再尤度推定量と混合ガウス分布

１．最尤推定量

二項分布$B(n,p)$の確率密度関数は，

$$f(x|n,p) = {}_nC_x p^x (1-p)^{n-x}$$

であった．ここで，定数と変数を入れ替えて，

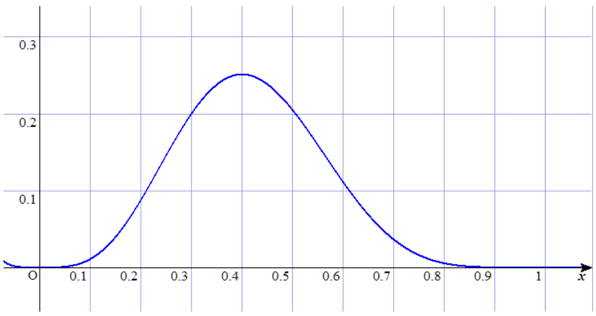
$$f(p|n,x) = {}_nC_x p^x (1-p)^{n-x} \qquad (1)$$

とする．(1)はデータn,x（xは yes が得られた回数）から，母比率pを得るための関数（**尤度関数**）を表している．最尤推定法では，母比率の最尤推定量を

$$\hat{p} = \frac{x}{n}$$

とする．この数学的意味を見ていく．

母比率の最尤推定量は，$f(p|n,x)$の最大値をとるpとする．(1)において，$n = 10, x = 4$の場合の尤度関数$f(p|n,x) = {}_{10}C_4 p^4 (1-p)^6$のグラフを描くと，以下となる．



最大値は，$f(p|n,x)$をpで微分して0となるpから得られる．$f(p|n,x)$の極大値を求めるために，あえて

$$y = \log f(p|n,x) = \log\left({}_{10}C_x p^x (1-p)^{n-x}\right) = x\log p + (n-x)\log(1-p) + A \qquad (2)$$

を考える．ここでAは定数である．これを**対数尤度関数**という．

対数尤度関数は尤度関数の単調性は保持するので，最大値をとるpは尤度関数のときと同じである．式(2)をpで微分すると，

$$y' = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p)}$$

である．したがって，$y' = 0$のとき，

$$\hat{p} = \frac{x}{n}$$

を得る．

このように，最尤推定量は，対数尤度関数を考えて，母数を推定することができる．

問題１．Yes と No で答えるアンケート 100 個において，Yse の回答が 33 個であった．このデータから得られるpに関する尤度関数$f(p|n,x)$について，以下の問いに答えよ．

(1) $f(p|n,x)$を求めよ．

(2) $f(p|n,x)$の対数尤度関数を求めよ．

(3) 母比率の最尤推定量$\hat{p}$を求めよ．

２．ガウス分布の最尤推定量

ガウス分布（正規分布）$N(\mu,\sigma^2)$の最尤推定量$\hat{\mu}$と$\hat{\sigma}$を見ていこう．$N(\mu,\sigma^2)$の確率密度関数は

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

であり，$x_1, x_2, \cdots . x_n$というn個の独立なデータが与えられたときの尤度関数は，

$$f(\mu,\sigma|x) = \prod_{i=1}^{n} f(\mu,\sigma|x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{3}$$

である．よって，対数尤度関数は

$$y = \log f(\mu,\sigma|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{4}$$

となる．これより，母平均$\hat{\mu}$と，母標準偏差$\hat{\sigma}$は

$$\frac{d}{d\mu}\log f(\mu,\sigma|x) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \tag{5}$$

$$\frac{d}{d\sigma}\log f(\mu,\sigma|x) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \tag{6}$$

を考えればよい．したがって，(5)と(6)の連立方程式を解いて，

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

を得る．

問題２．次のデータはガウス分布$N(\mu,\sigma^2)$に従っていると考える．得点の最尤推定量$\hat{\mu}$と$\hat{\sigma}^2$を求めよ．

得点	5	15	25	35	45	55	65	75	85	95
度数	0	0	1	2	7	13	9	7	1	0

３．混合ガウス分布と EM アルゴリズム

2 個の**混合ガウス分布**$\omega_1 N(\mu_1,\sigma_1^2)+\omega_2 N(\mu_2,\sigma_2^2)$の確率密度関数は，

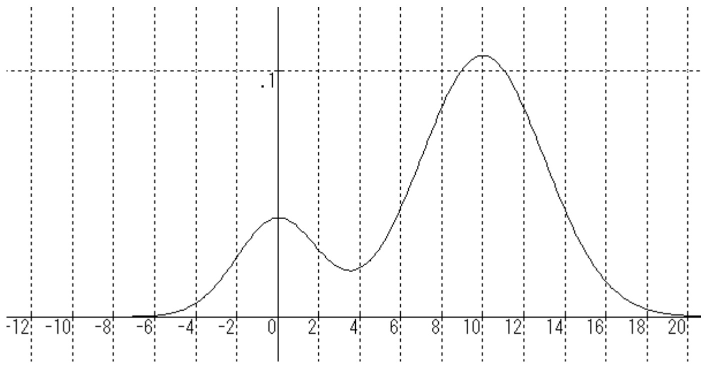
$$f(x|\omega_1,\omega_2,\mu_1,\mu_2,\sigma_1,\sigma_2)=\frac{\omega_1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}+\frac{\omega_2}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}},\quad \omega_1+\omega_2=1$$

で与えられる．ここで，ω_jをガウス分布$N\left(\mu_j,\sigma_j^2\right)$の**重み**という．そして，$x_1,x_2,\cdots.x_n$という$n$個の独立なデータが与えられたときの尤度関数は，

$$f(\omega_1,\omega_2,\mu_1,\mu_2,\sigma_1,\sigma_2|x_i)=\prod_{i=1}^{n}\left(\frac{\omega_1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}+\frac{\omega_2}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}\right) \tag{7}$$

となる．

以下は，混合ガウス分布$0.2N(0,2^2)+0.8N(10,3^2)$のグラフである．



さて，式(7)を最大にする$\omega_1,\omega_2,\mu_1,\mu_2,\sigma_1,\sigma_2$を求める必要がある．しかしこのままでは積の微分になるため最尤指定量を求めることが困難なので，対数尤度関数を考える． 以下，$\mu_1,\mu_2,\sigma_1,\sigma_2$をまとめて$\boldsymbol{\theta}$と書き，さらに，

$$\phi_j(x|\boldsymbol{\theta})\coloneqq\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{\left(x-\mu_j\right)^2}{2\sigma_j^2}}$$

と書くことにする．すると，確率密度関数は

$$f(x|\boldsymbol{\theta})=\omega_1\phi_1(x|\boldsymbol{\theta})+\omega_2\phi_2(x|\boldsymbol{\theta}) \tag{8}$$

となる．したがって，$x_1,x_2,\cdots.x_n$というn個の独立なデータが与えられたときの対数尤度関数は，

$$y=\log\prod_{i=1}^{n}f(\boldsymbol{\theta}|x_i)=\sum_{i=1}^{n}\log f(\boldsymbol{\theta}|x_i) \tag{9}$$

で，この最大値をとる$\boldsymbol{\theta}$を求めれば，それが最尤推定量となる．しかし，$f(\boldsymbol{\theta}|x_i)$は和の式であるため偏

微分して0となる$\boldsymbol{\theta}$がyの最大値をとるかどうかは怪しい．

そこで，以下において最尤推定量を求めるための**EMアルゴリズム**という方法について説明する．

EMアルゴリズムのポイントは，**負担率**γ_{ij}というものを定義して，$f(x|\boldsymbol{\theta})$の近似式をベイズ更新していくことにある．

（**Eステップ**）まず，ω_1, ω_2と$\boldsymbol{\theta}$，すなわち$\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2$を仮に設定する．そして，データ$x_i$について，$\phi_1(x_i|\boldsymbol{\theta})$と$\phi_2(x_i|\boldsymbol{\theta})$を尤度，$\omega_1, \omega_2$を事前分布と考えて（正しくはもう少し議論が必要であるが，とにかく）ベイズの定理から，事後分布

$$\gamma_{i1} = \phi_1(\boldsymbol{\theta}|x_i) = \frac{\omega_1 \phi_1(x_i|\boldsymbol{\theta})}{f(x_i|\boldsymbol{\theta})}, \qquad \gamma_{i2} = \phi_2(\boldsymbol{\theta}|x_i) = \frac{\omega_2 \phi_2(x_i|\boldsymbol{\theta})}{f(x_i|\boldsymbol{\theta})}$$

を考える．これを負担率γ_{ij}とよぶ．簡単に言えば，γ_{i1}は$f(\boldsymbol{\theta}|x_i)$における前半のガウス分布の確率$\phi_1(x_i|\boldsymbol{\theta})$の割合であり，$\gamma_{i2}$は$f(\boldsymbol{\theta}|x_i)$における後半のガウス分布の確率$\phi_2(x_i|\boldsymbol{\theta})$の割合である．しかし事後分布$\gamma_{ij}$も$\omega_1, \omega_2$と$\boldsymbol{\theta}$に依存しているので，これはまだ仮の負担率である．

（**Mステップ**）式(8)をそれぞれの変数$\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2$で偏微分して$0$と置くことで，

$$\omega_1 = \frac{1}{n}\sum_{i=1}^{n} \gamma_{i1}, \qquad \omega_2 = \frac{1}{n}\sum_{i=1}^{n} \gamma_{i2},$$

$$\mu_1 = \frac{\sum_{i=1}^{n} \gamma_{i1} x_i}{\sum_{i=1}^{n} \gamma_{i1}}, \qquad \mu_2 = \frac{\sum_{i=1}^{n} \gamma_{i2} x_i}{\sum_{i=1}^{n} \gamma_{i2}}$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{n} \gamma_{i1} (x_i - \mu_1)^2}{\sum_{i=1}^{n} \gamma_{i1}}, \qquad \sigma_2^2 = \frac{\sum_{i=1}^{n} \gamma_{i2} (x_i - \mu_2)^2}{\sum_{i=1}^{n} \gamma_{i2}}$$

が得られる．

そして，またEステップに戻り，Mステップを行うという操作を繰り返すことで，$\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2$は次第に収束していき，それを最尤推定値とする．

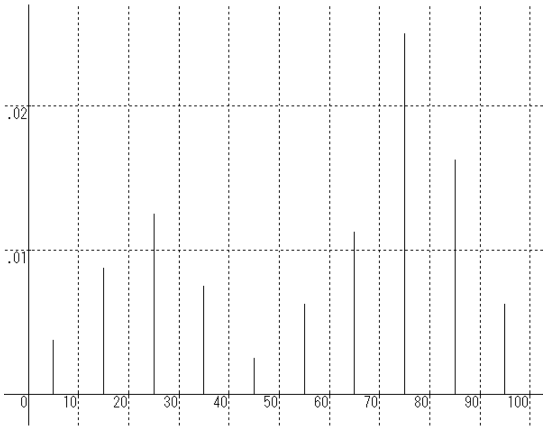
たとえば，次のデータは混合ガウス分布$\omega_1 N(\mu_1, \sigma_1^2) + \omega_2 N(\mu_2, \sigma_2^2)$に従っていると仮定しよう．

得点	5	15	25	35	45	55	65	75	85	95	計
度数	3	7	10	6	2	5	9	20	13	5	80

このデータの度数を連続分布の確率pに換える．それは$p=$度数$/800$と計算して以下のようになる．

得点	5	15	25	35	45	55	65	75	85	95
確率	0.004	0.009	0.013	0.008	0.003	0.006	0.011	0.025	0.016	0.006

以下のグラフが，上のデータ（得点，確率）のヒストグラムである．このデータをもとに，EM アルゴリズムを実行し，$\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2$を決める方法をもう一度説明する．



（Ｅステップ）グラフから，$\omega_1 = 0.4$，$\omega_2 = 0.6$，$\mu_1 = 25$，$\mu_2 = 75$，$\sigma_1^2 = 10^2$，$\sigma_2^2 = 10^2$を仮に設定する．これをもとに，

$$\gamma_{i1} = \phi_1(\boldsymbol{\theta}|x_i) = \frac{\omega_1 \phi_1(x_i|\boldsymbol{\theta})}{f(x_i|\boldsymbol{\theta})}, \qquad \gamma_{i2} = \phi_2(\boldsymbol{\theta}|x_i) = \frac{\omega_2 \phi_2(x_i|\boldsymbol{\theta})}{f(x_i|\boldsymbol{\theta})}$$

を求めておく．ただし，$i = 1,2,\cdots,80$である．

（Ｍステップ）それぞれの変数$\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2$を以下の式で，コンピュータを使って計算すると

$$\omega_1 = \frac{1}{n}\sum_{i=1}^{n} \gamma_{i1} = 0.34, \qquad \omega_2 = \frac{1}{n}\sum_{i=1}^{n} \gamma_{i2} = 0.66,$$

$$\mu_1 = \frac{\sum_{i=1}^{n} \gamma_{i1} x_i}{\sum_{i=1}^{n} \gamma_{i1}} = 23.3, \qquad \mu_2 = \frac{\sum_{i=1}^{n} \gamma_{i2} x_i}{\sum_{i=1}^{n} \gamma_{i2}} = 75.3$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{n} \gamma_{i1}(x_i - \mu_1)^2}{\sum_{i=1}^{n} \gamma_{i1}} = 107.5, \qquad \sigma_2^2 = \frac{\sum_{i=1}^{n} \gamma_{i2}(x_i - \mu_2)^2}{\sum_{i=1}^{n} \gamma_{i2}} = 132.3$$

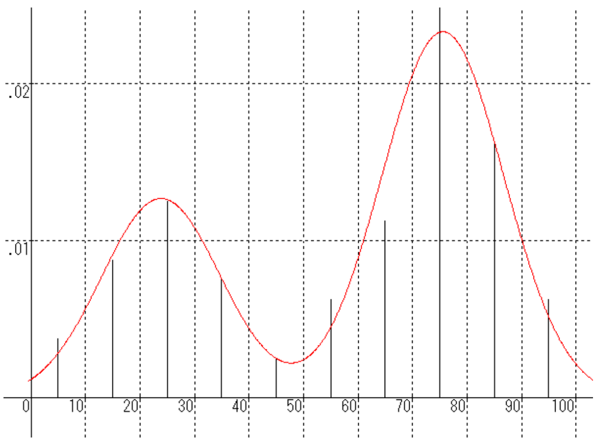
が得られる．このデータをもとに，再度Ｅステップとステップを実行すると，

$$\omega_1 = 0.34, \quad \omega_2 = 0.66, \quad \mu_1 = 23.5, \quad \mu_2 = 75.4, \quad \sigma_1^2 = 112.7, \quad \sigma_2^2 = 129.5$$

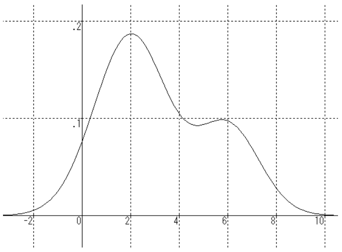
が得られる．ＥステップとＭステップを５０回繰り返して得られた結果が

$$\omega_1 = 0.35, \quad \omega_2 = 0.65, \quad \mu_1 = 23.9, \quad \mu_2 = 75.6, \quad \sigma_1^2 = 119.3, \quad \sigma_2^2 = 124.9$$

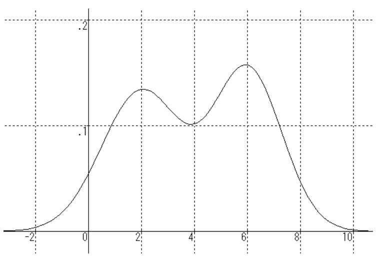
であり，これらから得られた混合ガウス分布のグラフによる近似が以下となる．



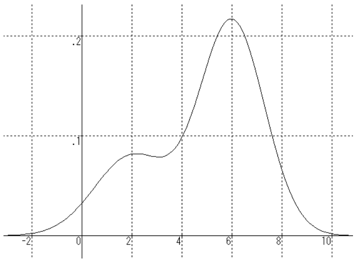
問題３．　２個の混合ガウス分布$0.7N(2, 1.5^2) + 0.3N(6, 1.3^2)$を考える．このグラフは，以下の(A),（B）,（C）の中のどれに対応するか適切なものを選べ.



（A）



（B）



（C）

解答

問題１．（1）$f(p|100{,}33) = {}_{100}C_{33}p^{33}(1-p)^{67}$

(2) $\log f(p|100{,}33) = 33\log p + 67\log(1-p) + A$　（3）$\hat{p} = 0.33$

問題２．$\hat{\mu} = 58$　$\hat{\sigma}^2 = 170.26$　問題３．(A)