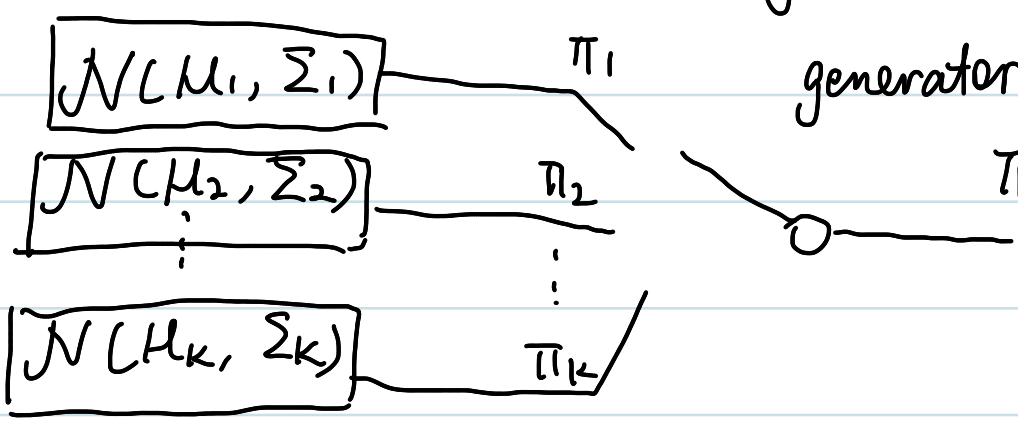
2. Mixture of Gaussian.

We will analyze Gaussian Mixture model.

The distribution of data point x is

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

There are k Gaussian R.V generators, a switch will choose generator i^{th} output with probability



Now, we will use $\vec{z}$ to determine which source produces current sample.

$\vec{z}_n = [z_1, z_2, \cdots z_K]^T$, $z_i \in \{0, 1\}$, $\sum z_i = 1$.

and $p(z_i = 1) = \pi_i$

So, $p(x \mid z) = \prod_{k=1}^{K} \mathcal{N}(x \mid \mu_k, \Sigma_k)^{z_k}$

$$p(x) = \sum_{z} p(z)\, p(x \mid z) = \sum_{z} \prod_{k=1}^{K} \mathcal{N}(x \mid \mu_k, \Sigma_k)^{z_k} \cdot p(z)$$

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

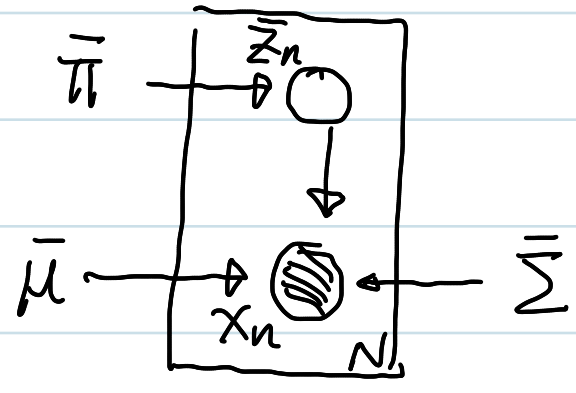
also, from Bayes' theorem we can have

$$\gamma(z_k) = p(z_k=1|x) = \frac{p(z_k=1)\,p(x|z_k=1)}{\sum_{j=1}^{k} p(z_j=1)\,p(x|z_j=1)}$$

$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{k} \pi_j \mathcal{N}(x|\mu_k, \Sigma_k)}$$

z_k here is called latent variable, sometimes also known as hidden variable.

Our target is to find out $\bar{\mu}$, $\bar{\pi}$, $\bar{\Sigma}$ from dataset $\{x_n\}$



1. Maximum Likelihood.

The likelihood function $p(\bar{X}|\bar{\pi}, \bar{\mu}, \bar{\Sigma})$ is

$$p(\bar{X}|\bar{\pi}, \bar{\mu}, \bar{\Sigma}) = \prod_{n=1}^{N} p(x|\bar{\pi}, \bar{\mu}, \bar{\Sigma})$$

$\ln(\cdot)$

$$= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

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

Before talking about how to do maximizating, let's look at a weird situation which could screw up the system.

Let's stant from a simple case where $\Sigma_k = \sigma_k^2 I$. If a data point x_n equals to one of Gaussian models' mean

and denote that Gaussian Distribution is $\mathcal{N}(\mu_k, \sigma_k^2 I)$. we will have

$$p(x_n \mid \mu_k, \sigma_k^2 I) \overset{x_n=\mu_k}{=\!=} \frac{1}{\sqrt{2\pi}\,\sigma_k} \cdot \exp\underbrace{\left\{\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right\}}_{=0}$$

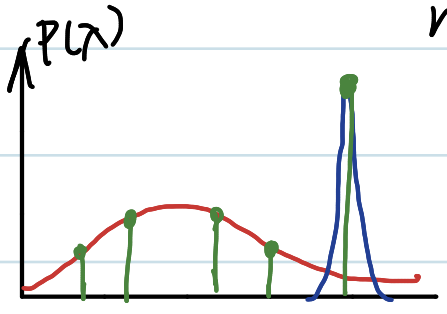
$$= \frac{1}{\sqrt{2\pi}\,\sigma_k}$$

So, if $\sigma_k \to 0$, $p(x_n \mid \mu_k, \sigma_k^2 I) \to \infty$. This is bad, because maximum likelihood may not work completely.

Even if there is only one sample $x_n = \mu_k$, this method will fail.

But we should also be clear that, we never face this problem when there is only one Gaussian model. Because if we only have one model, and we have likelihood $\to \infty$ for one data point, and contributions from other data point drop to 0, which lead the overall likelihood to zero (because other sample's likelihood is 0). This is the over-fit property of MLE. Bayesian method won't have this drawback.

Let's give an example. Samples are in green points, red & blue curves are real distribution. Once the $\mu_k \to$ one data point, no more update, because learning algorithm gets stuck.



2. EM for Gaussian Mixtures.

This would be the first time that we introduce E-M algorithm. Let's start with Gaussian Mixture problem and we will have a general analysis later.

EM algorithm is still a ML estimating algorithm, which is short for Expectation-Maximization.

From previous analyze, we know the logarithm of likelihood function $\ln p(X|\pi,\mu,\Sigma)$ is

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

First, let's compute the gradient w.r.t μ_k, we have

$$\frac{\partial}{\partial \mu_k} \ln p(\bar{X}|\bar{\pi},\bar{\mu},\bar{\Sigma}) = \sum_{n=1}^{N} \frac{\pi_k}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \cdot \frac{\partial}{\partial \mu_k} \mathcal{N}(x_n|\mu_k,\Sigma_k)$$

$$\frac{\partial}{\partial \mu_k} \mathcal{N}(x_n|\mu_k,\Sigma_k) = \frac{\partial}{\partial \mu_k} \cdot \frac{1}{\sqrt{2\pi^D|\Sigma_k|}} \exp\left\{-(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)\right\}$$

$$= \frac{1}{\sqrt{2\pi^D|\Sigma_k|}} \exp\left\{-(x_n-\mu_k)^T \Sigma^{-1}(x_n-\mu_k)\right\}\left\{-2\Sigma_k^{-1}(x_n-\mu_k)\right\}$$

$\therefore$ we have

$$0 = -2\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \Sigma^{-1}(x_n-\mu_k)$$

for simplicity, let

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

We have seen this notation on page 2

We have

$$0 = -\sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k^{-1} (x_n - \mu_k)$$

$$\Rightarrow \mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

← Remind you of k-mean method's μ_k ?

Similarly. set $\frac{\partial}{\partial \Sigma_k} \ln p(\bar{X} \mid \bar{\pi}, \bar{\mu}, \bar{\Sigma})$ to 0, we have

$$\frac{\partial}{\partial \Sigma_k} \ln p(\bar{X} \mid \bar{\pi}, \bar{\mu}, \bar{\Sigma}) = \sum_{n=1}^{N} \frac{\pi_k}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} \cdot \frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

$$\frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}} \cdot \frac{-1}{|\Sigma_k|^2} \cdot |\Sigma_k| (\Sigma_k^{-1})^T \exp\left(-(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right)$$

$\frac{\partial}{\partial x} \det(X) = \det(X) (X^{-1})^T$

$$+ \frac{1}{(2\pi)^{D/2} |\Sigma_k|} \cdot \exp\left\{-(x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k)\right\} \cdot \left(\Sigma_k^{-T} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-T}\right)$$

$\frac{\partial}{\partial X} a^T X^{-1} b = -X^{-T} a b^T X^{-T}$

$$= \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \left\{-\Sigma_k^{-T} + \Sigma_k^{-T} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-T}\right\}$$

we then have

$$\sum_{n=1}^{N} \gamma(z_{nk})\left(-\Sigma_k^{-T} + \Sigma_k^{-T}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-T}\right)=0$$

multiple both side by Σ_k^T twice, one on the left, onee on the right, and do transpose, we then have

$$\sum_{n=1}^{N} \gamma(z_{nk})\left(\Sigma_k - (x_n-\mu_k)(x_n-\mu_k)^T\right)=0$$

$$\Rightarrow \Sigma_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

So far, we know how to update μ & Σ based on give latent var. Our last task is to find the optimal solution for π. Note that $\pi^T\cdot\mathbb{1}=1$, so it's a constraint. We will make use of Lagrangian. The problem we are trying to solve is

$$\max_{\bar{\pi}} \ln p(\bar{X}\mid\bar{\pi},\bar{\mu},\bar{\Sigma})$$
$$\text{s.t. } \mathbb{1}^T\bar{\pi}=1$$

$$\Rightarrow L(\pi,\lambda) = \ln p(\bar{X}\mid\bar{\pi},\bar{\mu},\bar{\Sigma}) + \lambda\left(\sum_k \pi_k - 1\right)$$

$$\frac{\partial}{\partial\pi_k}L(\bar{\pi},\lambda) = \sum_{n=1}^{N}\frac{\mathcal{N}(x_n\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x_n\mid\mu_j,\Sigma_j)} + \lambda = 0$$

multiple bothside by π_k, we have.

$$\sum_{n=1}^{N}\gamma(z_{nk}) + \pi_k\lambda = 0 \Rightarrow \pi_k = -\frac{1}{\lambda}\sum_{n=1}^{N}\gamma(z_{nk})$$

also, $\sum_k \pi_k = 1$, we have

$$-\sum_k \frac{1}{\lambda} \sum_{n=1}^{N} \gamma(z_{nk}) = 1 \quad \Rightarrow \quad \lambda = -\sum_k \sum_{n=1}^{N} \gamma(z_{nk}) = -N$$

∴ the update for π :

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

Now, let's see the whole steps from a very high level.

E-Step:

E-step is to find the posterior probabilities. In this model, the posterior is $\gamma(z_{nk}) = p(z_k = 1 \mid x_n) = \underline{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}$

$$\sum_j \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)$$

The value is regarded as the best guess of x_n's true label π.

M-Step:

M-Step is to find out the maximum likelihood solution for all the parameters of the model.

Before giving out the whole algorithm, let's talk about termination condition. We can keep track the log likelihood function, until it's high enough

EM algorithm

① Initial μ, π, Σ

② E-step.

compute $\gamma(z_{nk}) = \dfrac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}$

③ M-step

$$\mu_k^{+} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

$$\Sigma_k^{+} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k^{+})(x_n - \mu_k^{+})^T$$

$$\pi_k^{+} = \frac{N_k}{N}$$

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

④ Compute log likelihood

$$\ln p(\bar{x} | \bar{\mu}, \bar{\Sigma}, \bar{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right\}$$

if it doesn't reach termination condition, refure to step ②