta)}{q(Y)}\right) \cdot q(Y)\, dY = \left\langle \log \frac{P(X,Y|\theta)}{q(Y)} \right\rangle_q$$

Hence: $\log P(X|\theta) \geq \mathcal{F}(q(Y), \theta) = \left\langle \log \frac{P(X,Y|\theta)}{q(Y)} \right\rangle_q$

① $\mathcal{F}(q(Y)|\theta) = \langle \log P(X,Y|\theta) \rangle_q + \mathcal{H}(q)$ The entropy of $q(Y)$

proof: $\mathcal{F}(q(Y)|\theta) = \left\langle \log \frac{P(X,Y|\theta)}{q(Y)} \right\rangle_q = \langle \log P(X,Y|\theta) \rangle - \underbrace{\langle \log q(Y) \rangle_q}_{= \mathcal{H}(q(Y)) = \left\langle \log \frac{1}{q(Y)} \right\rangle_q}$

② $\mathcal{F}(q(Y)|\theta) = \underbrace{\log P(X|\theta)}_{\ell(\theta)} - KL[q(Y) \| P(Y|X,\theta)]$

proof: $\mathcal{F}(q(Y)|\theta) = \left\langle \log \frac{P(X,Y|\theta)}{q(Y)} \right\rangle_q = \int \log\left(\frac{P(Y|X,\theta) \cdot P(X|\theta)}{q(Y)}\right) q(Y)\, dY$

$$= \int \log(P(X|\theta)) \cdot q(Y) \cdot dY + \int \log\left(\frac{P(Y|X,\theta)}{q(Y)}\right) q(Y)\, dY$$

$$= \log P(X|\theta) - KL[q(Y) \| P(Y|X,\theta)]$$

$$\leq \log P(X|\theta) \quad \text{as } KL[q(Y) \| P(Y|X,\theta)] \geq 0.$$

The lower bound for $\mathcal{F}(q(Y)|\theta)$ is $\ell(\theta)$ (See Appendix)

EM Steps:

E-step: $q^{(k)}(Y) := \underset{q(Y)}{\operatorname{argmax}}\, \mathcal{F}(q^{(k-1)}(Y), \theta^{(k-1)})$

$= \ell(\theta^{(k-1)}) - \underset{q(Y)}{\operatorname{argmin}} \{ KL[q(Y) \| P(Y|X,\theta^{(k-1)})] \}$

$= \ell(\theta^{(k-1)})$ $\quad = 0$ by choosing: $q(Y) = P(Y|X,\theta)$.

"maximise $\mathcal{F}(q(Y),\theta)$ with θ fixed w.r.t. latent variables Y"

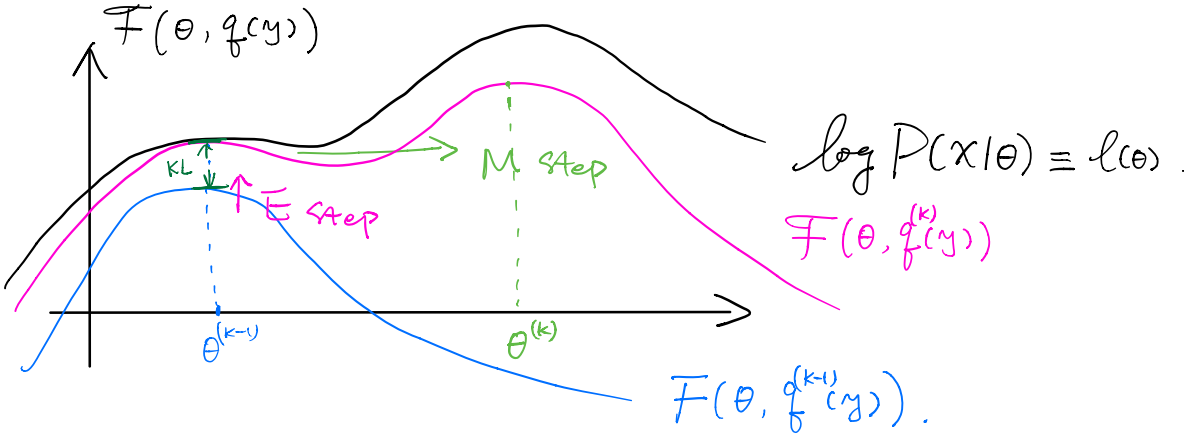
"After E step, $\mathcal{F}(q,\theta) = \ell(\theta) \Leftrightarrow$ maximum of free energy is maximum of likelihood".

M-step: $\theta^{(k)} := \underset{\theta}{\operatorname{argmax}}\, \mathcal{F}(q^{(k)}(Y), \theta^{(k-1)})$

$= \underset{\theta}{\operatorname{argmax}} \langle \log P(Y,X|\theta) \rangle_{q^{(k)}}$ $\quad + \mathcal{H}(q^{(k)}(Y))$

Ignore, as $\mathcal{H}$ does not depend on θ

"maximise $\mathcal{F}(q(Y),\theta)$ with latent Y fixed w.r.t. θ"



M step moves to a larger point at $\mathcal{F}(\theta, q^{(k)}(y))$

* E and M steps never decreases the log likelihood.

$$\ell(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)})$$

E-step (↓) M-step (↓) Jensen (↓)

Mixture of Gaussian

Data: $\underset{\sim}{X} = \{\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_N\}$, Latent process: $s_i \overset{iid}{\sim} \text{Disc}(\underset{\sim}{\pi})$.

Component distributions:

$$\underset{\sim}{x}_i \mid s_i = m \sim \mathcal{N}(\underset{\sim}{\mu}_m, \underline{\underline{\Sigma}}_m)$$

Marginal distribution:

$$P(\underset{\sim}{x}_i) = \sum_{m=1}^{K} \pi_m P_m(\underset{\sim}{x}; \theta_m).$$

Log-likelihood:

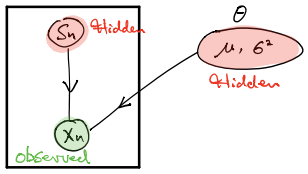
$$\ell(\{\underset{\sim}{\mu}_m\}, \{\underline{\underline{\Sigma}}_m\}, \underset{\sim}{\pi}) = \sum_{i=1}^{n} \log \sum_{m=1}^{K} \frac{\pi_m}{\sqrt{|2\pi \underline{\underline{\Sigma}}_m|}} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}_i - \underset{\sim}{\mu}_m)^T \underline{\underline{\Sigma}}_m^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_m)\right\}$$

1. Univariate Case. (Single Gaussian)

Density of a data point x:

$$p(x|\theta) = \sum_{m=1}^{K} \underbrace{P(x|s=m,\theta)}_{\propto \frac{1}{\sigma_m}\exp\left\{-\frac{1}{2\sigma_m^2}(x-\mu_m)^2\right\}} \cdot \underbrace{P(s=m|\theta)}_{\pi_m} \propto \frac{1}{\sigma_m}\exp\left\{-\frac{1}{2\sigma_m^2}(x-\mu_m)^2\right\} \cdot \pi_m$$

s_n: which component generated observation x_n



E-step: computes the posterior for s_i given the current parameters:

$$q(s_i) = P(s_i|x_i,\theta) \propto P(x_i|s_i,\theta) \cdot P(s_i|\theta)$$

Set this to minimise KL.

Define Responsibility: $r_{im} \equiv q(s_i = m)$

(which component best explains the data)

Here $r_{im} = q(s_i = m) \propto \frac{\pi_m}{\sigma_m}\exp\left\{-\frac{1}{2\sigma_m^2}(x_i-\mu_m)^2\right\}$ And that $\sum_m r_{im} = 1$ (normalisation)

M Step:

$$\mathcal{E} = \langle \log P(x, s|\theta) \rangle_{q(s)}$$

Discrete

$$= \sum q(s) \log \{ P(x|s,\theta) \cdot P(s|\theta) \}$$

$$= \sum_{i,m} r_{im} \left\{ \log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right\}.$$

Optimum is found by setting partial derivative of E to 0.

$$\frac{\partial E}{\partial \mu_m} = \sum_i r_{im} \frac{(x_i - \mu_m)}{\sigma_m^2} = 0$$

$$\frac{\partial E}{\partial \sigma_m} = \sum_i r_{im} \left(-\frac{1}{\sigma_m} + \frac{(x_i - \mu_m)^2}{\sigma_m^3} \right) = 0$$

$$\frac{\partial E}{\partial \pi_m} = \sum_i r_{im} \frac{1}{\pi_m}, \quad \frac{\partial E}{\partial \pi_m} + \lambda = 0$$

Then update the parameters

$$\Rightarrow \mu_m = \frac{\sum_i r_{im} x_i}{\sum_i r_{im}}$$

Weighted mean of data

$$\Rightarrow \sigma_m^2 = \frac{\sum_i r_{im} (x_i - \mu_m)^2}{\sum_i r_{im}}$$

weighted covariance of data

$$\Rightarrow \pi_m = \frac{1}{n} \sum_i r_{im}$$

Note that r_{im} explains how good the cluster m explains the data. if cluster m has good explaination, the corresponding mean or covariance has bigger contribution to μ_m and σ_m^2

Mixture of Gaussians (Bishop Book)

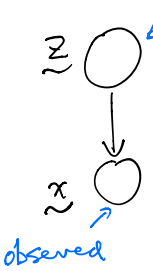
The Gaussian mixture can be written as a superposition of Gaussians in this form:

$$p(\underset{\sim}{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\underset{\sim}{x} | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k)$$

Let $\underset{\sim}{z} \in \mathbb{R}^{K\times 1}$, satisfying $z_k \in \{0,1\}$, and $\sum_k z_k = 1$.

The marginal distribution over $\underset{\sim}{z}$ is specified in terms of π_k

s.t. $p(z_k = 1) = \pi_k$ and $\pi_k \in [0,1]$, $\sum_{k=1}^{K} \pi_k = 1$



Graphical representation of a mixture model.

Joint: $p(\underset{\sim}{x}, \underset{\sim}{z}) = p(\underset{\sim}{x} | \underset{\sim}{z}) \cdot p(\underset{\sim}{z})$

$$p(\underset{\sim}{z}) = \prod_{k=1}^{K} \pi_k^{z_k} = \pi_j^{z_j} \quad \text{for some } j, \text{ and } z_j = 1.$$

$$p(\underset{\sim}{x} | z_k = 1) = \mathcal{N}(\underset{\sim}{x} | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k)$$

Hence $p(\underset{\sim}{x} | \underset{\sim}{z}) = \prod_{k=1}^{K} \mathcal{N}(\underset{\sim}{x} | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k)$

$$p(\underset{\sim}{x}) = \sum_{\underset{\sim}{z}} p(\underset{\sim}{x}, \underset{\sim}{z}) = \sum_{\underset{\sim}{z}} p(\underset{\sim}{x} | \underset{\sim}{z}) \cdot p(\underset{\sim}{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\underset{\sim}{x} | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k)$$

Also: $\gamma(z_k) \equiv p(z_k = 1 | \underset{\sim}{x})$ — Probability that it belongs to cluster k given data $\underset{\sim}{x}$

① posterior probability once we observed $\underset{\sim}{x}$

② Responsibility that component k takes for explaining the observation $\underset{\sim}{x}$.

$$= \frac{p(\underset{\sim}{x} | z_k = 1) \cdot p(z_k = 1)}{p(\underset{\sim}{x})}$$

$$= \frac{p(\underset{\sim}{x} | z_k = 1) \cdot p(z_k = 1)}{\sum_{j=1}^{K} p(\underset{\sim}{x} | z_j = 1) \cdot p(z_j = 1)}$$

$$= \frac{\mathcal{N}(\underset{\sim}{x} | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k) \cdot \pi_k}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\underset{\sim}{x} | \underset{\sim}{\mu}_j, \underline{\underline{\Sigma}}_j)}$$

regarded as prior probability of $z_k = 1$

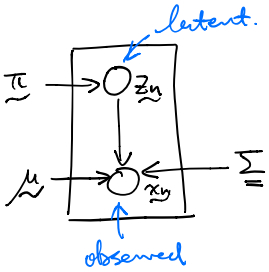
Maximum likelihood : $\mathcal{D} : \{x_1, \dots, x_N\}$. (CAN solve in closed form)

$$\underline{\underline{X}} = \begin{pmatrix} - - - \ \underset{\sim}{x}_1^T - - - \\ - - - \ \underset{\sim}{x}_2^T - - - \\ \vdots \\ - - - \ \underset{\sim}{x}_N^T - - - \end{pmatrix} \in \mathbb{R}^{N \times K} \quad ; \quad \underline{\underline{Z}} = \begin{pmatrix} - - - \ \underset{\sim}{z}_1^T - - - \\ - - - \ \underset{\sim}{z}_2^T - - - \\ \vdots \\ - - - \ \underset{\sim}{z}_N^T - - - \end{pmatrix} \in \mathbb{R}^{N \times K}$$

↑ observed ↑ latent

$$p(\underline{\underline{X}} | \underset{\sim}{\pi}, \underline{\underline{\mu}}, \underline{\underline{\Sigma}}) = \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \pi_k \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k) \right)$$

$$\log p(\underline{\underline{X}} | \underset{\sim}{\pi}, \underline{\underline{\mu}}, \underline{\underline{\Sigma}}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k) \right\}$$



Try normal maximise likelihood : (Doesn't provide a closed form solution for this)

* $\frac{\partial}{\partial \underset{\sim}{\mu}_k} \log p(\underline{\underline{X}} | \underset{\sim}{\pi}, \underline{\underline{\mu}}, \underline{\underline{\Sigma}}) = 0 \Rightarrow 0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_k, \underline{\underline{\Sigma}}_k)}{\sum_j \pi_j \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_j, \underline{\underline{\Sigma}}_j)}}_{\gamma(z_{nk})} \underline{\underline{\Sigma}}_k (\underset{\sim}{x}_n - \underset{\sim}{\mu}_k)$ (w.r.t. $\underset{\sim}{\mu}_k$)

$$\Rightarrow \underset{\sim}{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \cdot \underset{\sim}{x}_n}{\underbrace{\sum_{n=1}^{N} \gamma(z_{nk})}_{N_k}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \underset{\sim}{x}_n$$

* $\frac{\partial}{\partial \underline{\underline{\Sigma}}_k} \log p(\underline{\underline{X}} | \underset{\sim}{\pi}, \underline{\underline{\mu}}, \underline{\underline{\Sigma}}) = 0$ (w.r.t. $\underline{\underline{\Sigma}}_k$)

$$\Rightarrow \underline{\underline{\Sigma}}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (\underset{\sim}{x}_n - \underset{\sim}{\mu}_k)(\underset{\sim}{x}_n - \underset{\sim}{\mu}_k)^T}{\underbrace{\sum_{n=1}^{N} \gamma(z_{nk})}_{N_k}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\underset{\sim}{x}_n - \underset{\sim}{\mu}_k)(\underset{\sim}{x}_n - \underset{\sim}{\mu}_k)^T$$

* $\log p(\underline{\underline{X}} | \underset{\sim}{\pi}, \underline{\underline{\mu}}, \underline{\underline{\Sigma}}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$ (w.r.t. π_k, by Lagrange multiplier)

$$\Rightarrow \pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} = \frac{N_k}{N}.$$

← The mixing coefficient for the k^{th} component is given by average responsibility which that component takes for explaining the data points.

EM for Gaussian Mixtures.

Given a Gaussian Mixture model. Goal: maximise the log likelihood w.r.t. parameters

* mean, covariances of the components $\underset{\sim}{\mu}$, $\underset{=}{\Sigma}$
* mixing coefficients π_k

EM steps:

1. Initialize $\underset{\sim}{\mu}_k$, $\underset{=}{\Sigma}_k$ and π_k, the find log likelihood

2. E step: evaluate the responsibilities:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_k, \underset{=}{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_j, \underset{=}{\Sigma}_j)}$$

3. M step: Re-estimate parameters by current responsibilities.

$$\underset{\sim}{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \underset{\sim}{x}_n ,$$

$$\underset{=}{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \cdot (\underset{\sim}{x}_n - \underset{\sim}{\mu}_k^{new})(\underset{\sim}{x}_n - \underset{\sim}{\mu}_k^{new})^T$$

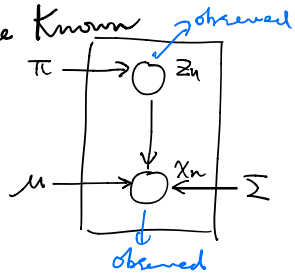
$$\pi_k^{new} = \frac{N_k}{N}.$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

4. Evaluate new log-likelihood:

$$\log p(\underset{\sim}{X} | \underset{\sim}{\mu}, \underset{=}{\Sigma}, \underset{\sim}{\pi}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k^{new} \mathcal{N}(\underset{\sim}{x}_n | \underset{\sim}{\mu}_k^{new}, \underset{=}{\Sigma}_k^{new}) \right\} \cdots$$

EM for Gaussian Mixture with Latent Variable Known



Consider maximising log likelihood of COMPLETE dataset $\{\underline{X}, \underline{Z}\}$.

$$p(\underline{X}, \underline{Z} \mid \underline{\mu}, \underline{\Sigma}, \underline{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\underline{x}_n \mid \underline{\mu}_k, \underline{\Sigma}_k)^{z_{nk}}$$

where z_{nk} is the k^{th} component of $\underline{z}_n$.

$$\log p(\underline{X}, \underline{Z} \mid \underline{\mu}, \underline{\Sigma}, \underline{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\log \pi_k + \log \mathcal{N}(\underline{x}_n \mid \underline{\mu}_k, \underline{\Sigma}_k)\}$$

joint

Compare with the ☺ equation, we can find closed form solution here, the same as single Gaussian case.

CAN Find closed form solution

$$p(\underline{Z} \mid \underline{X}, \underline{\mu}, \underline{\Sigma}, \underline{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \{\pi_k \mathcal{N}(\underline{x}_n \mid \underline{\mu}_k, \underline{\Sigma}_k)\}^{z_{nk}}$$

$\mathbb{E}(z_{nk}) = \gamma(z_{nk})$ is the responsibility

EM steps:

1. Initialise $\underline{\mu}^{old}$, $\underline{\Sigma}^{old}$, $\underline{\pi}^{old}$.
2. E step: Find the responsibility with $\underline{\mu}^{old}$, $\underline{\Sigma}^{old}$, $\underline{\pi}^{old}$.
3. M step: update parameters

$$\underline{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \underline{x}_n ,$$

$$\underline{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \cdot (\underline{x}_n - \underline{\mu}_k^{new})(\underline{x}_n - \underline{\mu}_k^{new})^T$$

$$\pi_k^{new} = \frac{N_k}{N}.$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

4. Evaluate new log-likelihood:

$$\log p(\underline{X}, \underline{Z} \mid \underline{\mu}^{new}, \underline{\Sigma}^{new}, \underline{\pi}^{new}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\log \pi_k^{new} + \log \mathcal{N}(\underline{x}_n^{new} \mid \underline{\mu}_k^{new}, \underline{\Sigma}_k^{new})\}$$

Gaussian Mixtures and K-means.

K-means: hard assignment of data points to clusters

EM : soft assignment based on posterior probabilities.

K-means:

* Objective function : $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$

  represents the sum of squares of the distances of each data point to its assigned vector μ_k.

* $r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$ ← similarly as $\gamma(z_{nk})$. but this posterior probability is certain

* $\frac{\partial J}{\partial \mu_k} = 0 \Leftrightarrow \mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$

✡ Connect Gaussian Mixture with EM with K-means.

Consider a Gaussian Mixture model where $\Sigma = \varepsilon I = \begin{pmatrix} \varepsilon & & 0 \\ & \ddots & \\ 0 & & \varepsilon \end{pmatrix}$

$$p(x|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi\varepsilon}} \cdot \exp\left\{-\frac{1}{2\varepsilon}\|x - \mu_k\|^2\right\} \quad \text{with } \varepsilon \text{ fixed.}$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} = \frac{\pi_k \cdot \exp\left\{-\frac{1}{2\varepsilon}\|x_n - \mu_k\|^2\right\}}{\sum_{j=1}^{K} \pi_j \exp\left\{-\frac{1}{2\varepsilon}\|x_n - \mu_j\|^2\right\}}.$$

and $\lim_{\varepsilon \to 0} \gamma(z_{nk}) = r_{nk}$ from the K-means

Let $\varepsilon \to 0$, consider the log-likelihood:

$$\mathbb{E}_Z\{\log p(X, Z|\mu, \Sigma, \pi)\} \to -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 + \text{constant.}$$

Hence maximizing the expected complete-data log-likelihood is equivalent to minimizing J for K-means.

Factor Analysis

$\underset{\sim}{x} \sim \mathcal{N}(0, \underset{=}{\Lambda}\underset{=}{\Lambda}^T + \underset{=}{\Psi})$, $\theta = \{\underset{=}{\Lambda}, \underset{=}{\Psi}\}$

E-step: for each $\underset{\sim}{x}_n$, compute posterior distribution of hidden factors given the observed data.

Set this in E-step

$$\uparrow \quad q(\underset{\sim}{y}_n) = p(\underset{\sim}{y}_n | \underset{\sim}{x}_n, \theta) = \frac{p(\underset{\sim}{y}_n, \underset{\sim}{x}_n | \theta)}{p(\underset{\sim}{x}_n | \theta)}$$

Calculate $p(\underset{\sim}{y}_n, \underset{\sim}{x}_n | \theta)$ as a function of $\underset{\sim}{y}_n$ and $\underset{\sim}{x}_n$ fixed.

$$p(\underset{\sim}{y}_n, \underset{\sim}{x}_n) = p(\underset{\sim}{x}_n | \underset{\sim}{y}_n) \cdot p(\underset{\sim}{y}_n)$$

$$\propto \exp\left\{-\frac{1}{2}\left[\underset{\sim}{y}_n^T \underset{\sim}{y}_n + (\underset{\sim}{x}_n - \underset{=}{\Lambda}\underset{\sim}{y}_n)^T \underset{=}{\Psi}^{-1}(\underset{\sim}{x}_n - \underset{=}{\Lambda}\underset{\sim}{y}_n)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\underset{\sim}{y}_n^T (\underset{=}{I} + \underset{=}{\Lambda}^T\underset{=}{\Psi}^{-1}\underset{=}{\Lambda})\underset{\sim}{y}_n - 2\underset{\sim}{y}_n^T \underset{=}{\Lambda}^T\underset{=}{\Psi}^{-1}\underset{\sim}{x}_n\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\underset{\sim}{y}_n^T \underset{=}{\Sigma}^{-1}\underset{\sim}{y}_n - 2\underset{\sim}{y}_n^T \underset{=}{\Sigma}^{-1}\underset{\sim}{\mu}_n + \underset{\sim}{\mu}_n^T \underset{=}{\Sigma}^{-1}\underset{\sim}{\mu}_n\right]\right\}$$

Hence $\underset{=}{\Sigma} = (\underset{=}{I} + \underset{=}{\Lambda}^T\underset{=}{\Psi}^{-1}\underset{=}{\Lambda})^{-1}$, $\underset{\sim}{\mu}_n = \underset{=}{\Sigma}\underset{=}{\Lambda}^T\underset{=}{\Psi}^{-1}\underset{\sim}{x}_n$

$\underset{\sim}{\mu}_n$ is a linear function of $\underset{\sim}{x}_n$

$\underset{=}{\Sigma}$ does not depend $\underset{\sim}{x}_n$.

M-step: find $\theta^{(t+1)}$ by maximizing

$$\theta^{(t+1)} = \arg\max_\theta \mathcal{F}(q(\underset{\sim}{Y}), \theta) = \arg\max_\theta \langle \log p(\underset{\sim}{X}, \underset{\sim}{Y} | \theta) \rangle + \text{const}$$

$$= \arg\max_\theta \sum_n \langle \log p(\underset{\sim}{y}_n | \theta) + \log p(\underset{\sim}{x}_n | \underset{\sim}{y}_n, \theta) \rangle_{q_n(\underset{\sim}{y}_n)} + \text{const}$$

$$\log p(\underset{\sim}{y}_n | \theta) + \log p(\underset{\sim}{x}_n | \underset{\sim}{y}_n, \theta)$$

$$= \underline{C} - \frac{1}{2}\underset{\sim}{y}_n^T \underset{\sim}{y}_n - \frac{1}{2}\log|\underset{=}{\Psi}| - \frac{1}{2}(\underset{\sim}{x}_n - \underset{=}{\Lambda}\underset{\sim}{y}_n)^T \underset{=}{\Psi}^{-1}(\underset{\sim}{x}_n - \underset{=}{\Lambda}\underset{\sim}{y}_n)$$

$$= \underline{C'} - \frac{1}{2}\log|\underset{=}{\Psi}| - \frac{1}{2}\left[\underset{\sim}{x}_n^T \underset{=}{\Psi}^{-1}\underset{\sim}{x}_n - 2\underset{\sim}{x}_n^T \underset{=}{\Psi}^{-1}\underset{=}{\Lambda}\underset{\sim}{y}_n + \underset{\sim}{y}_n^T \underset{=}{\Lambda}^T\underset{=}{\Psi}^{-1}\underset{=}{\Lambda}\underset{\sim}{y}_n\right]$$

$$= \underline{C'} - \frac{1}{2}\log|\underset{=}{\Psi}| - \frac{1}{2}\left[\underset{\sim}{x}_n^T \underset{=}{\Psi}^{-1}\underset{\sim}{x}_n - 2\underset{\sim}{x}_n^T \underset{=}{\Psi}^{-1}\underset{=}{\Lambda}\underset{\sim}{y}_n + \mathrm{Tr}\left(\underset{=}{\Lambda}^T\underset{=}{\Psi}^{-1}\underset{=}{\Lambda}\underset{\sim}{y}_n\underset{\sim}{y}_n^T\right)\right]$$

($\underset{\sim}{y}_n$ ↓ mean; $\underset{\sim}{y}_n\underset{\sim}{y}_n^T$ ↓ variance)

Take expectation w.r.t. $q_n(y_n)$ — Sufficient statistics of multivariate normal distribution.

$$\left\langle \log P(y_n|\theta) + \log P(x_n|y_n,\theta) \right\rangle_{q_n(y_n)}$$

$$= C' - \frac{1}{2}\log|\Psi| - \frac{1}{2}\left[x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \mathrm{Tr}\left(\Lambda^T \Psi^{-1} \Lambda \left(\mu_n \mu_n^T + \Sigma\right)\right)\right]$$

$$\mathcal{F} = \left\langle \log P(y_n|\theta) + \log P(x_n|y_n,\theta) \right\rangle_{q_n(y_n)} + \mathcal{H}$$

$$= C' - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \left[x_n^T \Psi^{-1} x_n - 2\underbrace{x_n^T \Psi^{-1} \Lambda \mu_n}_{\in \mathbb{R}} + \mathrm{Tr}\left(\Lambda^T \Psi^{-1} \Lambda \left(\mu_n \mu_n^T + \Sigma\right)\right)\right]$$

Partial derivatives: Note that $\frac{\partial}{\partial B}\mathrm{Tr}[AB] = A^T$, $\frac{\partial}{\partial A}\log|A| = A^{-T}$

$$\frac{\partial \mathcal{F}}{\partial \Lambda} = \Psi^{-1}\sum_n x_n \mu_n^T - \Psi^{-1}\Lambda\left(N\Sigma + \sum_n \mu_n \mu_n^T\right) = 0.$$

$$\Rightarrow \hat{\Lambda} = \left(\sum_n x_n \mu_n^T\right)\cdot\left(N\Sigma + \sum_n \mu_n \mu_n^T\right)^{-1}$$

$$\frac{\partial \mathcal{F}}{\partial \Psi^{-1}} = \frac{N}{2}\Psi - \frac{1}{2}\sum_n\left[x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda\left(\mu_n \mu_n^T + \Sigma\right)\Lambda^T\right] = 0$$

$$\Rightarrow \hat{\Psi} = \frac{1}{N}\sum_n\left[x_n x_n^T - \Lambda \mu_n x_n^T - x_n \mu_n^T \Lambda^T + \Lambda\left(\mu_n \mu_n^T + \Sigma\right)\Lambda^T\right]$$

$$= \underbrace{\Lambda \Sigma \Lambda^T}_{D\times D} + \frac{1}{N}\cdot\sum_n \underbrace{(x_n - \Lambda\mu_n)}_{D\times 1}\cdot\underbrace{(x_n - \Lambda\mu_n)^T}_{1\times D}$$

Note that if $\Sigma \to 0$, $\hat{\Psi}$ becomes the likelihood of ML linear regression.

Mixture of FA:

$$P(x|\theta) = \sum_k \pi_k \mathcal{N}(\mu_k, \Lambda_k \Lambda_k^T + \Psi),$$

- π_k: mixing proportion for FA.
- μ_k: centre
- $\theta = \left\{\{\pi_k, \mu_k, \Lambda_k\}_{k=1,\dots,K}, \Psi\right\}$.

Details see Hinton + Ghahramani's Paper.

EM for Exponential Families

For joint over $\underset{\sim}{z} = (\underset{\sim}{y}, \underset{\sim}{x})$ with exponential form:

$$P(\underset{\sim}{z}|\theta) = f(\underset{\sim}{z}) \exp\{\theta^T \cdot T(\underset{\sim}{z})\} / Z(\theta).$$

The free energy dependence on θ is given by:

$$\begin{aligned} F(q,\theta) &= \langle \log P(\underset{\sim}{y}, \underset{\sim}{x}|\theta) \rangle_q + H(q). \\ &= \int q(\underset{\sim}{y}) \log P(\underset{\sim}{y}, \underset{\sim}{x}|\theta) \, d\underset{\sim}{y} + H(q) \\ &\qquad \underset{\sim}{z} \\ &= \int q(\underset{\sim}{y}) \cdot \log \left[f(\underset{\sim}{z}) \exp\{\theta^T \cdot T(\underset{\sim}{z})\} / Z(\theta) \right] d\underset{\sim}{y} + H(q) \\ &= \int q(\underset{\sim}{y}) \cdot \left[\theta^T \cdot T(\underset{\sim}{z}) - \log Z(\theta) \right] d\underset{\sim}{y} + \text{constant} \\ &= \theta^T \langle T(\underset{\sim}{z}) \rangle_{q(\underset{\sim}{y})} - \log Z(\theta) + \text{constant}. \end{aligned}$$

E-step: Compute the expected sufficient statistics under q:

$$\text{i.e. } \frac{\partial}{\partial \theta} \log Z(\theta) = \langle T(\underset{\sim}{z}) | \theta \rangle$$

M-step: $\dfrac{\partial F}{\partial \theta} = \langle T(\underset{\sim}{z}) \rangle_{q(\underset{\sim}{y})} - \dfrac{\partial}{\partial \theta} \log Z(\theta)$

EM for exponential families Mixtures

note $s_i = m \Leftrightarrow \underset{\sim}{s}_i = (0, 0, \cdots, 0, \underset{m^{th} \text{ position}}{1}, 0, \cdots, 0)$

Let M component distributions' parameters into $\underset{=}{\Theta} = [\theta_m]$.

$$\log P(\underset{\sim}{x}, \underset{\sim}{s}) = \sum_i \left\{ (\log \underset{\sim}{\pi})^T \underset{\sim}{s}_i + \underset{\sim}{s}_i^T \underset{=}{\Theta}^T T(\underset{\sim}{x}_i) - \underset{\sim}{s}_i^T \log \underset{\sim}{Z}(\Theta) \right\} + \text{constant}.$$

where $\log Z(\Theta)$ collects log-normalisers $\forall$ components into an M-element vector.

E-step: Expected sufficient statistics

* $\sum_i \langle \underset{\sim}{s}_i \rangle_q$ responsibilities

* $\sum_i T(x_i) \langle \underset{\sim}{s}_i^T \rangle_q$ responsibility-weighted sufficient stats.

M-step: maximisation of expected log-joint

$$\underset{\sim}{\pi}^{(k+1)} \propto \sum_i \langle \underset{\sim}{s}_i \rangle_q .$$

$$\left\langle T(x) \middle| \underset{\sim}{\theta}_m^{\langle k+1 \rangle} \right\rangle = \frac{\sum_i T(x_i) \langle [\underset{\sim}{s}_i]_m \rangle_q}{\sum_i \langle [\underset{\sim}{s}_i]_m \rangle_q}$$

EM for MAP

$$\tilde{P}(\underset{\sim}{z}|\theta) = \frac{f(\underset{\sim}{z}) \cdot \exp\{\theta^T \cdot T(\underset{\sim}{z})\}}{Z(\theta)} ; \quad P(\theta) = \frac{F(\nu, \tau) \cdot \exp\{\theta^T \tau\}}{Z(\theta)^\nu}$$

$$\mathcal{F}_{MAP}(q, \theta) = \int q(\underset{\sim}{Y}) \log \tilde{P}(\underset{\sim}{Y}, \underset{\sim}{X}, \theta) \, d\underset{\sim}{Y} + \mathcal{H}(q) \leq \log P(\underset{\sim}{X}|\theta) + \log P(\theta).$$

$$= \int q(\underset{\sim}{Y}) \left\{ \theta^T \left(\sum_i T(\underset{\sim}{z}_i) + \tau \right) - (N + \nu) \log Z(\theta) \right\} d\underset{\sim}{Y} + \text{const}$$

$$= \theta^T \left(\langle T(\underset{\sim}{z}) \rangle_{q(\underset{\sim}{Y})} + \tau \right) - (N + \nu) \log Z(\theta) + \text{const}.$$

Hence expected sufficient stats in E step are unchanged.

Appendix :

* Positive Definiteness of KL.

Claim : $KL[q(x)\|p(x)] \geq$ with equality iff $\forall x : p(x) = q(x)$.

Pf : $KL(q\|p) = \sum_i q_i \log \frac{q_i}{p_i}$

$E \overset{\text{def}}{=} KL(q\|p) + \lambda\left(1 - \sum_i q_i\right) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda\left(1 - \sum_i q_i\right)$

$\frac{\partial E}{\partial q_i} = \log q_i - \log p_i + 1 - \lambda = 0 \quad \Rightarrow \quad q_i = p_i \exp(\lambda - 1)$.

$\frac{\partial E}{\partial \lambda} = 1 - \sum_i q_i = 0 \quad \Rightarrow \quad \sum_i q_i = 1$.

Check curvature (Hessian) :

$\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0 \; ; \qquad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0$.

Hence unique stationary point $q_i = p_i$ is a minimum. □

* Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM with θ^*, then :

$$\frac{\partial}{\partial \theta} \langle \log P(\underline{Y}, \underline{X} | \theta) \rangle_{P(\underline{Y}|\underline{X}, \theta^*)} \Big|_{\theta^*} = 0$$

$$\ell(\theta) = \log P(\underline{X}|\theta) = \langle \log P(\underline{X}|\theta) \rangle_{P(\underline{Y}|\underline{X},\theta^*)}$$

$$= \left\langle \log \frac{P(\underline{Y}, \underline{X}|\theta)}{P(\underline{Y}|\underline{X}, \theta)} \right\rangle_{P(\underline{Y}|\underline{X},\theta^*)}.$$

$$= \langle \log P(\underline{Y}, \underline{X}|\theta) \rangle_{P(\underline{Y}|\underline{X},\theta^*)} - \langle \log P(\underline{Y}|\underline{X}, \theta) \rangle_{P(\underline{Y}|\underline{X},\theta^*)}$$

Hence $\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta} \langle \log P(\underline{Y}, \underline{X}|\theta) \rangle_{P(\underline{Y}|\underline{X},\theta^*)} - \underbrace{\frac{d}{d\theta} \langle \log P(\underline{Y}|\underline{X}, \theta) \rangle_{P(\underline{Y}|\underline{X},\theta^*)}}_{= 0 \text{ if derivative exists. (min of KL)}}$

$$\Rightarrow \frac{d}{d\theta}\ell(\theta)\Big|_{\theta^*} = \frac{d}{d\theta} \langle \log P(\underline{Y}, \underline{X}|\theta) \rangle_{P(\underline{Y}|\underline{X},\theta^*)} \Big|_{\theta^*} = 0.$$

This term is not necessarily KL. But we can add some extra terms to make it KL.

$$\frac{d}{d\theta} KL\left[P(\underline{Y}|\underline{X},\theta^*) \,\|\, P(\underline{Y}|\underline{X},\theta)\right]$$

$$= \frac{d}{d\theta}\left\{-\left\langle \log P(\underline{Y}|\underline{X},\theta)\right\rangle_{P(\underline{Y}|\underline{X},\theta^*)} + \left\langle \log P(\underline{Y}|\underline{X},\theta^*)\right\rangle_{P(\underline{Y}|\underline{X},\theta^*)}\right\}$$