基于EM算法的GMM参数估计

汇报人：霍文君 1732976

议程

- 高斯混合分布(GMM)
- 期望最大化算法(EM)
- EM算法在GMM参数估计中的实现
- 程序展示

高斯混合分布(Gaussian Mixture Model)

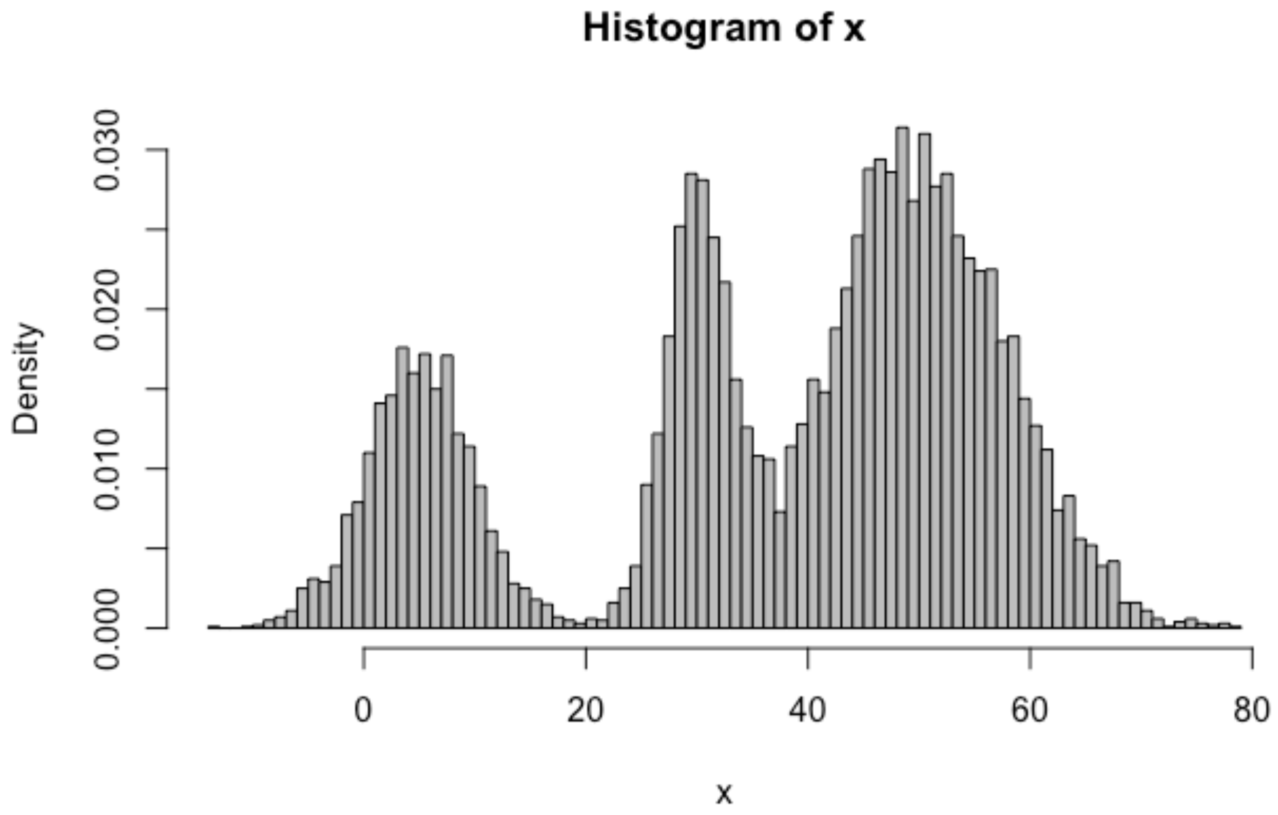
- 高斯混合模型是多个高斯模型的加权和，概率密度函数如下：

$$P(y) = \sum_{k=1}^{K} a_k \phi(y \mid \theta_k)$$

- 其中每个高斯模型的概率函数如下所示：

$$\phi(y \mid \theta_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu_k)^2}{2\sigma_k^2}}$$

高斯混合分布(Gaussian Mixture Model)



期望最大化算法(Expectation Maximization)

- 算法主要思想：利用样本数据递归估计模型参数，主要方法是求解参数使得似然函数最大化
- 算法步骤：
  - 重复以下过程直到收敛
    - E过程：根据参数初始值或者上一次迭代的模型参数计算后验概率
    - M过程：将似然函数最大化以求得新的参数
    - 具体证明过程请参考相关文献（比如 Andrew Ng《The EM algorithm》）

EM算法在GMM模型中的参数估计实现

- 对于N个训练样本，其中每个样本都服从混合高斯分布，概率密度函数为：

$$\sum_{j=1}^{K} \phi_j N(\mu_j, \sigma_j)$$

- 则对数似然函数为：

$$\begin{aligned}
L(\phi, \mu, \sigma) &= \log \prod_{i=1}^{N} p(x_i; \phi, \mu, \sigma) \\
&= \sum_{i=1}^{N} \log p(x_i; \phi, \mu, \sigma) \\
&= \sum_{i=1}^{N} \log \sum_{z_i=1}^{K} p(x_i, z_i; \phi, \mu, \sigma) \\
&= \sum_{i=1}^{N} \log \sum_{z_i=1}^{K} p(x_i | z_i; \mu, \sigma) p(z_i; \phi)
\end{aligned}$$

EM算法在GMM模型中的参数估计实现

- E过程：根据初始参数或者上一步迭代的模型参数求后验概率

$$
\begin{aligned}
w_i(j) &= Q(z_i = j; \theta) = p(z_i = j | x_i; \theta) \\
&= \frac{p(x_i, z_i = j; \theta)}{p(x_i; \theta)} \\
&= \frac{p(x_i | z_i = j; \mu, \sigma) p(z_i = j; \phi)}{\sum_{l=1}^{K} p(x_i | z_i = l; \mu, \sigma) p(z_i = l; \phi)} \\
&= \frac{\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right) \cdot \phi_j}{\sum\limits_{k=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \cdot \phi_k}
\end{aligned}
$$

EM算法在GMM模型中的参数估计实现

- M过程：根据E过程求得的后验概率，最大化对数似然函数，即求得参数 $\theta = (\phi_j, \mu_j, \sigma_j)$ 使得下面的函数最大：

$$
\begin{aligned}
J(\theta) &= \sum_{i=1}^{N} \sum_{z_i} Q(z_i) \log \frac{p(x_i, z_i; \theta)}{Q(z_i)} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} Q(z_i = j) \log \frac{p(x_i, z_i; \theta)}{Q(z_i = j)} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} w_i(j) \log \frac{\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right) \cdot \phi_j}{w_i(j)}
\end{aligned}
$$

EM算法在GMM模型中的参数估计实现

- 对数似然函数对每个参数求偏导并令其为零，求得参数为：

$$\mu_j = \frac{\sum_{i=1}^{N} w_i(j) x_i}{\sum_{i=1}^{N} w_i(j)}$$

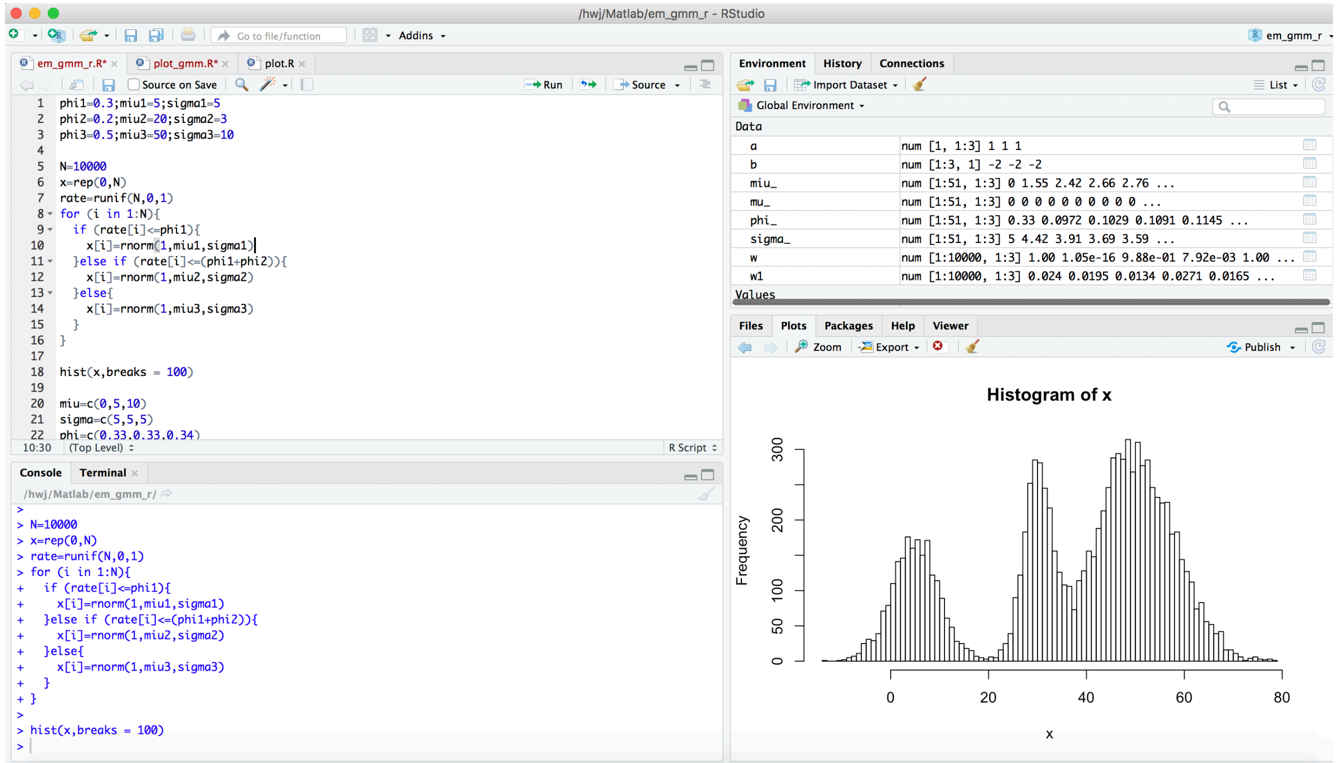
$$\sigma_j^2 = \frac{\sum_{i=1}^{N} w_i(j)(x_i - \mu_j)^2}{\sum_{i=1}^{N} w_i(j)}$$

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} w_i(j)$$

- 重复上述过程直至收敛

程序展示

- 实验平台：R语言
- 算法步骤：随机生成服从GMM的10000个样本点，并在样本数据集上调用EM算法求解模型参数



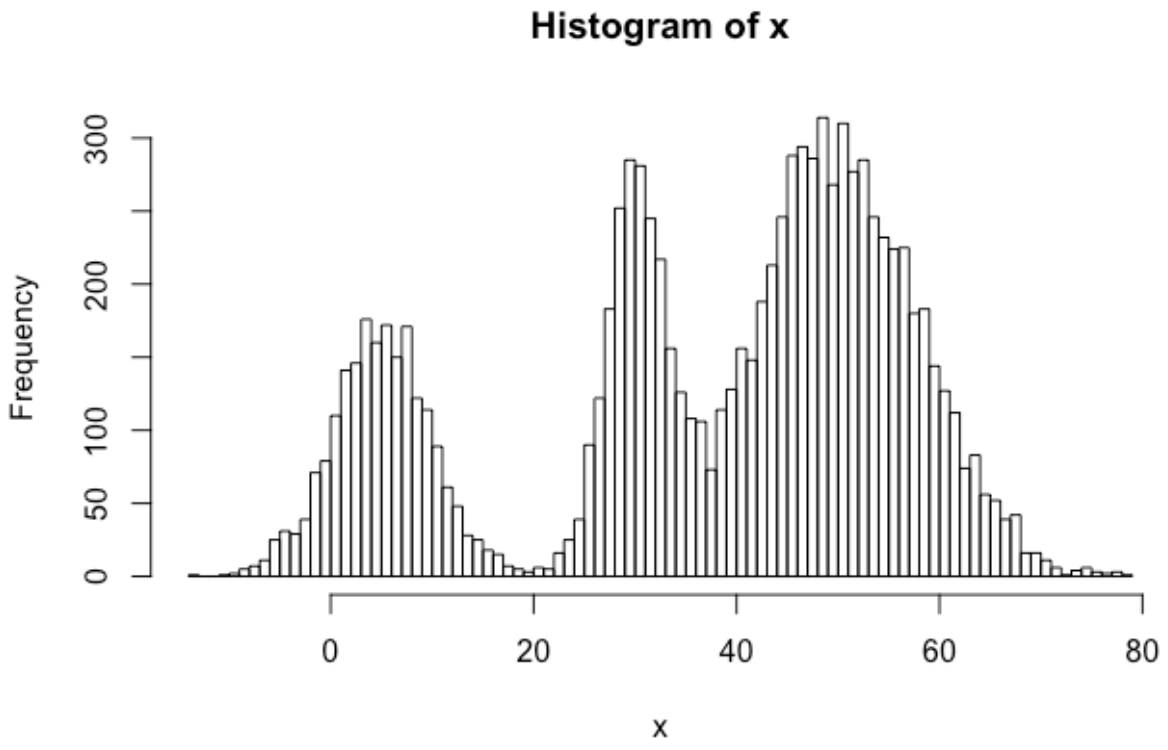
程序展示

- 随机生成数据

```
phi1=0.3;miu1=5;sigma1=5
phi2=0.2;miu2=20;sigma2=3
phi3=0.5;miu3=50;sigma3=10

N=10000
x=rep(0,N)
rate=runif(N,0,1)
for (i in 1:N){
  if (rate[i]<=phi1){
    x[i]=rnorm(1,miu1,sigma1)
  }else if (rate[i]<=(phi1+phi2)){
    x[i]=rnorm(1,miu2,sigma2)
  }else{
    x[i]=rnorm(1,miu3,sigma3)
  }
}

hist(x,breaks = 100)
```



程序展示

- EM算法

```
miu=c(0,5,10)
sigma=c(5,5,5)
phi=c(0.33,0.33,0.34)
w=matrix(0,N,3)

T=50
miu_=matrix(0,T+1,3)
sigma_=matrix(0,T+1,3)
phi_=matrix(0,T+1,3)
miu_[1,]=miu
sigma_[1,]=sigma
phi_[1,]=phi
```

```
for (t in 1:T){
  for (k in 1:3){
    w[,k]=phi[k]*dnorm(x,miu[k],sigma[k])
  }
  w1=matrix(1,N,3)
  for(i in 1:N){
    w1[i,1]=sum(w[i,])
    w1[i,2]=sum(w[i,])
    w1[i,3]=sum(w[i,])
  }
  w=w/w1

  for(k in 1:3){
    miu[k]=w[,k]%*%x/sum(w[,k])
    sigma[k]=(w[,k]%*%((x-miu[k])*(x-miu[k]))/sum(w[,k]))^(1/2)
    phi[k]=sum(w[,k])/N
  }
  miu_[t+1,]=miu
  sigma_[t+1,]=sigma
  phi_[t+1,]=phi
}
```

程序展示

- 绘制收敛图线

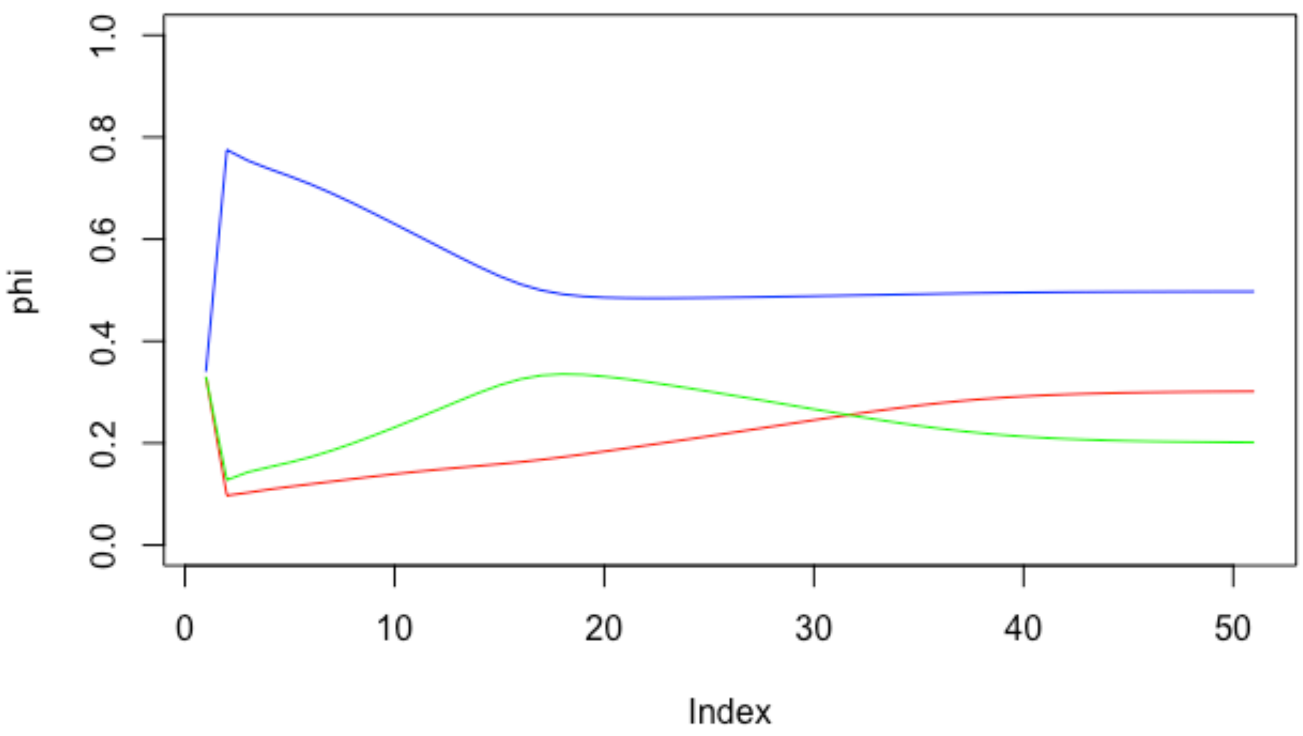
```
plot(phi_[,1],ylab='phi',ylim = c(0,1),col='red',type='l')
points(phi_[,2],col='green',type = 'l')
points(phi_[,3],col='blue',type='l')

plot(miu_[,1],ylab='miu',ylim = c(0,60),col='red',type='l')
points(miu_[,2],col='green',type = 'l')
points(miu_[,3],col='blue',type = 'l')

plot(sigma_[,1],ylab='sigma',ylim=c(0,20),col='red',type = 'l')
points(sigma_[,2],col='green',type = 'l')
points(sigma_[,3],col='blue',type = 'l')
```

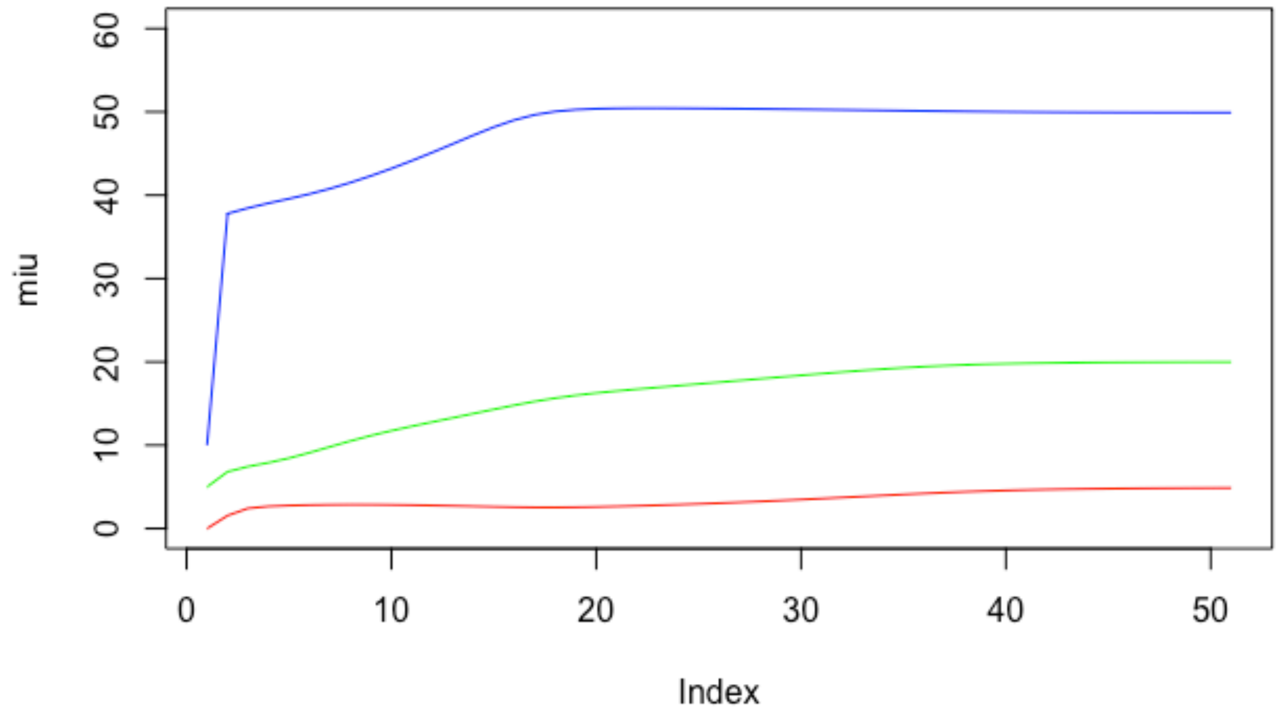
程序展示

- phi的初始参数分别为：0.3，0.2，0.5



程序展示

- Miu的初始参数为：5，20，50



程序展示

- Sigma初始参数为：5，3，10