EM Algorithm

EXPECTATION & MAXIMIZATION

응용수학과 최우빈

EM 알고리즘

관측되지 않는 잠재변수에 의존하는 확률 모델에서 maximum likelihood 나 maximum a posteriori(MAP)을 갖는 매개변수를 찾는 반복적인 알고리즘

1) 초기 추측 Θ(0) = (μ0 , μ0 , var0 , var0, p)

2) E – step : 매개변수에 관한 추정 값으로 log likelihood의 기대 값 계산

3)M – step : 기대 값을 최대화하는 변수 값 계산

(X,Z)의 확률 분포) : $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

Maximum Likelihood 함수 : $L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

X : 관측 가능 확률변수

Z : 관측 불가능 확률변수

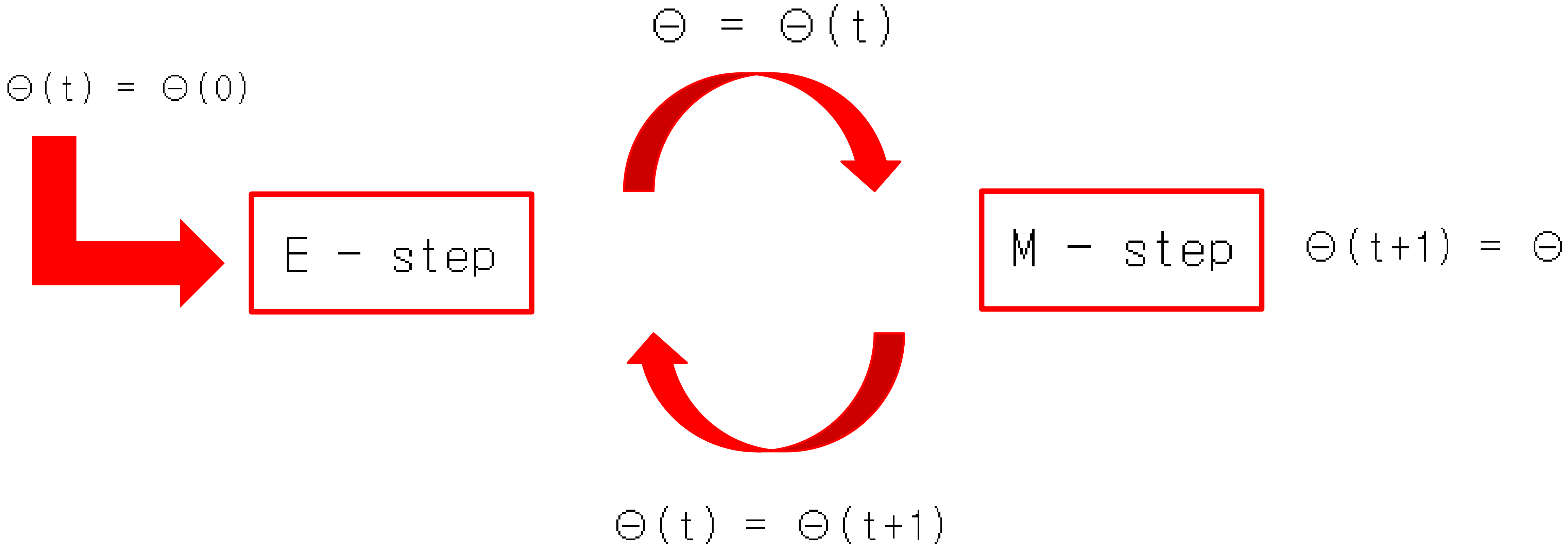
E – step : Θ(t) : 현재 Θ , Θ: 새로운 Θ , Q : likelihood의 기대 값

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$$

M – step : Q 최대화 → Θ(t+1) : 새로운 매개변수

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\arg\max}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

n번 실행 : log likelihood 수렴



Example 7.6.15

Table 7.1 Heights and weights for Example 7.6.15. The missing values are given random variable names.

Height	Weight
72	197
70	204
73	208
68	$X_{4,2}$
65	$X_{5,2}$
$X_{6,1}$	170

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)}, \rho^{(0)}) = (69.60, 194.75, 2.87, 14.82, 0.1764).$$

32 iterations,

$$\theta^{(32)} = (68.86, 189.71, 3.15, 15.03, 0.8965)$$

with log-likelihood -29.66.

Gauss_dist.m :

```
function [ y ] = gauss_dist(x,mu,var)
% GAUSS_DIST function for gaussian distribution
    y=(1/(sqrt(2*pi*var)))*exp((-(x-mu).^2)/(2*var));
end
```

```
% data=[72 197 70 204 73 208 68 null 65 null null 170];
data=[-0.39 0.12 0.94 1.67 1.76 2.44 3.72 4.28 4.92 5.53,...
        0.06 0.48 1.01 1.68 1.80 3.25 4.12 4.60 5.28 6.22];
```

```
% 초기 추측
temp=randperm(length(data));
pie(1)=0.5;
mu1(1)=data(temp(1)); % height 초기 평균
mu2(1)=data(temp(2)); % weight 초기 평균
var1(1)=var(data); % height 초기 분산
var2(1)=var(data); % weight 초기 분산
```

```
for i = 1:50 % 최대 50회 연산

% E - step : 매개변수의 추정 값으로 log likelihood 의 기대 값 계산
Qq1=gauss_dist(data,mu1(i),var1(i));
Qq2=gauss_dist(data,mu2(i),var2(i));
log_likelihood(i)=sum(log(((1-pie(i))*Qq1) + (pie(i)*Qq2)));

% responsibility : posterior distribution
responsibilities(i,:)=(pie(i)*Qq2)./(((1-pie(i))*Qq1)+(pie(i)*Qq2));

% M - step : 기대 값을 최대화 하는 변수 값 계산. 이 때, 변수 값은 다음 E - step 의 추정 값으로 쓰임
mu1(i+1)=sum((1-responsibilities(i,:)).*data)/sum(1-responsibilities(i,:));
mu2(i+1)=sum((responsibilities(i,:)).*data)/sum(responsibilities(i,:));

var1(i+1)=sum((1-responsibilities(i,:)).*((data-mu1(i)).^2))/sum(1-responsibilities(i,:));
var2(i+1)=sum((responsibilities(i,:)).*((data-mu2(i)).^2))/sum(responsibilities(i,:));

pie(i+1)=sum(responsibilities(i,:))/length(data);

end
```

열 1 ~ 8

-44.3359 -42.2845 -42.1152 -42.1016 -42.0847 -42.0605 -42.0240 -41.9665

열 9 ~ 16

-41.8719 -41.7103 -41.4328 -40.9907 -40.4395 -40.0043 -39.7723 -39.6262

열 17 ~ 24

-39.4936 -39.3597 -39.2297 -39.1152 -39.0274 -38.9706 -38.9396 -38.9248

열 25 ~ 32

-38.9183 -38.9155 -38.9143 -38.9138 -38.9135 $\boxed{-38.9134}$ -38.9134 -38.9134

열 33 ~ 40

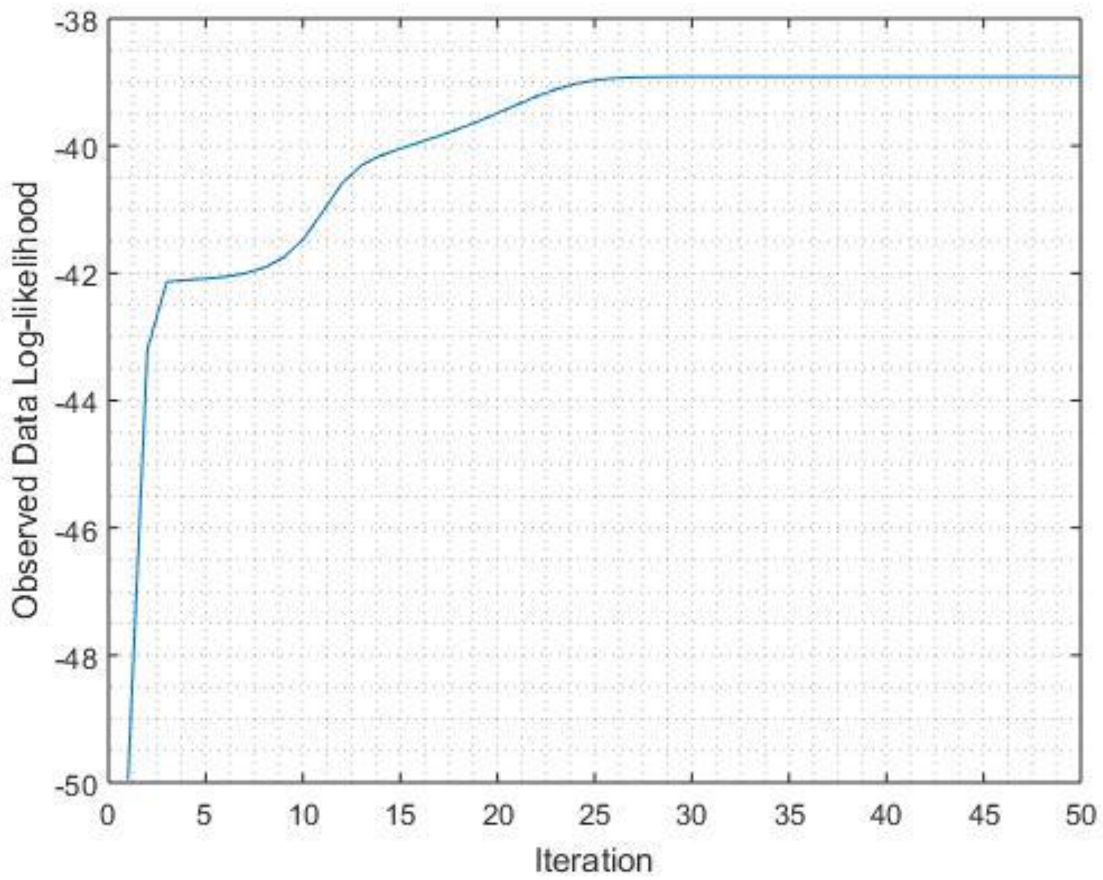
-38.9134 -38.9134 -38.9134 -38.9134 -38.9134 -38.9134 -38.9134 -38.9134

열 41 ~ 48

-38.9134 -38.9134 -38.9134 -38.9134 -38.9134 -38.9134 -38.9134 -38.9134

열 49 ~ 50

-38.9134 -38.9134



참고 문헌

(1) https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

(2) https://kr.mathworks.com/matlabcentral/fileexchange/45817-expectation-maximization-algorithm-with-gaussian-mixture-model