THU-70250043-0, Pattern Recognition (Spring 2021) **Homework: 3**

GMM and EM

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Hong Zhao vzhao@tsinghua.edu.cn

Student: Jingxuan Yang yangjx20@mails.tsinghua.edu.cn

EM and GD

1. In this problem you will see connections between the EM algorithm and gradient descent. Consider a GMM with known mixture weight π_k and spherical covariances (but the radius of spheres might be different). Its log likelihood is given by

$$l\left(\left\{\mu_k, \sigma_k^2\right\}_{k=1}^K\right) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k N(x_i|\mu_k, \sigma_k^2 I)\right). \tag{1}$$

A maximization algorithm based on gradient descent should be something like:

- Initialize μ_k and σ_k^2, $k \in \{1, 2, \cdots, K\}$. Set the iteration counter $t \leftarrow 1$.
- Repeat the following until convergence:
  - For $k = 1, 2, \cdots, K$,
$$\mu_k^{(t+1)} \leftarrow \mu_k^{(t)} + \eta_k^{(t)} \nabla_{\mu_k} l\left(\left\{\mu_k^{(t)}, (\sigma_k^2)^{(t)}\right\}_{k=1}^K\right) \tag{2}$$
  - For $k = 1, 2, \cdots, K$,
$$(\sigma_k^2)^{(t+1)} \leftarrow (\sigma_k^2)^{(t)} + s_k^{(t)} \nabla_{\sigma_k^2} l\left(\left\{\mu_k^{(t+1)}, (\sigma_k^2)^{(t)}\right\}_{k=1}^K\right) \tag{3}$$
  - Increase the iteration counter $t \leftarrow t + 1$.

Please prove that with properly chosen step size $\eta_k^{(t)}$ and $s_k^{(t)}$, the above gradient descent algorithm is essentially equivalent to the following *modified* EM algorithm:

- Initialize μ_k and σ_k^2, $k \in \{1, 2, \cdots, K\}$. Set the iteration counter $t \leftarrow 1$.
- Repeat the following until convergence:

– E-step:

$$\tilde{z}_{ik}^{(t+0.5)} \leftarrow P\left(x_i \in \text{cluster}_k \middle| \left\{\mu_j^{(t)}, (\sigma_j^2)^{(t)}\right\}_{j=1}^{K}, x_i\right) \tag{4}$$

– M-step:

$$\left\{\mu_k^{(t+1)}\right\}_{k=1}^{K} \leftarrow \underset{\{\mu_k\}_{k=1}^{K}}{\text{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{z}_{ik}^{(t+0.5)} \left[\log N\left(x_i \middle| \mu_k, (\sigma_k^2)^{(t)} I\right) + \log \pi_k\right] \tag{5}$$

– E-step:

$$\tilde{z}_{ik}^{(t+1)} \leftarrow P\left(x_i \in \text{cluster}_k \middle| \left\{\mu_j^{(t+1)}, (\sigma_j^2)^{(t)}\right\}_{j=1}^{K}, x_i\right) \tag{6}$$

– M-step:

$$\left\{(\sigma_k^2)^{(t+1)}\right\}_{k=1}^{K} \leftarrow \underset{\{\sigma_k^2\}_{k=1}^{K}}{\text{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{z}_{ik}^{(t+1)} \left[\log N\left(x_i \middle| \mu_k^{(t+1)}, \sigma_k^2 I\right) + \log \pi_k\right] \tag{7}$$

– Increase the iteration counter $t \leftarrow t+1$.

The main modification is inserting an extra E-step between the M-step for μ_k's and the M-step for σ_k^2's.

Hint: Find the exact algebraic form of step size $\eta_k^{(t)}$ and $s_k^{(t)}$ from M-step.

解: 设样本点 x_i 的维数为 d, 则正态分布概率密度为

$$N(x_i|\mu_k, \sigma_k^2 I) = \frac{1}{(2\pi\sigma_k^2)^{d/2}} \exp\left[-\frac{1}{2\sigma_k^2}(x_i-\mu_k)^\top (x_i-\mu_k)\right] \tag{8}$$

其对 μ_k 的梯度为

$$\nabla_{\mu_k} N(x_i|\mu_k, \sigma_k^2 I) = N(x_i|\mu_k, \sigma_k^2 I)\frac{x_i-\mu_k}{\sigma_k^2} \tag{9}$$

对 σ_k^2 的梯度为

$$\nabla_{\sigma_k^2} N(x_i|\mu_k, \sigma_k^2 I) = N(x_i|\mu_k, \sigma_k^2 I)\left[-\frac{d}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}(x_i-\mu_k)^\top (x_i-\mu_k)\right] \tag{10}$$

所以, 似然函数对 μ_k 的梯度为

$$\begin{aligned} \nabla_{\mu_k} l\left(\left\{\mu_k, \sigma_k^2\right\}_{k=1}^{K}\right) &= \nabla_{\mu_k} \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \sigma_k^2 I)\right) \\ &= \sum_{i=1}^{n} \frac{\pi_k N(x_i|\mu_k, \sigma_k^2 I)}{\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \sigma_k^2 I)} \frac{x_i-\mu_k}{\sigma_k^2} \end{aligned} \tag{11}$$

似然函数对 σ_k^2 的梯度为

$$\begin{aligned}\nabla_{\sigma_k^2} l\left(\left\{\mu_k, \sigma_k^2\right\}_{k=1}^K\right) &= \nabla_{\sigma_k^2} \sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k N(x_i|\mu_k, \sigma_k^2 I)\right) \\ &= \sum_{i=1}^n \frac{\pi_k N(x_i|\mu_k, \sigma_k^2 I)}{\sum_{k=1}^K \pi_k N(x_i|\mu_k, \sigma_k^2 I)} \left[-\frac{d}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}(x_i-\mu_k)^\top (x_i-\mu_k)\right]\end{aligned} \tag{12}$$

对改进 EM 算法, 第一个 E-step: 由 Bayes 公式可得

$$\begin{aligned}\tilde{z}_{ik}^{(t+0.5)} &= P\left(x_i \in \text{cluster}_k \middle| \left\{\mu_j^{(t)}, (\sigma_j^2)^{(t)}\right\}_{j=1}^K, x_i\right) \\ &= \frac{\pi_k N\left(x_i \middle| \mu_k^{(t)}, (\sigma_k^2)^{(t)} I\right)}{\sum_{k=1}^K \pi_k N\left(x_i \middle| \mu_k^{(t)}, (\sigma_k^2)^{(t)} I\right)}\end{aligned} \tag{13}$$

第一个 M-step: 令

$$Q^{(t+0.5)} \triangleq \sum_{i=1}^n \sum_{k=1}^K \tilde{z}_{ik}^{(t+0.5)} \left[\log N\left(x_i \middle| \mu_k, (\sigma_k^2)^{(t)} I\right) + \log \pi_k\right] \tag{14}$$

则其对 μ_k 的梯度为

$$\begin{aligned}\frac{\partial Q^{(t+0.5)}}{\partial \mu_k} &= \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \frac{N\left(x_i \middle| \mu_k, (\sigma_k^2)^{(t)} I\right)}{N\left(x_i \middle| \mu_k, (\sigma_k^2)^{(t)} I\right)} \frac{x_i - \mu_k}{(\sigma_k^2)^{(t)}} \\ &= \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \frac{x_i - \mu_k}{(\sigma_k^2)^{(t)}} \\ &= \frac{1}{(\sigma_k^2)^{(t)}} \left(\sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} x_i - \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \mu_k\right)\end{aligned} \tag{15}$$

令此梯度为 0, 则有

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} x_i}{\sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)}} \tag{16}$$

对梯度下降算法, $\mu_k^{(t+1)}$ 为

$$\begin{aligned}\mu_k^{(t+1)} &= \mu_k^{(t)} + \eta_k^{(t)} \nabla_{\mu_k} l\left(\left\{\mu_k^{(t)}, (\sigma_k^2)^{(t)}\right\}_{k=1}^K\right) \\ &= \mu_k^{(t)} + \eta_k^{(t)} \sum_{i=1}^n \frac{\pi_k N\left(x_i \middle| \mu_k^{(t)}, (\sigma_k^2)^{(t)} I\right)}{\sum_{k=1}^K \pi_k N\left(x_i \middle| \mu_k^{(t)}, (\sigma_k^2)^{(t)} I\right)} \frac{x_i - \mu_k^{(t)}}{(\sigma_k^2)^{(t)}} \\ &= \mu_k^{(t)} + \eta_k^{(t)} \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \frac{x_i - \mu_k^{(t)}}{(\sigma_k^2)^{(t)}}\end{aligned} \tag{17}$$

令梯度下降算法得到的 $\mu_k^{(t+1)}$ 与 EM 算法得到的结果相等, 则有

$$\frac{\sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} x_i}{\sum_{j=1}^n \tilde{z}_{jk}^{(t+0.5)}} = \mu_k^{(t)} + \eta_k^{(t)} \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \frac{x_i - \mu_k^{(t)}}{(\sigma_k^2)^{(t)}} \tag{18}$$

通分有

$$\sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} x_i = \sum_{j=1}^n \tilde{z}_{jk}^{(t+0.5)} \mu_k^{(t)} + \sum_{j=1}^n \tilde{z}_{jk}^{(t+0.5)} \eta_k^{(t)} \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \frac{x_i - \mu_k^{(t)}}{(\sigma_k^2)^{(t)}} \tag{19}$$

移项可得

$$\sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \left(x_i - \mu_k^{(t)}\right) = \frac{\eta_k^{(t)} \sum_{j=1}^n \tilde{z}_{jk}^{(t+0.5)}}{(\sigma_k^2)^{(t)}} \sum_{i=1}^n \tilde{z}_{ik}^{(t+0.5)} \left(x_i - \mu_k^{(t)}\right) \tag{20}$$

即

$$\frac{\eta_k^{(t)} \sum_{j=1}^n \tilde{z}_{jk}^{(t+0.5)}}{(\sigma_k^2)^{(t)}} = 1 \tag{21}$$

所以

$$\eta_k^{(t)} = \frac{(\sigma_k^2)^{(t)}}{\sum_{j=1}^n \tilde{z}_{jk}^{(t+0.5)}} \tag{22}$$

对改进 EM 算法, 第二个 E-step: 由 Bayes 公式可得

$$\begin{aligned}\tilde{z}_{ik}^{(t+1)} &= P\left(x_i \in \text{cluster}_k \middle| \left\{\mu_j^{(t+1)}, (\sigma_j^2)^{(t)}\right\}_{j=1}^K, x_i\right) \\ &= \frac{\pi_k N\left(x_i \middle| \mu_k^{(t+1)}, (\sigma_k^2)^{(t)} I\right)}{\sum_{k=1}^K \pi_k N\left(x_i \middle| \mu_k^{(t+1)}, (\sigma_k^2)^{(t)} I\right)}\end{aligned} \tag{23}$$

第二个 M-step: 令

$$Q^{(t+1)} \triangleq \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{z}_{ik}^{(t+1)} \left[\log N\left(x_i \middle| \mu_k^{(t+1)}, \sigma_k^2 I\right) + \log \pi_k\right] \tag{24}$$

则其对 σ_k^2 的梯度为

$$\begin{aligned}
\frac{\partial Q^{(t+1)}}{\partial \sigma_k^2} &= \sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \frac{N\left(x_i \middle| \mu_k^{(t+1)}, \sigma_k^2 I\right)}{N\left(x_i \middle| \mu_k^{(t+1)}, \sigma_k^2 I\right)} \left[-\frac{d}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}\left(x_i - \mu_k^{(t+1)}\right)^\top \left(x_i - \mu_k^{(t+1)}\right)\right] \\
&= \sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left[-\frac{d}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}\left(x_i - \mu_k^{(t+1)}\right)^\top \left(x_i - \mu_k^{(t+1)}\right)\right] \\
&= \frac{1}{2(\sigma_k^2)^2}\left[\sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left\|x_i - \mu_k^{(t+1)}\right\|^2 - \sigma_k^2 d \sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)}\right]
\end{aligned} \tag{25}$$

令此梯度为 0, 则有

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left\|x_i - \mu_k^{(t+1)}\right\|^2}{d\sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)}} \tag{26}$$

对梯度下降算法, $(\sigma_k^2)^{(t+1)}$ 为

$$\begin{aligned}
(\sigma_k^2)^{(t+1)} &= (\sigma_k^2)^{(t)} + s_k^{(t)} \nabla_{\sigma_k^2} l\left(\left\{\mu_k^{(t+1)}, (\sigma_k^2)^{(t)}\right\}_{k=1}^{K}\right) \\
&= (\sigma_k^2)^{(t)} + s_k^{(t)} \sum_{i=1}^{n} \frac{\pi_k N\left(x_i \middle| \mu_k^{(t+1)}, (\sigma_k^2)^{(t)} I\right)}{\sum_{k=1}^{K} \pi_k N\left(x_i \middle| \mu_k^{(t+1)}, (\sigma_k^2)^{(t)} I\right)} \left[-\frac{d}{2(\sigma_k^2)^{(t)}} + \frac{\left\|x_i - \mu_k^{(t+1)}\right\|^2}{2\left[(\sigma_k^2)^{(t)}\right]^2}\right] \\
&= (\sigma_k^2)^{(t)} + s_k^{(t)} \sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left[-\frac{d}{2(\sigma_k^2)^{(t)}} + \frac{1}{2\left[(\sigma_k^2)^{(t)}\right]^2}\left\|x_i - \mu_k^{(t+1)}\right\|^2\right]
\end{aligned} \tag{27}$$

令梯度下降算法得到的 $(\sigma_k^2)^{(t+1)}$ 与 EM 算法得到的结果相等, 则有

$$\frac{\sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left\|x_i - \mu_k^{(t+1)}\right\|^2}{d\sum_{j=1}^{n} \tilde{z}_{jk}^{(t+1)}} = (\sigma_k^2)^{(t)} + s_k^{(t)} \sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left[-\frac{d}{2(\sigma_k^2)^{(t)}} + \frac{1}{2\left[(\sigma_k^2)^{(t)}\right]^2}\left\|x_i - \mu_k^{(t+1)}\right\|^2\right] \tag{28}$$

通分并移项可得

$$\sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left(\left\|x_i - \mu_k^{(t+1)}\right\|^2 - d(\sigma_k^2)^{(t)}\right) = \frac{s_k^{(t)} d \sum_{j=1}^{n} \tilde{z}_{jk}^{(t+1)}}{2\left[(\sigma_k^2)^{(t)}\right]^2} \sum_{i=1}^{n} \tilde{z}_{ik}^{(t+1)} \left(\left\|x_i - \mu_k^{(t+1)}\right\|^2 - d(\sigma_k^2)^{(t)}\right) \tag{29}$$

即

$$\frac{s_k^{(t)} d \sum_{j=1}^{n} \tilde{z}_{jk}^{(t+1)}}{2\left[(\sigma_k^2)^{(t)}\right]^2} = 1 \tag{30}$$

所以

$$s_k^{(t)} = \frac{2\left[(\sigma_k^2)^{(t)}\right]^2}{d \sum_{j=1}^{n} \tilde{z}_{jk}^{(t+1)}} \tag{31}$$

综上所述, 若取步长分别为

$$\eta_k^{(t)} = \frac{(\sigma_k^2)^{(t)}}{\sum_{j=1}^{n} \tilde{z}_{jk}^{(t+0.5)}}, \quad s_k^{(t)} = \frac{2\left[(\sigma_k^2)^{(t)}\right]^2}{d \sum_{j=1}^{n} \tilde{z}_{jk}^{(t+1)}} \tag{32}$$

则梯度下降算法与改进 EM 算法本质上是等价的.

EM for MAP Estimation

2. The EM algorithm that we talked about in class was for solving a maximum likelihood estimation problem in which we wished to maximize

$$\prod_{i=1}^{m} p(x^{(i)}|\theta) = \prod_{i=1}^{m} \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}|\theta) \tag{33}$$

where $x^{(i)}$ were visible variables, $z^{(i)}$ were hidden variables and m was the number of samples. Suppose we are working in a Bayesian framework, and wanted to find the MAP estimate of the parameters θ by maximizing

$$\left(\prod_{i=1}^{m} p(x^{(i)}|\theta)\right) p(\theta) = \left(\prod_{i=1}^{m} \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}|\theta)\right) p(\theta) \tag{34}$$

Here, $p(\theta)$ is our prior on the parameters. Please generalize the EM algorithm to work for MAP estimation. You may assume that $\log p(x, z|\theta)$ and $\log p(\theta)$ are both concave in θ, so that the M-step is tractable if it requires only maximizing a linear combination of these quantities. (This roughly corresponds to assuming that MAP estimation is tractable when x, z is fully observed, just like in the frequentist case where we considered examples in which maximum likelihood estimation was easy if x, z was fully observed.)

Make sure your M-step is tractable, and also prove that $\left(\prod_{i=1}^{m} p(x^{(i)}|\theta)\right) p(\theta)$ (viewed as a function of θ) monotonically increases with each iteration of your algorithm.

解: 令 $X=\left\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\right\}$, 对式 (34) 取自然对数可得

$$\begin{aligned}\tilde{H}(\theta) &= \log\left[p(X|\theta)p(\theta)\right] \\ &= \log\left[\left(\prod_{i=1}^{m} p(x^{(i)}|\theta)\right)p(\theta)\right] \\ &= \log\left[\left(\prod_{i=1}^{m}\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}|\theta)\right)p(\theta)\right] \\ &= \sum_{i=1}^{m}\log\sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}|\theta\right) + \log p(\theta)\end{aligned} \tag{35}$$

引入隐变量分布 $q(\cdot)$, 由 Jensen 不等式可得

$$\tilde{H}(\theta) \geqslant \sum_{i=1}^{m}\sum_{z^{(i)}} q(z^{(i)})\log p\left(x^{(i)}, z^{(i)}|\theta\right) - \sum_{i=1}^{m}\sum_{z^{(i)}} q(z^{(i)})\log q(z^{(i)}) + \log p(\theta) \triangleq \tilde{F}(q,\theta) \tag{36}$$

直接优化 $\tilde{H}(\theta)$ 可能是很困难的, 所以我们转而优化 $\tilde{H}(\theta)$ 的下界函数 $\tilde{F}(q,\theta)$. 优化函数 $\tilde{F}(q,\theta)$ 可能也是比较困难的, 因此我们采用一种简单的迭代算法对 $\tilde{F}(q,\theta)$ 寻优, 首先对变量 q, θ 初始化, 然后固定变量 $\theta_{[k]}$ 寻找能够最大化函数 $\tilde{F}(q,\theta_{[k]})$ 的参数 $q_{[k+1]}$, 再固定参数 $q_{[k+1]}$ 寻找能够最大化函数 $\tilde{F}(q_{[k+1]},\theta)$ 的参数 $\theta_{[k+1]}$, 即反复执行下面的两个步骤:

$$q_{[k+1]} \leftarrow \underset{q}{\operatorname{argmax}}\,\tilde{F}(q,\theta_{[k]}) \tag{37}$$

$$\theta_{[k+1]} \leftarrow \underset{\theta}{\operatorname{argmax}}\,\tilde{F}(q_{[k+1]},\theta) \tag{38}$$

当 $q_{[k+1]}(z^{(i)}) = p(z^{(i)}|x^{(i)},\theta_{[k]})$ 时, 式 (37) 取到最大值, 因为此时有

$$\begin{aligned}\tilde{F}(q_{[k+1]},\theta_{[k]}) &= \sum_{i=1}^{m}\sum_{z^{(i)}} q_{[k+1]}(z^{(i)})\log\frac{p\left(x^{(i)}, z^{(i)}|\theta_{[k]}\right)}{q_{[k+1]}(z^{(i)})} + \log p(\theta_{[k]}) \\ &= \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)}|x^{(i)},\theta_{[k]})\log\frac{p(z^{(i)}|x^{(i)},\theta_{[k]})p(x^{(i)}|\theta_{[k]})}{p(z^{(i)}|x^{(i)},\theta_{[k]})} + \log p(\theta_{[k]}) \\ &= \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)}|x^{(i)},\theta_{[k]})\log p(x^{(i)}|\theta_{[k]}) + \log p(\theta_{[k]}) \\ &= \sum_{i=1}^{m}\log p(x^{(i)}|\theta_{[k]}) + \log p(\theta_{[k]}) \\ &= \log\left[\left(\prod_{i=1}^{m} p(x^{(i)}|\theta_{[k]})\right)p(\theta_{[k]})\right] \\ &= \tilde{H}(\theta_{[k]})\end{aligned} \tag{39}$$

当 $q_{[k+1]}(z^{(i)}) = p(z^{(i)}|x^{(i)},\theta_{[k]})$ 时,

$$\tilde{F}(q_{[k+1]},\theta) = \sum_{i=1}^{m}\sum_{z^{(i)}} q_{[k+1]}(z^{(i)})\log p\left(x^{(i)}, z^{(i)}|\theta\right) - \sum_{i=1}^{m}\sum_{z^{(i)}} q_{[k+1]}(z^{(i)})\log q_{[k+1]}(z^{(i)}) + \log p(\theta) \tag{40}$$

由于第二项不包含需要优化的变量 θ, 则可定义

$$\tilde{Q}(\theta_{[k]}, \theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta_{[k]}) \log p\left(x^{(i)}, z^{(i)}|\theta\right) + \log p(\theta) \tag{41}$$

假设 $\log p(x, z|\theta)$ 和 $\log p(\theta)$ 都是 θ 的凹函数, 则最大化函数 $\tilde{Q}$ 是一个无约束的凸优化问题, 即 M-step 是容易处理的 (tractable).

综上, 广义 EM 算法为:

- 初始化变量 q, θ
- E 步骤, 计算函数

$$\tilde{Q}(\theta_{[k]}, \theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta_{[k]}) \log p\left(x^{(i)}, z^{(i)}|\theta\right) + \log p(\theta) \tag{42}$$

- M 步骤,

$$\theta_{[k+1]} \leftarrow \operatorname*{argmax}_{\theta} \tilde{Q}(\theta_{[k]}, \theta) \tag{43}$$

- 如果算法收敛则停止, 否则回到 E 步骤.

下面证明在广义 EM 算法下, $\left(\prod_{i=1}^{m} p(x^{(i)}|\theta)\right) p(\theta)$ 作为 θ 的函数的单调性. 由式 (39) 并结合广义 EM 算法不断迭代求最大值可知

$$\tilde{H}(\theta_{[k]}) = \tilde{F}(q_{[k+1]}, \theta_{[k]}) \leqslant \tilde{F}(q_{[k+1]}, \theta_{[k+1]}) \leqslant \tilde{F}(q_{[k+2]}, \theta_{[k+1]}) = \tilde{H}(\theta_{[k+1]}), \quad \forall\ k \geqslant 1 \tag{44}$$

所以, 函数

$$\tilde{H}(\theta) = \log \left[\left(\prod_{i=1}^{m} p(x^{(i)}|\theta) \right) p(\theta) \right]$$

随着算法的迭代是单调递增的, 又自然对数函数是一一映射, 则 $\left(\prod_{i=1}^{m} p(x^{(i)}|\theta)\right) p(\theta)$ 作为 θ 的函数也是随着算法的迭代单调递增的.

Programming 1 (EM and GMM)

3. Consider the case that the hidden variable $y \in \{1, 2, \cdots, m\}$ is discrete while the visible variable $x \in R^d$ is continuous. In other words, we consider mixture models of the form

$$p(x) = \sum_{j=1}^{m} p(x|y = j)p(y = j) \tag{45}$$

We assume throughout that x is conditionally Gaussian in the sense that $x \sim \mathcal{N}(\mu_j, \Sigma_j)$ when $y = j$. We have provided you with an example EM code for mixture of Gaussians (with visualization) in MATLAB. The command to run is:

```
[param,history,ll] = em_mix(data,m,eps);
```

where the input points are given as rows of `data`, `m` is the number of components in the estimated mixture, and `eps` determines the stopping criteria of EM: the algorithm stops when the relative change in log-likelihood falls below `eps`. In the output, `param` is a cell array with `m` elements. Each element is a structure with the following fields:

`mean` - the resulting mean of the Gaussian component,

`cov` - the resulting covariance matrix of the component,

`p` - the resulting estimate of the mixing parameter.

The value of `param` is updated after every iteration of EM; the output argument `history` contains copies of these subsequent values of `param` and allows to analyze our experiments. Finally, `ll` is the vector where the t^{th} element is the value of the log-likelihood of the `data` after t iterations (i.e. the last element is the final log-likelihood of the fitted mixture of Gaussians).

Hint: For the following two questions you are encouraged to google "BIC (Bayesian Information Criterion)" to help you with the model selection process. Of course other criteria are welcomed as long as you give convincing reasons.

Hint: For this assignment, you are allowed to implement EM algorithm manually in python, and you can use scipy.io.loadmat to load the data.

3.1. Run the EM algorithm based on `data` provided by `emdata.mat` with `m = 2, 3, 4, 5` components. Select the appropriate model (number of components) and give reasons for your choice. Note that you may have to rerun the algorithm a few times (and select the model with the highest log-likelihood) for each choice of `m` as EM can sometimes get stuck in a local minimum. Is the model selection result sensible based on what you would expect visually? Why or why not?

解: 对每个 m 进行 10 次测试, 选择对数似然概率最大的那次测试结果作为该组的结果, 如图 1 所示.

选择 BIC 作为模型参数选择依据,

$$\text{BIC} = k \ln n - 2 \ln \hat{l} \tag{46}$$

其中, n 为数据点个数, $\hat{l}$ 为对数似然概率, k 为模型估计的变量个数, 其中分配概率为 1, 均值为 2, 协方差对称矩阵为 $4 - 1 = 3$, 又因为分配概率之和固定 $(= 1)$ 则最后需 -1,

$$k = (1 + d + d^2 - 1)m - 1 = 6m - 1 \tag{47}$$

其中, d 为数据点维数, 此处为 $d = 2$.

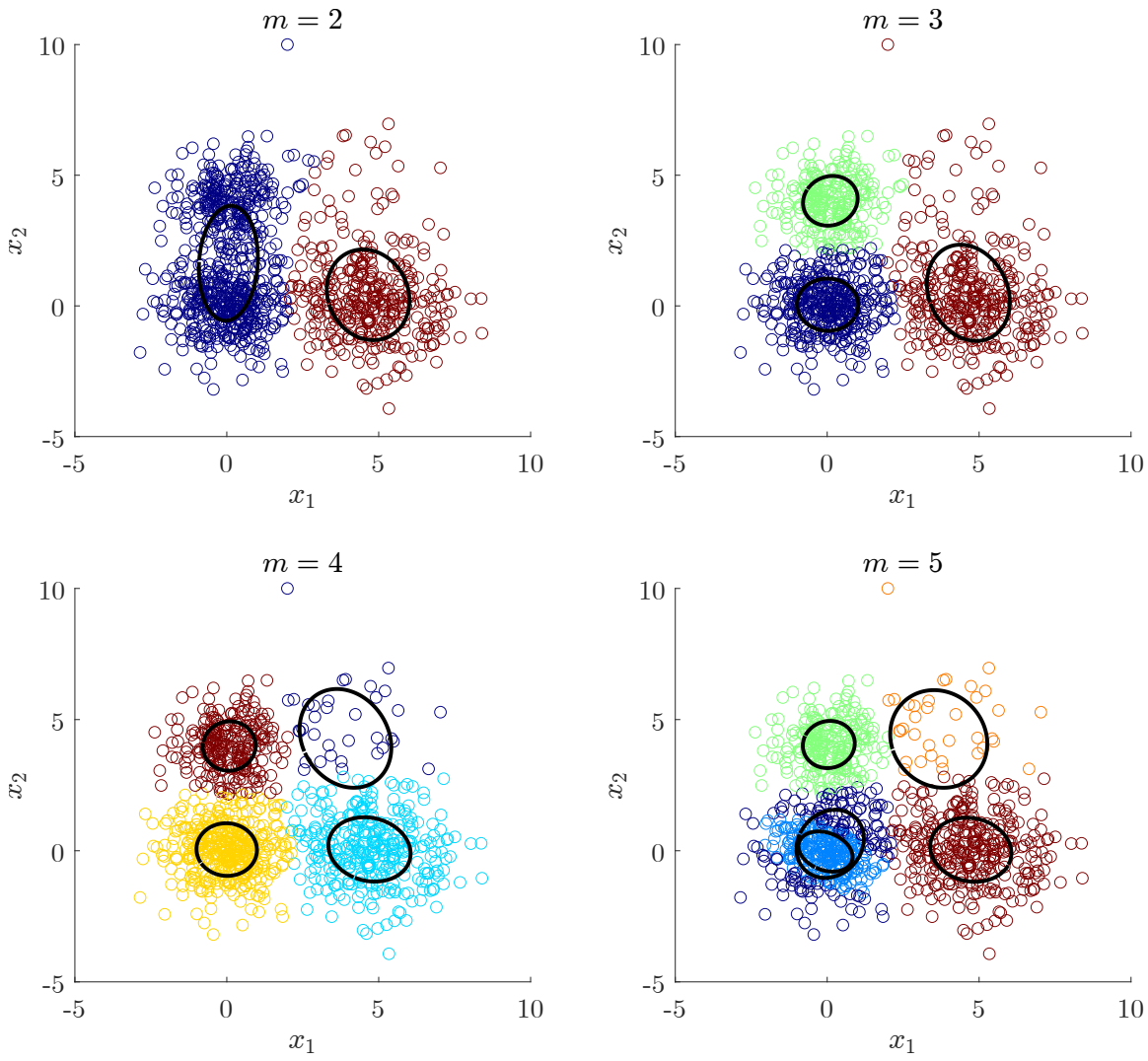

图 1: 使用 EM 算法对 GMM 模型进行估计

BIC 计算结果如表 1 所示, 可知当 $m=4$ 时, BIC 最小, 即 $m=4$ 是最恰当的模型参数.

表 1: 不同参数 m 取值对应的 BIC 数值

	$m=2$	$m=3$	$m=4$	$m=5$
BIC	8542.29	8371.61	8330.65	8372.01

此模型选择结果与视觉预估不符, 在初看数据时, 可以发现有 3 个明显的数据聚集中心, 因此猜测 $m=3$ 是最恰当的模型参数. 但是程序运行结果表明 $m=4$ 是最恰当的模型参数, 这是因为右上角那部分点虽然不够集中, 但是比较不容易被另外三个组分共同作用而产生, 相反若右上角存在一个单独的组分, 并且赋予其很小的比例系数, 则会使得模型拟合效果更好.

3.2. Modify the M-step of the EM code so that the covariance matrices of the Gaussian components are constrained to be equal. Give detailed derivation. Rerun the code and then select an appropriate model. Would we select a different number of components in this case?

解: EM 算法的 M-step 为

$$\Theta^{(i)} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta, \Theta^{(i-1)}) \tag{48}$$

设 GMM 模型为

$$p(\boldsymbol{x}|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(\boldsymbol{x}|\theta_i) \tag{49}$$

则有

$$\begin{aligned} Q(\Theta, \Theta^g) &= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_l p_l(x_i|\theta_l)) p(l|x_i, \Theta^g) \\ &= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \sum_{l=1}^{M}\sum_{i=1}^{N} \log(p_l(x_i|\theta_l)) p(l|x_i, \Theta^g) \end{aligned} \tag{50}$$

其中第一项不含 θ_l, 第二项不含 α_l, 则可在优化 $Q(\Theta, \Theta^g)$ 时分别进行优化.

由于 α_l 满足

$$\sum_{l=1}^{M} \alpha_l = 1 \tag{51}$$

则引入 Lagrange 乘子 λ 得到 Lagrange 函数

$$\mathcal{L}(\alpha_l, \lambda) = \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \lambda \left(\sum_{l=1}^{M} \alpha_l - 1 \right) \tag{52}$$

Lagrange 函数对 α_l 的偏导为

$$\frac{\partial \mathcal{L}(\alpha_l, \lambda)}{\partial \alpha_l} = \sum_{i=1}^{N} \frac{1}{\alpha_l} p(l|x_i, \Theta^g) + \lambda \tag{53}$$

令此偏导为 0 并对 l 求和, 有

$$\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g) + \sum_{l=1}^{M}\alpha_l\lambda = 0 \tag{54}$$

所以 $\lambda = -N$, 则

$$\alpha_l = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g) \tag{55}$$

定义

$$\begin{aligned} B &\triangleq \sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(x_i|\theta_l))p(l|x_i,\Theta^g) \\ &= \sum_{l=1}^{M}\sum_{i=1}^{N}\left[-\frac{\log|\Sigma|}{2} - \frac{(x_i-\mu_l)^\top\Sigma^{-1}(x_i-\mu_l)}{2}\right]p(l|x_i,\Theta^g) \\ &= \sum_{l=1}^{M}\left[\frac{1}{2}\log|\Sigma^{-1}|\sum_{i=1}^{N}p(l|x_i,\Theta^g) - \frac{1}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g)\mathrm{tr}\left(\Sigma^{-1}N_{l,i}\right)\right] \end{aligned} \tag{56}$$

其中 $N_{l,i} = (x_i-\mu_l)(x_i-\mu_l)^\top$.

B 对 μ_l 的偏导为

$$\frac{\partial B}{\partial \mu_l} = \sum_{i=1}^{N}\Sigma^{-1}(x_i-\mu_l)p(l|x_i,\Theta^g) \tag{57}$$

令此偏导为 0, 可得

$$\mu_l = \frac{\sum_{i=1}^{N}x_i p(l|x_i,\Theta^g)}{\sum_{i=1}^{N}p(l|x_i,\Theta^g)} \tag{58}$$

B 对 Σ^{-1} 的偏导为

$$\begin{aligned} \frac{\partial B}{\partial \Sigma^{-1}} &= \sum_{l=1}^{M}\left[\frac{1}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g)(2\Sigma - \mathrm{diag}(\Sigma)) - \frac{1}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g)(2N_{l,i} - \mathrm{diag}(N_{l,i}))\right] \\ &= \frac{1}{2}\sum_{l=1}^{M}\sum_{i=1}^{N}p(l|x_i,\Theta^g)(2M_{l,i} - \mathrm{diag}(M_{l,i})) \\ &= 2S - \mathrm{diag}(S) \end{aligned} \tag{59}$$

其中

$$M_{l,i} = \Sigma - N_{l,i}, \quad S = \frac{1}{2}\sum_{l=1}^{M}\sum_{i=1}^{N}p(l|x_i,\Theta^g)M_{l,i} \tag{60}$$

由 $2S - \mathrm{diag}(S) = 0$, 可知 $S = 0$, 即

$$\sum_{l=1}^{M}\sum_{i=1}^{N}p(l|x_i,\Theta^g)(\Sigma - N_{l,i}) = 0 \tag{61}$$

所以

$$\begin{aligned}\Sigma &= \frac{\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g)N_{l,i}}{\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g)} \\ &= \frac{\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g)(x_i-\mu_l)(x_i-\mu_l)^\top}{\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g)}\end{aligned} \tag{62}$$

则修改的 M-step 为

$$\begin{aligned}\alpha_l^{\text{new}} &= \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g) \\ \mu_l^{\text{new}} &= \frac{\sum_{i=1}^{N} x_i p(l|x_i,\Theta^g)}{\sum_{i=1}^{N} p(l|x_i,\Theta^g)} \\ \Sigma^{\text{new}} &= \frac{\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g)(x_i-\mu_l^{\text{new}})(x_i-\mu_l^{\text{new}})^\top}{\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g)}\end{aligned} \tag{63}$$

当每个组分的协方差矩阵都相同时, 测试过程中经常会出现 EM 算法收敛到局部最优解的情况, 为了尽可能避免局部最优解的影响, 对每个 m 进行 200 次测试, 选择对数似然概率最大的那次测试结果作为该组的结果, 如图 2 所示.

BIC 计算结果如表 2 所示, 可知当 $m=4$ 时, BIC 最小, 即 $m=4$ 仍为最恰当的模型参数.

表 2: 不同参数 m 取值对应的 BIC 数值

	$m=2$	$m=3$	$m=4$	$m=5$
BIC	8597.40	8463.17	8397.21	8415.93

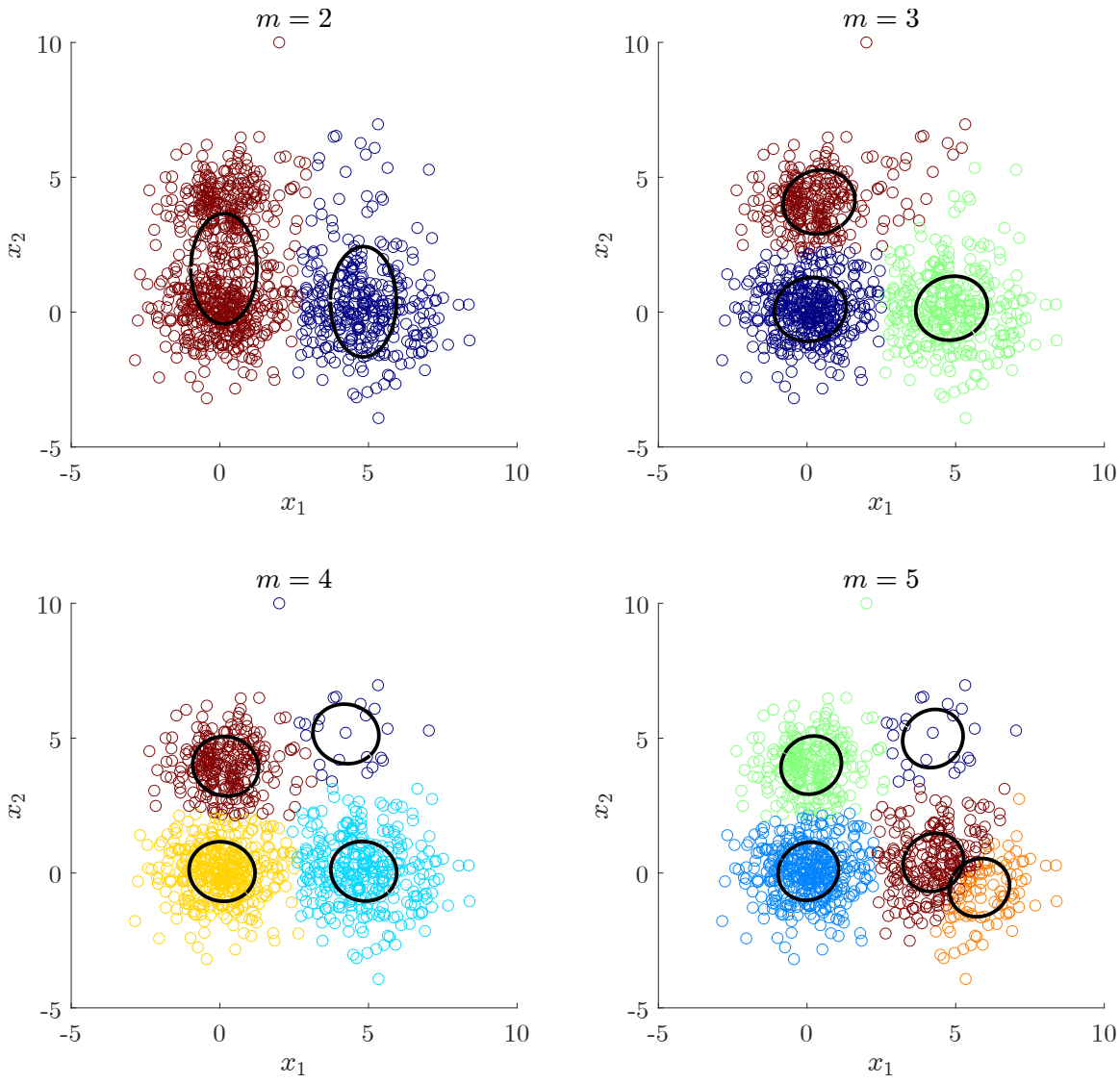

图 2: 使用修改的 EM 算法对 GMM 模型进行估计

Programming 2 (Missing Data)

4. Suppose we know that the ten data points in category ω_1 in Tab. 3 come from a three-dimensional Gaussian. Suppose, however, that we do not have access to the x_3 components for the even-numbered data points.

表 3: 样本数据点

	ω_1		
point	x_1	x_2	x_3
1	0.42	-0.087	0.58
2	-0.2	-3.3	-3.4
3	1.3	-0.32	1.7
4	0.39	0.71	0.23
5	-1.6	-5.3	-0.15
6	-0.029	0.89	-4.7
7	-0.23	1.9	2.2
8	0.27	-0.3	-0.87
9	-1.9	0.76	-2.1
10	0.87	-1.0	-2.6

4.1. Write an EM program to estimate the mean and covariance of the distribution. Start your estimate with $\boldsymbol{\mu}^0 = \mathbf{0}$ and $\boldsymbol{\Sigma}^0 = \boldsymbol{I}$, the three-dimensional identity matrix.

解: 记 $O = \{1, 3, 5, 7, 9\}$, $E = \{2, 4, 6, 8, 10\}$. 偶数点的 x_3 数据丢失, 则令 x_{i3}, $i \in E$ 为隐变量, 并令 $q(\cdot)$ 表示其概率分布, 则对数似然函数为

$$\begin{aligned} H(\theta) &= \sum_{i\in O} \ln p(x_{i1}, x_{i2}, x_{i3}|\theta) + \sum_{i\in E} \ln p(x_{i1}, x_{i2}|\theta) \\ &= \sum_{i\in O} \ln p(x_{i1}, x_{i2}, x_{i3}|\theta) + \sum_{i\in E} \ln \int_{-\infty}^{\infty} p(x_{i1}, x_{i2}, x_{i3}|\theta) \mathrm{d}x_{i3} \\ &= \sum_{i\in O} \ln p(x_{i1}, x_{i2}, x_{i3}|\theta) + \sum_{i\in E} \ln \int_{-\infty}^{\infty} q(x_{i3}) \frac{p(x_{i1}, x_{i2}, x_{i3}|\theta)}{q(x_{i3})} \mathrm{d}x_{i3} \\ &\geqslant \sum_{i\in O} \ln p(x_{i1}, x_{i2}, x_{i3}|\theta) + \sum_{i\in E} \int_{-\infty}^{\infty} q(x_{i3}) \ln \frac{p(x_{i1}, x_{i2}, x_{i3}|\theta)}{q(x_{i3})} \mathrm{d}x_{i3} \end{aligned} \tag{64}$$

由此可得 Q 函数为

$$Q(\theta^{(k)}, \theta) = \sum_{i\in O} \ln p(x_{i1}, x_{i2}, x_{i3}|\theta) + \sum_{i\in E} \int_{-\infty}^{\infty} p(x_{i3}|x_{i1}, x_{i2}, \theta^{(k)}) \ln p(x_{i1}, x_{i2}, x_{i3}|\theta) \mathrm{d}x_{i3} \tag{65}$$

令 $x_i = (x_{i1}, x_{i2}, x_{i3})^\top$, $\mu = (\mu_1, \mu_2, \mu_3)^\top$, 则

$$\ln p(x_{i1}, x_{i2}, x_{i3}|\theta) = \ln p(x_i|\theta) = -\frac{3}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma^{-1}| - \frac{1}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) \tag{66}$$

Q 对 μ 的偏导数为

$$\begin{aligned}\frac{\partial Q(\theta^{(k)},\theta)}{\partial \mu} &= \sum_{i\in O}\Sigma^{-1}(x_i-\mu)+\sum_{i\in E}\int_{-\infty}^{\infty}p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})\Sigma^{-1}(x_i-\mu)\mathrm{d}x_{i3}\\ &= \sum_{i\in O}\Sigma^{-1}\begin{pmatrix}x_{i1}-\mu_1\\ x_{i2}-\mu_2\\ x_{i3}-\mu_3\end{pmatrix}+\sum_{i\in E}\Sigma^{-1}\begin{pmatrix}x_{i1}-\mu_1\\ x_{i2}-\mu_2\\ \displaystyle\int_{-\infty}^{\infty}x_{i3}p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})\mathrm{d}x_{i3}-\mu_3\end{pmatrix}\\ &= \Sigma^{-1}\begin{pmatrix}\displaystyle\sum_{i=1}^{10}(x_{i1}-\mu_1)\\ \displaystyle\sum_{i=1}^{10}(x_{i2}-\mu_2)\\ \displaystyle\sum_{i\in O}x_{i3}+\sum_{i\in E}\mathbb{E}(x_{i3}|x_{i1},x_{i2},\theta^{(k)})-\sum_{i=1}^{10}\mu_3\end{pmatrix}\end{aligned}\tag{67}$$

令此偏导数为 0, 有

$$\begin{aligned}\mu_1 &= \frac{1}{10}\sum_{i=1}^{10}x_{i1}\\ \mu_2 &= \frac{1}{10}\sum_{i=1}^{10}x_{i2}\\ \mu_3 &= \frac{1}{10}\sum_{i\in O}x_{i3}+\frac{1}{10}\sum_{i\in E}\mathbb{E}(x_{i3}|x_{i1},x_{i2},\theta^{(k)})\end{aligned}\tag{68}$$

由高斯分布的性质可知, 高斯分布的条件分布仍为高斯分布, 记

$$\Sigma^{(k)}=\begin{bmatrix}\Sigma_{12,12}^{(k)} & \Sigma_{12,3}^{(k)}\\ \Sigma_{3,12}^{(k)} & \Sigma_{3,3}^{(k)}\end{bmatrix}\tag{69}$$

则由相关公式 [1–3] 可知

$$\mathbb{E}(x_{i3}|x_{i1},x_{i2},\theta^{(k)})=\mu_3^{(k)}+\Sigma_{3,12}^{(k)}\left[\Sigma_{12,12}^{(k)}\right]^{-1}\begin{pmatrix}x_{i1}-\mu_1^{(k)}\\ x_{i2}-\mu_2^{(k)}\end{pmatrix}\tag{70}$$

所以

$$\mu_3=\frac{1}{10}\sum_{i\in O}x_{i3}+\frac{1}{10}\sum_{i\in E}\left[\mu_3^{(k)}+\Sigma_{3,12}^{(k)}\left[\Sigma_{12,12}^{(k)}\right]^{-1}\begin{pmatrix}x_{i1}-\mu_1^{(k)}\\ x_{i2}-\mu_2^{(k)}\end{pmatrix}\right]\tag{71}$$

令

$$N_i^{(k)}=(x_i-\mu^{(k+1)})(x_i-\mu^{(k+1)})^{\top},\quad M_i=\Sigma-N_i^{(k)}\tag{72}$$

则 Q 对 Σ^{-1} 的偏导数为

$$\frac{\partial Q(\theta^{(k)},\theta)}{\partial \Sigma^{-1}} = \sum_{i\in O}\frac{1}{2}\big[2M_i - \mathrm{diag}(M_i)\big] + \sum_{i\in E}\int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})\frac{1}{2}\big[2M_i - \mathrm{diag}(M_i)\big]\mathrm{d}x_{i3} \tag{73}$$

令此偏导数为 0, 有

$$\sum_{i\in O} M_i + \sum_{i\in E}\int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})M_i\mathrm{d}x_{i3} = 0 \tag{74}$$

即

$$\sum_{i\in O}(\Sigma - N_i^{(k)}) + \sum_{i\in E}\int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})(\Sigma - N_i^{(k)})\mathrm{d}x_{i3} = 0 \tag{75}$$

由概率密度函数积分为 1, 可得

$$\sum_{i=1}^{10}\Sigma - \sum_{i\in O} N_i^{(k)} - \sum_{i\in E}\int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})N_i^{(k)}\mathrm{d}x_{i3} = 0 \tag{76}$$

所以

$$\begin{aligned}\Sigma &= \frac{1}{10}\sum_{i\in O} N_i^{(k)} + \frac{1}{10}\sum_{i\in E}\int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})N_i^{(k)}\mathrm{d}x_{i3}\\ &\triangleq \frac{1}{10}\sum_{i\in O} N_i^{(k)} + \frac{1}{10}\sum_{i\in E} W_i^{(k)}\end{aligned} \tag{77}$$

其中

$$\begin{aligned}W_i^{(k)} &= \int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})(x_i-\mu^{(k+1)})(x_i-\mu^{(k+1)})^\top\mathrm{d}x_{i3}\\ &= \int_{-\infty}^{\infty} p(x_{i3}|x_{i1},x_{i2},\theta^{(k)})\begin{bmatrix}(x_{i1}-\mu_1)^2 & (x_{i1}-\mu_1)(x_{i2}-\mu_2) & (x_{i1}-\mu_1)(x_{i3}-\mu_3)\\ (x_{i2}-\mu_2)(x_{i1}-\mu_1) & (x_{i2}-\mu_2)^2 & (x_{i2}-\mu_2)(x_{i3}-\mu_3)\\ (x_{i3}-\mu_3)(x_{i1}-\mu_1) & (x_{i3}-\mu_3)(x_{i2}-\mu_2) & (x_{i3}-\mu_3)^2\end{bmatrix}\mathrm{d}x_{i3}\end{aligned} \tag{78}$$

记

$$E_i^{(k)} \triangleq \mathbb{E}(x_{i3}|x_{i1},x_{i2},\theta^{(k)}) = \mu_3^{(k)} + \Sigma_{3,12}^{(k)}\left[\Sigma_{12,12}^{(k)}\right]^{-1}\begin{pmatrix}x_{i1}-\mu_1^{(k)}\\ x_{i2}-\mu_2^{(k)}\end{pmatrix} \tag{79}$$

以及

$$D_i^{(k)} \triangleq \mathrm{Var}(x_{i3}|x_{i1},x_{i2},\theta^{(k)}) = \Sigma_{3,3}^{(k)} - \Sigma_{3,12}^{(k)}\left[\Sigma_{12,12}^{(k)}\right]^{-1}\Sigma_{12,3}^{(k)} \tag{80}$$

则由期望和方差的定义得到

$$W_i^{(k)} = \begin{bmatrix}(x_{i1}-\mu_1)^2 & (x_{i1}-\mu_1)(x_{i2}-\mu_2) & (x_{i1}-\mu_1)\left(E_i^{(k)}-\mu_3\right)\\ (x_{i2}-\mu_2)(x_{i1}-\mu_1) & (x_{i2}-\mu_2)^2 & (x_{i2}-\mu_2)\left(E_i^{(k)}-\mu_3\right)\\ \left(E_i^{(k)}-\mu_3\right)(x_{i1}-\mu_1) & \left(E_i^{(k)}-\mu_3\right)(x_{i2}-\mu_2) & D_i^{(k)} + \left(E_i^{(k)}\right)^2 - 2\mu_3 E_i^{(k)} + \mu_3^2\end{bmatrix} \tag{81}$$

综上所述, 有

$$\begin{aligned}
\mu_1^{k+1} &= \frac{1}{10}\sum_{i=1}^{10} x_{i1} \\
\mu_2^{k+1} &= \frac{1}{10}\sum_{i=1}^{10} x_{i2} \\
\mu_3^{k+1} &= \frac{1}{10}\sum_{i\in O} x_{i3} + \frac{1}{10}\sum_{i\in E} E_i^{(k)} \\
\Sigma^{k+1} &= \frac{1}{10}\sum_{i\in O} N_i^{(k)} + \frac{1}{10}\sum_{i\in E} W_i^{(k)}
\end{aligned} \tag{82}$$

编写程序, 得到估计的均值向量和协方差矩阵分别为

$$\hat{\mu} = \begin{pmatrix} -0.0709 \\ -0.6047 \\ 0.7728 \end{pmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 0.9062 & 0.5678 & 0.8814 \\ 0.5678 & 4.2007 & 0.4621 \\ 0.8814 & 0.4621 & 1.7828 \end{bmatrix} \tag{83}$$

4.2. Compare your final estimation with the case when we remove all even-numbered data points (2, 4, 6, 8, 10).

解: 均值向量和协方差矩阵分别为

$$\hat{\mu} = \begin{pmatrix} -0.4020 \\ -0.6094 \\ 0.4460 \end{pmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 1.4563 & 0.9843 & 1.4148 \\ 0.9843 & 6.1061 & 0.8287 \\ 1.4148 & 0.8287 & 2.3009 \end{bmatrix} \tag{84}$$

全部偶数数据点的缺失导致均值和协方差矩阵的估计有了很大的变化, 除了 x_2 的均值估计基本保持不变外, 其余数值均发生了较大的变化, 与只缺失 x_3 的偶数数据点相比近乎是两个不同的模型了, 可以看出数据点的个数对模型参数的估计有较大的影响.

4.3. Compare your final estimation with the case when there are no missing data, namely we have access to all x_3.

解: 均值向量和协方差矩阵分别为

$$\hat{\mu} = \begin{pmatrix} -0.0709 \\ -0.6047 \\ -0.9110 \end{pmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 0.9062 & 0.5678 & 0.3941 \\ 0.5678 & 4.2007 & 0.7337 \\ 0.3941 & 0.7337 & 4.5419 \end{bmatrix} \tag{85}$$

由于只缺失第三维的部分数据, 所以前两维的均值和协方差估计都是准确的, 而第三维有明显的误差, 且误差较大. 这主要是因为缺失的五个数据, 恰好都是 x_3 比较小的数据, 导致估计有了较大的误差.

参考文献

[1] Flying pig. Deriving the conditional distributions of a multivariate normal distribution, 2012. https://stats.stackexchange.com/questions/30588

[2] Ruye Wang. Marginal and conditional distributions of multivariate normal distribution, 2006. http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html

[3] Steffen Lauritzen. The Multivariate Gaussian Distribution, 2009. http://www.stats.ox.ac.uk/ steffen/teaching/bs2HT9/gauss.pdf