

Expectation Maximization

A/Prof Richard Yi Da Xu

Yida.Xu@uts.edu.au

Wechat: aubedata

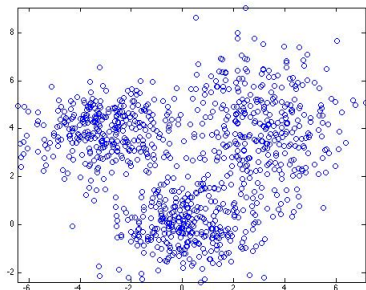
<https://github.com/roboticcam/machine-learning-notes>

University of Technology Sydney (UTS)

February 18, 2018

Motivation - Mixture Density models

When you have data that looks like:



Can you fit them using a single-mode Gaussian distribution, i.e.,:

$$\begin{aligned} p(X) &= \mathcal{N}(X|\mu, \Sigma) \\ &= (2\pi)^{-k/2} |\Sigma|^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \end{aligned}$$

Clearly NOT! This is typically modelling using Mixture Densities, in the case of Gaussian Mixture Model (k-mixture) (GMM):

$$p(X) = \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) \quad \sum_{l=1}^k \alpha_l = 1$$

Gaussian Mixture model result

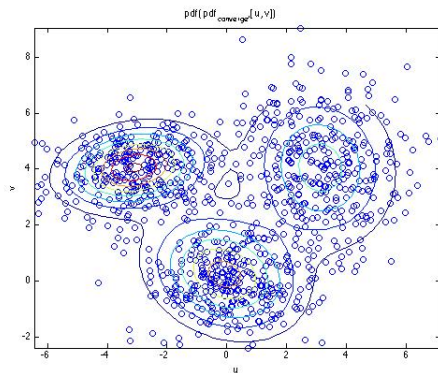


Figure: gmm fitting result

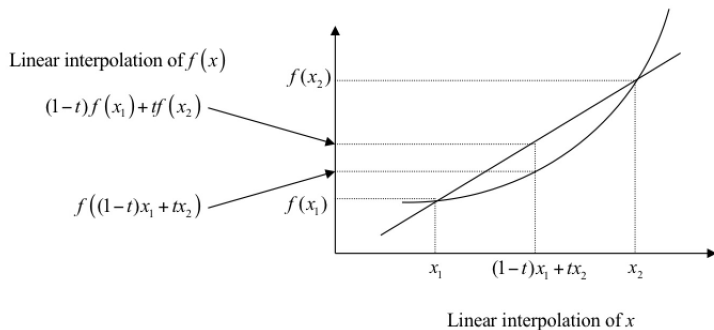
Let $\Theta = \{\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}$

$$\Theta_{\text{MLE}} = \arg \max_{\Theta} \mathcal{L}(\Theta|X)$$

$$= \arg \max_{\Theta} \left(\sum_{i=1}^n \log \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) \right)$$

- ▶ Unlike single mode Gaussian, we can't just take derivatives and let it equal zero easily.
- ▶ We need to use Expectation-Maximization to help us solving this

Convex function



$$f((1-t)x_1 + tx_2) \leq (1-t)f(x_1) + tf(x_2) \quad t \in (0 \dots 1)$$

Jensens inequality

Using notation Φ instead of f :

$$\Phi((1-t)x_1 + tx_2) \leq (1-t)\Phi(x_1) + t\Phi(x_2) \quad t \in (0 \dots 1)$$

Can be generalised further, let $\sum_{i=1}^n p_i = 1$:

$$\Phi(p_1x_1 + p_2x_2 + \dots p_nx_n) \leq p_1\Phi(x_1) + p_2\Phi(x_2) \dots p_n\Phi(x_n) \quad \sum_{i=1}^n p_i = 1$$

$$\Rightarrow \Phi\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i \Phi(x_i)$$

$$\Rightarrow \Phi\left(\sum_{i=1}^n p_i f(x_i)\right) \leq \sum_{i=1}^n p_i \Phi(f(x_i)) \quad \text{by replacing } x_i \text{ with } f(x_i)$$

Can also generalised to the continous case, by letting $\int_{x \in \mathbb{S}} p(x) = 1$:

$$\Phi\left(\int_{x \in \mathbb{S}} f(x)p(x)\right) \leq \int_{x \in \mathbb{S}} \Phi(f(x_i))p(x) \Rightarrow \Phi\mathbb{E}[f(x)] \leq \mathbb{E}[\Phi(f(x_i))]$$

Jensens inequality: $-\log(x)$

$\Phi(x) = -\log(x)$ is a convex function:

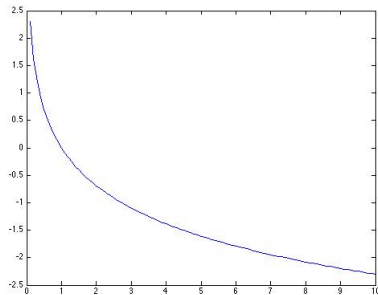


Figure: plot of $\Phi(x) = -\log(x)$

when $\Phi(\cdot)$ is convex

$$\begin{aligned}\Phi\mathbb{E}[f(x)] &\leq \mathbb{E}[\Phi(f(x_i))] \\ \Rightarrow -\log \mathbb{E}[f(x)] &\leq \mathbb{E}[-\log(f(x_i))]\end{aligned}$$

when $\Phi(\cdot)$ is concave

$$\begin{aligned}\Phi\mathbb{E}[f(x)] &\geq \mathbb{E}[\Phi(f(x_i))] \\ \Rightarrow -\log \mathbb{E}[f(x)] &\geq \mathbb{E}[-\log(f(x_i))]\end{aligned}$$

The Expectation-Maximization Algorithm

Instead of perform:

$$\theta^{\text{MLE}} = \arg \max_{\theta} (\mathcal{L}(\theta)) = \arg \max_{\theta} (\log[p(X|\theta)])$$

- ▶ **The trick** is to assume some “latent” variable Z to the model.
- ▶ such that we generate a series of $\Theta = \{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(t)}\}$

For each iteration of the E-M algorithm, we perform:

$$\Theta^{(g+1)} = \arg \max_{\theta} \left(\int_Z \log(p(X, Z|\theta)) p(Z|X, \Theta^{(g)}) dz \right)$$

However, we must ensure convergence:

$$\log[p(X|\Theta^{(g+1)})] = \mathcal{L}(\Theta^{(g+1)}) \geq \mathcal{L}(\Theta^{(g)}) \quad \forall i$$

$$\begin{aligned}\mathcal{L}(\theta|X) &= \ln(p(X|\theta)) \\&= \ln\left(\frac{p(X, Z|\theta)}{p(Z|X, \theta)}\right) = \ln\left(\frac{\frac{p(X, Z|\theta)}{Q(Z)}}{\frac{p(Z|X, \theta)}{Q(Z)}}\right) \\&= \ln\left(\frac{p(X, Z|\theta)}{Q(Z)} \times \frac{Q(Z)}{p(Z|X, \theta)}\right) \\&= \ln\left(\frac{p(X, Z|\theta)}{Q(Z)}\right) + \ln\left(\frac{Q(Z)}{p(Z|X, \theta)}\right) \\ \implies \ln(p(X|\theta)) &= \int_Z \ln\left(\frac{p(X, Z|\theta)}{Q(Z)}\right) Q(Z) + \int_Z \ln\left(\frac{Q(Z)}{p(Z|X, \theta)}\right) Q(Z) \\&= \int_Z \ln\left(\frac{p(X, Z|\theta)}{Q(Z)}\right) Q(Z) + \underbrace{\text{KL}(Q(Z) \| p(Z|X, \theta))}_{\geq 0} \\&= F(\theta, Q) + \int_Z \ln\left(\frac{Q(Z)}{p(Z|X, \theta)}\right) Q(Z)\end{aligned}$$

Proof of convergence: using M-M (2)

Another way of knowing:

$$\mathcal{L}(\theta|X) = \ln(p(X|\theta)) \geq \int_Z \ln\left(\frac{p(X, Z|\theta)}{Q(Z)}\right) Q(Z)$$

is to use Jensen's inequality:

$$\begin{aligned}\mathcal{L}(\theta|X) &= \ln p(X|\theta) = \ln \int_Z p(X, Z|\theta) \\ &= \underbrace{\ln \left(\int_Z \frac{p(X, Z|\theta)}{Q(Z)} Q(Z) \right)}_{\ln \mathbb{E}_{Q(Z)}[f(Z)]} \\ &\geq \underbrace{\int_Z \ln \left(\frac{p(X, Z|\theta)}{Q(Z)} \right) Q(Z)}_{\mathbb{E}_{Q(Z)} \ln[f(Z)]}\end{aligned}$$

Proof of convergence: using M-M (3)

E-M becomes a M-M algorithm

$$\begin{aligned}\mathcal{L}(\Theta|X) &= \int_Z \ln \left(\frac{p(X, Z|\Theta)}{Q(Z)} \right) Q(Z) + \int_Z \ln \left(\frac{Q(Z)}{p(Z|X, \Theta)} \right) Q(Z) \\ &= F(\Theta, Q) + \text{KL}(Q(Z) \| p(Z|X, \Theta))\end{aligned}$$

STEP 1 Fix $\Theta = \Theta^{(g)}$, maximize $Q(Z)$

- ▶ $\mathcal{L}(\Theta|X)$ is fixed, i.e., independent of $Q(Z)$. Therefore, $\mathcal{L}(\Theta|X)$ is the upper bound of $F(\Theta, Q)$.
- ▶ To make $\mathcal{L}(\Theta|X) = F(\Theta, Q)$, i.e., $\text{KL}(\cdot) = 0$, we choose $Q(Z) = p(Z|X, \Theta^{(g)})$. Therefore:

$$\mathcal{L}(\Theta|X) = \int_Z \ln \left(\frac{p(X, Z|\Theta)}{p(Z|X, \Theta^{(g)})} \right) p(Z|X, \Theta^{(g)})$$

STEP 2 Fix $Q(Z)$, maximize Θ

$$\Theta^{(g+1)} = \arg \max_{\Theta} \left(\int_Z \log (p(X, Z|\Theta)) p(Z|X, \Theta^{(g)}) \right) dz$$

Proof of convergence: “Tagare” approach (1)

$$\begin{aligned}\mathcal{L}(\theta|X) &= \ln[p(X|\theta)] = \ln[p(Z, X, \theta)] - \ln[p(Z|X, \theta)] \\ \Rightarrow \int_{z \in \mathbb{S}} \ln[p(X|\theta)] p(z|X, \Theta^{(g)}) dz \\ &= \int_{z \in \mathbb{S}} \ln[p(Z, X, \theta)] p(z|X, \Theta^{(g)}) dz - \int_{z \in \mathbb{S}} \ln[p(Z|X, \theta)] p(z|X, \Theta^{(g)}) dz \\ \Rightarrow \ln[p(X|\theta)] &= \underbrace{\int_{z \in \mathbb{S}} \ln[p(Z, X, \theta)] p(z|X, \Theta^{(g)}) dz}_{Q(\theta, \Theta^{(g)})} - \underbrace{\int_{z \in \mathbb{S}} \ln[p(Z|X, \theta)] p(z|X, \Theta^{(g)}) dz}_{H(\theta, \Theta^{(g)})}\end{aligned}$$

In E-M, we only maximise, i.e., $\Theta^{(g+1)} = \arg \max_{\theta} Q(\theta, \Theta^{(g)})$. Why? **a trick** If we can prove:

$$\arg \max_{\theta} \left[\int_{z \in \mathbb{S}} \ln[p(Z|X, \theta)] p(z|X, \Theta^{(g)}) dz \right] = \Theta^{(g)} \Rightarrow H(\Theta^{(g+1)}, \Theta^{(g)}) \leq H(\Theta^{(g)}, \Theta^{(g)})$$

Then

$$\mathcal{L}(\Theta^{(g+1)}) = \underbrace{Q(\Theta^{(g+1)}, \Theta^{(g)})}_{\geq Q(\Theta^{(g)}, \Theta^{(g)})} - \underbrace{H(\Theta^{(g+1)}, \Theta^{(g)})}_{\leq H(\Theta^{(g)}, \Theta^{(g)})} \geq Q(\Theta^{(g)}, \Theta^{(g)}) - H(\Theta^{(g)}, \Theta^{(g)}) = \mathcal{L}(\Theta^{(g)})$$

The “Tagare” approach (2)

$$\text{To prove} \quad \arg \max_{\theta} [H(\theta, \Theta^{(g)})] = \arg \max_{\theta} \left[\int_{z \in \mathbb{S}} \ln[p(Z|X, \theta)] p(z|X, \Theta^{(g)}) dz \right] = \Theta^{(g)}$$

$$\implies \text{To prove} \quad H(\Theta^{(g)}, \Theta^{(g)}) - H(\theta, \Theta^{(g)}) \geq 0 \quad \forall \theta$$

$$\begin{aligned} H(\Theta^{(g)}, \Theta^{(g)}) - H(\theta, \Theta^{(g)}) &= \int_{z \in \mathbb{S}} \ln[p(Z|X, \Theta^{(g)})] p(z|X, \Theta^{(g)}) dz - \int_{z \in \mathbb{S}} \ln[p(Z|X, \theta)] p(z|X, \Theta^{(g)}) dz \\ &= \int_{z \in \mathbb{S}} \ln \left[\frac{p(Z|X, \Theta^{(g)})}{p(Z|X, \theta)} \right] p(z|X, \Theta^{(g)}) dz = \int_{z \in \mathbb{S}} -\ln \left[\frac{p(Z|X, \theta)}{p(Z|X, \Theta^{(g)})} \right] p(z|X, \Theta^{(g)}) dz \\ &\geq -\ln \left[\int_{z \in \mathbb{S}} \frac{p(Z|X, \theta)}{p(Z|X, \Theta^{(g)})} p(z|X, \Theta^{(g)}) dz \right] = 0 \end{aligned}$$

Since $\Phi(\cdot) = -\ln$ is a convex uncton:

- ▶ Gaussian Mixture Model
- ▶ Probabilistic Latent Semantic Analysis (PLSA)

E-M Example: Gaussian Mixture Model

Gaussian Mixture Model (k-mixture) (GMM):

$$p(X|\Theta) = \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) \quad \sum_{l=1}^k \alpha_l = 1$$
$$\text{and } \theta = \{\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}$$

For data $X = \{x_1, \dots, x_n\}$ we introduce “latent” variable $Z = \{z_1, \dots, z_n\}$, each z_i indicates which mixture component x_i belong to.

Looking at the E-M algorithm:

$$\Theta^{(g+1)} = \arg \max_{\Theta} [Q(\Theta, \Theta^{(g)})] = \arg \max_{\Theta} \left(\int_Z \log(p(X, Z|\Theta)) p(Z|X, \Theta^{(g)}) dz \right)$$

We need to define both $p(X, Z|\Theta)$ and $p(Z|X, \Theta)$

$$p(X|\Theta) = \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) = \prod_{i=1}^n \sum_{l=1}^k \alpha_l \mathcal{N}(X_i|\mu_l, \Sigma_l)$$

How to define $p(X, Z|\Theta)$

$$p(X, Z|\Theta) = \prod_{i=1}^n p(x_i, z_i|\Theta) = \prod_{i=1}^n \underbrace{p(x_i|z_i, \Theta)}_{\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})} \underbrace{p(z_i|\Theta)}_{\alpha_{z_i}} = \prod_{i=1}^n \alpha_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i})$$

Notice that $p(X, Z|\Theta)$ is actually simpler than $p(X|\Theta)$.

How to define $p(Z|X, \Theta)$

$$p(Z|X, \Theta) = \prod_{i=1}^n p(z_i|x_i, \Theta) = \prod_{i=1}^n \frac{\alpha_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i})}{\sum_{l=1}^k \alpha_l \mathcal{N}(x_i|\mu_l, \Sigma_l)}$$

$$\begin{aligned} Q(\Theta, \Theta^{(g)}) &= \int_{\mathbf{Z}} \ln(p(\mathbf{X}, \mathbf{Z}|\Theta)) p(\mathbf{Z}|\mathbf{X}, \Theta^{(g)}) d\mathbf{z} \\ &= \int_{z_1} \dots \int_{z_n} \left(\sum_{i=1}^n \ln p(z_i, x_i|\Theta) \prod_{i=1}^n p(z_i|x_i, \Theta^{(g)}) \right) dz_1, \dots, dz_n \end{aligned}$$

- ▶ Let $P(Y)$ be the joint pdf: $P(y_1, \dots, y_n)$
- ▶ also let $F(Y)$ be a linear function, where each term involves only one variable y_i , i.e., $F(Y) = f_1(x_1) + \dots + f_n(x_n) = \sum_{i=1}^n f_i(y_i)$

Theorem:

$$\int_{y_1} \dots \int_{y_n} \left(\sum_{i=1}^n f_i(y_i) \right) P(Y) dY = \sum_i^N \left(\int_{y_i} f_i(y_i) P_i(y_i) dy_i \right)$$

$$\int_Y (F(Y)) P(Y) dY = \int_{y_1} \int_{y_2} \dots \int_{y_N} \left(\sum_{i=1}^N (f_i(y_i)) \right) P(Y) dy_1, \dots, dy_N$$

Expand it out, this equation has N sum terms. The first term is:

$$= \int_{y_1} \int_{y_2} \dots \int_{y_N} f_1(y_1) P(y_1, \dots, y_N) \prod_{i=1}^N (dy_i) = \int_{y_1} f_1(y_1) \left(\int_{y_2} \dots \int_{y_N} P(y_1, \dots, y_N) \prod_{i=2}^N (dy_i) \right) dy_1$$

What's inside the big bracket becomes the marginal probability density of $P(y_1)$, therefore, the first term becomes:

$$= \int_{y_1} f_1(y_1) p(y_1) dy_1$$

Apply this to each of the N terms, therefore:

$$\int_Y (F(Y)) P(Y) dY = \int_{y_1} f_1(y_1) P_1(y_1) dy_1 + \dots + \int_{y_N} f_N(y_N) P_N(y_N) dy_N$$

Knowing,

$$\int_{y_1} \dots \int_{y_n} \left(\sum_{i=1}^n f_i(y_i) \right) P(Y) dY = \sum_i^N \left(\int_{y_i} f_i(y_i) P_i(y_i) dy_i \right)$$

$$\begin{aligned} Q(\Theta, \Theta^{(g)}) &= \int_{z_1} \dots \int_{z_n} \left(\sum_{i=1}^n \ln p(z_i, x_i | \Theta) \prod_{i=1}^n p(z_i | x_i, \Theta^{(g)}) \right) dz_1, \dots, dz_n \\ &= \sum_{i=1}^n \left(\int_{z_i} \ln p(z_i, x_i | \Theta) p(z_i | x_i, \Theta^{(g)}) dz_i \right) \quad z_i \in \{1, \dots, k\} \\ &= \sum_{z_i=1}^k \sum_{i=1}^n \ln p(z_i, x_i | \Theta) p(z_i | x_i, \Theta^{(g)}) \quad \text{swap the summation terms} \\ &= \sum_{l=1}^k \sum_{i=1}^n \ln [\alpha_l \mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \quad \text{substitute Gaussian and replace } z_i \rightarrow l \end{aligned}$$

The M-Step objective function

$$\begin{aligned} Q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \ln[\alpha_l \mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \\ &= \sum_{l=1}^k \sum_{i=1}^n \ln(\alpha_l) p(l | x_i, \Theta^{(g)}) + \sum_{l=1}^k \sum_{i=1}^n \ln[\mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \end{aligned}$$

The first term contains only α and the second term contains only μ, Σ . So we can maximize both terms independently.

The M-Step: maximizing α

Maximizing α means that:

$$\frac{\partial \sum_{l=1}^k \sum_{i=1}^n \ln(\alpha_l) p(l|x_i, \Theta^{(g)})}{\partial \alpha_1, \dots, \partial \alpha_k} = [0 \dots 0] \quad \text{subject to } \sum_{l=1}^k \alpha_l = 1$$

This is to be solved using Lagrange Multiplier

$$\begin{aligned} \text{LM}(\alpha_1, \dots, \alpha_k, \lambda) &= \sum_{l=1}^k \ln(\alpha_l) \underbrace{\left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right)}_{\text{contains no } \alpha} - \lambda \left(\sum_{l=1}^k -1 \right) \\ \Rightarrow \frac{\partial \text{LM}}{\partial \alpha_l} &= \frac{1}{\alpha_l} \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) - \lambda = 0 \\ \Rightarrow \alpha_l &= \frac{1}{N} \sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \end{aligned}$$

The M-Step: maximizing μ, Σ

Maximizing μ, Σ means that:

$$\frac{\partial \sum_{l=1}^k \sum_{i=1}^n \ln(\alpha_l) p(l|x_i, \Theta^{(g)})}{\partial \mu_1, \dots, \partial \mu_k, \partial \Sigma_1, \dots, \partial \Sigma_k} = [0 \dots 0]$$

- ▶ You will need some linear algebra identities to solve this. It's quite involved. For details, please refer:
- ▶ J. Bilmes. "A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models"

Some formulas to remember

- ▶ derivatives of log of determinant (**with** determinant)

$$\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^\top$$

- ▶ Derivatives of Traces

$$\frac{\partial \text{tr}(F(\mathbf{X}))}{\partial \mathbf{X}} = (f(\mathbf{X}))^\top$$

where $f(\cdot)$ is the **scalar derivative** of $F(\cdot)$

- ▶ Derivatives of Traces of inverse, fact 1

$$\frac{\partial \text{tr}(\mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}} = \mathbf{A}^\top \mathbf{B}^\top$$

- ▶ Derivatives of Traces of inverse, fact 2

$$\frac{\partial \text{tr}((\mathbf{X} + \mathbf{A})^{-1})}{\partial \mathbf{X}} = -((\mathbf{X} + \mathbf{A})^{-1}(\mathbf{X} + \mathbf{A})^{-1})^\top$$

- ▶ Derivatives of Traces of inverse, fact 3

$$\frac{\partial \text{tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})}{\partial \mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1})^\top$$

$$\begin{aligned} \text{second part of } Q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \ln[\mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \\ &= \sum_{i=1}^n \sum_{l=1}^k \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} \exp \left(-\frac{1}{2} (x_i - \mu)^{\top} \Sigma^{-1} (x_i - \mu) \right) \right) p(l | x_i, \Theta^{(g)}) \end{aligned}$$

Let \mathbf{Y} be zero-meaned data matrix, where each **column** of \mathbf{Y} is $y_i - \mu_l$:

$$\mathcal{L} \equiv \mathcal{L}(p(\mathbf{Y} | \mathbf{K})) = -\frac{DN}{2} \ln(2\pi) - \frac{D}{2} \ln |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^{\top})$$

$$\begin{aligned} \text{second part of } Q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \ln[\mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \\ &= \sum_{i=1}^n \sum_{l=1}^k \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} \exp \left(-\frac{1}{2} (x_i - \mu)^{\top} \Sigma^{-1} (x_i - \mu) \right) \right) p(l | x_i, \Theta^{(g)}) \end{aligned}$$

$$\Rightarrow S(\mu_l, \Sigma_l) = \sum_{i=1}^n -\frac{1}{2} \ln(|\Sigma_l|) p(l | x_i, \Theta^{(g)}) - \sum_{i=1}^n \frac{1}{2} (x_i - \mu_l)^{\top} \Sigma^{-1} (x_i - \mu_l) p(l | x_i, \Theta^{(g)})$$

$$\Rightarrow S(\mu_l, \Sigma_l^{-1}) = -\text{Tr} \left(\frac{\Sigma_l^{-1}}{2} \sum_{i=1}^n (x_i - \mu_l)(x_i - \mu_l)^{\top} p(l | x_i, \Theta^{(g)}) \right) + \text{Constant}$$

$$\Rightarrow \frac{\partial S(\mu_l, \Sigma_l^{-1})}{\partial \mu_l} = \frac{\Sigma_l^{-1}}{2} \sum_{i=1}^n 2(x_i - \mu_l) p(l | x_i, \Theta^{(g)}) = \mathbf{0} \quad \square$$

$$\begin{aligned} \text{second part of } Q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \ln[\mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \\ &= \sum_{i=1}^n \sum_{l=1}^k \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} \exp \left(-\frac{1}{2} (x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \right) \right) p(l | x_i, \Theta^{(g)}) \end{aligned}$$

- ▶ let \mathbf{Y} be zero-meaned data matrix, where each **column** of \mathbf{Y} is $x_i - \mu_l$
- ▶ let \mathbf{P} be diagonal matrix in which \mathbf{P}_{ii} correspond to $p(l | x_i, \Theta^{(g)})$

$$\mathcal{L} \equiv \mathcal{L}(p(\mathbf{Y} | \mu_l, \Sigma_l)) = -\frac{d \times \text{tr}(\mathbf{P})}{2} \ln(2\pi) - \frac{\text{tr}(\mathbf{P})}{2} \ln |\Sigma_l| - \frac{1}{2} \text{tr}(\Sigma_l^{-1} \mathbf{Y} \mathbf{P} \mathbf{Y}^\top)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Sigma_l} &= \Sigma_l^{-1} \mathbf{Y} \mathbf{P} \mathbf{Y}^\top \Sigma_l^{-1} - \text{tr}(\mathbf{P}) \Sigma_l^{-1} = \mathbf{0} \\ \Rightarrow \Sigma_l^{-1} \mathbf{Y} \mathbf{P} \mathbf{Y}^\top \Sigma_l^{-1} &= \text{tr}(\mathbf{P}) \Sigma_l^{-1} \\ \Rightarrow \mathbf{Y} \mathbf{P} \mathbf{Y}^\top \Sigma_l^{-1} &= \text{tr}(\mathbf{P}) \Rightarrow \Sigma_l^{-1} = \text{tr}(\mathbf{P}) (\mathbf{Y} \mathbf{P} \mathbf{Y}^\top)^{-1} \\ \Rightarrow \Sigma_l &= \text{tr}(\mathbf{P})^{-1} (\mathbf{Y} \mathbf{P} \mathbf{Y}^\top) = \frac{(\mathbf{Y} \mathbf{P} \mathbf{Y}^\top)}{\text{tr}(\mathbf{P})} \\ &= \frac{\sum_{i=1}^n (x_i - \mu_l)(x_i - \mu_l)^\top p(l | x_i, \Theta^{(g)})}{\sum_{i=1}^n p(l | x_i, \Theta^{(g)})} \end{aligned}$$

$$S(\mu_l, \Sigma_l^{-1}) = \sum_{i=1}^n \left(-\frac{1}{2} \ln(|\Sigma_l|) - \frac{1}{2} (x_i - \mu_l)^T \Sigma_l^{-1} (x - \mu_l) \right) p(l|x_i, \Theta^{(g)})$$

Change Σ to Σ^{-1} , this is so that after taking derivative of $\ln(X)$, the result is in terms of X^{-1}

$$\begin{aligned} &= \left(\sum_{i=1}^n \ln(|\Sigma_l^{-1}|) p(l|x_i, \Theta^{(g)}) - \frac{1}{2} \text{tr} \left(\underbrace{\Sigma_l^{-1} \sum_{i=1}^n (x_i - \mu_l)(x - \mu_l)^T p(l|x_i, \Theta^{(g)})}_{M_l} \right) \right) \\ \Rightarrow \frac{\partial S(\mu_l, \Sigma_l^{-1})}{\partial \Sigma_l^{-1}} &= \frac{2 \sum_{i=1}^n \Sigma_l p(l|x_i, \Theta^{(g)}) - \sum_{i=1}^n \text{diag}(\Sigma) p(l|x_i, \Theta^{(g)})}{2} - \frac{2M_l - \text{diag}(M_l)}{2} = 0 \\ \Rightarrow 2 \left(\sum_{i=1}^n \Sigma p(l|x_i, \Theta^{(g)}) - M_l \right) - \sum_{i=1}^n \text{diag}(\Sigma p(l|x_i, \Theta^{(g)}) - M_l) &= 0 \\ \Rightarrow \sum_{i=1}^n \Sigma p(l|x_i, \Theta^{(g)}) - M_l &= 0 \\ \Rightarrow \Sigma &= \frac{\sum_{i=1}^n M_l}{\sum_{i=1}^n p(l|x_i, \Theta^{(g)})} = \frac{\sum_{i=1}^n (x_i - \mu_l)(x - \mu_l)^T p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^n p(l|x_i, \Theta^{(g)})} \end{aligned}$$

Summary of Gaussian Mixture Model

Maximizing μ, Σ means that to update $\Theta^{(g)} \rightarrow \Theta^{(g+1)}$:

$$\alpha_l^{(g+1)} = \frac{1}{N} \sum_{i=1}^N p(l|x_i, \Theta^{(g)})$$

$$\mu_l^{(g+1)} = \frac{\sum_{i=1}^N x_i p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^N p(l|x_i, \Theta^{(g)})}$$

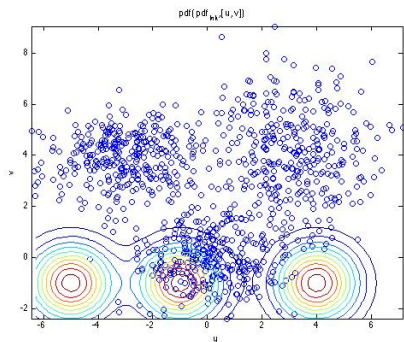
$$\Sigma_l^{(g+1)} = \frac{\sum_{i=1}^N [x_i - \mu_l^{(i+1)}][x_i - \mu_l^{(i+1)}]^T p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^N p(l|x_i, \Theta^{(g)})}$$

To program it to MATLAB, note that we need to compute the responsibility probability

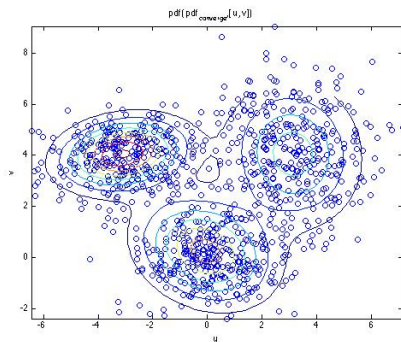
$$p(l|x_i, \Theta^{(g)}) = \frac{\mathcal{N}(x_i|\mu_l, \Sigma_l)}{\sum_{s=1}^k \mathcal{N}(x_i|\mu_s, \Sigma_s)}$$

To show the diagram again

This shows $\Theta^{(1)}$:

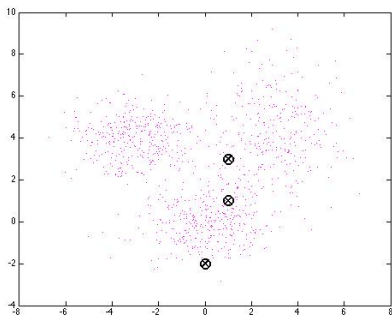


This shows $\Theta^{(\text{Converge})}$:



Other clustering methods: K-means

This shows the data and the initial “means”:

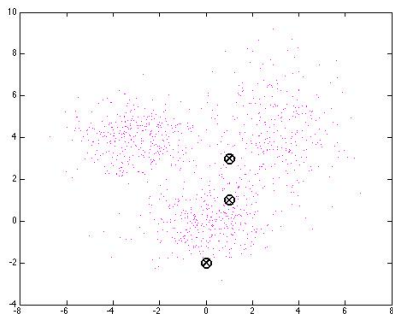


- ▶ Imagine we know that there are K types of data, and we have N data.
- ▶ How do we cluster these N data into K types automatically?
- ▶ Like GMM, this is unsupervised, clustering algorithm

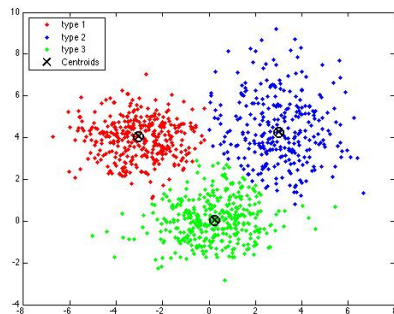
- ▶ **STEP 1:** Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- ▶ **STEP 2:** Assign each object to the group that has the closest centroid.
- ▶ **STEP 3:** When all objects have been assigned, recalculate the positions of the K centroids. Repeat Steps 2 and 3 until the centroids no longer move.

K-means

The data and the initial K “means”:



The final K “means”:



See the MATLAB Demos

Let \mathbf{Y} be zero-mean data matrix, where each column of \mathbf{Y} is y_j :

$$\begin{aligned} p(\mathbf{Y}|\mathbf{K}) &= \frac{1}{(2\pi)^{\frac{DN}{2}} |\mathbf{K}|^{\frac{N}{2}}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^\top)\right) \\ \mathcal{L} \equiv \mathcal{L}(p(\mathbf{Y}|\mathbf{K})) &= -\frac{DN}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^\top) \\ \frac{\partial \mathcal{L}}{\partial \mathbf{K}} &= -\frac{N}{2} ((\mathbf{K}^{-1})^\top) - \frac{1}{2} (-(\mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^\top \mathbf{K}^{-1})^\top) \end{aligned}$$

when \mathbf{K} is symmetric:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{K}} &\propto -N\mathbf{K}^{-1} + \mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^\top \mathbf{K}^{-1} \\ &= \mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^\top \mathbf{K}^{-1} - N\mathbf{K}^{-1} \\ &= (\mathbf{K}^{-1} \mathbf{Y})(\mathbf{K}^{-1} \mathbf{Y})^\top - N\mathbf{K}^{-1} \\ &= (\mathbf{K}^{-1} \mathbf{Y})(\mathbf{K}^{-1} \mathbf{Y})^\top - N\mathbf{K}^{-1} \quad \text{because } \mathbf{K} \text{ is symmetric} \end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}} = (\mathbf{K}^{-1} \mathbf{Y})(\mathbf{K}^{-1} \mathbf{Y})^\top - N \mathbf{K}^{-1}$$

since \mathbf{K} is symmetric, we can equivalently write the above into:

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = \text{tr} \left(\left[\underbrace{(\mathbf{K}^{-1} \mathbf{Y})(\mathbf{K}^{-1} \mathbf{Y})^\top - N \mathbf{K}^{-1}}_{\mathbf{A}} \right] \underbrace{\frac{\partial \mathbf{K}}{\partial \theta_j}}_{\frac{\partial \mathbf{A}}{\partial \theta}} \right)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{A}} &= \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{1,2} & a_{2,2} \end{bmatrix} & \frac{\partial \mathbf{A}}{\partial \theta} &= \begin{bmatrix} \frac{\partial a_{1,1}}{\partial \theta} & \frac{\partial a_{1,2}}{\partial \theta} \\ \frac{\partial a_{1,2}}{\partial \theta} & \frac{\partial a_{2,2}}{\partial \theta} \end{bmatrix} \\ \implies \text{tr} \left(\frac{\partial \mathcal{L}}{\partial \mathbf{A}} \times \frac{\partial \mathbf{A}}{\partial \theta} \right) &= a_{1,1} \frac{\partial a_{1,1}}{\partial \theta} + a_{1,2} \frac{\partial a_{1,2}}{\partial \theta} + a_{1,2} \frac{\partial a_{1,2}}{\partial \theta} + a_{2,2} \frac{\partial a_{2,2}}{\partial \theta} \end{aligned}$$

If we let:

$$\begin{aligned}\text{cov}(y_p, y_q) &= k(x_p, x_q) + \sigma_n^2 \delta_{pq} \\ &= \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}\end{aligned}$$

Parameter of the model include $\Theta = \{\sigma_f^2, l^2, \sigma_n^2\}$

$$\begin{aligned}\frac{\partial \mathbf{K}}{\partial \sigma_f^2} &= \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) \\ \frac{\partial \mathbf{K}}{\partial l} &= \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) \frac{(x_p - x_q)^2}{l^3} \\ \frac{\partial \mathbf{K}}{\partial \sigma_n^2} &= \delta_{pq}\end{aligned}$$