7 算例变化：《混合高斯模型 (GMM)》

7.1 模型定义与生成过程

高斯混合模型（Gaussian Mixture Model, GMM）是一种典型的分层概率模型，它假设观测数据是由多个具有不同参数的高斯分布混合而成的。在聚类任务中，GMM 被广泛用于描述数据的“软分配”逻辑。

设观测数据集为 $\mathcal{X} = \{X^1, X^2, \ldots, X^N\}$。假设这些数据源自 K 个不同的高斯分量，每个分量由均值 μ_k 和标准差 σ_k 刻画。模型的生成逻辑可以描述为：

1. **选择簇标签**：首先根据先验概率分配，从 K 个可选类别中抽取一个类别标签 Z^i。
2. **生成观测值**：在给定类别 $Z^i = k$ 的条件下，从对应的高斯分布 $\mathcal{N}(\mu_k, \sigma_k)$ 中采样得到观测数据 X^i。

在该模型中，所有待优化的参数集定义为 $\Theta = \{\alpha_{1:K}, \mu_{1:K}, \sigma_{1:K}\}$，其中 α_k 代表第 k 个簇的先验概率，且满足 $\sum_{k=1}^{K} \alpha_k = 1$。

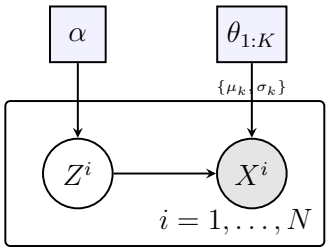

图 5: 高斯混合模型 (GMM) 的概率图表示。α 为混合权重，θ 为各簇的高斯参数。实线表示变量间的生成依赖关系。

7.2 隐变量 Z 的引入与物理意义

在 GMM 中，每个观测样本 X^i 背后都隐藏着一个不可见的类别归属信息。我们引入隐变量（Latent Variable）Z^i 来描述这一信息。

为了数学处理的方便，我们通常采用“One-hot”编码形式来表示 Z^i。定义指示变量 $Z_k^i \in \{0, 1\}$：

$$Z_k^i = \begin{cases} 1, & \text{若样本} X^i \text{ 属于第} k \text{ 个簇} \\ 0, & \text{否则} \end{cases}$$

隐变量 Z 的引入将复杂的混合分布问题转化为了一个包含“缺失数据”的简单概率问题：

- **先验分布**：$P(Z_k^i = 1|\Theta) = \alpha_k$。这反映了在没有任何观测数据时，我们对样本归属的初步判断。
- **条件似然**：$P(X^i|Z_k^i = 1, \Theta) = \mathcal{N}(X^i|\mu_k, \sigma_k)$。这描述了在已知类别标签时，数据点的分布形态。

通过引入 Z，我们建立起了观测数据 X 与模型参数 Θ 之间的桥梁。变分推断的目标，本质上就是寻找这些隐变量 Z^i 在给定观测 X^i 下的后验分布。

7.3 高斯混合模型 (GMM) 的证据下界 (ELBO) 推导

7.3.1 联合概率分布的分解

在 GMM 中，观测数据 X^i 与隐变量 Z^i 的联合概率可以分解为：

$$P(X, Z|\Theta) = \prod_{i=1}^{N} P(X^i|Z^i, \Theta)P(Z^i|\Theta)$$

利用指示变量 Z_k^i，对数联合似然函数可以写为：

$$\begin{aligned}\log P(X, Z|\Theta) &= \sum_{i=1}^{N}\sum_{k=1}^{K} Z_k^i \left[\log \alpha_k + \log \mathcal{N}(X^i|\mu_k, \sigma_k)\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} Z_k^i \left[\log \alpha_k - \frac{1}{2}\log(2\pi\sigma_k^2) - \frac{(X^i - \mu_k)^2}{2\sigma_k^2}\right]\end{aligned} \tag{15}$$

7.3.2 推导最优变分分布 $q^*(Z^i)$

根据坐标上升变分推断 (CAVI) 的一般理论，隐变量 Z^i 的最优变分后验分布 $q^*(Z^i)$ 应当满足其对数形式等于联合似然函数在其他变量分布下的期望。

CAVI 基本定理

$$\log q_j^*(Z_j) = \mathbb{E}_{q_{-j}}[\log p(X, Z)] + \text{const} \tag{16}$$

在 GMM 的语境下，我们将对数联合似然函数代入上式。由于我们采用均值场假设且样本间独立，对于特定的样本 i，只需提取联合概率中包含 Z^i 的项进行处理：

$$\begin{aligned}\log q^*(Z^i) &= \mathbb{E}_{q(\Theta)}[\log p(X^i, Z^i|\Theta)] + \text{const} \\ &= \mathbb{E}\left[\textstyle\sum_{k=1}^{K} Z_k^i \left(\log \alpha_k + \log \mathcal{N}(X^i|\mu_k, \sigma_k)\right)\right] + \text{const}\end{aligned}$$

由于指示变量 Z_k^i 与关于模型参数 Θ 的期望算子无关，我们可以将其移出期望符号：

$$\log q^*(Z^i) = \sum_{k=1}^{K} Z_k^i \cdot \underbrace{\mathbb{E}_\Theta\left[\log \alpha_k + \log \mathcal{N}(X^i|\mu_k, \sigma_k)\right]}_{\text{记为}\eta_{ik}} + \text{const}$$

对等式两边取指数 (Exponential)，即可得到变分分布的解析形式：

$$q^*(Z^i) \propto \exp\left(\sum_{k=1}^{K} Z_k^i \cdot \eta_{ik}\right) = \prod_{k=1}^{K} (e^{\eta_{ik}})^{Z_k^i}$$

对等式两边取指数，并利用 Z^i 的 One-hot 约束 ($\sum_k Z_k^i = 1$)，得到：

$$q^*(Z^i) \propto \prod_{k=1}^{K} \left[\alpha_k \mathcal{N}(X^i|\mu_k, \sigma_k)\right]^{Z_k^i}$$

分布识别 观察上式可知，$q^*(Z^i)$ 的数学形式与参数为 γ_i 的 **Categorical 分布**完全一致。[9] 因此，我们推导出最优变分分布具有分类分布的形式，其变分参数（Responsibility）满足：

$$\gamma_{ik} \propto \alpha_k \mathcal{N}(X^i|\mu_k, \sigma_k)$$

归一化处理后，即得到我们在 E-step 中使用的解析更新公式：

$$q(Z) = \prod_{i=1}^{N} q(Z^i), \quad \text{其中} q(Z^i) = \text{Categorical}(\gamma_{i1}, \ldots, \gamma_{iK}) \tag{17}$$

这里，变分参数 γ_{ik} 代表了样本 i 属于第 k 个簇的后验概率（即 Responsibility），且满足 $\sum_{k=1}^{K} \gamma_{ik} = 1$。

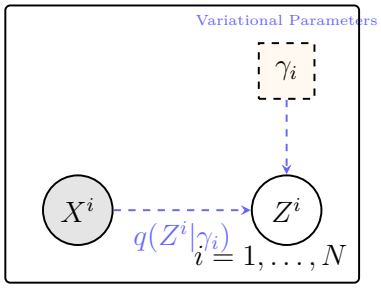

图 6: GMM 变分推断路径图。蓝色虚线代表推断过程，γ_i 是通过坐标上升或神经网络输出的局部变分参数。

7.3.3 ELBO 的全展开式

证据下界 $\mathcal{L}(q, \Theta)$ 定义为对数联合似然的期望与变分熵之和。利用线性期望性质 $\mathbb{E}_q[Z_k^i] = \gamma_{ik}$，展开如下：

$$\mathcal{L}(q, \Theta) = \mathbb{E}_q[\log P(X, Z|\Theta)] - \mathbb{E}_q[\log q(Z)]$$

代入具体形式，得到 GMM 的全展开 ELBO 表达式：

GMM 全展开 ELBO

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik} \left[\log \alpha_k - \frac{1}{2}\log(2\pi\sigma_k^2) - \frac{(X^i - \mu_k)^2}{2\sigma_k^2}\right] - \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik} \log \gamma_{ik}}_{\text{(变分熵项 Entropy)}} \tag{18}$$

[9]附注：Categorical (分类) 分布的定义：

分类分布（Categorical Distribution）是伯努利分布向多维情况的泛化，用于描述一个随机变量在 K 个互斥类别中取值的概率。在本模型中，对于隐变量 Z^i，其概率质量函数（PMF）定义为：$q(Z^i|\gamma_i) = \prod_{k=1}^{K} \gamma_{ik}^{Z_k^i}$. 其中，参数需满足以下物理约束：**非负性**：$\gamma_{ik} \geq 0$，代表样本 i 属于第 k 个簇的概率。**归一性**：$\sum_{k=1}^{K} \gamma_{ik} = 1$。

在此定义下，隐变量指示器 Z_k^i 的期望值具有极其简洁的形式：

$$\mathbb{E}_q[Z_k^i] = 1 \cdot P(Z_k^i = 1) + 0 \cdot P(Z_k^i = 0) = \gamma_{ik}$$

这一性质是后续将对数联合似然的期望 $\mathbb{E}_q[\log P]$ 展开为包含 γ_{ik} 的线性求和式的关键数学基础。

7.3.4 ELBO 关于各参数的梯度推导

为了通过梯度下降或坐标上升优化目标函数，我们需要计算 $\mathcal{L}$ 关于各项参数的偏导数。基于前述全展开式，计算结果汇总如下：

1. 关于变分参数 γ_{ik} 的梯度 在计算 γ_{ik} 的梯度时，需注意每个样本的变分分布满足约束 $\sum_{k=1}^{K}\gamma_{ik}=1$。其偏导数为：

$$\frac{\partial\mathcal{L}}{\partial\gamma_{ik}}=\log\alpha_k-\frac{1}{2}\log(2\pi\sigma_k^2)-\frac{(X^i-\mu_k)^2}{2\sigma_k^2}-\log\gamma_{ik}-1$$

令该导数为 0（并引入拉格朗日乘子处理约束），即可导出变分后验的指数对齐形式。

2. 关于模型参数 Θ 的梯度 (M-step) 对于模型参数，梯度仅涉及期望对数似然项。

- **均值 μ_k 的梯度**：$\frac{\partial\mathcal{L}}{\partial\mu_k}=\sum_{i=1}^{N}\gamma_{ik}\frac{(X^i-\mu_k)}{\sigma_k^2}$. 该梯度为 0 时，$\mu_k$ 表现为所有样本以 γ_{ik} 为权重的加权平均值。
- **方差 σ_k^2 的梯度**：$\frac{\partial\mathcal{L}}{\partial\sigma_k^2}=\sum_{i=1}^{N}\gamma_{ik}\left[-\frac{1}{2\sigma_k^2}+\frac{(X^i-\mu_k)^2}{2(\sigma_k^2)^2}\right]$
- **先验权重 α_k 的梯度**（需满足 $\sum\alpha_k=1$）：$\frac{\partial\mathcal{L}}{\partial\alpha_k}=\sum_{i=1}^{N}\frac{\gamma_{ik}}{\alpha_k}$

综合在一起，我们得到以下梯度计算的结果:

GMM 梯度汇总

$$\begin{cases}\nabla_{\mu_k}\mathcal{L}=\dfrac{1}{\sigma_k^2}\displaystyle\sum_{i=1}^{N}\gamma_{ik}(X^i-\mu_k)\\[2ex]\nabla_{\sigma_k^2}\mathcal{L}=\dfrac{1}{2(\sigma_k^2)^2}\displaystyle\sum_{i=1}^{N}\gamma_{ik}\left[(X^i-\mu_k)^2-\sigma_k^2\right]\\[2ex]\nabla_{\gamma_{ik}}\mathcal{L}=\text{const}+\log\dfrac{\alpha_k\mathcal{N}(X^i|\mu_k,\sigma_k)}{\gamma_{ik}}\\[2ex]\nabla_{\alpha_k}=\displaystyle\sum_{i=1}^{N}\frac{\gamma_{ik}}{\alpha_k}\end{cases}\tag{19}$$

7.4 GMM 的变分迭代算法 (CAVI)

通过令各参数的偏导数为 0，我们可以推导出如下交替迭代过程。在变分推断的语境下，这被称为坐标上升变分推断 (Coordinate Ascent Variational Inference, CAVI)。

7.4.1 E-step：更新变分参数（后验分配 γ_{ik}）

在此步骤中，我们固定当前的模型参数 $\{\alpha_k,\mu_k,\sigma_k^2\}$。目标是为每个样本 X^i 计算其属于各个簇的概率。根据 $\nabla_{\gamma_{ik}}\mathcal{L}=0$ 的推导，更新公式如下：

E-step: Responsibility Update

$$\gamma_{ik}^{(t+1)} = \frac{\alpha_k^{(t)} \mathcal{N}(X^i | \mu_k^{(t)}, \sigma_k^{(t)})}{\sum_{j=1}^{K} \alpha_j^{(t)} \mathcal{N}(X^i | \mu_j^{(t)}, \sigma_j^{(t)})} \tag{20}$$

物理直观 此公式具有清晰的贝叶斯解释：样本 i 属于簇 k 的“软分配”权重，取决于该簇的先验强度 α_k 与该簇分布对数据点拟合程度（似然）的乘积。分母起到了归一化作用，确保了同一样本对所有簇的分配概率之和为 1。

7.4.2 M-step：更新模型参数（极大似然估计）

在此步骤中，我们固定变分参数 γ_{ik}。通过令模型参数的梯度为 0，我们可以一次性得到所有模型参数的解析更新公式。这些公式本质上是在当前的“软分配”权重下，对各簇特征进行的加权统计。

M-step: Parameter Updates

$$\begin{cases} \alpha_k^{(t+1)} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \gamma_{ik}^{(t+1)} \\ \mu_k^{(t+1)} = \dfrac{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)} X^i}{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)}} \\ (\sigma_k^2)^{(t+1)} = \dfrac{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)} (X^i - \mu_k^{(t+1)})^2}{\sum_{i=1}^{N} \gamma_{ik}^{(t+1)}} \end{cases} \tag{21}$$

我们会惊奇地发现，这个结果和我们之前在课堂讲解 EM 算法的时候得到的结果是一致的。[10]

物理意义总结

- **先验** α_k：等于该簇分配到的“有效样本数”占总样本数的比例。
- **均值** μ_k：由所有样本根据归属概率 γ_{ik} 加权平均决定。
- **方差** σ_k^2：反映了样本点偏离新均值的加权平方和。

[10]还有一个区别是 Z 的定义存在细微差异，你发现了没?

7.5 回忆——GMM 的课堂黑板推导 EM 算法（注：不考虑变分）

7.5.1 主要迭代公式的阐述

Theorem 1 (最大化 GMM 的对数似然). *对于引入了隐变量 $Z = \{z^1, \cdots, z^N\}$ 的概率图模型 (PGM), EM 算法 [a] [b] 为极大似然估计 (MLE) $\theta^* = \arg\max_\theta\{\mathcal{L}(\theta|\bar{X})\}$ 提供了一种迭代方案：*

$$\theta^{(g+1)} = \arg\max_\theta \int \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ \tag{22}$$

[a] 为了简便起见，参数集记为 θ，请勿混淆。此外，X 和 z 是特定样本 X^i 和 z^i 的一般形式。

[b] 应满足不等式 $\log P(X|\theta^{(g+1)}) \geq \log P(X|\theta^{(g)})$。它确保了期望迭代公式 (22) 的收敛性。

为了计算上述项，使用了包含隐变量 Z 的对数形式：

$$\log P(X|\theta) = \log P(X, Z|\theta) - \log P(Z|X, \theta). \tag{23}$$

由于 $P(X) = \frac{P(X,Z)}{P(Z|X)}$，这是显而易见的。现在的技巧是计算 $\log P(X|\theta)$ 关于 $P(Z|X, \theta^{(g)})$ 的期望。

$$\begin{aligned} \log P(X|\theta) &= E_{P(Z|X,\theta^{(g)})}[\log P(X|\theta)] \\ &= E_{P(Z|X,\theta^{(g)})}[\log P(X, Z|\theta) - \log P(Z|X, \theta)] \\ \int_Z \log P(X|\theta) \cdot P(Z|X, \theta^{(g)}) dZ &= \int_Z \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ \\ &\quad - \int_Z \log P(Z|X, \theta) \cdot P(Z|X, \theta^{(g)}) dZ \end{aligned} \tag{24}$$

然后我们得到：

$$\begin{aligned} \log P(X|\theta) &= Q(\theta, \theta^{(g)}) - H(\theta, \theta^{(g)}) \\ \text{其中} \quad Q(\theta, \theta^{(g)}) &:= \int_Z \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ \\ H(\theta, \theta^{(g)}) &:= \int_Z \log P(Z|X, \theta) \cdot P(Z|X, \theta^{(g)}) dZ \end{aligned} \tag{25}$$

为了继续推导，我们需要了解一些关于 H 的知识：

Theorem 2 (一个引理). $H(\theta^{(g)}, \theta^{(g)}) \geq H(\theta, \theta^{(g)})$

证明：

$$\begin{aligned} & H(\theta^{(g)}, \theta^{(g)}) - H(\theta, \theta^{(g)}) \\ &= \int_Z \log P(Z|X, \theta^{(g)}) \cdot P(Z|X, \theta^{(g)}) dZ - \int_Z \log P(Z|X, \theta) \cdot P(Z|X, \theta^{(g)}) dZ \\ &= \int_Z (-\log \tfrac{P(Z|X,\theta)}{P(Z|X,\theta^{(g)})}) \cdot P(Z|X, \theta^{(g)}) dZ \\ &\text{(利用 Jensen 不等式。此处 } -\log(x) \text{ 是一个凸函数。)} \\ &(E(-\log(x)) \geq -\log(E(x))) \\ &\geq -\log \int_Z \tfrac{P(Z|X,\theta)}{P(Z|X,\theta^{(g)})} \cdot P(Z|X, \theta^{(g)}) dZ \\ &= -\log \int_Z P(Z|X, \theta) dZ \\ &= -\log 1 = 0 \end{aligned}$$

现在我们把课堂所推演的结果整理，重述为：

Theorem 3 (最大化 GMM 的对数似然). 迭代公式 $\theta^{(g+1)} = \arg\max_{\theta} \int \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(g)}) dZ$ (公式 *22*) 足以解决高斯混合模型问题):

$$\theta_{MLE} = \arg\max_{\theta}\{\mathcal{L}(\theta|\bar{X})\} = \arg\max_{\theta}\left\{\sum_{i=1}^{N}\log\left[\sum_{k=1}^{K}\alpha_k\mathcal{N}(X_i|\mu_k,\sigma_k)\right]\right\} \tag{26}$$

7.5.2 EM 算法的实现逻辑

E-步：估计隐变量的分布

$$P(Z|X, \theta^{(old)}).$$

M-步：关于隐变量最大化联合对数似然的期望

$$\theta^{(new)} = \arg\max_{\theta} \int \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(old)}) dZ$$

$$= \arg\max_{\theta} E_Z[\log P(X, Z|\theta)]$$

7.5.3 具体迭代式的计算

首先，我们对 X 的似然有如下设定:

$$P(X^i|\theta) = \sum_{k=1}^{K}\alpha_k\mathcal{N}(X^i|\mu_k,\sigma_k),$$
$$P(\bar{X}|\theta) = \prod_{i=1}^{N}\sum_{k=1}^{K}\alpha_k\mathcal{N}(X^i|\mu_k,\sigma_k)$$

在这个基础上，对于 EM 迭代公式 (22)，我们还需要同时计算 $P(X, Z|\theta)$ 和 $P(Z|X, \theta)$。

如何计算全概率 $P(X, Z|\theta)$： [11]

$$P(X, Z|\theta) = \prod_{i=1}^{N} P(X^i, z^i|\theta) = \prod_{i=1}^{N} \underbrace{P(X^i|z^i, \theta))}_{\mathcal{N}(\mu_{z^i},\sigma_{z^i})} \underbrace{P(z^i|\theta))}_{\alpha_{z^i}} = \prod_{i=1}^{N} \alpha_{z^i}\mathcal{N}(X^i|\mu_{z^i},\sigma_{z^i})$$

如何计算后验 $P(Z|X, \theta)$： [12]

$$P(Z|X, \theta) = \prod_{i=1}^{N} P(z^i|X^i, \theta) = \prod_{i=1}^{N} \frac{\alpha_{z^i}\mathcal{N}(X^i|\mu_{z^i},\sigma_{z^i})}{\sum_{k=1}^{K}\alpha_k\mathcal{N}(X^i|\mu_k,\sigma_k)} \tag{27}$$

因此，我们有:

[11]注：这个结论是自然的——得益于隐变量 Z 的设定。

[12]注：这个公式是可以自行设定的，因为有一定的物理意义，所以，这在课堂讲授的时候是直接给出的。但是实际上，这个结果是完全可以通过变分 CAVI 的过程求出来的，见（eq. 20)。简言之，在第一次讲授 GMM 的时候，我们只使用了隐变量模型的思想，我们引入了 EM，但是我们没有设定变分函数 $q(z)$ 以及对应的变分参数 γ。

$$
\begin{aligned}
&Q(\theta,\theta^{(g)}) = \int_Z \log P(X,Z|\theta)\cdot P(Z|X,\theta^{(g)})dZ \\
&= \int_{z_1}\cdots\int_{z_N}(\sum_{i=1}^{N}\log P(X^i,z^i|\theta)\cdot\prod_{i=1}^{N}P(z^i|X^i,\theta^{(g)}))dz_1\cdots dz_N \\
&= \sum_{z_1=1}^{K}\cdots\sum_{z_N=1}^{K}(\sum_{i=1}^{N}\underbrace{(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))}_{f_i(z^i)}\cdot\underbrace{\prod_{i=1}^{N}P(z^i|X^i,\theta^{(g)}))}_{\tilde{P}(z_1,\cdots,z_N)} \\
&= \sum_{z_1=1}^{K}\cdots\sum_{z_N=1}^{K}((f_1(z_1)+\cdots+f_N(z_N))\cdot\tilde{P}(z_1,\cdots,z_N)) \\
&= \sum_{z_1=1}^{K}\cdots\sum_{z_N=1}^{K}f_1(z_1)\cdot\tilde{P}(z_1,\cdots,z_N)+\cdots \\
&= \sum_{z_1=1}^{K}f_1(z_1)\cdot\sum_{Z_2=1}^{K}\cdots\sum_{z_N=1}^{K}\tilde{P}(z_1,\cdots,z_N)+\cdots \\
&= \sum_{z_1=1}^{K}f_1(z_1)\cdot\tilde{P}(z_1)+\cdots
\end{aligned}
$$

$\tilde{P}(z_1)$ 是 $\tilde{P}(z_1,\cdots,z_N)$ 关于 z_1 的边缘概率密度。

$$
\begin{aligned}
&= \sum_{z_1=1}^{K}f_1(z_1)\cdot\tilde{P}(z_1)+\cdots+\sum_{z_N=1}^{K}f_N(z_N)\cdot\tilde{P}(z_N) \\
&= \sum_{i=1}^{N}\sum_{z^i=1}^{K}f_i(z^i)\cdot\tilde{P}(z^i) \\
&= \sum_{i=1}^{N}\sum_{z^i=1}^{K}(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))P(z^i|X^i,\theta^{(g)}))
\end{aligned}
$$

因此我们有

M-步

$$
\begin{aligned}
&\int\log P(X,Z|\theta)\cdot P(Z|X,\theta^{(g)})dZ \\
&= \sum_{i=1}^{N}\sum_{z^i=1}^{K}(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))P(z^i|X^i,\theta^{(g)})) \\
&= \sum_{i=1}^{N}\sum_{z^i=1}^{K}(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))\frac{\alpha_{z^i}\mathcal{N}(\mu_{z^i},\sigma_{z^i})}{\sum_{k=1}^{K}\alpha_k\mathcal{N}(\mu_k,\sigma_k)}
\end{aligned}
$$

为了推导 M-步中的迭代公式，需要计算：

$$
\frac{\partial\sum_{i=1}^{N}\sum_{z^i=1}^{K}(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))P(z^i|X^i,\theta^{(g)}))}{\partial\alpha_1,\cdots,\partial\alpha_K} = [0,\cdots,0],
$$

约束条件为 $\sum_{k=1}^{K}\alpha_k = 1$。

以及

$$
\begin{aligned}
&\frac{\partial\sum_{i=1}^{N}\sum_{z^i=1}^{K}(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))P(z^i|X^i,\theta^{(g)}))}{\partial\mu_1,\cdots,\partial\mu_K} = [0,\cdots,0], \\
&\frac{\partial\sum_{i=1}^{N}\sum_{z^i=1}^{K}(\log\alpha_{z^i}+\log\mathcal{N}(\mu_{z^i},\sigma_{z^i}))P(z^i|X^i,\theta^{(g)}))}{\partial\sigma_1,\cdots,\partial\sigma_K} = [0,\cdots,0],
\end{aligned}
$$

在上述计算中，第一个是最简单的。建议尝试一下。

(... ... 此处省略了若干计算步骤！)

最后，我们得到 M-步中的迭代公式: [13]

$$
\begin{aligned}
\alpha_k^{(g+1)} &= \frac{1}{N}\sum_{i=1}^{N} P(z^i = k|X^i, \theta^{(g)}) \\
\mu_k^{(g+1)} &= \frac{\sum_{i=1}^{N} X^i P(z^i{=}k|X^i,\theta^{(g)})}{\sum_{i=1}^{N} P(z^i{=}k|X^i,\theta^{(g)})} \\
\sigma_k^{(g+1)} &= \frac{\sum_{i=1}^{N} [X^i{-}\mu_k^{(g+1)}][X^i{-}\mu_k^{(g+1)}]^T P(z^i{=}k|X^i,\theta^{(g)})}{\sum_{i=1}^{N} P(z^i{=}k|X^i,\theta^{(g)})}
\end{aligned}
\tag{28}
$$

[13]注：请与（eq. 21）比较。他们实际是一样的，为什么呢?