EM 算法用于混合高斯模型的参数估计

中国民航大学计算机学院，张志远
2010 年 2 月 25 日

1．极大似然估计

假定随机变量X服从某一个参数为θ的分布，其概率密度为$P(x;\theta)$, $\theta\in\Theta$，其中θ为带估参数，Θ是θ的可能取值范围。设X_1, X_2，……, X_n是来自X的样本，则X_1, X_2，……, X_n的联合分布概率为

$$\prod_{i=1}^{n} p(x_i;\theta)\text{。}$$

又设x_1, x_2, ……, x_n是相应于样本X_1, X_2，……, X_n的一个样本值。则事件$\{X_1 = x_1$, $X_2 = x_2$, ……, $X_n = x_n\}$发生的概率为

$$L(\theta)=L(x_1,x_2,\ldots\ldots,x_n;\theta)=\prod_{i=1}^{n} p(x_i;\theta),\theta\in\Theta$$

这一概率随 θ 的取值变化而变化，称 $L(\theta)$为样本的似然函数。

最大似然估计法就是根据固定样本观察值x_1, x_2, ……, x_n，在θ可能的取值范围Θ内挑选使似然函数$L(\theta)$达到最大的参数$\hat{\theta}$，作为参数θ的估计值。即取$\hat{\theta}$使

$$L(x_1,x_2,\ldots\ldots,x_n;\hat{\theta})=\max_{\theta\in\Theta} L(x_1,x_2,\ldots\ldots,x_n;\theta)$$

称 $Log(L(\theta))$为对数似然估计。

对于最大似然估计的求解一般通过微分学中的极值问题求解。$\hat{\theta}$一般可从方程

$$\frac{d}{d\theta}\ln(L(\theta))=0$$

求得。

2．EM 算法的一般描述

当有部分数据缺失或者无法观察到时，EM算法提供了一个高效的迭代程序用来计算这些数据的最大似然估计。在每一步迭代分为两个步骤：期望（Expectation）步骤和最大化（Maximization）步骤，因此称为EM算法。

假设全部数据Z是由可观测到的样本$X=\{X_1, X_2，……, X_n\}$和不可观测到的样本$Y=\{Y_1, Y_2，……, Y_n\}$组成的，则$Z = X\cup Y$。EM算法通过搜寻使全部数据的似然

函数Log(L(Z; θ))的期望值最大来寻找极大似然估计$\hat{\theta}$，注意此处的θ不是一个变量，而是多个变量组成的参数集合。此期望值是在Z所遵循的概率分布上计算，此分布由未知参数θ确定。然而Z所遵循的分布是未知的。EM算法使用其当前的假设θ_{t-1}代替实际参数θ，以估计Z的分布。

$$Q(\theta,\theta_{t-1}) = E[\ln L(Z\,|\,\theta)\,|\,X,\theta_{t-1}]$$

EM算法重复以下两个步骤直至收敛。

步骤1：E步骤：使用当前假设θ_{t-1}和观测到的数据X来估计Z上的概率分布以计算Q(θ, θ_{t-1})。

$$Q(\theta,\theta_{t-1}) \leftarrow E[\ln L(Z\,|\,\theta)\,|\,X,\theta_{t-1}]$$

步骤2：M步骤：求使函数Q(θ, θ_{t-1})最大化的θ并将其作为下一次循环的假设θ_t。

$$\theta_t \leftarrow \arg\max_{\theta\in\Theta} Q(\theta,\theta_{t-1})$$

3．EM 算法用于 K 均值问题的参数估计

为简单起见，假设 k=2，然后再扩展到任意个 k 的情况。

问题描述：假设X是由 2 个高斯分布**均匀混合**而成的，这两个高斯分布的均值不同，但是具有相同的方差。设样本值为x_1, x_2, ……, x_n，其中每一个x_i可以表示为一个三元组< x_i, z_{i1}, z_{i2}>，其中z_{i1}和z_{i2}中只能有一个取 1，其余的为 0。此处的z_{i1}和z_{i2}为隐藏变量，是未知的。此处的**均匀混合**是指x_i可能来自于第一个分布，也可能来自于第二个分布，其概率相等。

步骤 1：假定当前的假设h=<μ_1, μ_2 >成立，计算每个z_{ij}的期望值，即实例x_i由第j个正态分布生成的概率的期望。

$$E[Z_{ij}] = \frac{p(X = x_i\,|\,\mu = \mu_j)}{p(X = x_i)} = \frac{Exp(-\frac{1}{2\sigma^2}(x_i-\mu_j)^2)}{\sum_{k=1}^{2} Exp(-\frac{1}{2\sigma^2}(x_i-\mu_k)^2)}$$

步骤 2：使用第一步中计算出来的E [Z_{ij}]来导出新的极大似然假设h'=<μ_1', μ_2'>。

$$\mu_j \leftarrow \frac{\sum_{i=1}^{n} E[z_{ij}]x_i}{\sum_{i=1}^{n} E[z_{ij}]}$$

3.1 对于任意个 K 的扩展

问题描述：假设X是由K个高斯分布**均匀混合**而成的，这K个高斯分布的均值不同，但是具有相同的方差。设样本值为$x_1, x_2, \ldots\ldots, x_n$，其中每一个$x_i$可以表示为一个K+1 元组< $x_i, z_{i1}, z_{i2}, \ldots, z_{ik}$>，其中只有一个取 1，其余的为 0。此处的$z_{i1}$到$z_{ik}$为隐藏变量，是未知的。且任意$z_{ij}$被选择的概率相等，即

$$P\{z_{ij}=1, j=1,2,\ldots,K\}=\frac{1}{K}\text{。}$$

步骤 1：求E[lnP(Y|h')]。注意每个实例y_i=< $x_i, z_{i1}, z_{i2}, \ldots, z_{ik}$>的概率$p(y_i \mid h')$为：

$$p(y_i \mid h')=p(x_i, z_{i1},\ldots,z_{ik} \mid h')=\frac{1}{\sqrt{2\pi\sigma^2}}Exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{k}z_{ij}(x_i-u'_j)^2)$$

注意只有一个z_{ij}值为 1，其余的为 0。因此：

$$LnP(Y \mid h')=Ln\prod_{i=1}^{n}p(y_i \mid h')=\sum_{i=1}^{n}Lnp(y_i \mid h')=\sum_{i=1}^{n}\left(Ln\frac{1}{\sqrt{2\pi\sigma^2}}-\frac{1}{2\sigma^2}\sum_{j=1}^{k}z_{ij}(x_i-u'_j)^2\right)$$

因此有：

$$E\left[LnP(Y \mid h')\right]=E\left[\sum_{i=1}^{n}\left(Ln\frac{1}{\sqrt{2\pi\sigma^2}}-\frac{1}{2\sigma^2}\sum_{j=1}^{k}z_{ij}(x_i-u'_j)^2\right)\right]$$
$$=\sum_{i=1}^{n}\left(Ln\frac{1}{\sqrt{2\pi\sigma^2}}-\frac{1}{2\sigma^2}\sum_{j=1}^{k}E[z_{ij}](x_i-u'_j)^2\right)$$

而：

$$E[Z_{ij}]=\frac{p(X=x_i \mid \mu=\mu_j)}{p(X=x_i)}=\frac{Exp(-\frac{1}{2\sigma^2}(x_i-\mu_j)^2)}{\sum_{k=1}^{K}Exp(-\frac{1}{2\sigma^2}(x_i-\mu_k)^2)}$$

步骤 2：求使E[lnP(Y|h')]期望最大的假设h'=<$\mu'_1, \ldots \mu'_k$>

对E[lnP(Y|h')]求μ'_j的偏导数：

$$\frac{\partial}{\partial\mu'_j}E[LnP(Y \mid h')]=\sum_{i=1}^{n}\frac{1}{\sigma^2}E[z_{ij}](\mu'_j-x_i)=\sum_{i=1}^{n}\frac{1}{\sigma^2}E[z_{ij}]\mu'_j-\sum_{i=1}^{n}\frac{1}{\sigma^2}E[z_{ij}]x_i=0$$

因此，得：

$$u'_j=\frac{\sum_{i=1}^{n}E[z_{ij}]x_i}{\sum_{i=1}^{n}E[z_{ij}]}$$

3.2 Matlab 实现

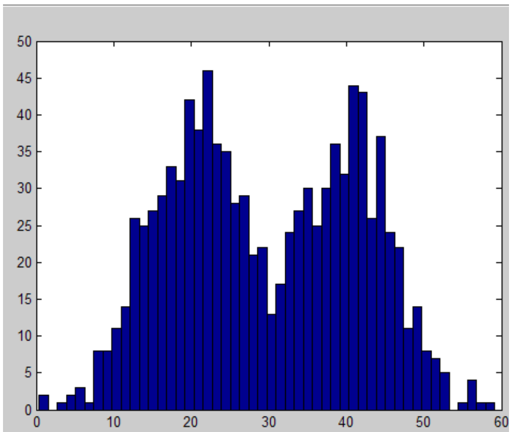

图 1 测试样本的分布情况

```
% EM 测试用于模拟 2 个正态分布的均值估计
% 中国民航大学计算机学院，张志远
% 2010 年 2 月 25 日

% ----------------------------------------------------
% 生成符合条件的随机样本
% ----------------------------------------------------
% 首先指定两个高斯分布的参数，注意这两个高斯分布的方差相同
Sigma = 6;
Miu1 = 40;
Miu2 = 20;
% 然后随机均匀选择两种高斯分布，并生成样本值
N = 1000;
X = zeros(1, N);
for i=1:N
    if rand > 0.5     % 选择第一个高斯分布
        X(1,i) = randn * Sigma + Miu1;
    else              % 选择第二个高斯分布
        X(1,i) = randn * Sigma + Miu2;
    end
end

% ----------------------------------------------------
% EM 算法，步骤 1，计算 E[Zij]
% ----------------------------------------------------
% 首先假设两个高斯分布的均值为随机的两个值
k = 2;
Miu = rand(1,k);
Expectations = zeros(N, k);

for step = 1 : 1000
    % 计算 E[Zij]
```

```
    for i = 1 : N
        % 计算分母
        Denominator = 0;
        for j = 1 : k
            Denominator = Denominator + exp(-1/(2*Sigma^2) * (X(1,i)-Miu(1,j))^2);
        end
        for j = 1 : k
            % 计算分子
            Numerator = exp(-1/(2*Sigma^2) * (X(1,i)-Miu(1,j))^2);
            Expectations(i,j) = Numerator / Denominator;
        end
    end

    % --------------------------------------------------
    % EM 算法，步骤 2，求最大化 E[Zij]的参数 Miu
    % --------------------------------------------------
    OldMiu = Miu;
    for j = 1 : k
        Numerator = 0;
        Denominator = 0;
        for i = 1 : N
            Numerator = Numerator + Expectations(i,j) * X(1,i);
            Denominator = Denominator + Expectations(i,j);
        end
        Miu(1,j) = Numerator / Denominator;
    end

    % --------------------------------------------------
    % 判断假设是否发生了足够的变化
    % --------------------------------------------------
    Epsilon = 0.0001;
    if sum(abs(Miu - OldMiu)) < Epsilon
        break;
    end
    step
    Miu
end
disp('============Miu===================');
Miu
hist(X,50)
```

输出结果：

在第 13 步时迭代终止，得均值估计为：

Miu = [19.9747　　40.2692]

讨论，如果初始值设置的两个 Miu 值相同，例如 Miu=[0, 0]，则迭代一次即终止，始终得不到正确的结果。

4. EM 算法用于混合高斯模型

混合模型是指随机变量 X 的概率密度为：

$$p(x|\Theta)=\sum_{k=1}^{M}\alpha_k p_k(x|\theta_k),且\sum_{\mathrm{k}}\alpha_{\mathrm{k}}=1$$

此处的$\Theta=(\alpha_1,\ldots,\alpha_M,\theta_1,\ldots,\theta_M)$，即混合模型由M个分量组成，每个分量的权重系数为$\alpha_k$。如果每一个分量都满足高斯分布，则称其为混合高斯模型。由此可知，上面的K均值问题是混合高斯分布的一个特例。

例如一个班级学生的身高为X，假设男生和女生的身高分别服从$N\sim(\mu_1,\sigma_1^2)$和$N\sim(\mu_2,\sigma_2^2)$，则$X\sim\alpha_1N(\mu_1,\sigma_1^2)+\alpha_2N(\mu_2,\sigma_2^2)$，且$\alpha_1+\alpha_2=1$，$0\leqslant\alpha_1,\alpha_2\leqslant1$。

对数似然函数：

$$Ln(L(\Theta|X))=Ln\prod_{i=1}^{N}p(x_i|\Theta)=\sum_{i=1}^{N}Ln\left(\sum_{k=1}^{M}\alpha_k p_k(x_i|\theta_k)\right)$$

这个函数因为含有加法的对数，不太好直接计算极值，因此将其改写为含有引变量Y的函数。设$Y=(y_1,y_2,\ldots,y_n)$，且$y_i\in\{1,2,\ldots,M\}$, i=1, 2, …, N。当$y_i=k$时表示第i个样本观测值x_i是由第k个分量产生的。因此，引入隐含变量Y后的对数似然函数为：

$$Ln(L(\Theta|X,Y))=Ln\prod_{i=1}^{N}p(x_i,y_i|\Theta)=\sum_{i=1}^{N}Ln\left(\alpha_{yi}p_{yi}(x_i|\theta_{yi})\right)$$

步骤 1：求完整数据的对数似然函数的期望。注意变量是Y，且X和Θ^{t-1}已知。此处的$Y=(y_1,\ldots,y_N),\Theta^{t-1}=\left(\alpha_1^{t-1},\ldots,\alpha_M^{t-1},\theta_1^{t-1},\ldots,\theta_M^{t-1}\right)$是上次推出的假设参数值（初始参数可以随机指定）。

$$Q(\Theta,\Theta^{t-1})=E\left[Ln(L(\Theta|X,Y)\right]=\sum_{y}Ln(L(\Theta|X,y)p(y|X,\Theta^{t-1})$$

$$=\sum_{y}\sum_{i=1}^{N}Ln\left(\alpha_{yi}p_{yi}(x_i|\theta_{yi})\right)p(y|x_i,\Theta^{t-1})$$

第i个观测样本x_i取第k个分量的概率密度为$p(y_i=k\,|\,x_i,\Theta^{t-1})$。因此上式

$$=\sum_{k=1}^{M}\sum_{i=1}^{N}Ln(\alpha_k p_k(x_i\mid\theta_k))p(k\mid x_i,\Theta^{t-1})$$
$$=\sum_{k=1}^{M}\sum_{i=1}^{N}Ln(\alpha_k)p(k\mid x_i,\Theta^{t-1})+\sum_{k=1}^{M}\sum_{i=1}^{N}Ln(p_k(x_i\mid\theta_k))p(k\mid x_i,\Theta^{t-1})$$

由于每一个成分都满足高斯分布，因此

$$p_k(x_i\mid\theta_k)=\frac{1}{\sqrt{2\pi\sigma_k{}^2}}Exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right)$$

而由贝叶斯公式，可得：

$$p(k\mid x_i,\Theta^{t-1})=\frac{p(k,x_i\mid\theta_k^{t-1})}{\sum_{k'=1}^{M}p(k',x_i\mid\theta_{k'}^{t-1})}=\frac{p_k(x_i\mid\theta_k^{t-1})}{\sum_{k'=1}^{M}p_k{}'(x_i\mid\theta_{k'}^{t-1})}$$

步骤二：求满足期望最大的参数Θ^t。

【1】求μ_k

对μ_k求偏导数得：

$$\frac{\partial}{\partial\mu_k}Q(\Theta,\Theta^{t-1})=\frac{\partial}{\partial\mu_k}\left(\sum_{i=1}^{N}Ln(p_k(x_i\mid\theta_k))p(k\mid x_i,\Theta^{t-1})\right)$$
$$=\frac{\partial}{\partial\mu_k}\left(\sum_{i=1}^{N}\left(Ln\frac{1}{\sqrt{2\pi\sigma_k{}^2}}-\frac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right)p(k\mid x_i,\Theta^{t-1})\right)$$
$$=\sum_{i=1}^{N}\left(\frac{1}{\sigma_k{}^2}(x_i-\mu_k)p(k\mid x_i,\Theta^{t-1})\right)$$
$$=\frac{1}{\sigma_k{}^2}\sum_{i=1}^{N}\left(x_i p(k\mid x_i,\Theta^{t-1})-\mu_k p(k\mid x_i,\Theta^{t-1})\right)$$
$$=\frac{1}{\sigma_k{}^2}\sum_{i=1}^{N}\left(x_i p(k\mid x_i,\Theta^{t-1})\right)-\frac{\mu_k}{\sigma_k{}^2}\sum_{i=1}^{N}p(k\mid x_i,\Theta^{t-1})$$
$$=0$$

由此，我们得到：

$$\mu_k=\frac{\sum_{i=1}^{N}x_i p(k\mid x_i,\Theta^{t-1})}{\sum_{i=1}^{N}p(k\mid x_i,\Theta^{t-1})}$$

【2】求$\sigma_k{}^2$

对$\sigma_k{}^2$求偏导数得：

$$\frac{\partial}{\partial \sigma_k{}^2} Q(\Theta,\Theta^{t-1}) = \frac{\partial}{\partial \sigma_k{}^2}\left(\sum_{i=1}^{N} Ln\big(p_k(x_i \mid \theta_k)\big) p(k \mid x_i,\Theta^{t-1})\right)$$
$$= \frac{\partial}{\partial \sigma_k{}^2}\left(\sum_{i=1}^{N}\left(Ln\frac{1}{\sqrt{2\pi\sigma_k{}^2}} - \frac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right) p(k \mid x_i,\Theta^{t-1})\right)$$
$$= \frac{\partial}{\partial \sigma_k{}^2}\left(\sum_{i=1}^{N}\left(-\frac{1}{2}Ln(2\pi) - \frac{1}{2}Ln(\sigma_k{}^2) - \frac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right) p(k \mid x_i,\Theta^{t-1})\right)$$
$$= \sum_{i=1}^{N}\left(-\frac{1}{2\sigma_k{}^2} + \frac{(x_i-\mu_k)^2}{2(\sigma_k{}^2)^2}\right) p(k \mid x_i,\Theta^{t-1})$$
$$= \frac{1}{2(\sigma_k{}^2)^2}\sum_{i=1}^{N}\left((x_i-\mu_k)^2 - \sigma_k{}^2\right) p(k \mid x_i,\Theta^{t-1})$$
$$= \frac{1}{2(\sigma_k{}^2)^2}\sum_{i=1}^{N}(x_i-\mu_k)^2\, p(k \mid x_i,\Theta^{t-1}) - \frac{1}{2(\sigma_k{}^2)^2}\sum_{i=1}^{N}\sigma_k{}^2\, p(k \mid x_i,\Theta^{t-1})$$
$$= 0$$

由此，我们得到：

$$\sigma_k{}^2 = \frac{\sum_{i=1}^{N}(x_i-\mu_k)^2\, p(k \mid x_i,\Theta^{t-1})}{\sum_{i=1}^{N} p(k \mid x_i,\Theta^{t-1})}$$

【3】求α_k

求α_k需要借助于拉格朗日乘子，不能单纯使用偏导数求解，因为所有的α_k加起来必须等于 1。

$$\frac{\partial}{\partial \alpha_k}\left[Q(\Theta,\Theta^{t-1}) + \lambda(\sum_{k=1}^{M}\alpha_k - 1)\right] = 0$$
$$\frac{\partial}{\partial \alpha_k}\left[\sum_{k=1}^{M}\sum_{i=1}^{N} Ln(\alpha_k) p(k \mid x_i,\Theta^{t-1}) + \lambda(\sum_{k=1}^{M}\alpha_k - 1)\right] = 0$$
$$\frac{1}{\alpha_k}\sum_{i=1}^{N} p(k \mid x_i,\Theta^{t-1}) + \lambda = 0$$
$$\sum_{i=1}^{N} p(k \mid x_i,\Theta^{t-1}) + \lambda\alpha_k = 0$$

将所有的M个α_k式子加起来，得：

$$\sum_{k=1}^{M}\left[\sum_{i=1}^{N} p(k \mid x_i,\Theta^{t-1}) + \lambda\alpha_k\right] = 0$$
$$\sum_{i=1}^{N}\sum_{k=1}^{M} p(k \mid x_i,\Theta^{t-1}) + \lambda\sum_{k=1}^{M}\alpha_k = 0$$

注意到：$\sum_{k=1}^{M} p(k \mid x_i, \Theta^{t-1}) = 1, \sum_{k=1}^{M} \alpha_k = 1$。

因此有：$\lambda = -N$，得

$$\alpha_k = -\frac{1}{N} \sum_{i=1}^{N} p(k \mid x_i, \Theta^{t-1})$$

4.1 Matlab 实现

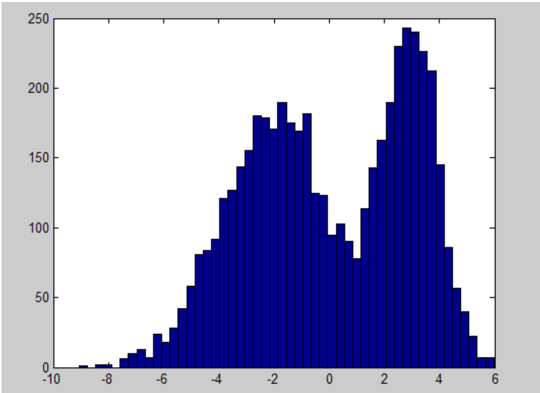

图 2 测试样本的分布情况

```
% EM 测试用于模拟 k 个正态分布的混合模型的参数估计
% 中国民航大学计算机学院，张志远
% 2010 年 2 月 27 日

% ----------------------------------------------------
% 生成符合条件的随机样本
% ----------------------------------------------------
% 首先指定两个高斯分布的参数
Miu1 = 3;
Miu2 = -2;
Sigma1 = 1;
Sigma2 = 2;
Alpha1 = 0.4;
Alpha2 = 0.6;
% 然后 alpha1 和 alpha2 的概率选择两种高斯分布，并生成样本值
N = 5000;
X = zeros(1, N);
N1 = floor(N * Alpha1);            %第一种分布的样本数
N2 = N - N1;                           %第二种分布的样本数
X(1:N1) = randn(1, N1) * Sigma1 + Miu1;
X(N1+1:N) = randn(1, N2) * Sigma2 + Miu2;
hist(X,50)
% ----------------------------------------------------
% EM 算法，步骤 1，计算均值
```

```
% ------------------------------------------------------
% 首先假设两个高斯分布的均值为随机的两个值,注意此处的 Sigma 是 Sigma 的平方。
M = 2;
Miu = rand(1,M);
Sigma = rand(1,M);
Alpha = [0.2, 0.8];

PkMatrix = zeros(N, M);

for step = 1 : 100
    % 计算均值
    for i = 1 : N
        for k = 1 : M
            PkMatrix(i,k) = exp(-1*(X(1,i)-Miu(1,k))^2/(2*Sigma(1,k))) / sqrt(2*pi*Sigma(1,k));
        end
    end
    % SumPkMatrixI 是 N*1 的矩阵，其每一行的值等于 P1i + P2i + P3i + ... + PMi
    % 即 SumPkMatrixI 是对 PkMatrix 按行相加的结果
    SumPkMatrixI = sum(PkMatrix, 2);
    for i = 1 : N
        PkMatrix(i,:) = PkMatrix(i,:) / SumPkMatrixI(i,1);
    end

    % ------------------------------------------------------
    % EM 算法，步骤 2，求最大化的参数 Miu,Sigma 和 Alpha
    % ------------------------------------------------------
    OldMiu = Miu;
    OldSigma = Sigma;
    OldAlpha = Alpha;
    for k = 1 : M
        P1 = 0;
        P2 = 0;
        for i = 1 : N
            P1 = P1 + PkMatrix(i,k);
            P2 = P2 + PkMatrix(i,k) * X(1,i);
        end
        Miu(1,k) = P2/P1;
        Alpha(1,k) = P1 / N;

        % 注意必须先计算好 Miu 才能计算 P3，
        %因为要计算 Sigma 用到的 P3 依赖于的是新的 Miu
        P3 = 0;
        for i = 1 : N
            P3 = P3 + PkMatrix(i,k) * (X(1,i)-Miu(1,k))^2;
```

```
        end
        Sigma(1,k) = P3/P1;
    end

    % ---------------------------------------------------
    % 判断假设是否发生了足够的变化
    % ---------------------------------------------------
    Epsilon = 0.0001;
    if sum(abs(Miu - OldMiu)) < Epsilon &&
    sum(abs(Sigma - OldSigma)) < Epsilon &&
    sum(abs(Alpha - OldAlpha)) < Epsilon
        break;
    end
    disp('=============================');
    step
    Miu
    Sigma
    Alpha
end
```

输出结果：

在第 33 步时迭代终止，得参数估计值为（注意此处的Sigma是指σ^2）：

Miu = [2.8246   -2.2027]

Sigma = [1.2007   3.2662]

Alpha = [0.4421   0.5579]

讨论，如果初始值设置的 Alpha 不满足和等于 1 的条件，也不影响最后的结果。

参考文献

【1】盛骤等，概率论与数理统计[M]，浙江大学出版社，高等教育出版社

【2】Tom M. Mitchell著，曾华军等译，机器学习，机械工业出版社

【3】A. P. DEMPSTER, N. M. LAIRD and D. B. RDIN, Maximum Likelihood from Incomplete Data via the EM Algorithm

【4】J. A. Bilmes (1998), A General Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.

【5】Jiangsheng Yu, Expectation Maximization-An Approach to Parameter Estimation, http://icl.pku.edu.cn/yujs/papers/pdf/EM.pdf

【6】李昌利，沈玉利，期望最大算法及其应用(2)，计算机工程与应用，2008年第44卷30期