机器学习
Machine Learning

第10讲：无监督学习

Part-1 聚类和混合模型

1. K均值聚类算法
K-means Clustering

将未标注的数据点集分簇（或分组）
每一簇具有一定的聚集特性，
是无监督学习的一种基本算法

D维样本集合： $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

将N个样本聚类成K簇，每一簇有D维特征向量 $\boldsymbol{\mu}_k$

需学习参数： $\boldsymbol{\mu}_k$： $k = 1, \ldots, K$

对每一个样本 $\mathbf{x}_n$， 定义标识变量 $r_{nk} \in \{0, 1\}$

$\mathbf{x}_n$ 属于k簇，则 $r_{nk} = 1$ $r_{nj} = 0 \text{ for } j \neq k$

K均值聚类算法（续）

为了导出有效算法，定义如下目标函数

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

求 $\{r_{nk}\}$ 和 $\{\boldsymbol{\mu}_k\}$ 使得目标函数J最小

算法分成两步：选择初始 $\{\boldsymbol{\mu}_k\}$

第一步：固定 $\{\boldsymbol{\mu}_k\}$ ，确定 $\{r_{nk}\}$ 使 J最小

第二步：固定 $\{r_{nk}\}$，确定 $\{\boldsymbol{\mu}_k\}$ 使 J最小

反复迭代，直到算法收敛

K均值聚类算法（续）

第一步：由（#1），显然，对于固定 $\{\boldsymbol{\mu}_k\}$ 和n

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

（#2）

第二步，对于固定 $\{r_{nk}\}$ ，（#1）对 $\{\boldsymbol{\mu}_k\}$ 导数为0，有

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

K均值聚类算法（续）

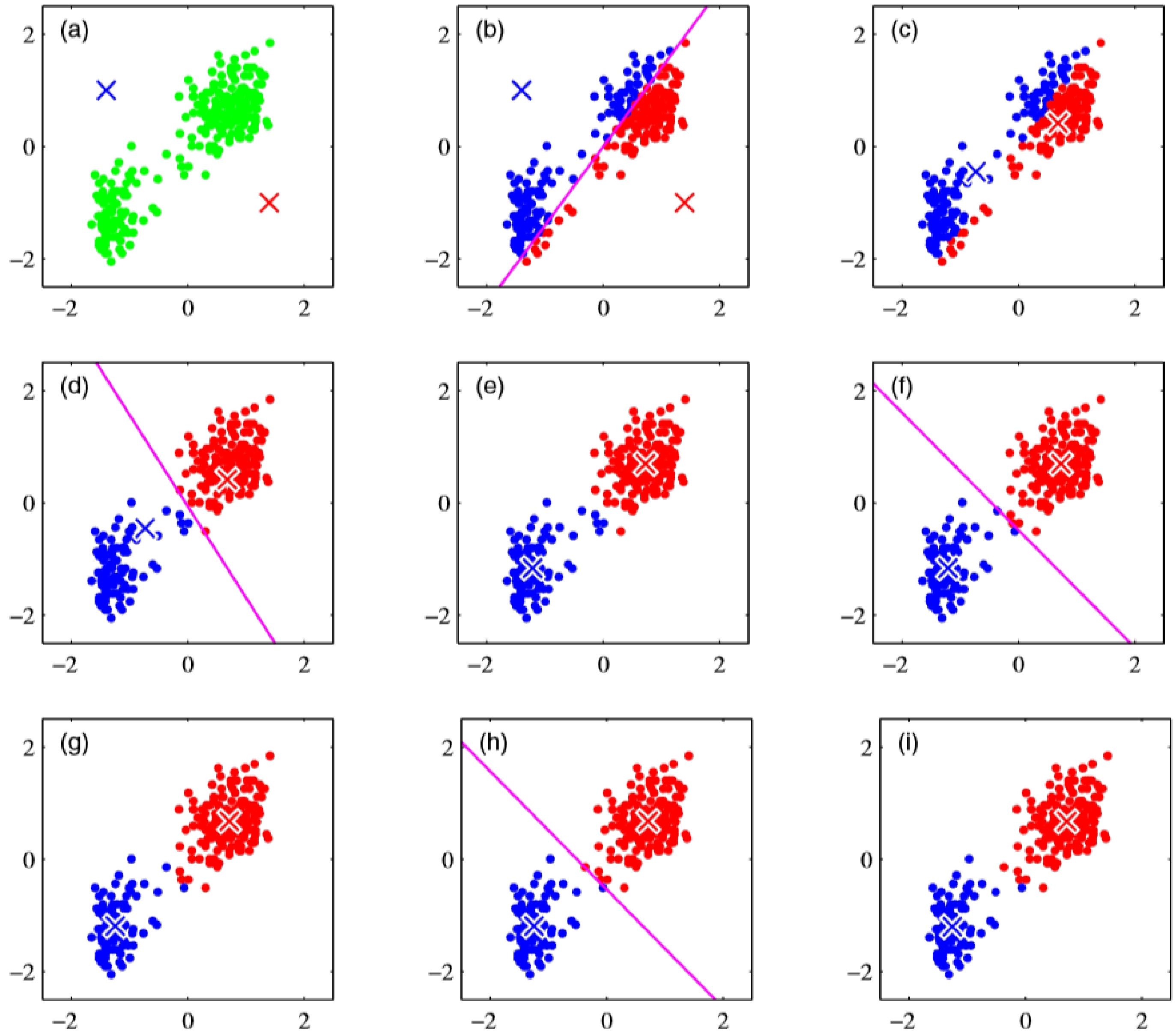
求各簇特征向量为

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}. \quad （\#3）$$

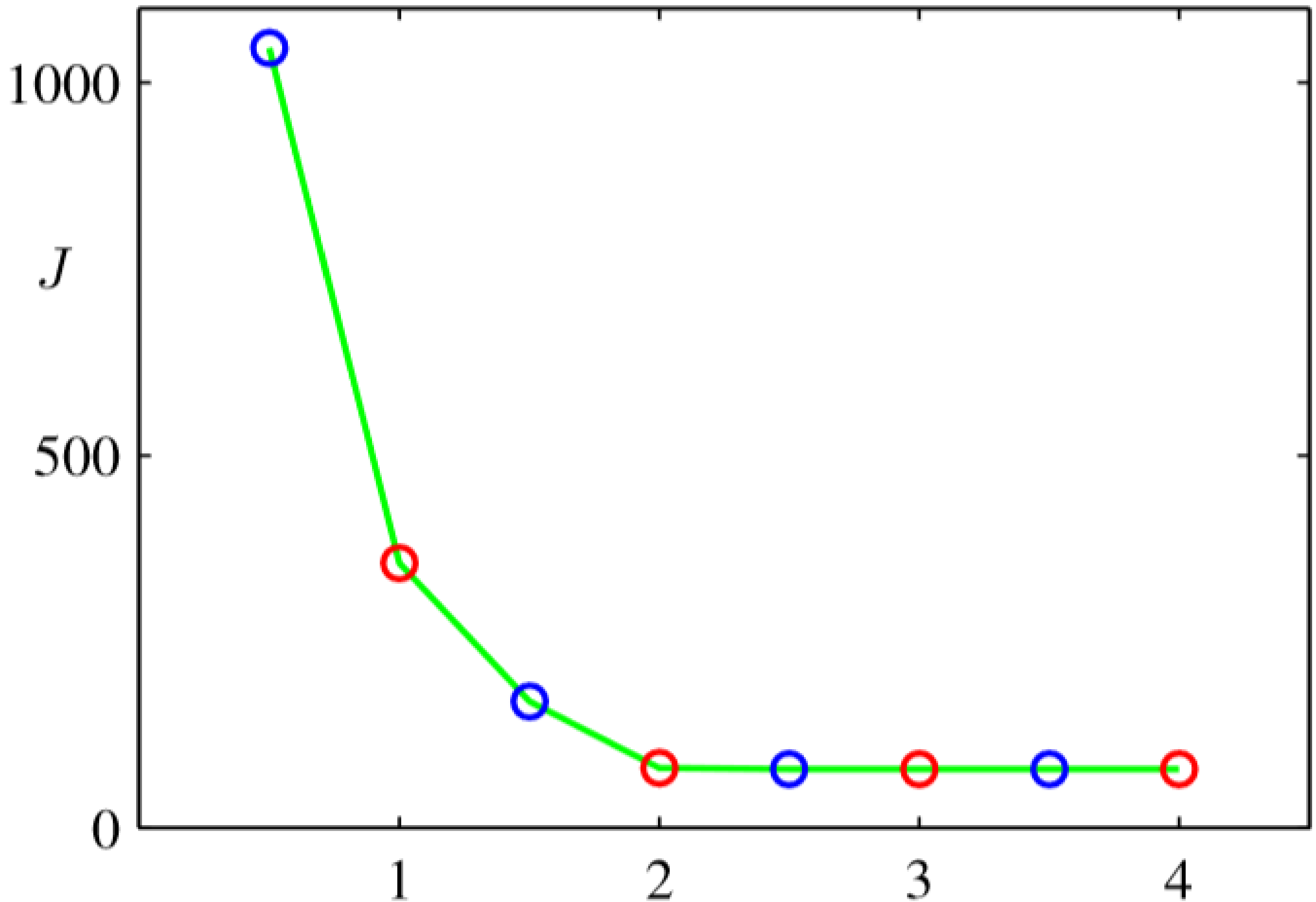
注：通过反复迭代（#2）和（#3），其中（#2）确定每个样本属于离特征向量 $\boldsymbol{\mu}_k$ 最近的簇，（#3）重新计算每个簇的均值向量为特征向量。

该算法称为K均值聚类算法。

例子：Old Faithful 数据集聚类，二类的简单例子



例子：Old Faithful 数据集聚类的收敛曲线



K均值聚类算法

1. K均值聚类算法与其他方法有密切联系，例如与图像压缩的矢量量化算法。

2. 可以定义更一般的相似性度量，$\mathcal{V}(\mathbf{x}, \mathbf{x}')$
由此扩展成更一般的K-中心点算法（K-medoid）。

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

3. K均值方法与混合高斯模型有紧密联系，可以用K均值算法为混合高斯模型的EM算法提供初值。同时，K均值算法也可看作一个EM算法的实例。

2. 混合高斯模型（GMM）：隐变量观点

GMM

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

引入一个隐变量（假设存在，但观测不到）

K维，

$$z_k \in \{0, 1\} \text{ and } \sum_k z_k = 1$$

$z_k = 1$ 表示 $\mathbf{x}$ 取自第k个高斯分布 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

故： $p(z_k = 1) = \pi_k$

GMM：隐变量观点（续）

先验概率参数 $\{\pi_k\}$ 满足

$$0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

$\mathbf{z}$ 是 1-of-K ,概率表示为

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

给出 $\mathbf{z}$ 的一个特殊值，则 $\mathbf{x}$ 表示为

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

GMM：隐变量观点（续）

$\mathbf{x}$ 的条件分布为

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

联合分布为 $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$

对联合分布求边际分布，得到存在隐变量时 $\mathbf{x}$ 的PDF为

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

GMM：隐变量观点（续）

π_k 是 $z_k = 1$ 的先验分布，求 $z_k = 1$ 的后验分布

并表示为

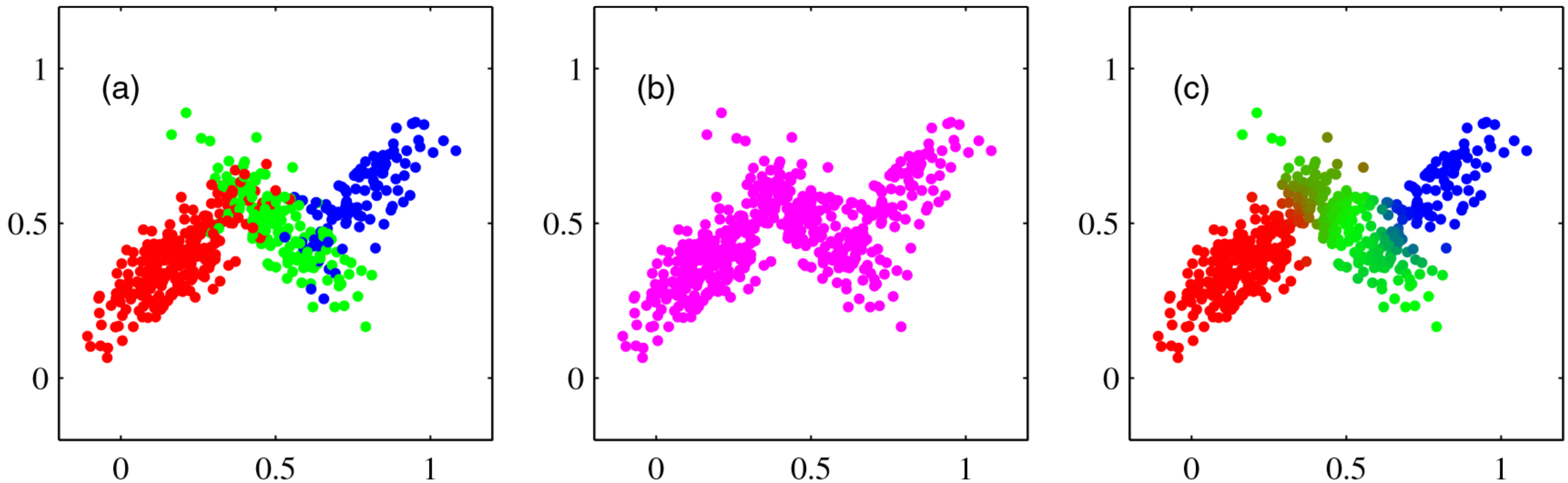
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})$$

由贝叶斯公式得：

$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.\end{aligned}$$

这个隐变量的后验概率将起很关键作用！

GMM：隐变量观点示例



左：已知 π_k 和参数 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ ，仿真产生若干数据，

并记下 $z_k = 1$ ，用红绿蓝表示样本点产生那个k分量，

故：左图是联合分布

中：去掉颜色，即去掉隐变量信息，实际样本是不知隐变量

右：已知 π_k $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ ，用中图的各样本点坐标估计：

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})$$

<u>实际中，只有中图的样本点集：估计</u> π_k $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$

3. GMM参数估计：MLE的EM算法
EM：Expectation-Maximization

给出数据集： $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
假定K（超参数）

估计GMM的所有参数集 π_k $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$

$\mathbf{X}$ 是 $N \times D$ 维样本矩阵， $\mathbf{x}_n^{\mathrm{T}}$ 是第n行

$\mathbf{Z}$ 是 $N \times K$ 维隐变量矩阵

对数似然函数为：

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

GMM参数估计：MLE的EM算法（续）

由于对数中的GMM各分量求和，对数并不能直接作用到高斯函数的指数项，无法获得二次函数和的形式，带来计算的困难。

利用隐变量的后验概率，通过迭代解该问题

对数似然函数对 $\boldsymbol{\mu}_k$ 求导为0，得：

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

以上方程是高度非线性的，无解析解。若用迭代解，
先假设可用旧参数值（第一次迭代时用初始猜测值）计算出 $\gamma(z_{nk})$ ，然后用 $\gamma(z_{nk})$ 表示 $\boldsymbol{\mu}_k$ 的解。

GMM参数估计：MLE的EM算法（续）

均值向量的解为

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

其中

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

对数似然函数对 $\boldsymbol{\Sigma}_k$ 求导为0，类似地得：

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

GMM参数估计：MLE的EM算法（续）

为求 π_k ，除对数似然函数，还要加上约束项，即

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

上式对 π_k 求导为0，得：

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

上式分子分母同乘 π_k ，引入 $\gamma(z_{nk})$，并用 $\lambda = -N$

得：

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

GMM参数估计：MLE的EM算法（续）

这是EM算法解GMM的思路

1. E步，利用旧参数值，计算隐变量后验概率$\gamma(z_{nk})$
2. M步，利用$\gamma(z_{nk})$计算新的参数集 π_k $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$

GMM参数估计的的EM算法描述

1. 初始化 π_k ,$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ ，并计算初始对数似然函数
2. E步：计算隐变量的后验概率

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

GMM参数估计的的EM算法描述（续）

3. M步：更新计算GMM的各项参数

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

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

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

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

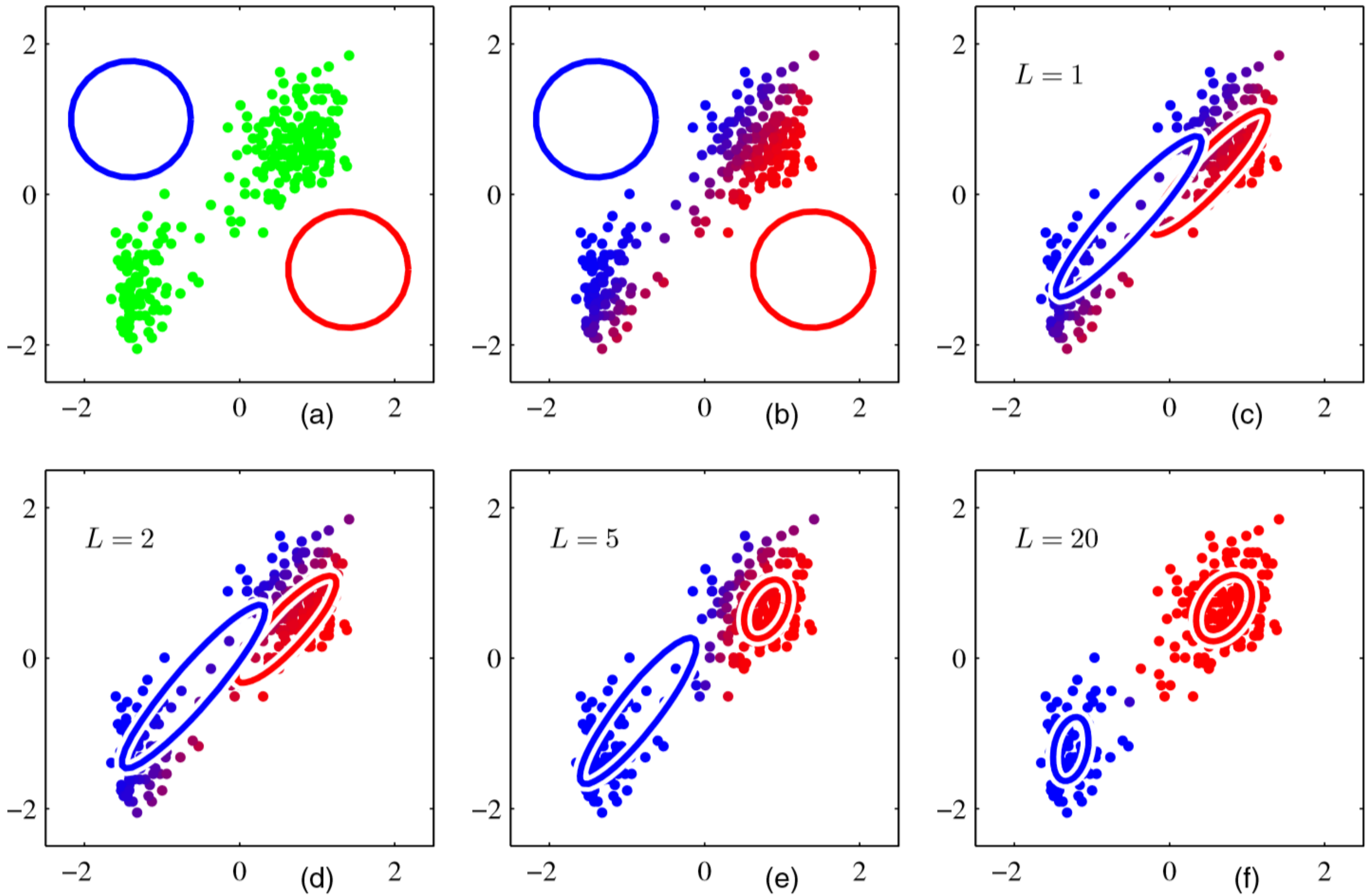
4. 用新参数计算对数似然函数，

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

若满足条件则停止，否则，转2继续迭代

GMM参数估计的EM算法实例

Old Faithful 数据集



4. EM算法

由数据集 $\mathbf{X}$ ，通过MLE估计参数 $\boldsymbol{\theta}$

对数似然函数

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

可能难以处理，例如GMM情况。
定义隐变量集 $\mathbf{Z}$ ，联合分布 $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$

$\{\mathbf{X},\mathbf{Z}\}$ 称为完整数据集， $\mathbf{X}$ 为不完整数据集

用联合分布更易于求解 $\boldsymbol{\theta}$ ，但是 $\mathbf{Z}$ 是未观测到的量

<u>**替代方法（EM）**</u>：定义隐变量的后验分布 $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$

求 $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ 在 $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$ 下的**条件期望**，作为优化函数

EM算法（续）

EM算法采用迭代方法，并分为E步和M步，依次迭代。

初始时取 $\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}_0$

E步： 参数固定为 $\boldsymbol{\theta}^{\text{old}}$，并确定隐变量后验概率为

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$

计算 $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ 的条件期望

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \qquad \text{（\#1）}$$

注：该式放在E步或M步均可

M步： 更新参数，得到参数更新值 $\boldsymbol{\theta}^{\text{new}}$

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \qquad \text{（\#2）}$$

迭代： $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ E步和M步依次迭代直到收敛

以EM标准步骤，重新考察GMM参数估计

由隐变量和完整数据集 $\{\mathbf{X}, \mathbf{Z}\}$ 的表示，得联合概率分布

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

联合对数似然函数

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

需要求 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ 的条件期望，

只需要求对 z_{nk} 的条件期望

用EM重新考察GMM参数估计（续）

为了得到 $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$， 这里只需要 $p(z_{nk}=1|x_n,\mu_k^{old},\Sigma_k^{old})$

$$\mathbb{E}[z_{nk}] = p(z_{nk}=1|\boldsymbol{x}_n,\boldsymbol{\mu}_k^{old},\Sigma_k^{old})$$

$$= \frac{\pi_k^{old} N(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{old},\Sigma_k^{old})}{\sum_{j=1}^{K}\pi_j^{old} N(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{old},\Sigma_j^{old})} = \gamma(z_{nk})$$

联合分布的条件期望为（即：$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ ）

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

以下求 $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$最大，结果同前。

用EM收敛的一种解释

利用完整数据集$\{\mathbf{X},\mathbf{Z}\}$，似然函数写为

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

通过推导，得对数似然函数的一种分解为

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

其中

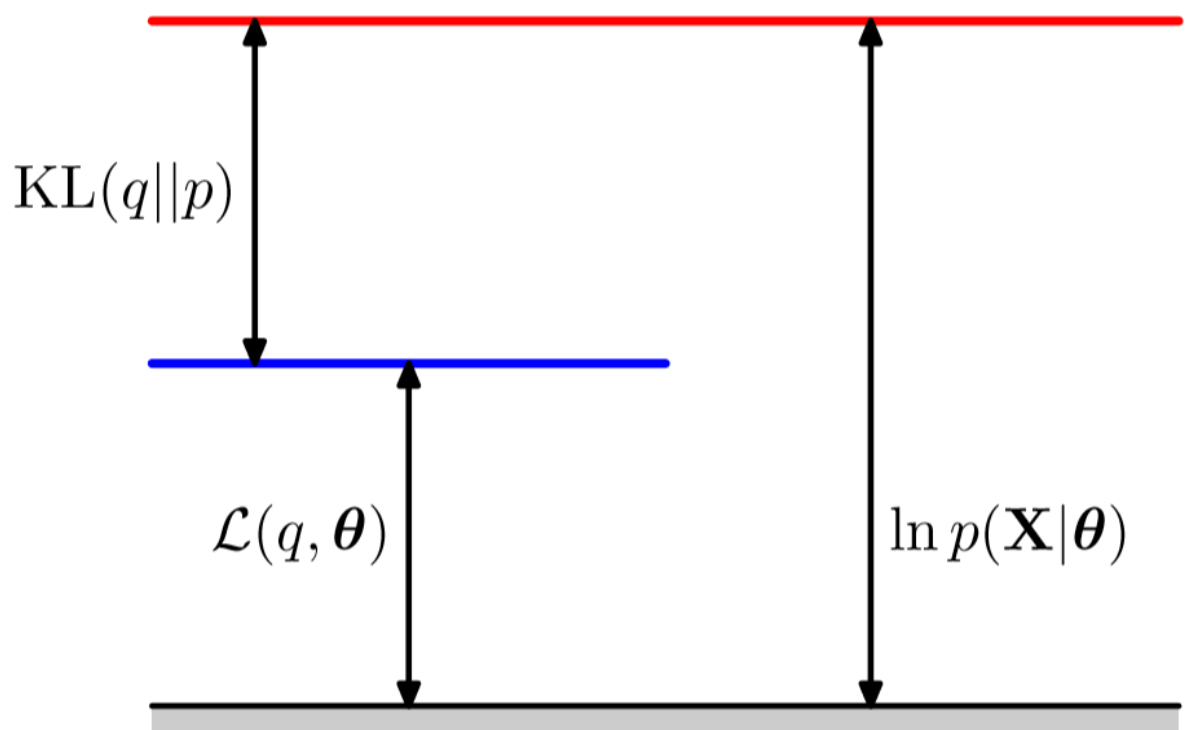
$$\begin{aligned}
\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\
\mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}
\end{aligned}$$

用EM收敛的一种解释（续）

$\mathrm{KL}(q\|p)$ KL散度，$\mathrm{KL}(q\|p) \geqslant 0$

$\mathcal{L}(q, \boldsymbol{\theta})$ 对数似然函数的下界

三个量的关系示意图，EM算法中，其值变化



用EM收敛的一种解释（续）

一般情况下， $\mathrm{KL}(q\|p) \geqslant 0$.

故 $\mathcal{L}(q, \boldsymbol{\theta}) \leqslant \ln p(\mathbf{X}|\boldsymbol{\theta})$ 为对数似然函数下界

只有当 $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ 时，$\mathrm{KL}(q\|p)$ =0

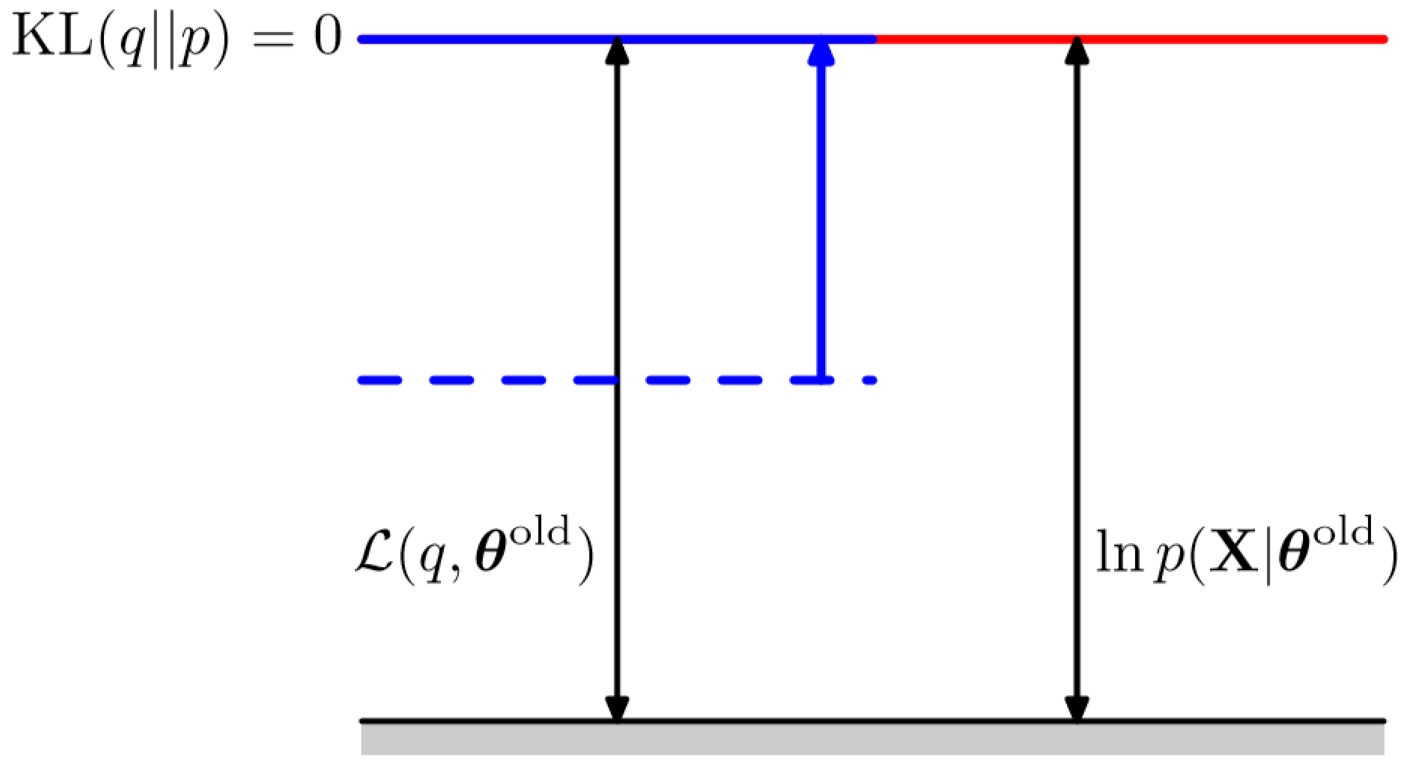
E步： 参数固定为 $\boldsymbol{\theta}^{\text{old}}$

若取： $q(\mathbf{Z})$ =$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$，KL散度为0

下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ = $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ 为最大。

用EM收敛的一种解释（续）

E步以后的各量示意图



用EM收敛的一种解释（续）

M步： $q(\mathbf{Z})$ （$= p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$ ）固定不变，则可有

$$
\begin{aligned}
\mathcal{L}(q,\boldsymbol{\theta}) &= \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \\
&= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \text{const}
\end{aligned}
$$

求 $\boldsymbol{\theta}^{\text{new}}$ ，使 $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ 最大，即 $\mathcal{L}(q,\boldsymbol{\theta})$ 最大，即

$\mathcal{L}(q,\boldsymbol{\theta}^{\text{new}})$ 增加，同时，由于新参数 $\boldsymbol{\theta}^{\text{new}}$ ，KL也不再为0

故：对数似然函数的增加，可能大于 $\mathcal{L}(q,\boldsymbol{\theta})$ 的增加

用EM收敛的一种解释（续）

M步，各量的变化

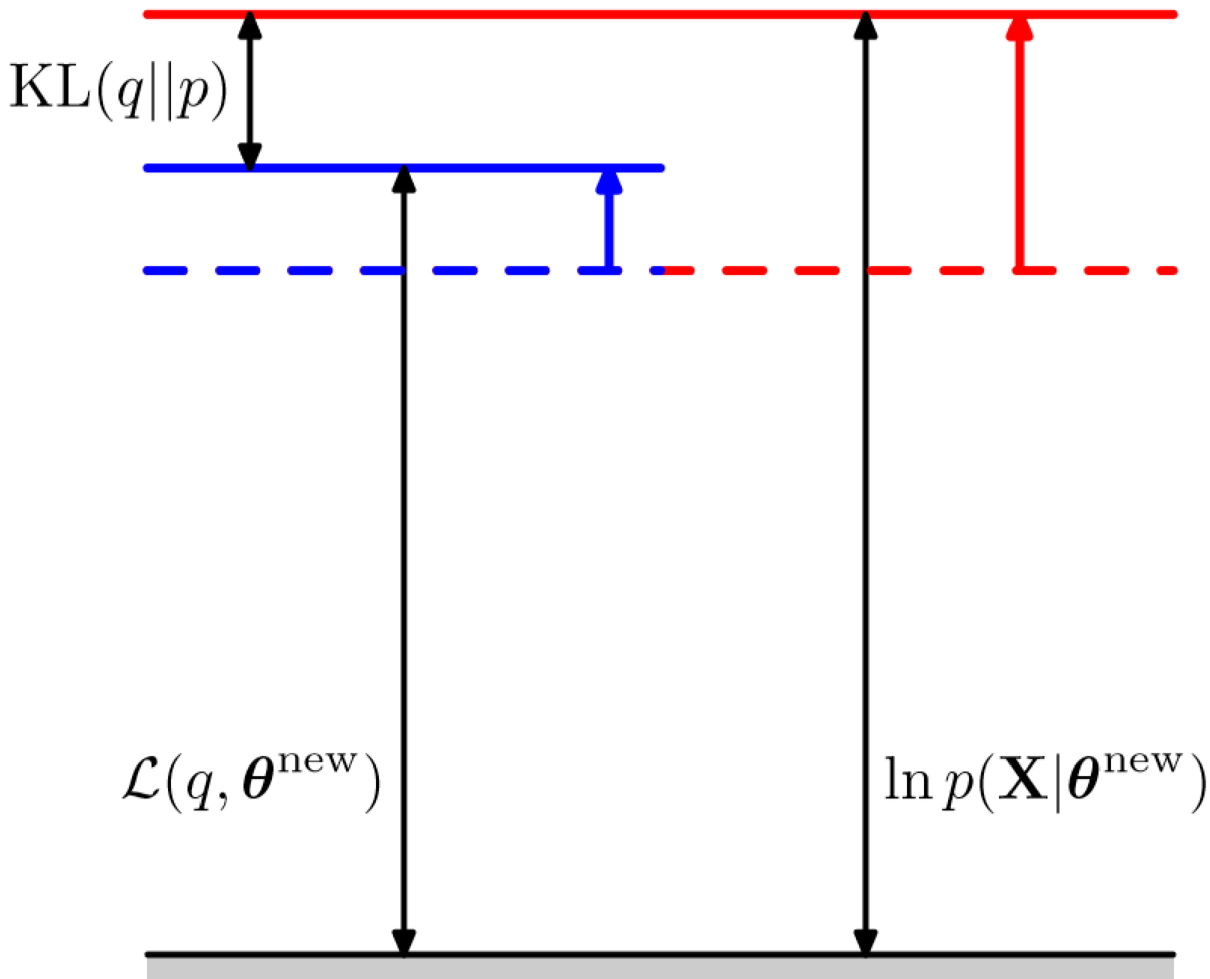
$\mathcal{L}(q, \boldsymbol{\theta}^{\text{new}})$ 增大

$\text{KL}(q||p)$ 不再为0

$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{new}})$

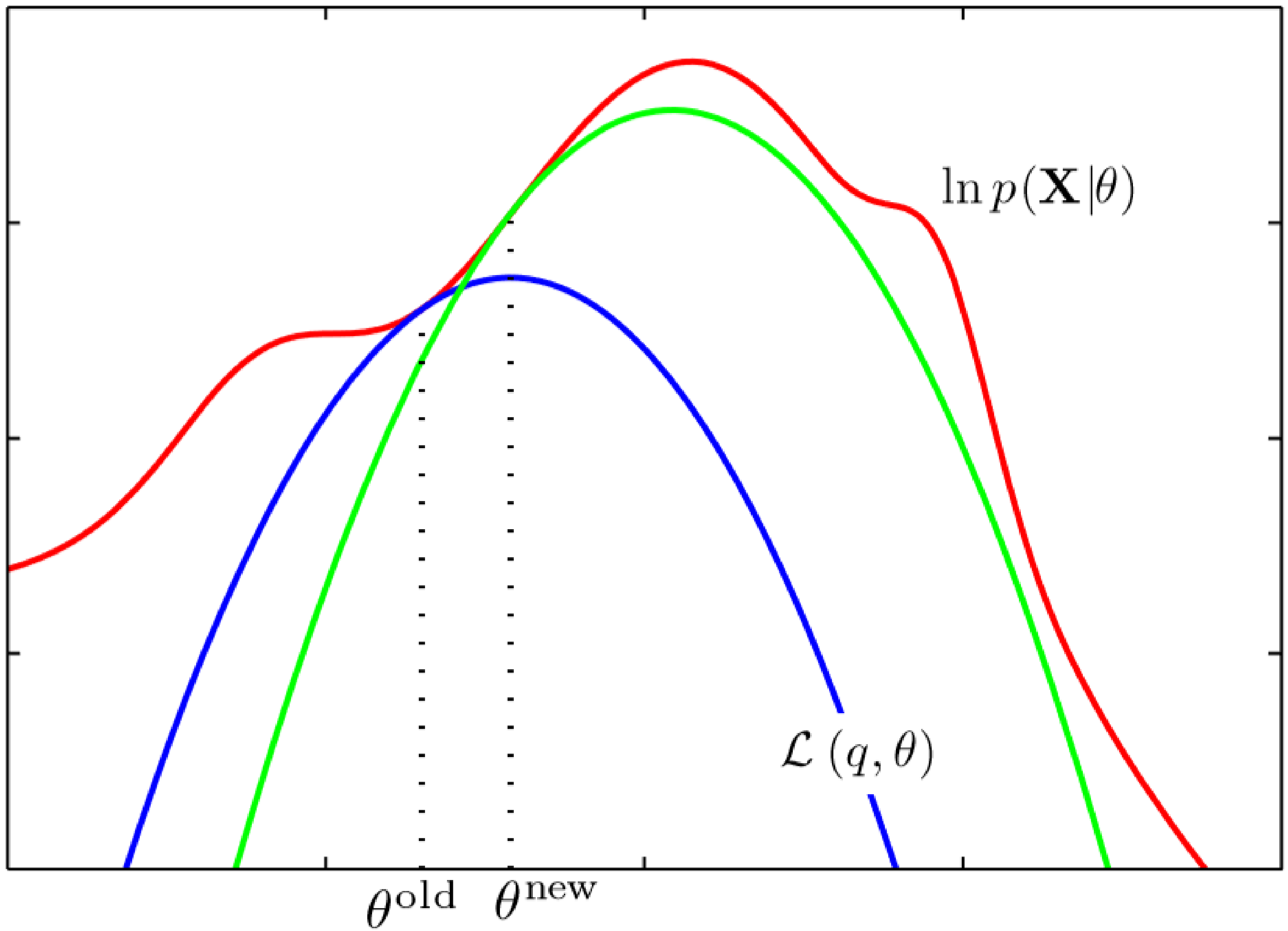
为两增量和

每个E步和M步，保证对数似然函数单调增保证收敛。



用EM收敛的一种解释（续）

EM算法收敛的变化示意图。



说明：按照EM的算法，每一步都在增加 $\ln p(\mathbf{X}|\theta)$ 直到收敛

EM算法的一点说明

- EM算法是现代统计学中的一种有效计算最大似然的算法，有更一般的形式（完整数据集概念），机器学习的无监督学习中主要利用了隐变量这种形式;
- EM算法也可有效的计算MAP问题。
- EM算法在现代统计学、信号处理和机器学习等领域都有应用。