Chapter 9 Mixture Models and Expectation Maximization

目录

- K-means聚类算法
- 高斯混合模型
- EM的另一种观点
- 通用的EM算法

聚类

- 输入：数据集$\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$，由一个随机的$D$维欧式变量$\mathbf{x}$的$N$条观测数据所构成

- 目标：将数据集划分成K个聚簇

- 如何做?
  - 聚簇由**一组数据点**构成的，聚簇内的数据点之间距离小，聚簇间数据点之间的距离大
  - 引入一组D维向量$\boldsymbol{\mu}_k$（$k=1,\ldots,K$），$\boldsymbol{\mu}_k$是与第k个聚簇关联的**原型(prototype)**
  - 目标就是找出数据点到聚簇的一种分配，以及一组向量$\{\boldsymbol{\mu}_k\}$，使得**每个数据点到它最近向量的距离的平方和最小**

目标函数

- 对于每个数据点$\mathbf{x}_n$，我们引入一个相应的二值指示变量r_{nk}描述了数据点将被分配给K个聚簇中的哪一个
  - 如果数据点$\mathbf{x}_n$分配到了聚簇k，那么$r_{nk}=1$，而对于$j \neq k$则有$r_{nk}=0$
- 目标函数（失真度量）：

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2$$

  - 这表示了每个数据点$\mathbf{x}_n$到它被分配的向量$\boldsymbol{\mu}_k$的距离平方和
  - 任务是找出$\{r_{nk}\}$和$\{\boldsymbol{\mu}_k\}$的取值，以便最小化J

K-means算法
迭代求解

- 为了完成此项任务，我们使用了一个迭代过程，每次迭代都包括了两个连续的步骤，对应于针对r_{nk}和$\boldsymbol{\mu}_k$的相继优化
  - 首先，为$\boldsymbol{\mu}_k$选择某个初始值
  - 在第一阶段，保持$\boldsymbol{\mu}_k$固定，针对r_{nk}来最小化J
  - 在第二阶段，我们保持r_{nk}固定，针对$\boldsymbol{\mu}_k$来最小化J。
  - 这两个阶段一直重复，直到收敛。
- 这两个阶段分别对应于EM算法的E步骤和M步骤

K-means算法
E步骤（重新分配）

- 问题：如何确定r_{nk}?

- J是r_{nk}的一个线性函数，且涉及不同n的项相互独立

- 闭合式解：将第n个数据点付给最近的聚簇中心

$$r_{nk} = \begin{cases} 1 & if\ k = \arg\min_j \left\| \mathbf{x}_n - \boldsymbol{\mu}_j \right\|^2 \\ 0 & otherwise \end{cases}$$

K-means算法
M步骤（聚簇中心的更新）

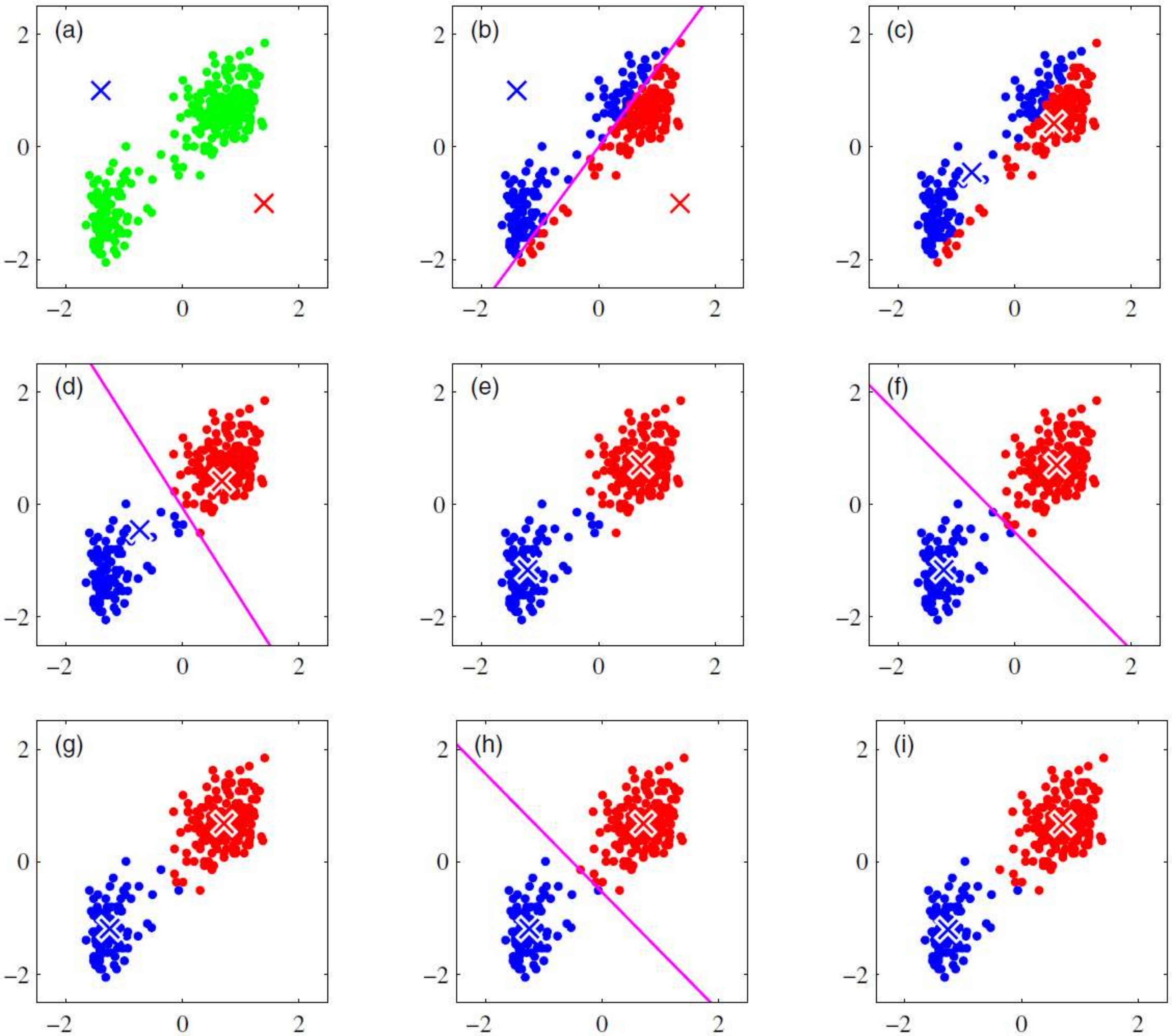
- 问题：如何在保持r_{nk}固定时来优化$\boldsymbol{\mu}_k$?

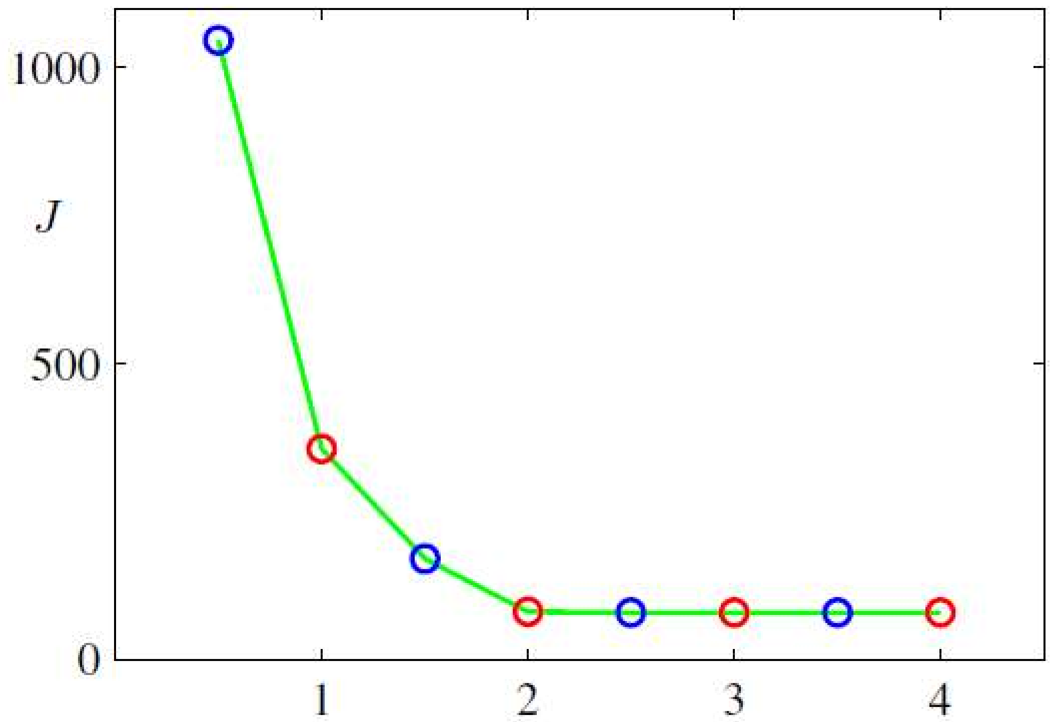
- 目标函数J是$\boldsymbol{\mu}_k$的二次函数，通过将其对于$\boldsymbol{\mu}_k$的导数设为0就可以对其进行最小化

- 闭合式解：

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \Rightarrow \boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

即：将$\boldsymbol{\mu}_k$设置为所有分配给聚簇k的数据点的均值


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
2
0
−2
−2
0
2



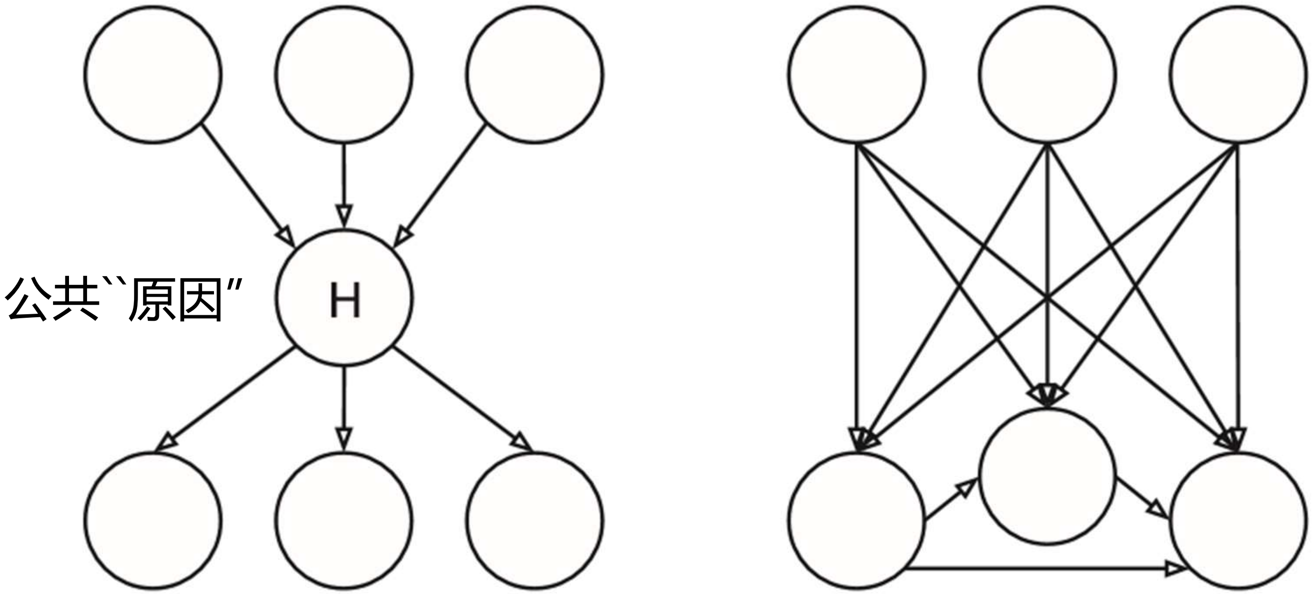
1000
500
0
J
1
2
3
4

存在的问题

- 限制了**数据变量的类型**→K-modes算法
- 对**离群点过于敏感**→ K-medoids算法

目录

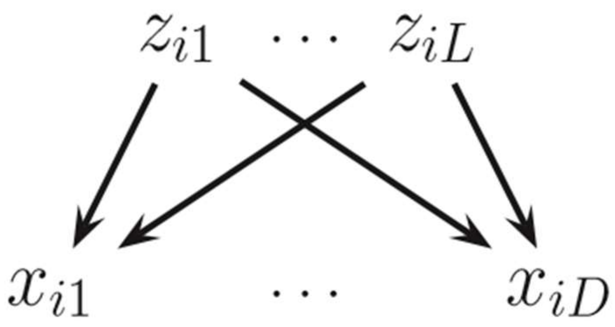
潜变量模型
Latent Variable Model (LVM)



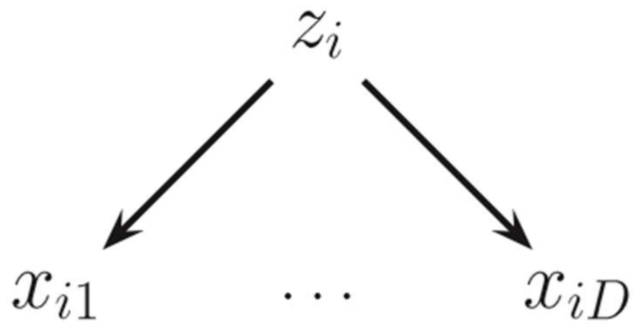
17 parameters　　　59 parameters

潜变量模型

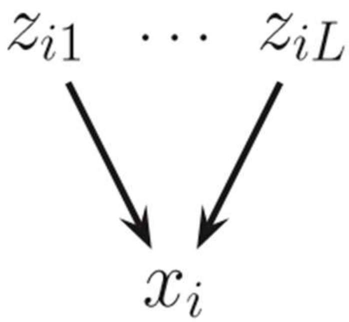
多对多、多对一、一对多、一对一



(a)



(b)



(c)

(d)

具体模型：似然函数与先验概率

$p(\mathbf{x}_i \mid \mathbf{z}_i)$	$p(\mathbf{z}_i)$	Name
MVN	Discrete	Mixture of Gaussians
Prod. Discrete	Discrete	Mixture of multinomials
Prod. Gaussian	Prod. Gaussian	Factor analysis/ probabilistic PCA
Prod. Gaussian	Prod. Laplace	Probabilistic ICA/ sparse coding
Prod. Discrete	Prod. Gaussian	Multinomial PCA
Prod. Discrete	Dirichlet	Latent Dirichlet allocation
Prod. Noisy-OR	Prod. Bernoulli	BN20/ QMR
Prod. Bernoulli	Prod. Bernoulli	Sigmoid belief net

混合模型
Mixture Models

- 混合模型：最简单的潜变量模型——$z_i \in \{1, \ldots, K\}$表示一个离散潜状态
  - 离散先验$p(z_i) = Cat(\boldsymbol{\pi})$
  - 似然函数$p(\mathbf{x}_i | z_i = k) = p_k(\mathbf{x}_i)$，这里$p_k$是第$k$个基分布(base distribution)

- 混合模型：

$$p(\mathbf{x}_i | \theta) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i | \theta)$$

高斯混合分布
Gaussian Mixture Distribution

- 高斯混合：使用最广泛的混合模型，每个基分布都是一个高斯分布
- 高斯分布的线性叠加(linear superposition)

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

- 我们引入一个K维二值随机变量$\mathbf{z}$，它具有“K中选1 （1 of K）”的表示（即只有一个特殊的元素z_k等于1，而其它的元素都为0）

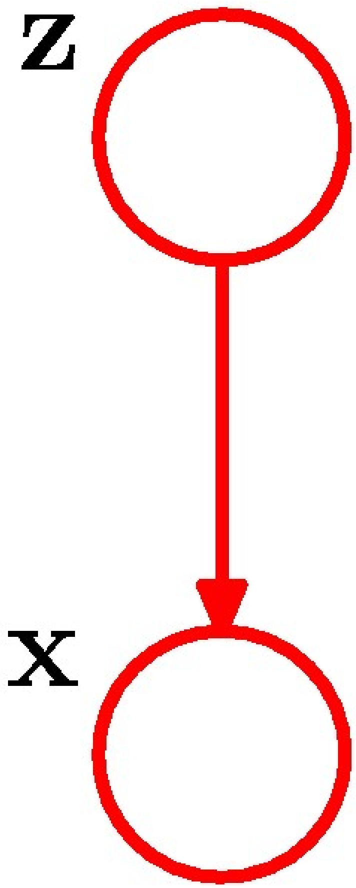
联合分布与图模型

- 联合概率的分解：

$$p(\mathbf{x},\mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$$

边缘分布 $p(\mathbf{z})$ ？

条件分布$p(\mathbf{x}|\mathbf{z})$？



$\mathbf{z}$上的边缘分布
Marginal Distribution over $\mathbf{z}$

- z上的边缘概率分布被指定为混合系数

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

- $\mathbf{z}$使用"K中择一"表示法，因此

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

x的条件分布

- 给定z的具体值，x的条件概率分布是一个高斯分布

$$p(\mathrm{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

这也可以写成如下形式：

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

$\mathbf{x}$的边缘分布

- **联合分布**$p(\mathbf{x},\mathbf{z})$:

$$p(\mathbf{x},\mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

- **于是，**$\mathbf{x}$的边缘分布就具有**高斯混合**的形式

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

响应度
Responsibility

- 给定x时**z的条件概率**：

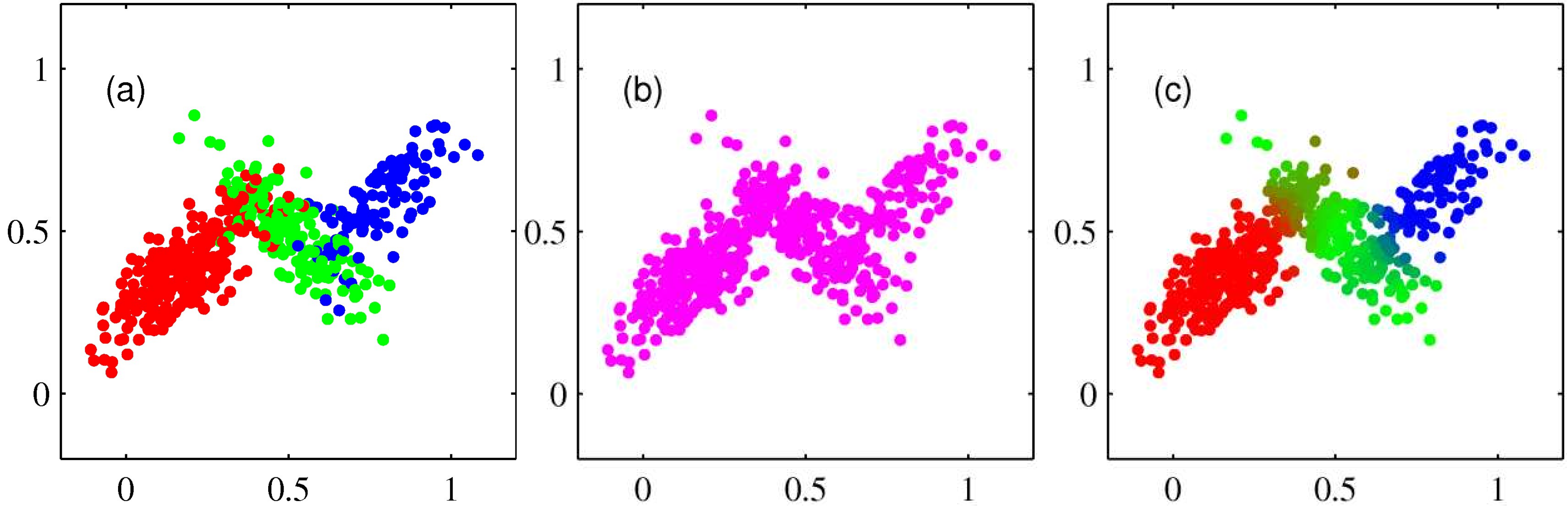
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

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

$\gamma(z_k)$也可以被看作成员k对于解释观察数据x所采取的**响应度**（或者所承担的**责任**）

数据生成过程

- 生成依照高斯混合模型分布的随机样本：
  - **首先**根据边缘分布$p(\mathbf{z})$来生成一个$\mathbf{z}$值（记为$\hat{\mathbf{z}}$），
  - **而后**根据条件分布$p(\mathbf{x}|\hat{\mathbf{z}})$来生成一个$\mathbf{x}$值。



(a)
(b)
(c)
1
0.5
0
0
0.5
1

最大似然
Maximum Likelihood

- 给定观察数据的集合$\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$，我们希望用高斯混合对该数据进行建模
  - 数据集的表示：$N \times D$的矩阵$\mathbf{X}$。
  - 相应潜变量的表示： $N \times K$的矩阵$\mathbf{Z}$。

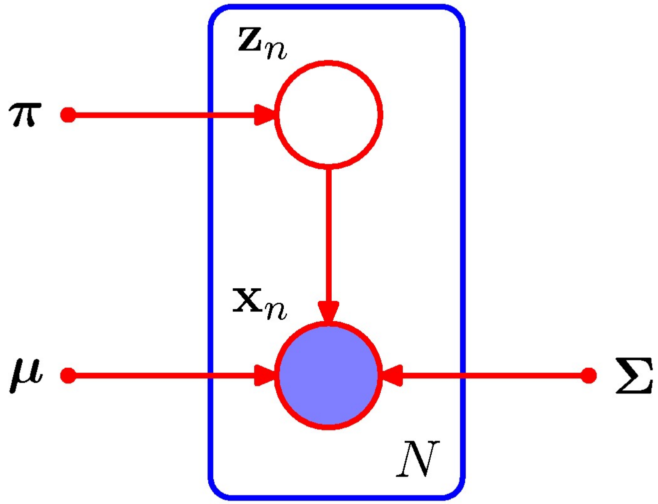
- 对数似然为：

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

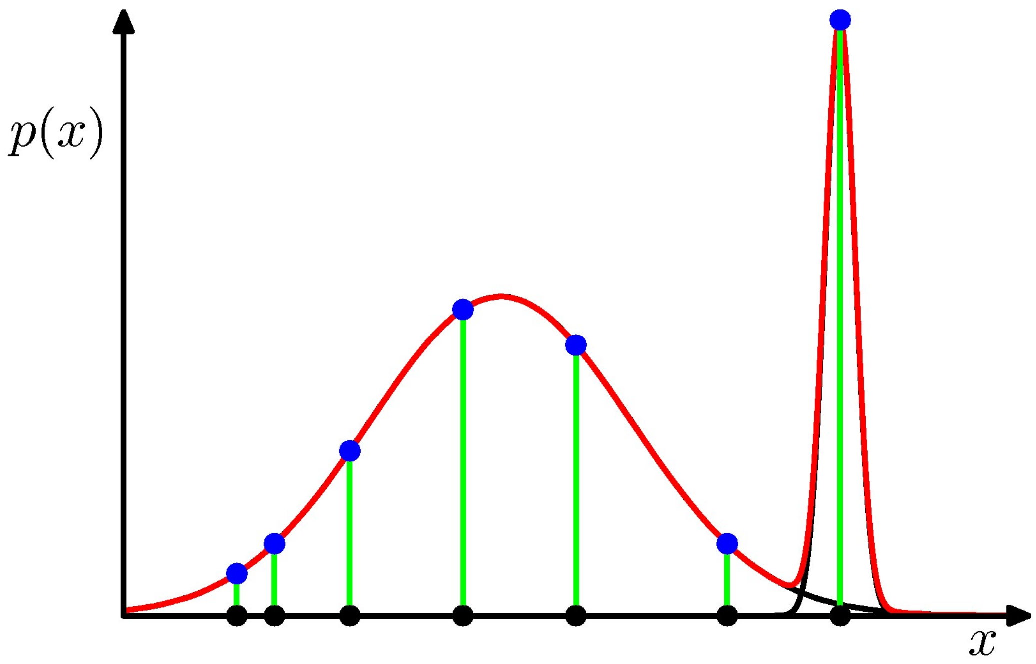
  - 这里我们假定：数据点是独立同分布的

高斯混合的图表示

N个数据点是独立同分布的



高斯混合中的奇异性
Singularity



最大似然难以直接求解

- 困难出现的原因：对数中出现了 k 上的累加，因此对数函数不再直接作用在高斯分布上

- EM算法：
  - 具有**广泛的可用性**（适用于具有潜变量的概率模型）
  - 可以为**变分推理**技术奠定基础

EM算法

对数似然对于均值的导数

- 对数似然函数

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

- 将$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$相对于高斯成员的均值$\boldsymbol{\mu}_k$的导数置为0，我们可以得到

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}}_{\gamma(z_{nk})} \Sigma_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

EM算法
对数似然对于均值的导数

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

其中我们定义

- $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$
- 这可以被解释为分配给聚簇k的样本点的有效数目。

EM算法
协方差矩阵

- 设$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})$对于$\boldsymbol{\Sigma}_k$的导数为0，可得：

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

EM算法

混合系数π_k

- 最后，我们要针对混合系数π_k来最大化$\ln p(X|\pi,\mu,\Sigma)$。这里我们必须要把约束(9.9)考虑进来，即要求所有混合系数累加为1。

$$\ln p(X|\pi,\mu,\Sigma)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)$$

即：

$$0=\sum_{n=1}^{N}\frac{\mathcal{N}\left(x_n\middle|\mu_k,\Sigma_k\right)}{\sum_j\pi_j\mathcal{N}\left(x_n\middle|\mu_j,\Sigma_j\right)}+\lambda$$

- 如果我们在两边都乘以π_k，并且使用约束(9.9)在k上累加，我们可以得到$\lambda=-N$。

EM算法
混合系数π_k

- 使用它去消去λ，并重新组织，我们得到：

$$0 = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}\left(x_n \middle| \mu_j, \Sigma_j\right)} + \lambda \pi_k$$

$$\Leftrightarrow 0 = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}\left(x_n \middle| \mu_j, \Sigma_j\right)} - N\pi_k$$

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

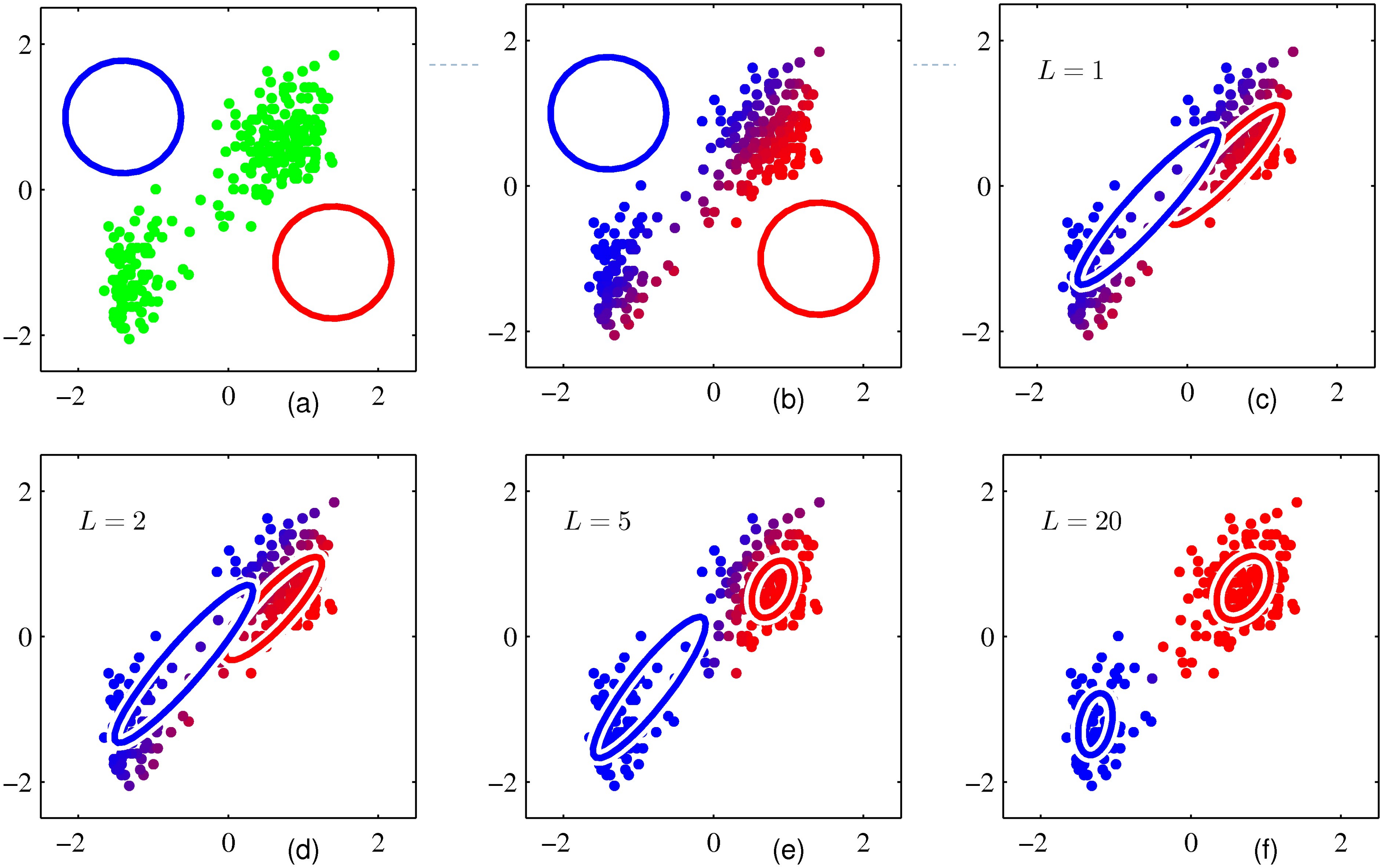
EM算法

- 上述结论并未构成混合模型参数的闭合式解，
  - 因为响应度(responsibilities)γ_{nk}以一种复杂的方式反过来依赖于那些参数

- 以一种简单的迭代方式来寻找最大似然解！

EM算法
迭代过程

- 首先为均值、协方差、混合系数选择某些初始值；然后迭代E步骤和M步骤
  - E步骤：使用当前的参数值来计算后验概率（或响应度）；
  - M步骤：使用这些概率来重新估计均值、协方差以及混合系数
- 如果对数似然或者参数的变化量小于某个设定阈值，则我们认为它已经收敛。



L = 1
L = 2
L = 5
L = 20
(a)
(b)
(c)
(d)
(e)
(f)

高斯混合的EM算法

步骤1： 初始化均值μ_k，协方差Σ_k以及混合系数π_k，并评估对数似然的初始值。

步骤2（E-步骤）： 使用当前参数值来评估责任度(responsibilities)

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

步骤3（M-步骤）： 使用当前责任度来重新估计参数：

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

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

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

其中

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

步骤4： 评估对数似然，并检查参数或对数似然的收敛性

目录

- K-means聚类算法
- 高斯混合模型
- EM的另一种观点
- 通用的EM算法

EM的另一种观点
An Alternative View of EM

- **EM算法的目标：为具有潜变量的模型找出最大似然解**

- 对数似然函数由下式给出：

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln\left\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\right\}$$

对数似然难以优化：**潜变量上的累加求和出现在对数的里面**

- 称$\{\mathbf{X},\mathbf{Z}\}$为完全数据集(complete dataset)，而实际的观测数据$\mathbf{X}$为不完全的(incomplete)

假定：完全数据对数似然的最大化很简单

- 潜变量Z取值的知识状态：仅仅是由后验分布 $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$ 所给出的
  - E步骤：求取完全数据对数似然在后验分布下的期望值

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})=\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$$

- M步骤：最大化完全数据对数似然的期望值

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

   where

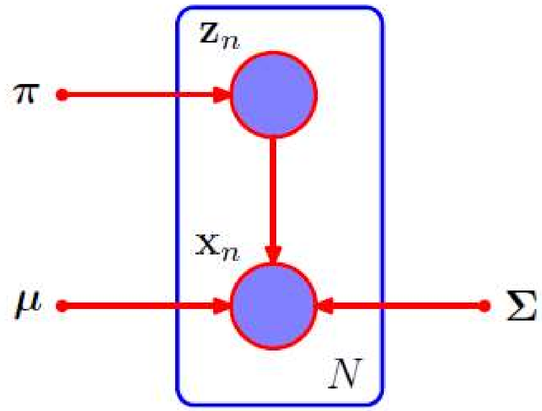
$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

   and return to step 2.

重访：高斯混合



完全数据集$\{\mathbf{X},\mathbf{Z}\}$的似然函数：

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

取对数，我们得到：

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

对数现在直接作用在高斯分布上，高斯分布自身隶属于指数族，这就得到了一个简单得多的最大似然问题。

重访：高斯混合
完全数据对数似然的最大化

- 针对均值和协方差的优化：等同于单个高斯的优化

- 混合系数的优化：

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

潜变量的后验分布

- **潜变量的后验分布**

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

- **潜变量的期望值**

$$\begin{aligned} \mathbb{E}[z_{nk}] &= \frac{\sum_{z_{nk}} z_{nk} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}}{\sum_{z_{nj}} [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk}) \end{aligned}$$

与K-means的关系

目录

- K-means聚类算法
- 高斯混合模型
- EM的另一种观点
- 通用的EM算法

一般意义下的EM算法

- EM算法：针对具有潜变量的概率模型，寻求其极大似然解的一种通用技术
  - $\mathbf{X}$：所有的**观察变量**
  - $\mathbf{Z}$：所有的**潜变量**
  - **参数**$\boldsymbol{\theta}$控制了联合概率分布$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$
  - 目标：最大化似然函数

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

这里，我们假定$\mathbf{Z}$是离散的。

推导
Derivation

- 假定：
  - 直接优化$p(\mathbf{X}|\boldsymbol{\theta})$的难度很大或者不可行
  - 但是优化完全数据似然函数$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$却容易得多
- 如何去做?
  - 我们首先引入一个分布$q(\mathbf{Z})$
  - 则我们有：

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

其中，

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

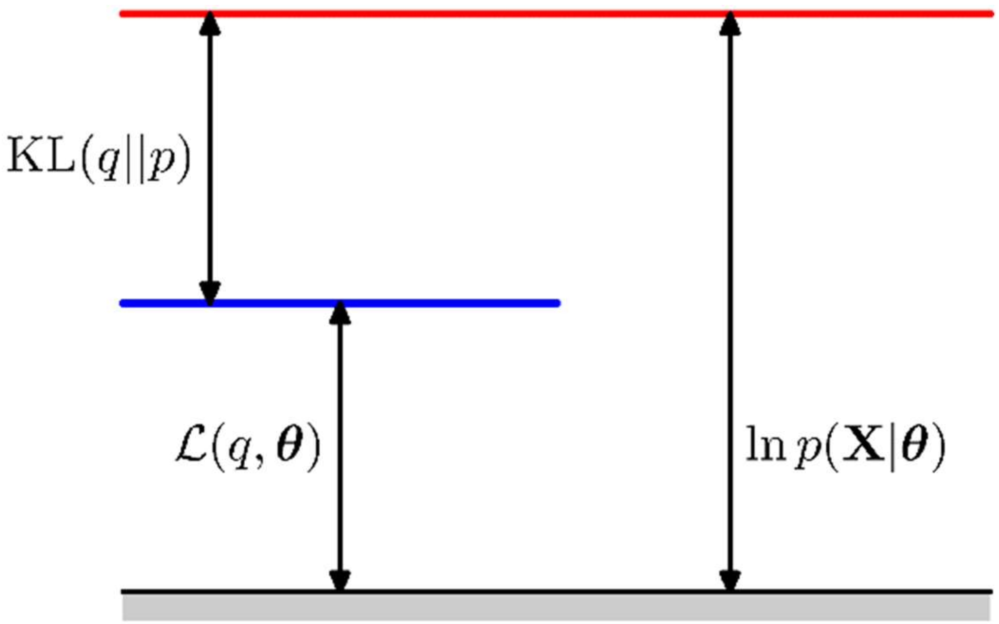
$$KL(q \parallel p) = -\sum_{Z} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \right\}$$

KL散度
KL-Divergence

- $KL(q \parallel p)$是$q(\mathbf{Z})$与$p(\mathbf{Z}|\mathbf{X},\theta)$之间的KL散度，因此$KL(q \parallel p) \geq 0$
  - 等号当且仅当$q(Z) = p(Z|X,\theta)$时成立。

- 于是，$\mathcal{L}(q,\boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$，换言之$\mathcal{L}(q,\boldsymbol{\theta})$是$\ln p(\mathbf{X}|\boldsymbol{\theta})$的一个下界。

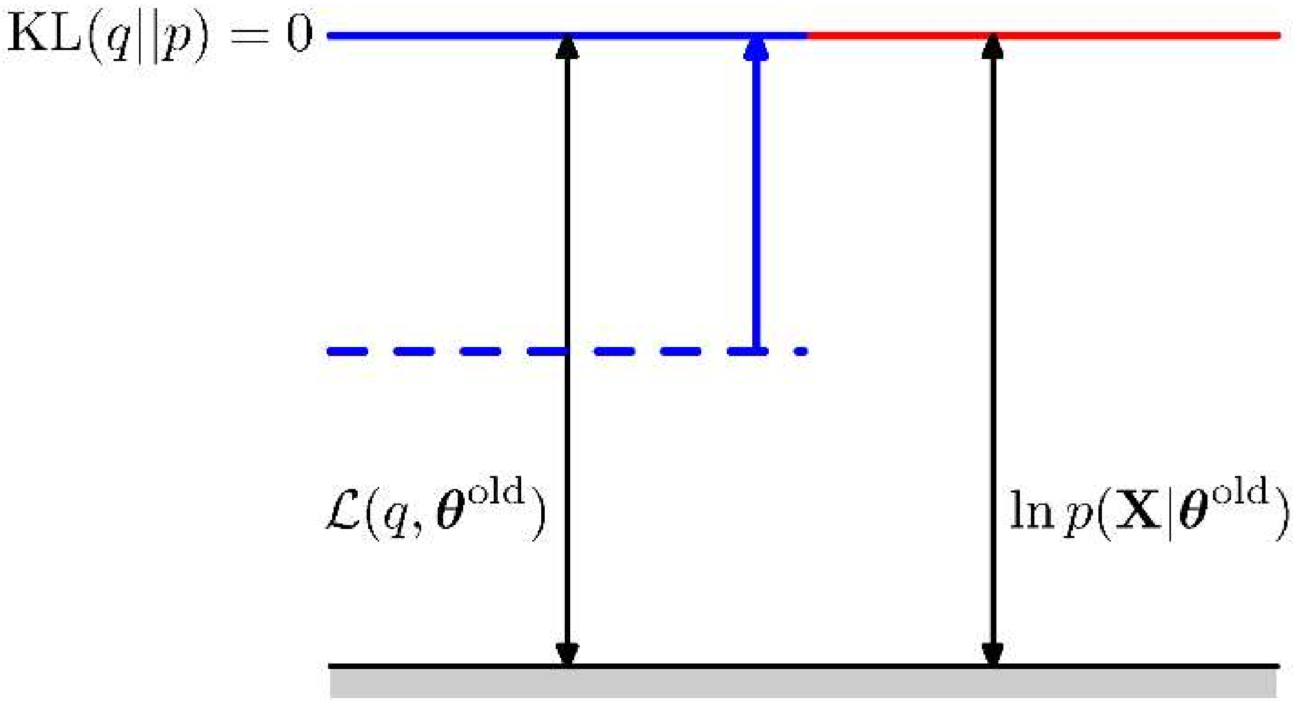
图示：对数似然的分解


$\mathrm{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

EM算法：E-步骤

- 在E-步骤中，保持$\boldsymbol{\theta}^{old}$固定不变，针对$q(\mathbf{Z})$来最大化下界$\mathcal{L}(q,\boldsymbol{\theta}^{old})$
  - 由于$\ln p(\mathbf{X}|\boldsymbol{\theta}^{old})$ 并不依赖于$q(\mathbf{Z})$，
  - 因此：当KL-Divergence消失时（即当 $q(\mathbf{Z})$等于后验分布 $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$的时候），$\mathcal{L}(q,\boldsymbol{\theta}^{old})$的最大值就出现了。

图示：E-步骤

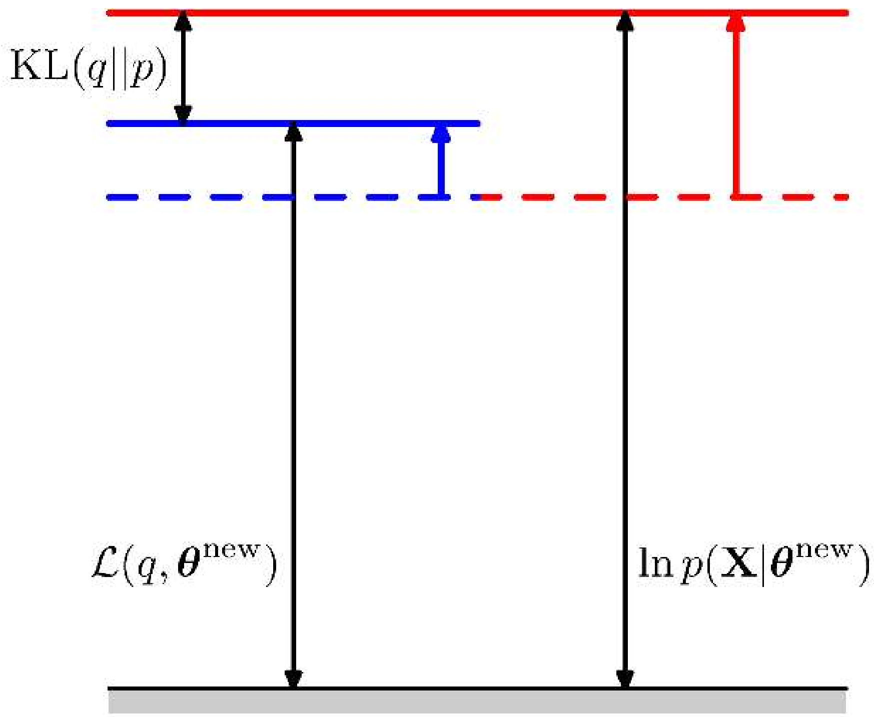


$\mathrm{KL}(q\|p) = 0$
$\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$

EM算法：M-步骤

- 在M-步骤中，分布$q(\mathbf{Z})$被保持不变，针对$\boldsymbol{\theta}$来最大化下限$\mathcal{L}(q,\boldsymbol{\theta}^{old})$，从而得到某个新的值$\boldsymbol{\theta}^{new}$。
  - 这将使得下限$\mathcal{L}$增大 (除非它已经到达最大值)，这必然导致相应对数似然的增大
  - 由于分布q是使用旧参数值来确定的（而不是新参数值）且它在M-步骤中保持不变，它并不等于新的后验分布，因此将有一个非0的KL散度。
  - 对数似然的增加量因此将大于下界的增加量

图示：M-步骤



$\mathrm{KL}(q||p)$
$\mathcal{L}(q,\boldsymbol{\theta}^{\mathrm{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{new}})$

- 将$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$代入

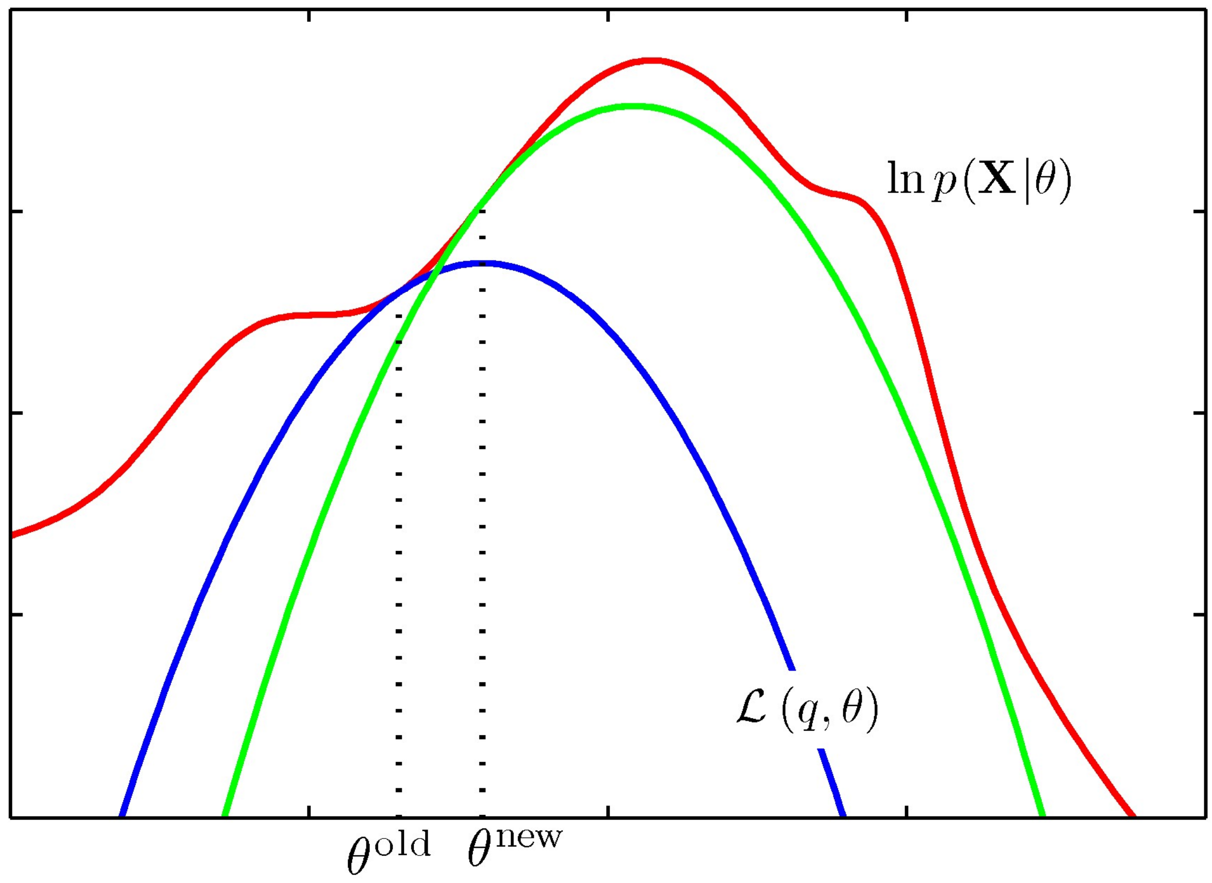
$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

可得:

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p\left(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}\right) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p\left(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}\right) \ln p\left(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}\right)$$

$$= \mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) + const$$

- 这里的常量是q分布的负熵，因此是独立于θ的。因此，在M-步骤中需要最大化的量是**完全数据对数似然**的期望值。


$\ln p(\mathbf{X}|\theta)$
$\mathcal{L}(q, \theta)$
θ^{old}
θ^{new}

概率的因子分解

- 独立同分布数据集：$\mathbf{X}$由N个数据点$\{\mathbf{x}_n\}$构成，而$\mathbf{Z}$则由N个相应的潜变量$\{\mathbf{z}_n\}$构成，这里$n = 1,2,\dots,N$
- 我们有：

$$p(\mathbf{X},\mathbf{Z}) = \prod_n p(\mathbf{x}_n,\mathbf{z}_n)$$

- 在$\{\mathbf{z}_n\}$上进行边缘化，有：

$$p(\mathbf{X}) = \prod_n p(\mathbf{x}_n)$$

- E步骤中计算出来的后验概率

$$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) = \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} = \frac{\prod_{n=1}^{N} p(\mathbf{x}_n,\mathbf{z}_n|\boldsymbol{\theta})}{\sum_{\mathbf{Z}}\prod_{n=1}^{N} p(\mathbf{x}_n,\mathbf{z}_n|\boldsymbol{\theta})} = \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\theta})$$

最大后验的EM算法

- 在模型中引入了参数上的先验$p(\boldsymbol{\theta})$，使用EM算法来最大化后验分布$p(\boldsymbol{\theta}|\mathbf{X})$

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) = \ln p(\boldsymbol{\theta}, \mathbf{X}) - \ln p(\mathbf{X})$$

- 我们有：

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

- 交替地针对q和$\boldsymbol{\theta}$来右手边
  - 针对q的优化就得出了与标准EM算法相同的E步骤的方程
  - M步骤的方程只要对标准最大似然的M步骤方程进行很小的修改

广义EM算法
Generalized EM Algorithm

- EM算法将潜在困难的最大化似然函数问题分解成了两个阶段：E步骤和M步骤
- 对于复杂的模型，有可能出现：E步骤或M步骤或者两者都是不易处理的
- 广义EM（Generalized EM或GEM）算法解决了难以求解的M步骤的问题
  - 一种方法就是在M步骤期间使用一种非线性优化策略，诸如共轭梯度方法。
  - 另一种形式则被称为**期望条件最大化(Expectation Conditional Maximization, ECM)**算法