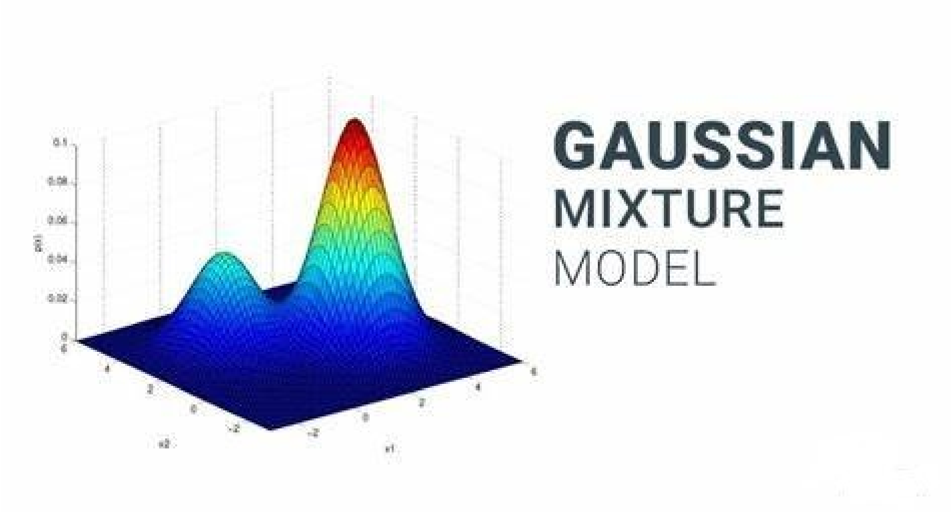
# Gaussian mixture model



### Modeling probability density functions for data that fit a Gaussian distribution

**Assuming that the data is one-dimensional**

$$p(X) = \frac{1}{2\pi} exp\{-\frac{(X-\mu)^2}{2\sigma^2}\}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$$

**Assume that the data is high-dimensional (dimension is )**

$$p(X) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} exp\{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^T$$

Under the assumption that the data conforms to a Gaussian distribution, its mean and covariance matrix can be estimated by the above equation, thus completing the modeling of the probability density function of the data.

However, in most cases, the actual data distribution is not a Gaussian distribution, but may be a linear superposition of two or more Gaussian distributions, which is called Gaussian mixture distribution, and the corresponding modeling process is called Gaussian mixture model (GMM).

# Gaussian mixture model

假设数据是由 $K$ 个高斯概率密度函数混合而成，则高斯混合分布如下所示：

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

其中 $\pi_k$ 表示第 $k$ 个高斯密度函数在整个高斯混合分布中所占的比例，区间为 $[0, 1]$ 且 $\sum_{k=1}^{K} \pi_k = 1$ ，$N(X, \mu_k, \Sigma_k)$ 表示第 $k$ 个高斯概率密度函数：

$$N(X, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} exp\{-\frac{1}{2}(X - \mu_k)^T \Sigma^{-1} (X - \mu_k)\}$$

高斯混合模型要做的事情，是通过一组训练数据 $\{X_i\}_{i=1}^{n}$ ，估计 $K$ 个三元组参数$\{\pi_k, \mu_k, \Sigma_k\}$ 的值。我们即不知道每一个数据属于哪一个高斯密度函数，也不知道三元组参数的具体取值，那么如何进行求解呢？

## Solution: EM algorithm

这里就要搬出EM算法（Expectation-maximization），固定一个求另一个，不断循环迭代。

使用EM算法求解高斯混合模型的流程如下:

>> 输入: $\{X_i\}_{i=1}^n$

>> 输出: $K$ 个三元组$\{\pi_k, \mu_k, \Sigma_k\}$

**Step 1**. 随机选取 $k$ 个三元组$\{\pi_k, \mu_k, \Sigma_k\}$;

**Step 2**. 求 $\gamma_{ik} = \frac{\pi_k N(X_i, \mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k N(X_i, \mu_k, \Sigma_k)}$ ，表示第 $i$ 个数据 $X_i$ 属于第 $k$ 个高斯密度函数的概率;

**Step 3**: 重新计算 $k$ 个三元组$\{\pi_k, \mu_k, \Sigma_k\}$:

$$\pi_k = \frac{1}{n}\sum_{i=1}^n \gamma_{ik}$$

$$\mu_k = \frac{\sum_{i=1}^n \gamma_{ik} X_i}{\sum_{i=1}^n \gamma_{ik}}$$

$$\Sigma_k = \frac{\sum_{i=1}^n \gamma_{ik}(X_i - \mu_k)(X_i - \mu_k)^T}{\sum_{i=1}^n \gamma_{ik}}$$

> 相比高斯概率分布的 $\mu = \frac{1}{n}\sum_{i=1}^n X_i$ 和 $\Sigma = \frac{1}{n-1}\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T$，高斯混合模型用 $\gamma_{ik}$ 对 $K$ 个概率密度函数进行加权的归一化。

**Step 4**. 回到2循环直至收敛。