

# EM Algorithm

Xiao Han, HP Labs

[artex.xh@gmail.com](mailto:artex.xh@gmail.com)

Ver.5.2008.10.31

Reference: Christopher M.Bishop, *Pattern Recognition and Machine Learning*

For latest version, please visit <http://glatteis.spaces.live.com>

# 预备知识

- 概率（加法、乘法、条件概率、i.i.d.、多维随机变量、高斯分布、贝叶斯、Maximum log-likelihood）
- 求导（偏导、向量求导、矩阵求导、拉格朗日乘法）

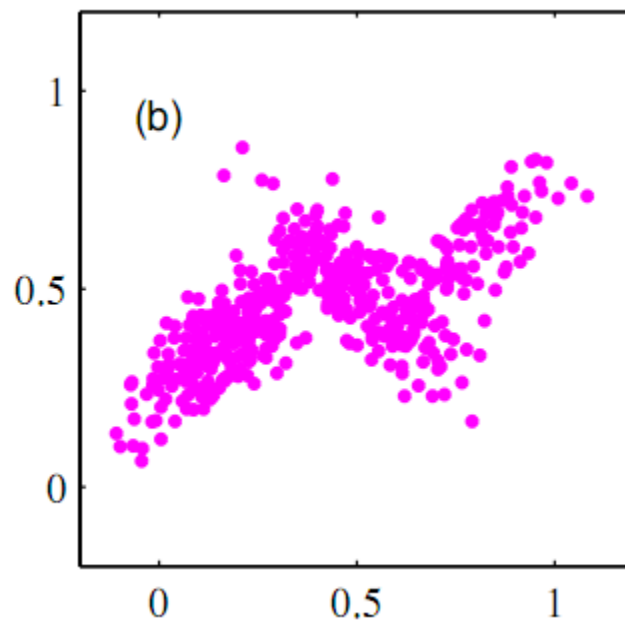
# 问题的来源

- 给定一些观察数据 $\mathbf{x}$ , 假设 $\mathbf{x}$ 符合如下的混合高斯分布

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

我们要求混合高斯分布的三组参数  $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$

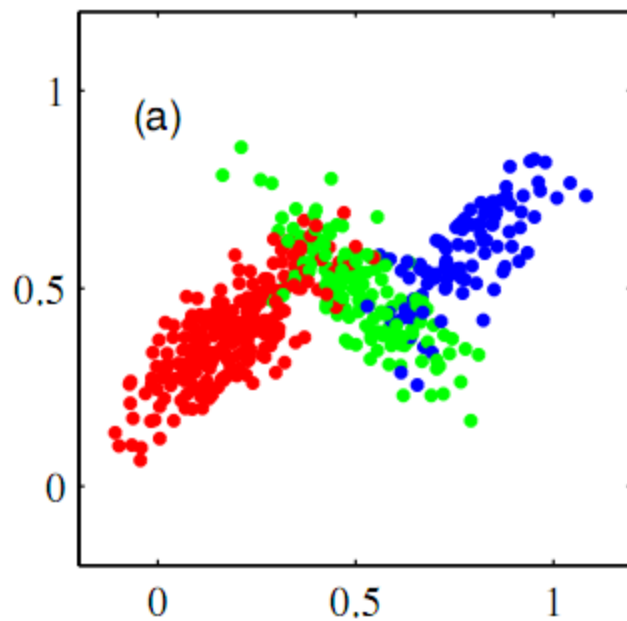
# 问题图示



# 简化的问题

- 该混合高斯分布一共有 $k$ 个分布，并且对于每一个观察到的 $x$ ，如果我们同时还知道它是属于 $k$ 中哪一个分布的，则求各个参数并不是件难事
- 比如用 $z$ 来表示每一个高斯分布，那么我们的观察集不仅仅是 $\{x_1, x_2, x_3 \dots\}$ ，而是 $\{(x_1, z_2), (x_2, z_3), (x_3, z_1) \dots\}$

# 简化问题的图示

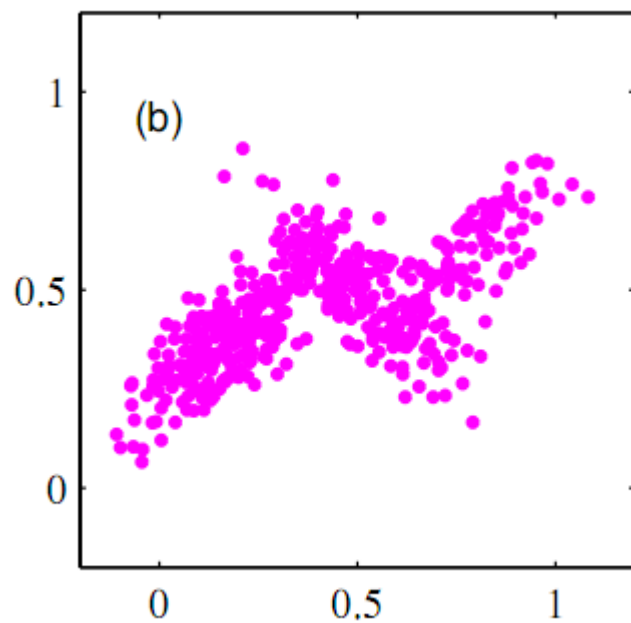


# 实际问题

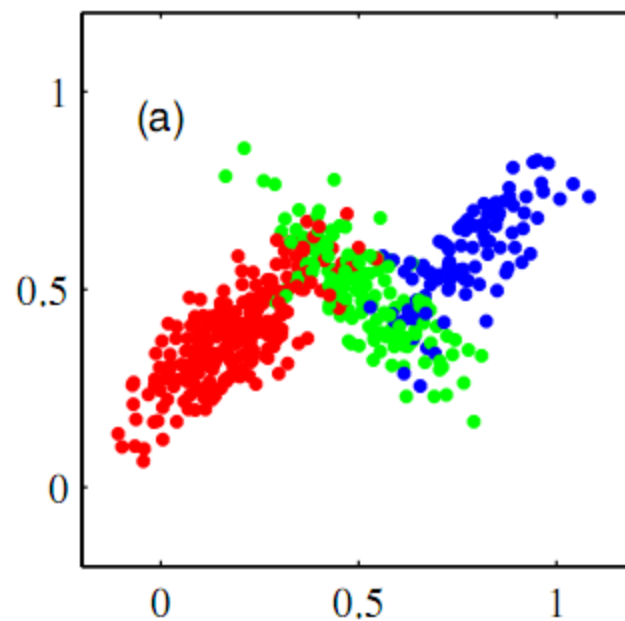
- 而现实往往是：  
我们不知道每个 $x$ 属于哪个分布，也就是说 $z$ 是我们观察不到的， **$z$ 是隐藏变量(latent variable)**

# 引入两个概率

- $p(x)$



- $p(x,z)$



# 隐藏变量Z

- 为了将k个高斯分布用一个随机变量表示
- 可以采用1-of-K的表示法，例如k=3时：
- $z_1=1$ 表示(1 0 0)，并 $p(z_1=1)=\pi_1$ ,
- $z_2=1$ 表示(0 1 0)，并 $p(z_2=1)=\pi_2$ ,
- $z_3=1$ 表示(0 0 1)，并 $p(z_3=1)=\pi_3$

$$p(z_k = 1) = \pi_k$$

$$0 \leq \pi_k \leq 1$$

$$\sum_{k=1}^K \pi_k = 1$$

- 于是  $p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$ .

- 这里的粗体z表示的是形如(1 0 0)这样的向量

# 隐藏变量与混合高斯分布

- 将 $\mathbf{z}$ 引入后

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- 最终得到

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

# 约定

对于观察集 $\{X\}$ 中的各个观察值 $x_i$ , 我们认为相互之间独立。特别的, 如果 $x_1, x_5, x_9$ 来自于同一高斯分布, 我们认为他们满足i.i.d.(独立同分布)

# 再看简化的问题

- 前面说过，在简化问题中我们观察到的是 $\{\mathbf{X}, \mathbf{Z}\}$ ，因此根据以下两个式子

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}. \quad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

- 可以得到

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

N是数据集X的大小

# 两个问题的比较

- 回忆我们的最终目标是：找一组合适的 $\pi, \mu, \Sigma$ ，满足数据集 $\{\mathbf{X}\}$ 的分布。
- 即：maximum log-likelihood
- 对原始问题，我们要找 $\pi, \mu, \Sigma$ ，使下式最大

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

- 对简化问题，同样要找 $\pi, \mu, \Sigma$ ，使下式最大

$$\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \}.$$

$Z_{nk}$  表示 $Z_n$ 的第 $k$ 个元素

# 计算复杂度

- 后者的ln直接作用于正态分布，使正态分布由**乘**的**e指数**形式变为**加**的简单形式

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

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}.$$

# 简化问题的计算1

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}.$$

- 为了最大化上式，由于 $\mathbf{Z}_{nk}$ 已知，我们可以把上式按观察到的 $(\mathbf{x}, \mathbf{z})$ 分为 $k$ 组：

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &= \sum_{n \in C_1} (\ln \pi_1 + \ln N(x_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) + \sum_{n \in C_2} (\ln \pi_2 + \ln N(x_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \\ &+ \dots + \sum_{n \in C_k} (\ln \pi_k + \ln N(x_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \end{aligned}$$

由于这 $k$ 组分布相互独立，我们只需要分别最大化每一组 $\boldsymbol{\Sigma}$

## 简化问题的计算2

- 而最大化  $\sum_{n \in C_k} (\ln \pi_k + \ln N(x_n | \mu_k, \Sigma_k))$

实际变成一个单高斯分布最大化参数的问题，因为：

$$\begin{aligned} \sum_{n \in C_k} (\ln \pi_k + \ln N(x_n | \mu_k, \Sigma_k)) &= n \ln \pi_k + \sum_{n \in C_k} \ln N(x_n | \mu_k, \Sigma_k) \\ &= n \ln \pi_k + \ln [N(x_1 | \mu_k, \Sigma_k) \times N(x_2 | \mu_k, \Sigma_k) \times \cdots \times N(x_n | \mu_k, \Sigma_k)] \\ &= n \ln \pi_k + \ln N(X | \mu_k, \Sigma_k) \end{aligned}$$

- 其中 $X$ 是所有属于第 $k$ 个分布的观察值
- $n$ 是指有多少个观察值属于第 $k$ 个分布

# 简化问题的计算3——计算单一高斯分布的参数

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}).$$

- 先对求 $\boldsymbol{\mu}$ 偏导（此处用到  $\frac{\partial(\mathbf{x}^T \mathbf{a})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}^T \mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$ ）

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})$$

- 令上式等于0则，

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$$

- 同理可得

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^T$$

# 简化问题的结论

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

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

其中 
$$N_k = \sum_{n=1}^N z_{nk}$$

这是上页 $\mu$ 和 $\Sigma$ 两式的一般式，注意 $z_{nk}$ 只能取0或1此式便不难理解

## 关于参数 $\pi$

- 由于要使  $\ln p(X, Z | \mu, \Sigma, \pi)$  达到最大，同时参数 $\pi$

必须满足  $\sum_{k=1}^K \pi_k = 1$ ，运用拉格朗日乘法可得

$$\pi_k = \frac{1}{N} \sum_{n=1}^N z_{nk} = \frac{N_k}{N}$$

# 实际问题

- 至此我们已经解决了简化问题的参数求解。但是，实际上我们往往不知道 $Z_{nk}$ ，即 $Z$ 往往是隐藏变量。也就无法运用前面简化问题的算法
- 虽然不知道 $Z_{nk}$ ，但是我们可以用它的期望 $E[Z_{nk}]$ 去估计 $Z_{nk}$

## $Z_{nk}$ 的期望估计1

- 根据前面提到的这两个公式，及贝叶斯公式

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k} \quad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

- 可以得到 $\mathbf{Z}$ 的后验概率

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^N \prod_{k=1}^K [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}.$$

## $z_{nk}$ 的期望估计2

$$\begin{aligned} E(z_{nk} | x_n) &= \sum_{z_{nk}} z_{nk} \cdot p(z_{nk} | x_n) \\ &= \sum_{z_{nk}} z_{nk} \cdot \frac{p(z_{nk}) p(x_n | z_{nk})}{p(x_n)} \\ &= \frac{1 \cdot p(z_{nk} = 1) p(x_n | z_{nk} = 1) + 0 \cdot p(z_{nk} = 0) p(x_n | z_{nk} = 0)}{p(x_n)} \\ &= \frac{p(z_{nk} = 1) p(x_n | z_{nk} = 1)}{p(x_n)} \\ &= \frac{\pi_k \cdot N(x_n | \mu_k, \Sigma_k)}{\sum_j \pi_j \cdot N(x_n | \mu_j, \Sigma_j)} = \gamma(z_{nk}) \end{aligned}$$

## 用 $E(Z_{nk})$ 代替 $Z_{nk}$

- 代入简化问题中的  $\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}.$$

现在我们要使该式最大，也就是期望值最大  
(Expectation Maximum-EM)

# 简化问题->实际问题

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}.$$

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}.$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N z_{nk} \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N z_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

$$N_k = \sum_{n=1}^N z_{nk}$$

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

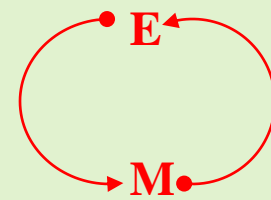
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

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

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

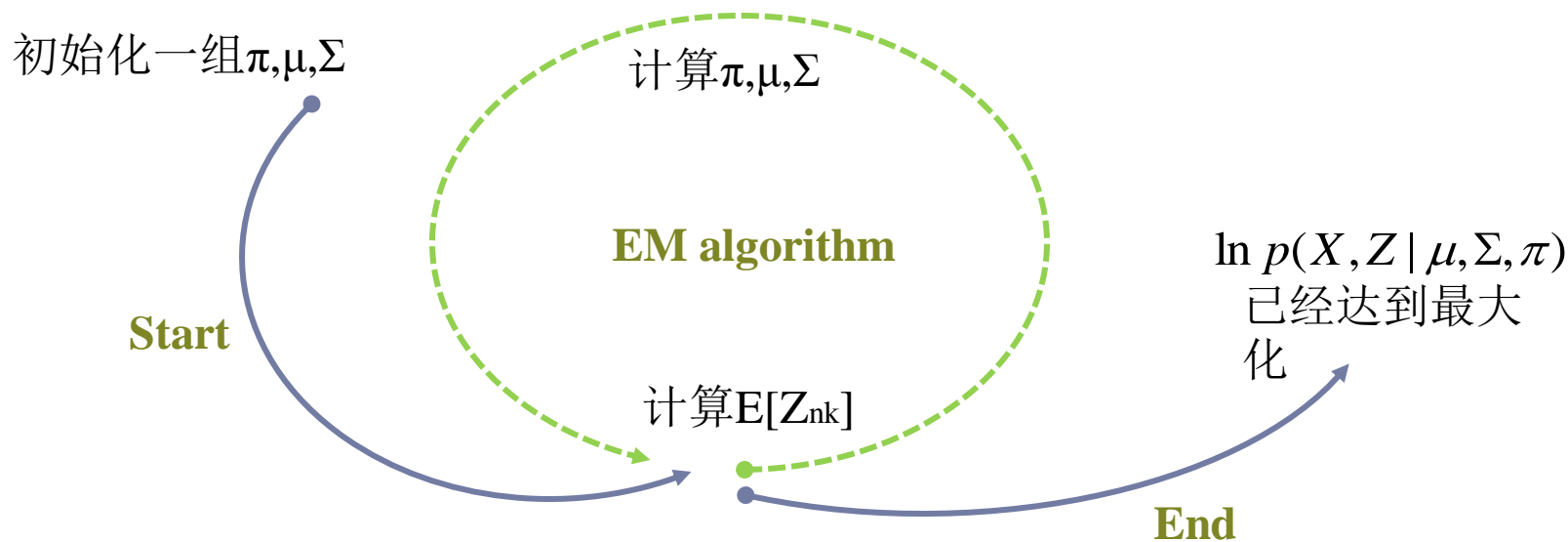
$$\pi_k = \frac{N_k}{N}$$

只有M-Step

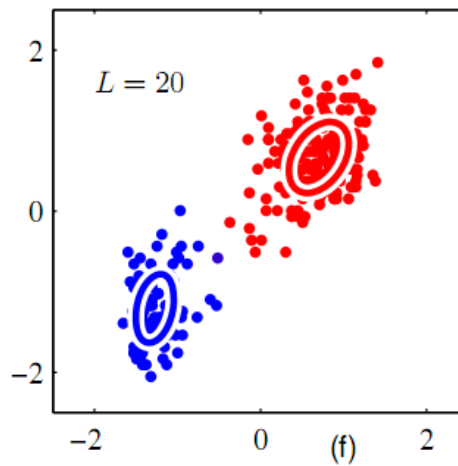
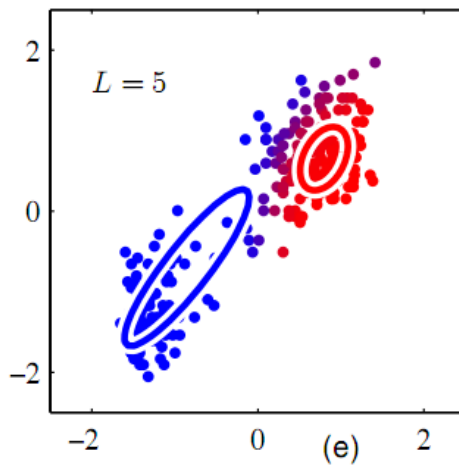
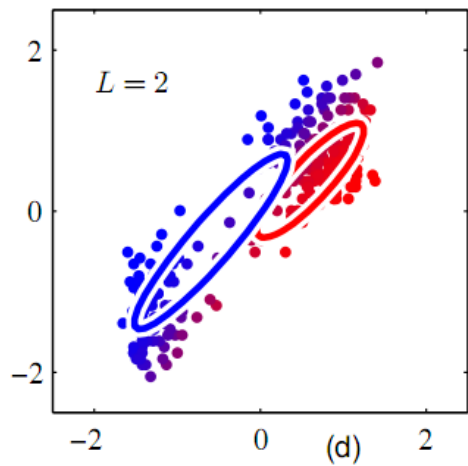
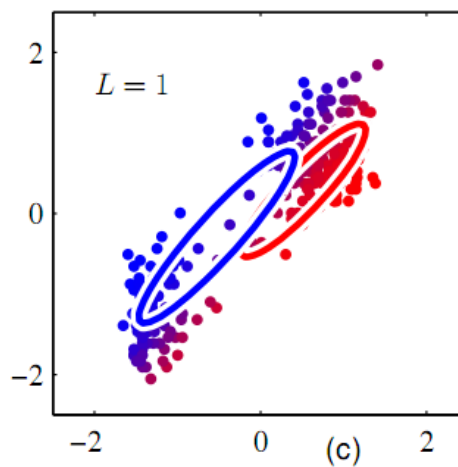
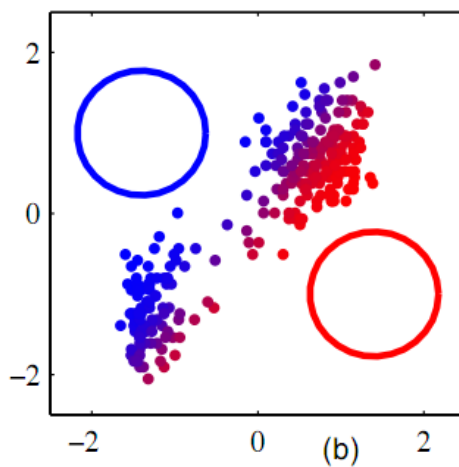
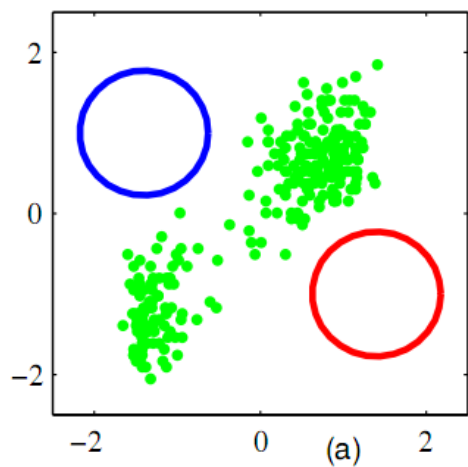


# 迭代

- 注意实际问题与简化问题的不同，我们在用期望  $E[Z_{nk}]$  去估计  $Z_{nk}$ ，因此我们需要不断的根据后验概率  $p(Z|X, \mu, \Sigma, \pi)$  去更新  $E[Z_{nk}]$



# 问题的解决



# 为什么要用 $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$ 去求 $E[\mathbf{Z}]$ ?

- Q:我们的目标是用 $E[\mathbf{Z}]$ 去估计隐藏变量 $\mathbf{Z}$ ，那么我们是否可以随便假设一个分布 $p(\mathbf{Z})$ ，就用它去计算 $E[\mathbf{Z}]$ 呢，为什么非要用  $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$  呢？
- A:可以的，但随便假设一个分布不是最优的方案，它可能会使EM算法的迭代次数大大增加。直觉上讲，我们拥有一定知识(比如 $\mathbf{X}, \pi, \mu, \Sigma$ )后的推断 $\mathbf{Z}$ 会准确些。我们稍后会给出严格的却不晦涩的证明。

# 一般化的EM算法

- 令 $\mathbf{X}$ 为所有的观察变量
- 令 $\mathbf{Z}$ 为所有的隐藏变量
- $\Theta$ 为所有的参数
- 我们的目的是要最大化 $\ln p(\mathbf{X} | \Theta)$ ,并求出此时的参数 $\Theta$
- 并且, 我们引入 $q(\mathbf{Z})$ 来描述隐藏变量的分布

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) = 1$$

# 拆分 $\ln p(X | \Theta)$

$$\begin{aligned} \ln p(X | \theta) &= \ln p(X, Z | \theta) - \ln p(Z | X, \theta) \\ &= \ln p(X, Z | \theta) - \ln q(Z) - [\ln p(Z | X, \theta) - \ln q(Z)] \\ &= \ln \frac{p(X, Z | \theta)}{q(Z)} - \ln \frac{p(Z | X, \theta)}{q(Z)} \end{aligned}$$

- 等式两边同乘以  $q(Z)$ , 并对  $Z$  求和

$$\sum_z q(Z) \ln p(X | \theta) = \sum_z [q(Z) (\ln \frac{p(X, Z | \theta)}{q(Z)} - \ln \frac{p(Z | X, \theta)}{q(Z)})]$$

- 由于  $Z$  与  $p(X | \Theta)$  独立且  $\sum_z q(Z) = 1$  于是,

$$\ln p(X | \theta) = \sum_z q(Z) \ln \frac{p(X, Z | \theta)}{q(Z)} - \sum_z q(Z) \ln \frac{p(Z | X, \theta)}{q(Z)}$$

# 引入两个函数

$$\ln p(X | \theta) = \sum_z q(Z) \ln \frac{p(X, Z | \theta)}{q(Z)} - \sum_z q(Z) \ln \frac{p(Z | X, \theta)}{q(Z)}$$

- 写作
- 其中：

$$\ln p(\mathbf{X} | \theta) = \mathcal{L}(q, \theta) + \text{KL}(q \| p)$$

$$\mathcal{L}(q, \theta) = \sum_z q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z} | \theta)}{q(\mathbf{Z})} \right\}$$

这部分之所以不写成KL距离是因为其意义不够明显

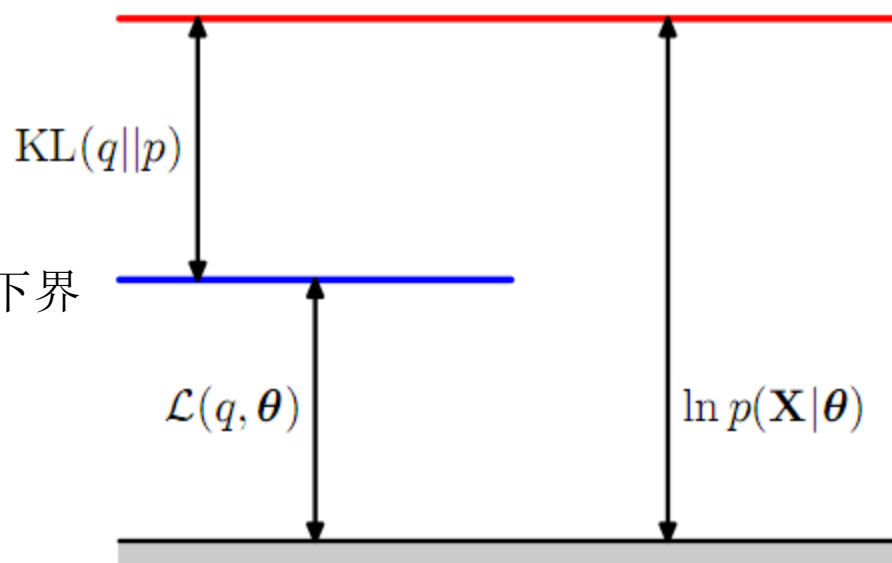
$$\text{KL}(q \| p) = \sum_z q(Z) \ln \frac{q(Z)}{p(Z | X, \theta)}$$

# KL-divergence

- Kullback–Leibler divergence (also information divergence, information gain, or relative entropy)
- KL距离、信息散度、信息增益、相对熵、交叉熵
- 表示两个概率分布之间差异程度
- 性质
  - $KL(q||p) \neq KL(p||q)$
  - $KL(q||p) \geq 0$ , 等号在 $p=q$ 时成立
  - 虽然是“距离”但不满足三角不等式

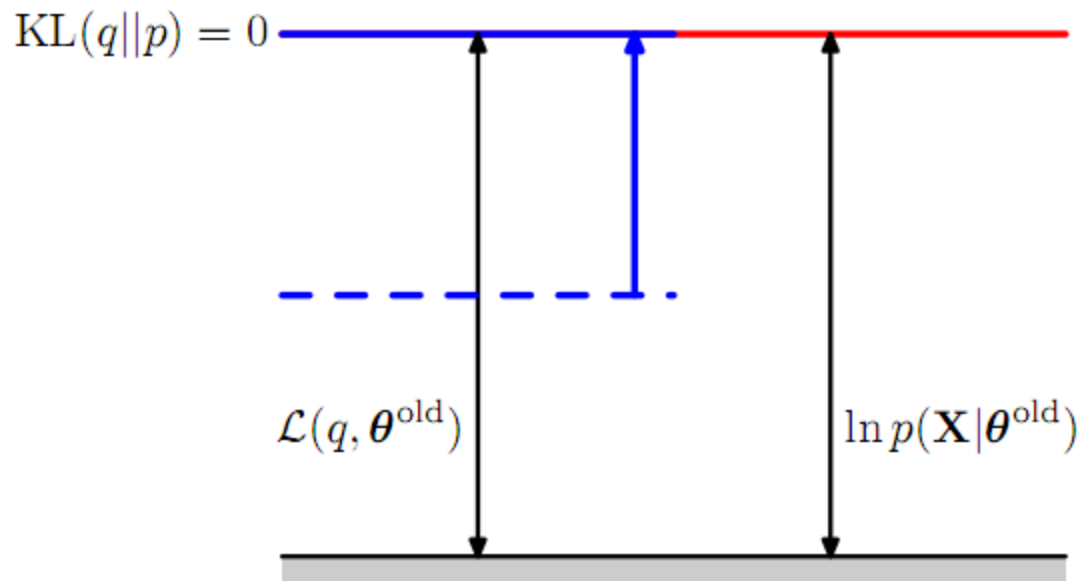
# 图示

$L(q, \Theta)$  是  $\ln p(\mathbf{X} | \Theta)$  的下界



# E-step

- 假设当前的参数为 $\Theta_{\text{old}}$ , 则E-step可以被描述为: **固定 $\Theta_{\text{old}}$  找一个分布 $q(\mathbf{Z})$ , 使得 $L(q, \Theta_{\text{old}})$ 最大。**
- 由于 $\ln p(\mathbf{X} | \Theta_{\text{old}})$ 与 $\mathbf{Z}$ 无关, 则使 $L(q, \Theta_{\text{old}})$ 最大 **即: 使 $KL(q||p)$ 最小(=0), 也就是说 $q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X}, \Theta_{\text{old}})$**



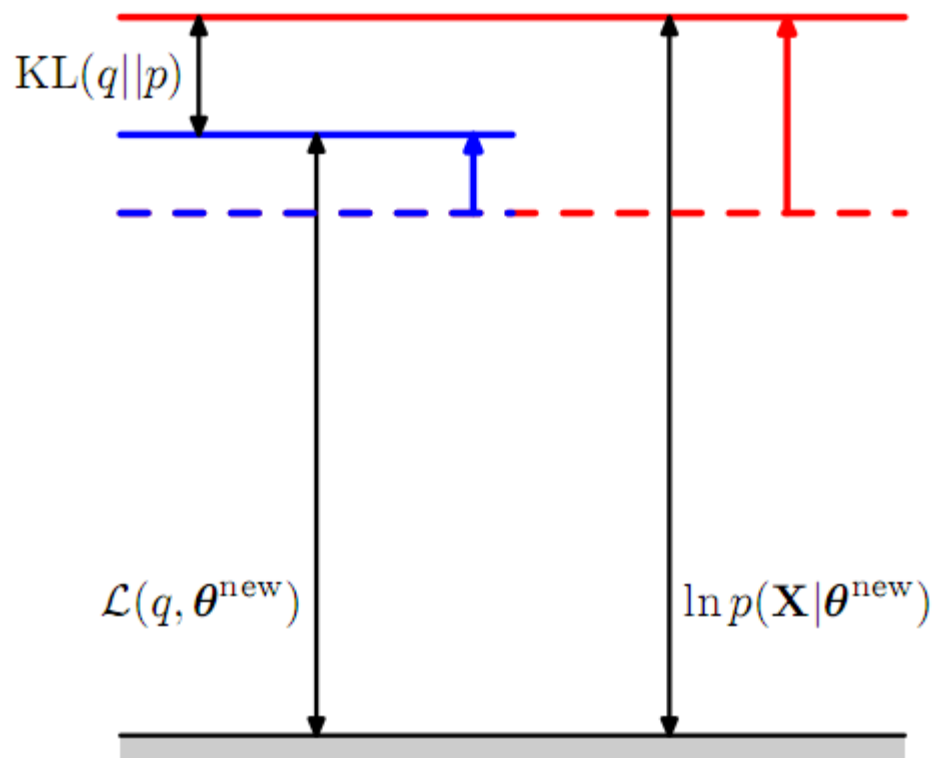
# Revisit

为什么要用  $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$  去求  $E[\mathbf{Z}]$ ?

- A: 容易看出选这个分布可以使  $KL(q||p)=0$

# M-Step

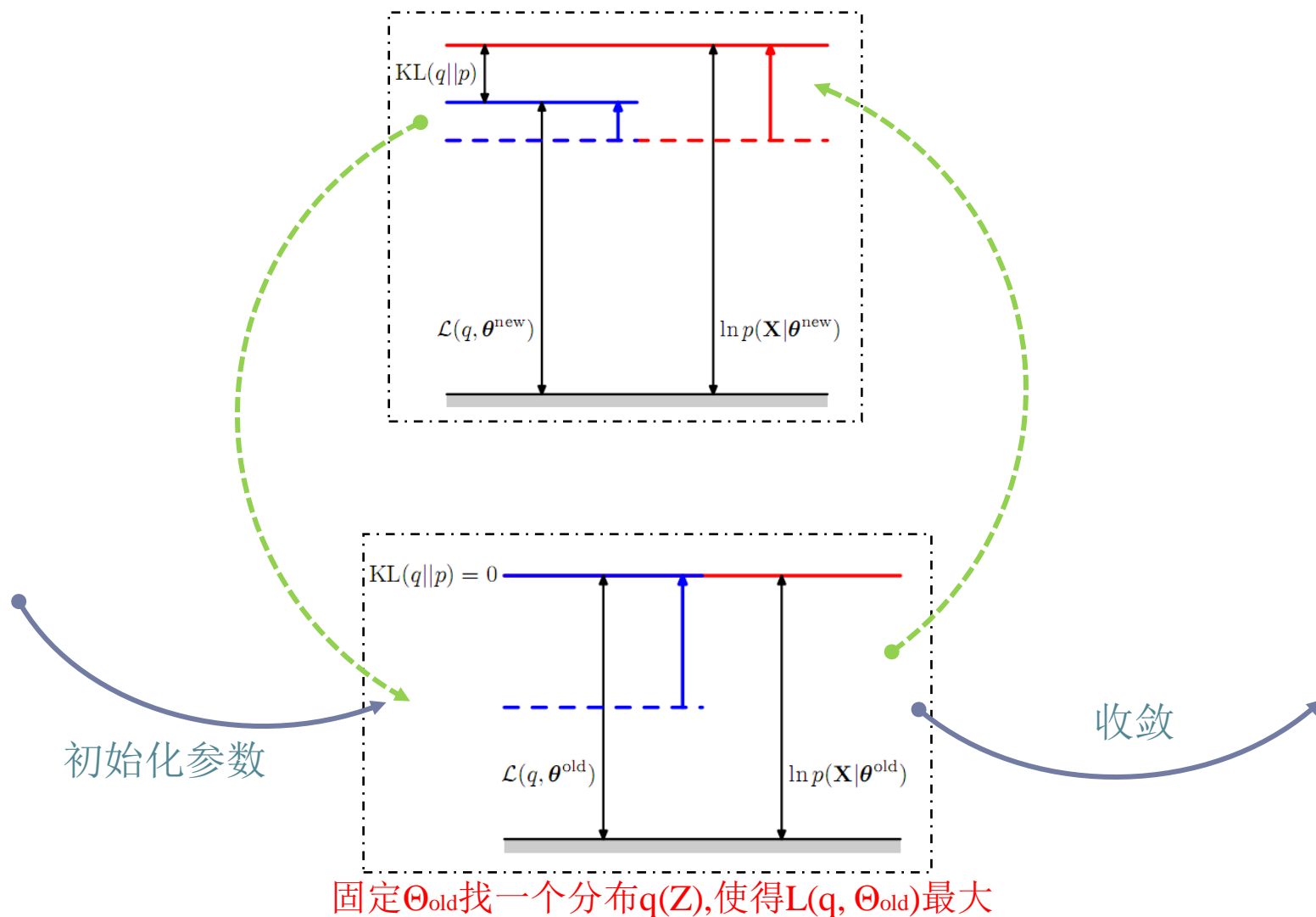
- M-step可以被描述为:固定 $q(Z)$  找一个组参数 $\Theta_{\text{new}}$ , 使得 $L(q, \Theta_{\text{new}})$ 最大
- $\ln p(X | \theta)$  的增大可能来自于两部分:  $L(q, \Theta_{\text{new}})$ 和 $\text{KL}(q||p)$  (因为此时 $p(Z|X, \Theta_{\text{new}})$ 而 $q(Z|X, \Theta_{\text{old}})$ ,  $p \neq q$ 所以 $\text{KL}(q||p) \geq 0$ )



# EM 算法

Xiao Han 2008/10/20

固定 $q(Z)$  找一个组参数 $\Theta_{\text{new}}$ ,使得 $L(q, \Theta_{\text{new}})$ 最大



# 对于最大化目标的解释

- Q:既然要最大化 $p(\mathbf{X}|\Theta)$ ,为什么不直接最大化 $\ln p(\mathbf{X}|\Theta)$
- A:正如前面混合高斯分布的引例中看到的,直接最大化 $\ln p(\mathbf{X}|\Theta)$ (比如利用求导的方法)往往是件困难/复杂的事情

# 对于最大化目标的解释

- Q1:怎么想到要将 $\ln p(X|\Theta)$ 拆成这样两项的?  
Q2:为什么要最大化 $L(q, \Theta)$ ?
- A: 如前所述求 $\ln p(X|\Theta)$ 的最大值是件困难的事情, 我们想能不能通过逐步提升 $\ln p(X|\Theta)$ 的下界, 即 $L(q, \Theta)$ , 来找到 $\ln p(X|\Theta)$ 的最大值。因为 $L(q, \Theta)$ 是比较容易求最大值的。

## \*对于“提高下界”的说明

- Q: “提高下界”找最大值的方法比较抽象，如何理解？
- A: 好比我们为一个建筑照相，我们需要找到一个最佳的**照相模式**，使得照片的效果最好，可是这个模式我们是不知道的。
- 但是我们知道：照相模式由两部分组成：1.照相人所站的位置， 2.照相机的参数。
  - a. 我们拿着相机随便站在一个地方
  - b. 先固定住照相机的参数，找一个照相的位置，使得照相效果**最好**(相当于我们提升了照相效果的下界)
  - c. 再在找到的位置上，调整照相机的参数，使得照相效果**更好**
- 重复b,c两个过程，我们就可以找到最佳的照相模式

# 对于最大化目标的解释

- Q:混合高斯分布的例子中，我们一直在最大化 $\ln p(\mathbf{X}, \mathbf{Z} | \Theta)$ ，在EM一般性描述中我们却一直在最大化一个 $L(q, \Theta)$ 函数，这个怎么解释呢？
- A:根据定义

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{z} | \theta)}{q(\mathbf{z})} \right\}$$

- 在E-step后，我们将 $q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X}, \Theta_{\text{old}})$ 代入上式，可得

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{z}} \underline{p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})} \ln p(\mathbf{X}, \mathbf{z} | \theta) - \sum_{\mathbf{z}} \underline{p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})} \ln p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$$

- 注意到M-step是固定 $q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X}, \Theta_{\text{old}})$ ，寻找 $\Theta_{\text{new}}$ ，因此红线处在M-Step中都是常数。因此，**最大化 $L(q, \Theta)$** ，就是在**最大化 $\ln p(\mathbf{X}, \mathbf{Z} | \Theta)$**

# 对于最大化目标的解释

- Q:我们可不可以利用EM算法最大化 $\ln p(\Theta | X)$
- A:可以。但是我们需要先验概率 $p(\Theta)$

$$\begin{aligned}\ln p(\theta | X) &= \ln p(X, \theta) - \ln p(X) \\ &= \ln p(X | \theta) + \ln p(\theta) - \ln p(X) \\ &= L(q, \theta) + KL(q \| p) + \ln p(\theta) - \ln p(X)\end{aligned}$$

- E-step时， $\ln p(\Theta)$ 和 $\ln p(X)$ 均为常数，因此我们做同样的处理：固定 $\Theta_{old}$ 找一个分布 $q(Z)$ ,使得 $L(q, \Theta_{old})$ 最大
- M-step时，固定 $q(Z)$  找一个组参数 $\Theta_{new}$ ,使得 $L(q, \Theta_{new}) + \ln p(\Theta_{new})$ 最大。

# Thinking in EM

- **K-means**: 一种使用EM思想求混合“高斯分布”参数方法，却往往被简单描述成“一种聚类方法”
- **K-means**忽略了高斯分布的方差影响
- **K-means**被用在图像压缩，Image Segmentation

# 简化的高斯混合模型

- 在这里，我们仍然认为观察到的数据符合以下混合高斯模型

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- 但此时，我们认为这k个高斯分布的协方差矩阵是相等的，均记做 $\epsilon \mathbf{I}$ ，其中 $\epsilon$ 为已知的常数， $\mathbf{I}$ 单位矩阵。
- 我们要求的参数只有 $\pi, \mu$

# 简化的高斯混合模型

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- 代入 $\boldsymbol{\Sigma}_k = \epsilon \mathbf{I}$  可得

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp \left\{ -\frac{1}{2\epsilon} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 \right\}.$$

- 和在简化问题中一样，目的找一组参数（ $\boldsymbol{\pi}, \boldsymbol{\mu}$ ）使得下式最大

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}.$$

## 简化的期望E[Z]

- 由于Z是隐藏变量，我们需要用E[Z]去估计Z的值
- 正如ppt第22页中推导E[Z]的方法，只不过在最后我们可以写出一个更简单的式子

$$E(z_{nk} | x_n) = \dots = \frac{\pi_k \cdot N(x_n | \mu_k, \Sigma_k)}{\sum_j \pi_j \cdot N(x_n | \mu_j, \Sigma_j)}$$

$$= \frac{\pi_k \cdot \exp\left(-\frac{\|x_n - \mu_k\|^2}{2\varepsilon}\right)}{\sum_j \pi_j \cdot \exp\left(-\frac{\|x_n - \mu_j\|^2}{2\varepsilon}\right)} = \gamma(z_{nk})$$

ε就是前面所说的ε

# $\varepsilon$ 趋于0时

$$E(z_{nk} | x_n) = \frac{\pi_k \cdot \exp\left(\frac{-\|x_n - \mu_k\|^2}{2\varepsilon}\right)}{\sum_j \pi_j \cdot \exp\left(\frac{-\|x_n - \mu_j\|^2}{2\varepsilon}\right)} = \gamma(z_{nk})$$

- 可以想象，对于一个 $x_n$ ，我们遍历所有的 $\mu_j$ ，只有对于那些使得 $\|x_n - \mu_j\|^2$ 也趋于0的项，

$\pi_j \cdot \exp\left(\frac{-\|x_n - \mu_j\|^2}{2\varepsilon}\right)$  才有值，其余的都为0

$$E(z_{nk} | x_n) = \frac{\pi_k}{0 + \dots + 0 + \pi_k + 0 + \dots + 0} = 1$$

# E[Z<sub>nk</sub>]的进一步简化

在此再做一个转化，使得E[Z<sub>nk</sub>]进一步简化

对于一个 $\mathbf{x}_n$ ，遍历所有的 $\mu_j$ ，找使得  $\|\mathbf{x}_n - \mu_j\|^2$  趋于0的项

对于一个 $\mathbf{x}_n$ ，遍历所有的 $\mu_j$ ，找使得  $\|\mathbf{x}_n - \mu_j\|^2$  最小的项

最终的E[Z<sub>nk</sub>]可以被简化为

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

# K-means的EM方法

- 得到 $E[Z_{nk}]$ 之后，代入ppt24页各式中，但这次，我们不用再估计 $\Sigma_k$ 了

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

所有属于第k个高斯分布的观察点的均值

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

当前有多少个观察点属于第k个高斯分布

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

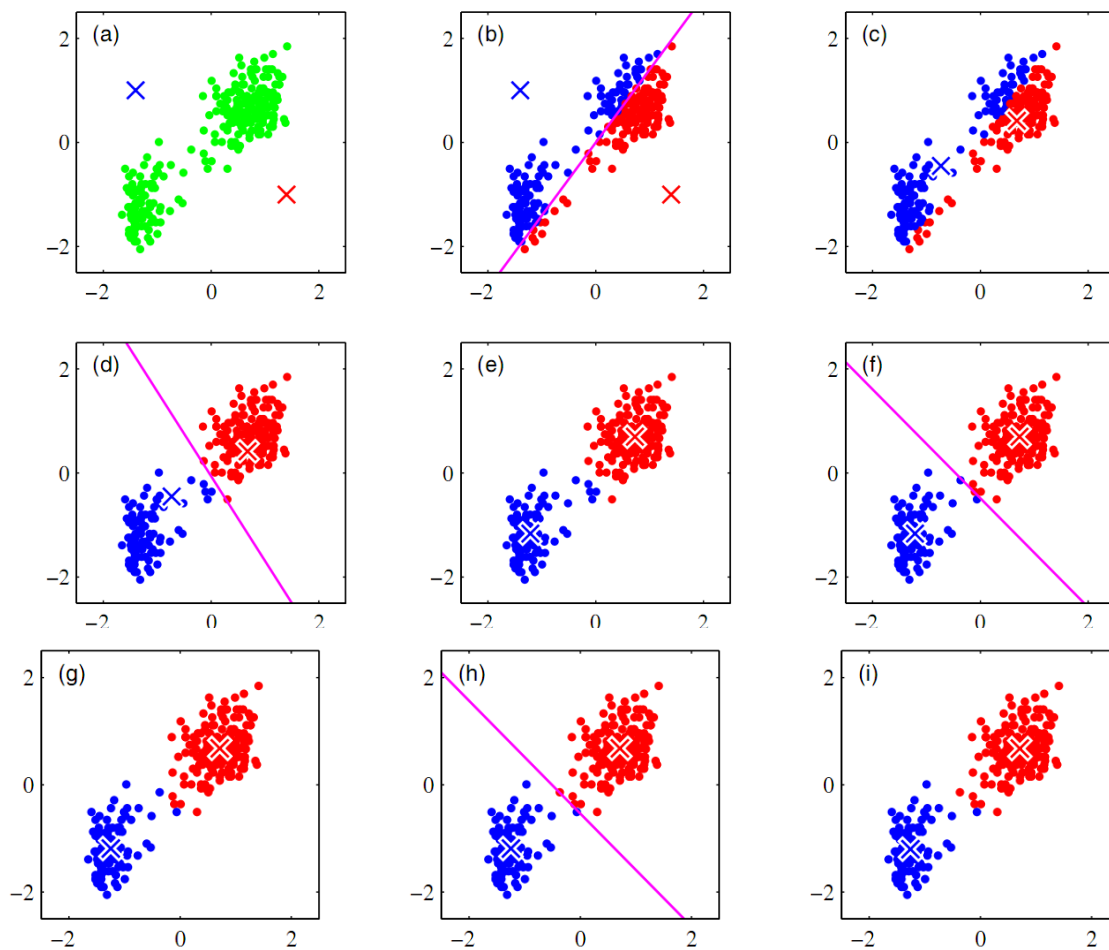
- 由ppt15-17页各式，得到此时的最大化目标为

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \rightarrow -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 + \text{const.}$$

E-Step

M-Step

# K-means的步骤图示



# 使用K-means算法进行Image Segmentation

 $K = 2$  $K = 3$  $K = 10$ 

Original image



# FAQ

Thanks!

Please contact me when you find mistakes in the slides.

[artex.xh@gmail.com](mailto:artex.xh@gmail.com)