Chapter 10 가우시안 혼합 모델

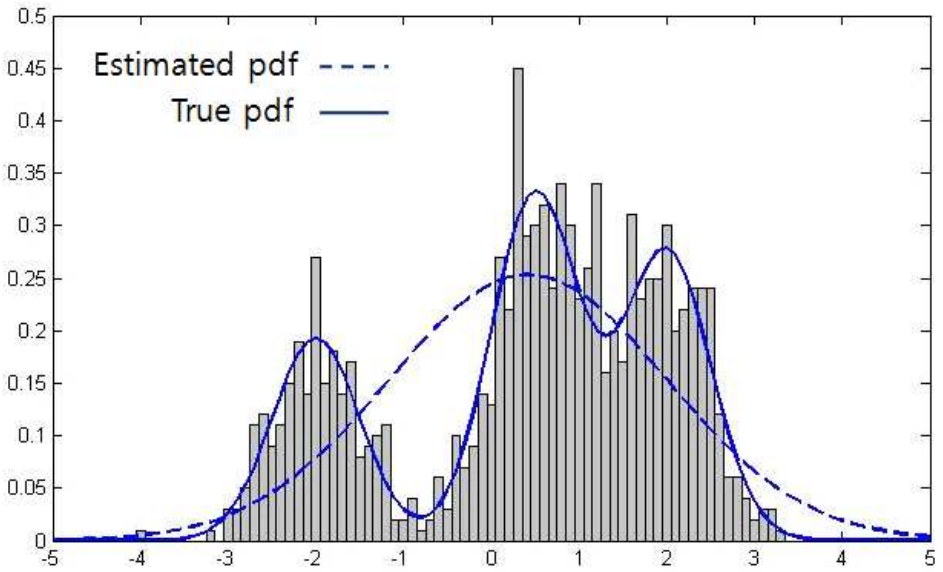
[학습목표]
이 장에서는 3장에서 살펴본 확률밀도함수의 추정 방법의 심화된 형태에 대해서 알아본다. 확률밀도함수를 추정하기 위해 모수적 방법을 쓰는 경우 가우시안과 같은 간단한 함수를 사용하는 대신 여러 개의 가우시안을 합하여 만들어지는 가우시안 혼합 모델을 사용하는 방법에 대해 알아본다. 또한 이 모델의 파라미터를 추정하기 위한 학습법인 EM 알고리즘에 대해서도 알아본다.

10. 가우시안 혼합 모델

10.1. 가우시안 혼합 모델

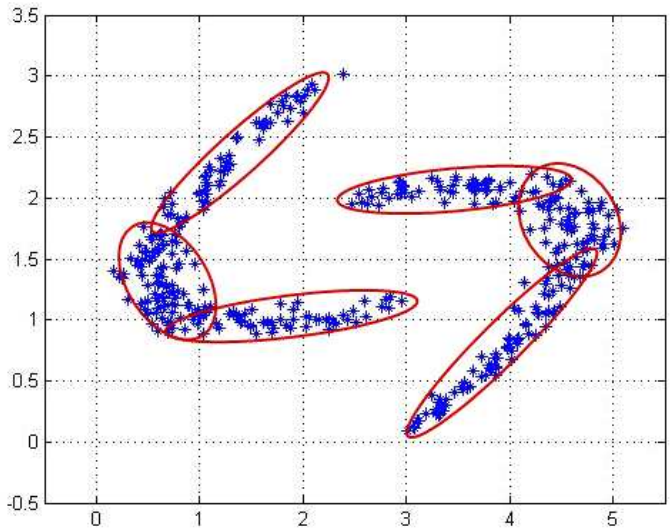
10.1.1 가우시안 혼합 모델의 필요성

지금까지 앞 장에서 살펴본 바와 같이 패턴을 분류함에 있어서 데이터의 분포 특성을 분석하는 것은 매우 중요하다. 데이터의 분포 특성을 알기 위해서 적절한 확률밀도함수를 가정하여 데이터 분포에 대한 모델을 만드는 것을 확률 모델이라 한다. 가장 대표적으로 사용되는 확률 모델로 가우시안 확률 모델이 있다. 이는 하나의 클래스 혹은 관찰된 전체 데이터 집합이 평균을 중심으로 하여 뭉쳐져 있는 분포 형태를 표현하는데 적합한 확률 모델이다. 앞서 3장에서 가우시안 확률밀도함수를 확률 모델로 하여 주어진 데이터로부터 그 파라미터(평균과 공분산)를 추정하는 방법에 대하여 알아보았다. 또한 파라미터인 공분산 행렬의 형태에 따라 데이터의 분포 형태가 어떻게 달라지는지에 대해서도 알아보았다.
주어진 데이터에 대하여 가우시안 확률분포를 이용하여 모델을 설정하는 것은 가장 널리 사용되는 방법이다. 그러나 가우시안 확률분포는 기본적으로 데이터들이 평균을 중심으로 하나의 그룹으로 뭉쳐있는 유니모달(unimodal) 형태를 가진다는 것을 가정하고 있어서, 복잡한 분포 형태를 가지는 데이터의 확률밀도함수를 표현하기는 힘들다는 문제점이 있다. 따라서 복잡한 데이터 분포를 추정하기 위해서는 보다 일반적인 형태를 표현할 수 있는 확률 모델이 필요하며, 이때 가장 손쉽게 생각해 볼 수 있는 것이 여러 개의 가우시안을 합하여 만들어지는 모델이다. 이를 <가우시안 혼합 모델 (Gaussian Mixture Model)>이라고 한다.



[그림 10-1] 단일 가우시안 분포로 추정 불가능한 분포의 예

간단한 예로 [그림10-1]의 히스토그램으로 나타난 데이터 분포를 표현하기 위한 확률 모델을 생각해 보자. 이 데이터를 만약 평균과 공분산만을 이용하여 가우시안 확률밀도함수로 추정한다면 그림에서 점선과 같은 형태가 되어 데이터의 분포와는 상당히 거리가 먼 결과를 얻게 될 것이다. 실제로 데이터를 생성하는데 사용한 확률밀도함수는 그림에서 실선으로 나타난 곡선 형태를 가지는데, 이것은 3개의 가우시안 분포의 가중합으로 근사할 수 있다. 이와 같이 복수개의 가우시안 밀도함수들의 가중합된 형태의 밀도함수를 정의하여 사용하면 매우 다양한 형태의 밀도함수들을 충분히 비슷하게 근사해 낼 수 있다.
또 한 가지 예로 [그림10-2]에 주어진 2차원 데이터를 생각해 보자. 이 데이터의 경우 이차곡선을 따라서 그 주변에 어느 정도의 노이즈를 가지고 분포되도록 데이터를 생성하였다. 따라서 데이터 생성에 사용된 밀도함수는 가우시안 분포와는 무관하지만, 그림에서 타원으로 표시된 형태의 공분산을 가지는 가우시안 분포 여섯 개를 적절히 배합함으로써 데이터의 분포 특성을 잘 설명해 줄 수 있는 확률밀도함수를 추정할 수 있다. 이에 대해서는 5절에서 실제로 추정을 시도해 보고 그 결과를 평가해 보겠다.



[그림 10-2] 여섯 개의 가우시안으로 이루어진 데이터 집합

이와 같이 복수개의 가우시안 분포들의 합으로 새로운 확률분포를 나타내는 가우시안 혼합 모델을 사용하면, 하나의 가우시안 분포함수로 나타낼 수 없었던 분포 특성 뿐 아니라, 아무리 복잡한 형태의 함수라도 충분한 개수의 가우시안 함수를 사용하기만 하면 원하는 만큼 정확하게 근사해 낼 수 있다. 가우시안 혼합 모델을 이용하여 밀도함수가 추정되면, 이를 이용하여 베이즈 분류기를 설계할 수 도 있으며, 또한 K-means 클러스터링과 같은 군집화에도 적용할 수 있을 것이다.

10.1.2 가우시안 혼합 모델의 정의

이제 가우시안 혼합 모델을 수학적으로 정의하고 그 특성을 알아보자. 먼저 좀 더 일반적인 경우로, 간단한 확률밀도함수(혹은 성분(component))의 선형결합으로 정의되는 일반적인 혼

합 모델 (Mixture Model)에 대하여 먼저 정의한다. M개의 성분(간단한 밀도함수)의 혼합으로 정의되는 전체 확률밀도함수는 다음 식과 같이 표현된다.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{M} p(\boldsymbol{x}|C_i, \boldsymbol{\theta_i}) p(C_i) \qquad \text{[식 10-1]}$$

여기서 $p(\boldsymbol{x}|C_i, \boldsymbol{\theta_i})$는 혼합 모델의 기본 성분을 이루는 간단한 확률밀도함수로, 가우시안 혼합 모델의 경우는 가우시안 확률밀도함수가 되지만, 문제의 특성에 맞추어 다양한 함수들이 사용될 수 있다. $\boldsymbol{\theta}_i$는 i번째 성분이 되는 간단한 확률밀도함수를 정의하는 파라미터로, 가우시안 확률밀도함수의 경우는 평균 $\boldsymbol{\mu_i}$과 공분산 행렬 Σ_i이 된다. C_i는 i번째 성분임을 나타내는 확률변수이고, $p(C_i)$는 i번째 성분이 전체 혼합 확률밀도함수에서 차지하는 상대적인 중요도를 의미하는 것이다. $p(C_i)$는 파라미터 α_i로 표시하기도 하며, 다음 [식 10-2]와 같은 성질을 만족해야한다.

$$0 \le \alpha_i \le 1, \quad \sum_{i=1}^{M} \alpha_i = 1 \qquad \text{[식 10-2]}$$

이 혼합 확률 모델을 사용하여 데이터의 분포를 나타내기 위해 추정해야하는 파라미터 전체를 $\boldsymbol{\theta}$로 나타내면, M개의 성분을 가지는 가우시안 혼합 모델의 파라미터는 다음과 같이 나타낼 수 있다.

$$\boldsymbol{\theta} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \dots, \boldsymbol{\mu_M}, \Sigma_1, \Sigma_2, \dots \Sigma_M, \alpha_1, \alpha_2, \dots, \alpha_M) \qquad \text{[식 10-3]}$$

이 때 만약 각각의 i번째 성분의 가우시안 함수의 공분산 행렬이 $\sigma_i^2 \mathrm{I}$인 경우를 생각하면 파라미터는 다음과 같이 정의될 수 있다.

$$\boldsymbol{\theta} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \dots, \boldsymbol{\mu_M}, \sigma_1^2, \sigma_2^2, \dots \sigma_M^2, \alpha_1, \alpha_2, \dots, \alpha_M) \qquad \text{[식 10-4]}$$

이 책에서는 주로 [식 10-4]의 경우를 사용하여 설명할 것이나, 같은 계산이 보다 일반적인 [식 10-3]의 경우에도 그대로 적용된다.

10.2 가우시안 혼합 모델의 학습

10.2.1 최우추정법

가우시안 혼합 모델에서의 학습이란 데이터를 이용하여 파라미터를 추정하는 것을 의미한다. 결국 가우시안 혼합 모델이라고 하는 특정한 확률 모델을 가정하고, 그 파라미터를 추정

함으로써 확률밀도함수를 얻어내는 것이므로, 모수적 확률밀도 추정 방법의 일종으로 볼 수 있겠다. 따라서 3장에서 살펴본 우도함수를 최대로 하는 파라미터를 찾는 최우추정법이 가우시안 혼합 모델에서도 적용될 수 있다. 그러나 가우시안 혼합 모델의 경우 최우추정법에 의해 찾아지는 해는 단일 가우시안을 사용하는 경우와는 조금 다른 양상을 띠게 되어, 단순한 최우추정법으로 한 번에 최적의 추정치를 찾아가는 것은 불가능하다. 이에 대해서는 이 절의 마지막에 다시 살펴보도록 하고, 우선 최우추정법에 의해 추정치를 찾아보도록 하자.
데이터 집합 $X=\{x_1,x_2,...,x_N\}$ 가 주어졌을 때, i번째 데이터 x_i의 확률밀도 $p(x_i)$를 가우시안 혼합 모델로 나타내고자 한다. 문제를 간단히 하기 위해 여기서는 일단 데이터가 1차원인 경우를 생각하면, j번째 성분을 이루는 밀도함수는 평균 μ_j과 분산 σ_j^2을 파라미터로 가지는 단변량 가우시안 밀도함수가 된다. 즉, j번째 가우시안 확률밀도는 다음과 같이 명시적으로 표현된다.

$$p(x_i|C_j,\boldsymbol{\theta_j}) = p(x_i|\mu_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\} \quad \text{[식 10-5]}$$

이러한 개별 성분 M개를 결합한 혼합 확률밀도는 다음과 같이 쓸 수 있다.

$$\begin{aligned} p(x_i|\boldsymbol{\theta}) &= \sum_{j=1}^{M} p(x|C_j,\boldsymbol{\theta}_j)P(C_j|\boldsymbol{\theta}_j) \\ &= \sum_{j=1}^{M} \alpha_j p(x_i|\mu_j,\sigma_j^2) \end{aligned} \quad \text{[식 10-6]}$$

각 데이터 $x_i(i=1,...,N)$ 에 대해 정의된 혼합 확률밀도함수를 이용하여 학습 데이터 전체에 대한 로그우도(log-likelihood)를 정의하면 다음과 같이 쓸 수 있다.

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \sum_{j=1}^{M} \alpha_j p(x_i|\mu_j,\sigma_j^2) \quad \text{[식 10-7]}$$

데이터 집합에 대한 로그우도를 최대로 하는 최우추정량 $\hat{\boldsymbol{\theta}}$은 다음과 같이 정의된다.

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \operatorname{argmax}\left[\sum_{i=1}^{N} \ln p(x_i|\boldsymbol{\theta})\right] \\ &= \operatorname{argmax}\left[\sum_{i=1}^{N} \ln \sum_{j=1}^{M} \alpha_j p(x_i|\mu_j,\sigma_j^2)\right] \end{aligned} \quad \text{[식 10-8]}$$

최우추정량은 위의 로그우도를 파라미터에 대해 미분하여 0이 되는 값을 구함으로써 얻어진다. 평균 μ_j의 최우추정치를 계산하는 과정을 알아보면, 먼저 로그우도를 μ_j에 대해 편미분한 식을 다음과 같이 구한다.

$$\frac{\partial}{\partial \mu_j}\ln L(\boldsymbol{\theta}) = \frac{\partial}{\partial \mu_j}\sum_{i=1}^{N}\ln p(x_i|\boldsymbol{\theta}) \qquad \text{[식 10-9]}$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}\frac{\partial}{\partial \mu_j}p(x_i|\boldsymbol{\theta})$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}\frac{\partial}{\partial \mu_j}\left\{\sum_{k=1}^{M}\alpha_k\, p(x_i|\mu_k,\sigma_k^2)\right\}$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}\frac{\partial}{\partial \mu_j}\left\{\frac{1}{\sqrt{2\pi}\,\sigma}\sum_{k=1}^{M}\alpha_k \exp\left\{-\frac{(x_i-\mu_k)^2}{2\sigma^2}\right\}\right\}$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}\alpha_j\, p(x_i|\mu_j,\sigma_j^2)\frac{(x_i-\mu_j)}{\sigma_j^2}$$

$$= \sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta_j})\frac{(x_i-\mu_j)}{\sigma_j^2}$$

위 식에서 마지막 항의 $p(C_j|x_i,\boldsymbol{\theta_j})$는 데이터 x_i가 주어졌을 때, 그것이 j번째 성분으로부터 나왔을 확률값으로 다음과 같이 나타낼 수 있음을 이용하여 계산되었다.

$$P(C_j|x_i,\boldsymbol{\theta}) = \frac{p(x_n|C_j,\boldsymbol{\theta})p(C_j|\boldsymbol{\theta})}{p(x_i|\boldsymbol{\theta})} \qquad \text{[식 10-10]}$$

$$= \frac{\alpha_j\, p(x_i|\mu_j,\sigma_j^2)}{p(x_i|\boldsymbol{\theta})}$$

[식 10-9]의 마지막 항을 0으로 두고 로그우도의 극대치를 가지는 최우추정량 $\hat{\mu}_j$의 값을 찾으면 다음과 같이 계산할 수 있다.

$$0 = -\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta})x_i + \mu_j\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta}) \qquad \text{[식 10-11]}$$

$$\hat{\mu_j} = \frac{\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta})x_i}{\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta})} \qquad \text{[식 10-12]}$$

마찬가지로 나머지 파라미터 σ_j^2에 대한 추정치도 계산할 수 있는데, 이 과정은 [식 10-9]와 유사하므로 생략하겠다. 얻어지는 추정치는 다음과 같다.

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N} p(C_j|x_i,\boldsymbol{\theta})(x_i-\hat{\mu}_j)^2}{\sum_{i=1}^{N} p(C_j|x_i,\boldsymbol{\theta})} \qquad \text{[식 10-13]}$$

마지막으로 파라미터 α_j에 대한 추정을 생각해보자. α_j의 경우는, 단순히 로그우도를 최대화하는 것이 아니라 [식 10-2]에서 제시한 조건을 함께 만족해야 한다. 따라서 조건을 가진 최대화 문제가 되어 8장에서 잠깐 소개한 라그랑제승수를 이용하여 해를 찾는다. 최적화 대상이 되는 목적함수는 라그랑제승수 λ를 포함한 형태로 다음과 같이 정의할 수 있다.

$$\tilde{L}(\boldsymbol{\theta},\lambda) = \ln L(\boldsymbol{\theta}) + \lambda\left(\sum_{j=1}^{M}\alpha_j - 1\right) \qquad \text{[식 10-14]}$$

이 식을 α_j에 대하여 미분하면 다음과 같은 전개과정을 거치게 된다.

$$\frac{\partial}{\partial\alpha_j}\tilde{L}(\boldsymbol{\theta},\lambda) = \frac{\partial}{\partial\alpha_j}\sum_{i=1}^{N}\ln p(x_i|\boldsymbol{\theta}) + \frac{\partial}{\partial\alpha_j}\left\{\lambda\left(\sum_{k=1}^{M}\alpha_k - 1\right)\right\} \qquad \text{[식 10-15]}$$

$$= \frac{\partial}{\partial\alpha_j}\sum_{i=1}^{N}\ln p(x_i|\boldsymbol{\theta}) + \lambda$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}\frac{\partial}{\partial\alpha_j}p(x_i|\boldsymbol{\theta}) + \lambda$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}\frac{\partial}{\partial\alpha_j}\left\{\sum_{k=1}^{M}\alpha_k\, p(x_i|\mu_k,\sigma_k^2)\right\} + \lambda$$

$$= \sum_{i=1}^{N}\frac{1}{p(x_i|\boldsymbol{\theta})}p(x_i|\mu_k,\sigma_k^2) + \lambda$$

$$= \sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta_j})\frac{1}{\alpha_j} + \lambda$$

이 식의 유도과정에서 마지막 항을 얻기 위해서 [식 10-10]의 관계가 사용되었다. 마지막 항을 0으로 놓고 α_j에 대한 방정식을 풀면 다음과 같은 식을 얻을 수 있다.

$$\alpha_j = \frac{1}{\lambda}\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta}) \qquad \text{[식 10-16]}$$

그런데 $\sum_{j=1}^{M}\alpha_j = 1$ 의 조건식을 만족하여야 하므로, 이 조건식에 [식 10-16]을 대입하면 다음과 같은 관계식을 얻을 수 있다.

$$\sum_{j=1}^{M}\alpha_j = \frac{1}{\lambda}\sum_{j=1}^{M}\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta}) = \frac{1}{\lambda}\sum_{i=1}^{N}\sum_{j=1}^{M}p(C_j|x_i,\boldsymbol{\theta}) = 1 \qquad \text{[식 10-17]}$$

여기서 다시 $\sum_{j=1}^{M}p(C_j|x_i,\boldsymbol{\theta}) = 1$이 되는 당연한 사실을 이용하면, [식 10-17]을 만족하는 라그랑제상수 λ는 N이 되어, 최종적으로 얻어지는 α_j의 추정치는 다음과 같다.

$$\hat{\alpha_j} = \frac{1}{N}\sum_{i=1}^{N}p(C_j|x_i,\boldsymbol{\theta}) \qquad \text{[식 10-18]}$$

이상과 같이 주어진 데이터 $X=\{x_1,x_2,...,x_N\}$ 에 대한 파라미터의 최우추정량은 [식 10-12], [식 10-13], 그리고 [식 10-18]로 계산될 수 있다.

10.2.2 최우추정법의 문제점

이제 얻어진 최우추정량에 대한 식을 활용하여 실제로 파라미터의 추정값을 얻어낸다고 생각해 보자. 각 식의 우변의 값을 계산함으로써 각 파라미터의 추정치가 얻어지게 되는데, 이 우변의 식들을 자세히 들여다보면, 우리가 추정해야 하는 파라미터 $\boldsymbol{\theta} = (\mu_1,...,\mu_M,\sigma_1^2,...,\sigma_M^2,\alpha_1,...,\alpha_M)$가 여전히 들어가 있음을 알 수 있다. 이는 3장의 가우시안 분포함수에 대한 최우추정량 식과는 근본적으로 다른 성질을 가진다. 3장에서 다룬 최우추정량은 데이터 집합만 있으면 간단히 계산될 수 있었으나, 앞 절에서 얻은 최우추정량은 이 자체로서는 계산이 불가능한 형태를 가지고 있다.

이러한 문제를 해결하기 위하여 사용할 수 있는 방법으로 반복 알고리즘을 생각해 볼 수 있다. 반복 알고리즘은 우선 반복의 시작단계에서 임의의 값으로 파라미터의 추정량을 초기화한다. 이렇게 초기화된 파라미터 $\boldsymbol{\theta}^{(0)}$를 [식 10-12], [식 10-13], [식 10-18]의 우변에 대입하면 새로운 파라미터의 추정치를 얻어낼 수 있는데, 이것을 $\boldsymbol{\theta}^{(1)}$이라고 두자. 그러면 다음 단계에서는 다시 $\boldsymbol{\theta}^{(1)}$을 식의 우변에 넣어 다시 수정된 새로운 추정치 $\boldsymbol{\theta}^{(2)}$를 얻을 수 있을 것이다. 이와 같은 과정을 계속하여 τ번째 반복에서 얻어진 추정치 $\boldsymbol{\theta}^{(\tau)}$를 이용하여 다음 단계의 추정치 $\boldsymbol{\theta}^{(\tau+1)}$를 얻는 과정을 반복하게 된다. 이렇게 함으로써 반복할 때마다 목적함수의 값을 조금씩 증가시켜서 최종적으로 목적함수를 최대화하는 파라미터를 얻어가는 과정을 수행하게 된다.

이러한 반복 알고리즘은 최적화 문제에 있어서 널리 사용되는 방법으로, 이 장에서 살펴볼 가우시안 혼합 모델의 최우추정치를 얻기 위한 반복 알고리즘은 EM 알고리즘의 특별한 경우에 해당된다. EM 알고리즘은 가우시안 혼합 모델을 비롯하여 어떤 특정 성질을 가지는 목적함수를 최적화 하는 방법으로 사용되는 대표적인 최적화 방법 중의 하나이다. 다음 절에서는 이 EM 알고리즘을 가우시안 혼합 모델에 적용한 경우를 통해 그 특성을 이해하고, 가우시안 혼합 모델의 학습 알고리즘을 정리해 보겠다.

10.3 EM 알고리즘

10.3.1 가우시안 혼합 모델과 은닉변수

EM 알고리즘을 이해하기 위하여 간단한 예를 들어서 살펴보도록 하자. 남, 여 두 집단이 함께 속한 하나의 그룹으로부터 한 사람씩 뽑아서 신장을 측정하여 얻은 데이터를 $\{x_1,x_2,...,x_N\}$라고 하자. 신장에 대한 확률변수를 x 라고 했을 때, 확률밀도함수 $p(x)$를 추정하고자 한다([그림 10-3] 참조).

우선 확률밀도함수 $p(x)$에 대한 확률 모델을 설정해야 하므로, 관찰된 데이터 $X=\{x_1,x_2,...,x_N\}$ 가 추출된 표본 공간의 특성에 대해 생각해 보아야 할 것이다. 표본 공간 안에는 여자들의 집단과 남자들의 집단이 존재하고, 이들 두 집단의 평균 신장은 일반적으로 다르다고 예상할 수 있다. 따라서 $p(x)$는 두 개의 가우시안이 합쳐진 가우시안 혼합 모델을 이용하여 다음 식과 같이 표현해 볼 수 있겠다.

$$p(x)=\sum_{j=1}^{2}\alpha_j\,p(x|\mu_j,\sigma_j^2)=\alpha_1\,p(x|\mu_1,\sigma_1^2)+\alpha_2\,p(x|\mu_2,\sigma_2^2) \quad \text{[식 10-19]}$$

이 식에서 첫 번째 가우시안이 여자 그룹의 확률분포를, 두 번째 가우시안이 남자 그룹의 확률분포를 나타낸다고 가정한다. 이렇게 정의된 확률 모델에서 존재하는 파라미터는 μ_j, σ_j^2, α_j $(j=1,2)$의 총 여섯 개로, 데이터 집합 $X=\{x_1,x_2,...,x_N\}$ 를 이용하여 이 파라미터들을 추정함으로써 확률밀도 추정이 이루어진다.

[식 10-19]는 두 개의 가우시안 분포를 이루고 있으므로, 만약 각 데이터 x_i가 여자 그룹과 남자 그룹 중 어디에 속하는 것인지를 먼저 알고 있다면, 전체 데이터 집합을 두 그룹으로 나눈 후 각각에 대해 3장에서 배운 방법으로 추정이 가능할 것이다. 이 때 필요한 데이터의 클래스 라벨에 대한 정보를 나타내기 위해 새로운 변수 $\boldsymbol{z}=[C_1,C_2]$를 정의해 보자. 이때, C_1, C_2는 각각 1 아니면 0의 값을 가진다. 즉, 현재 관찰된 데이터 x 가 여자로부터 온 것이라면 C_1는 1의 값, C_2는 0의 값을 가지고, 남자로부터 얻어진 것이라면 반대로 C_2는 1의 값, C_1는 0의 값을 가지는 변수로 정의한다. 그런데 이 변수 $\boldsymbol{z}$는 우리가 데이터를 획득할 때에는 주어지지 않는 값으로, 외부에서 값이 측정될 수 없는 변수라 하여 <은닉변수(latent variable)>라고 한다. 이제 [식 10-19]에서 정의한 x 에 대한 확률 모델 $p(x)$에 은닉 확률변수 $\boldsymbol{z}$까지 함께 포함하여 $p(x,\boldsymbol{z})$와 같이 나타낼 수 있고, 이로부터 $p(x)$는 주변확률분포의 정의로부터 다음과 같이 나타낼 수 있다.

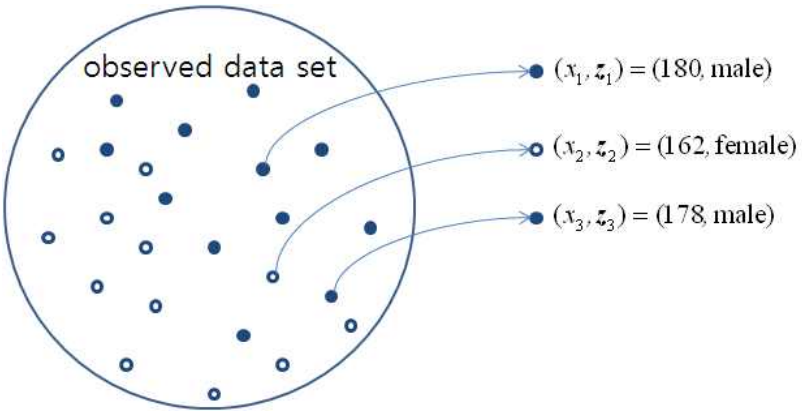
$$p(x)=\sum_{\boldsymbol{z}}p(x,\boldsymbol{z})=\sum_{\boldsymbol{z}}p(x|\boldsymbol{z})p(\boldsymbol{z})=\sum_{\boldsymbol{z}}\sum_{j=1}^{2}\alpha_j\,p(x|\mu_j,\sigma_j^2)p(\boldsymbol{z}) \quad \text{[식 10-20]}$$

이렇게 은닉 확률변수를 추가적으로 정의하여 포함시킴으로써 파라미터 추정을 위한 새로운

접근법인 EM 알고리즘을 알아볼 준비가 되었다. EM 알고리즘은 이와 같이 은닉변수를 가지는 확률 모델에서 파라미터를 반복적으로 추정해 가는 알고리즘이다. 다음 절에서 이에 대해 자세히 알아보겠다.

10.3.2 은닉변수를 가진 확률 모델을 위한 EM 알고리즘

앞 절에서 소개한 예를 계속 사용하여 숨겨진 변수를 가진 가우시안 혼합 모델에서 관찰된 데이터만을 가지고 어떻게 파라미터를 추정하는지에 대해 알아보자. 앞 절에서 주어진 확률 모델의 파라미터를 추정하는 경우에, 그림 10-3]과 같이 두 확률변수에 대한 관찰 데이터가 모두 함께 주어졌다면 간단히 추정이 가능하다. 그러나 현재 관찰된 데이터 집합은 확률 변수 x에 대한 데이터 $X=\{x_1,x_2,...,x_N\}$ 만 존재하여 $\boldsymbol{z}$에 대한 값은 관찰되지 않는다는 것이 문제의 핵심이 된다. 이러한 상황에서 우리는 모델을 위한 파라미터들, 즉 두 그룹의 평균 (μ_1,μ_2)과 분산 (σ_1^2,σ_2^2), 그리고 두 그룹이 전체에 미치는 영향을 나타내는 혼합계수(α_1,α_2)를 추정해야 한다.



[그림 10-3] 숨겨진 변수를 가진 확률 실험

이 문제에 대한 해결책을 찾기 위하여 우선 [그림 10-3]에서처럼 데이터 $X=\{x_1,x_2,...,x_N\}$와 함께 각 데이터들이 어떤 그룹으로부터 얻어진 것인지를 안다고 가정해보자. 즉, 각 데이터 x_i에 대해 숨겨진 확률변수의 값 $\boldsymbol{z_i}$도 관찰된 경우를 생각한다. 만약 이와 같이 숨겨진 변수에 대한 값까지 모두 관찰되었다고 가정하면, 우리가 추정하고자 하는 파라미터는 다음과 같은 단계를 통해 쉽게 얻어질 수 있다. 첫째로, 두 그룹의 평균과 표준 편차를 추정한다. 이는 앞서도 잠깐 언급한 바와 같이, 주어진 $\boldsymbol{z}$값에 따라 전체 데이터 집합을 여자 그룹과 남자 그룹으로 각각 나눔으로써 쉽게 해결된다. 분리된 두 그룹에 대해 따로따로 가우시안 분포의 파라미터를 추정하는 문제와 같아지므로, 3장에서 배운 바와 같이 표본 평균과 표본 분산으로 추정할 수 있다. 추정량을 식으로 표현하면 다음과 같다.

$$\mu_j=\frac{1}{N_j}\sum_{i=1}^{N}C_{ij}x_i\,,\ \ \sigma_j^2=\frac{1}{N_j}\sum_{i=1}^{N}C_{ij}(x_i-\mu_j)^2 \qquad \text{[식 10-21]}$$

이 식에서 $C_{ij}(i=1,...,N; j=1,2)$는 i번째 데이터 x_i에 대한 은닉변수 C_j의 값을 나타내며, $N_j(j=1,2)$는 각 그룹에 속한 데이터의 수를 나타낸다. 이어서 두 번째로 각 그룹의 중요도를 나타내는 혼합계수 파라미터 α_j도 추정해야 한다. 이 α_j는 전체 데이터에서 j번째 그룹이 차지하는 비율을 의미하므로, 각 그룹에 속한 데이터의 수 N_j를 이용하여 다음과 같이 간단히 추정할 수 있다.

$$\alpha_j = \frac{N_j}{N} \qquad \text{[식 10-22]}$$

그런데 우리가 추정하고자 하는 문제의 경우에는 $\boldsymbol{z_i} = [C_{i1}, C_{i2}]$ 의 값은 알 수 없으며, 단지 x_i의 값만 관찰된다. 이 문제를 해결하기 위해서 EM 알고리즘에서는 은닉변수 $\boldsymbol{z}$의 값에 대한 추정과 파라미터 $\boldsymbol{\theta} = [\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha_1, \alpha_2]$의 추정을 반복적으로 수행하는 전략을 취한다. 첫 번째 단계를 E-step, 두 번째 단계를 M-step이라고 한다. 각 단계에 대하여 구체적인 방법을 기술하면 다음과 같다.

(1) E-step : 은닉변수 $\boldsymbol{z}$의 기대치를 계산

이 단계에서는 파라미터의 값이 주어졌다고 가정하고 은닉변수 $\boldsymbol{z}$의 기대치를 계산한다. 실제로는 파라미터 값이 주어지지 않았으므로, 처음에는 임의의 값으로 정하고 이후에는 M-step을 수행하여 얻어지는 파라미터를 사용한다. 파라미터의 값이 주어졌다고 하면 이에 의해 확률밀도함수가 결정될 수 있으므로, 이것을 이용하여 확률변수 $\boldsymbol{z}$의 기대치 E[$\boldsymbol{z}$]를 계산할 수 있다. 구체적으로는, 확률변수 x에 대하여 데이터 집합 $X = \{x_1, x_2, ..., x_N\}$ 가 관찰되어 있고 파라미터 $\boldsymbol{\theta}$에 대한 값이 주어져 있는 상태에서, 각 데이터 x_i에 대응되는 $\boldsymbol{z}$의 기대치 E[$\boldsymbol{z}|x_i,\boldsymbol{\theta}$]를 계산하여 이것을 $\boldsymbol{z_i}$의 추정치로 사용하고자 하는 것이다.

앞 절에서 사용한 예에서 E[$\boldsymbol{z}|x_i,\boldsymbol{\theta}$]를 풀어쓰면 [E[$C_1|x_i,\boldsymbol{\theta}$], E[$C_2|x_i,\boldsymbol{\theta}$]]로 나타나고 C_1과 C_2는 각각 1 혹은 0의 값을 가지므로, 결국 기대치 E[$C_j|x_i,\boldsymbol{\theta}$]는 C_j가 1이 될 확률값 $p(C_j=1|x_i,\boldsymbol{\theta})$ 을 추정함으로써 얻을 수 있다. [식 10-19]에서 정의된 가우시안 혼합 모델에서 E[$C_j|x_i,\boldsymbol{\theta}$]의 값은 다음과 같이 계산할 수 있다.

$$E[C_j|x_i,\boldsymbol{\theta}] = P(C_j=1|x_i,\mu_j,\sigma_j^2) = \frac{p(C_j=1,x_i|\boldsymbol{\theta})}{p(x_i|\boldsymbol{\theta})} = \frac{\alpha_j\, p(x_i|\mu_j,\sigma_j^2)}{p(x_i|\boldsymbol{\theta})} \qquad \text{[식 10-23]}$$

이와 같이 E-step은 현재의 파라미터를 이용하여 은닉변수의 기대치를 계산하는 과정으로, 기대치(Expectation)의 앞 글자를 따서 E-step이라고 불린다.

(2) M-step : 은닉변수 z의 기대치를 이용하여 파라미터를 추정

E-step에서 우리는 관찰되는 $\boldsymbol{z_i}$ 값을 대신할 수 있는 기대치 E[$\boldsymbol{z}|x_i,\boldsymbol{\theta}$]를 추정하였다. 이

값을 활용하면 최우추정법에 따라 확률밀도함수의 로그우도를 최대화하는 파라미터를 추정할 수 있다. 즉, 우리가 얻고자 하는 파라미터를 수식으로 표현하면 다음과 같다.

$$\begin{aligned}\boldsymbol{\theta} &= \mathrm{argmax}\{\ln p(x_1,...,x_n, \boldsymbol{z_1},...,\boldsymbol{z_n})\} \\ &= \mathrm{argmax}\left\{\sum_{i=1}^{N}\ln p(x_i, \boldsymbol{z_i})\right\} \\ &= \mathrm{argmax}\left\{\sum_{i=1}^{N}\ln p(x_i, E[\boldsymbol{z}|x_i, \boldsymbol{\theta}])\right\}\end{aligned} \qquad \text{[식 10-24]}$$

앞 절에서 사용된 예제에서 추정해야 하는 파라미터는 평균(μ_1, μ_2)과 분산 (σ_1^2,σ_2^2), 그리 혼합계수(α_1,α_2)이다. 각각에 대해 로그우도를 최대화하는 값을 얻기 위하여 10.2.1절에서 계산된 최우추정량에 대한 결과로 돌아가 보자. 최우추정 방법에서 얻어진 [식 10-12], [식 10-13], [식 10-18]이 바로 사용될 수 없었던 것은 식의 우변에 추정해야 하는 파라미터를 이용해서 계산되어지는 값 $p(C_j|x_i,\boldsymbol{\theta})$가 존재했기 때문이었다. 그런데 [식 10-23]을 살펴보면, $p(C_j|x_i,\boldsymbol{\theta})$은 E-step에서 얻어진 기대치와 동일한 값이 된다. 따라서 M-step에서는 E-step에서 얻어진 기대치 E[$\boldsymbol{z}|x_i,\boldsymbol{\theta}$]를 이용하여 다음과 같이 로그우도를 최대화하는 파라미터를 추정할 수 있다.

$$\mu_j = \frac{\sum_{i=1}^{N} p(C_j|x_i,\boldsymbol{\theta})x_i}{\sum_{i=1}^{N} p(C_j|x_i,\boldsymbol{\theta})} = \frac{\sum_{i=1}^{N} \mathrm{E}[C_j|x_i,\boldsymbol{\theta}]x_i}{\sum_{i=1}^{N} \mathrm{E}[C_j|x_i,\boldsymbol{\theta}]} \qquad \text{[식 10-25]}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{N} p(C_j|x_i,\boldsymbol{\theta})(x_i-\mu_j)^2}{\sum_{i=1}^{N} p(C_j|x_i,\boldsymbol{\theta})} = \frac{\sum_{i=1}^{N} \mathrm{E}[C_j|x_i,\boldsymbol{\theta}](x_i-\mu_j)^2}{\sum_{i=1}^{N} \mathrm{E}[C_j|x_i,\boldsymbol{\theta}]} \qquad \text{[식 10-26]}$$

$$\alpha_j = \frac{1}{N}\sum_{i=1}^{N} p(C_j=1|x_i,\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}[C_j|x_i,\boldsymbol{\theta}] \qquad \text{[식 10-27]}$$

이 결과는 보다 직관적으로 이해할 수 있도록 다음과 같이 생각해 보자. 두 가우시간 성분으로 나타나는 각 그룹의 평균과 분산을 추정함에 있어서, 완전히 관찰된(이진값으로 주어지는) $\boldsymbol{z}$값이 존재한다면 [식 10-21]과 같이 각 그룹별로 나누어 표본평균과 표본분산을 계산함으로써 최우추정치를 얻을 수 있을 것이다. 그러나 지금의 경우에는 완전히 관찰된 $\boldsymbol{z}$값(이진값)을 대신하여 실수로 나타나는 평균값을 이용하여야 하므로, 이것을 각 데이터가 각 그룹에 속하는 소속도라고 생각하여 평균과 분산을 계산하는 것으로 볼 수 있다. 마찬가지 개념이 α_j를 추정할 때에도 적용될 수 있다. 다시 말하면 [식 10-25], [식 10-26], [식 10-27]에서 사용된 기대치 E[$C_j|x_i,\boldsymbol{\theta}$]가 실수값으로 주어지는 대신 x_i가 어떤 그룹에 속하는 지 여부에 따라 이진값으로 주어진다면, 이 추정치 계산식들은 [식 10-21], [식 10-22]에서 사용한 추정식과 동일해 진다.

이와 같이 M-step은 E-step에서 찾아진 기대치를 은닉변수의 관찰값으로 간주하여, 로그우도를 최대화 하는 파라미터를 찾는 과정으로, 최대화(Maximization)의 첫 글자를 따서

M-step이라고 불린다.

M-step을 수행하고 나면, 이전의 파라미터로부터 수정된 새로운 파라미터를 얻게 되고, 이것을 이용하여 다시 숨겨진 변수 z의 기대치 혹은 확률값을 계산하는 E-step을 수행할 수 있다. 이렇게 M-step과 E-step을 반복하면서 보다 정확한 파라미터를 추정하게 된다. 앞서 살펴본 예에서와 같이, 기본 확률변수 x의 값이 관찰되면서 숨겨진 확률변수 z의 값도 함께 관찰된다면 EM과 같은 반복적인 학습에 의한 추정은 필요하지 않다. 즉, EM 알고리즘은 숨겨진 확률변수 z를 가지고 있는 확률 모델의 파라미터를 추정하기 위해 사용되는 방법이다.

10.3.3 가우시안 혼합 모델을 위한 EM 알고리즘

지금까지 간단한 1차원 데이터를 위한 가우시안 혼합 모델에 대해 알아보았다. 이 절에서는 이를 일반화하여 n차원 입력벡터와 일반적인 행렬을 공분산으로 가지는 가우시안 혼합 모델에 대한 EM 알고리즘을 제시한다. 가우시안 혼합 모델을 위한 EM 알고리즘의 수행단계는 다음과 정리할 수 있다.

[가우시안 혼합 모델을 위한 EM 알고리즘의 수행 단계]

① 가우시안 성분의 수 M을 정하고, 주어진 데이터 집합 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ 을 설명할 수 있는 가우시안 혼합 모델을 정의한다.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{M} \alpha_i \, p(\boldsymbol{x}|\boldsymbol{\mu_i}, \Sigma_i)$$

② 각 파라미터의 초기치를 임의로 설정한다.

$\boldsymbol{\theta^{(0)}} = [\boldsymbol{\mu_1^{(0)}}, \boldsymbol{\mu_2^{(0)}}, ..., \boldsymbol{\mu_M^{(0)}}, \Sigma_1^{(0)}, \Sigma_2^{(0)}, ..., \Sigma_M^{(0)}, \alpha_1^{(0)}, \alpha_2^{(0)}, ..., \alpha_M^{(0)}]$

③ **[E-step]** τ번째 반복 단계에서 주어진 파라미터 $\boldsymbol{\theta^{(\tau)}}$를 이용하여 기대치를 계산한다.

$$r_{ij}^{(\tau)} = E[C_j|\boldsymbol{x}, \boldsymbol{\theta^{(\tau)}}] = \frac{\alpha_j^{(\tau)} p(\boldsymbol{x_i}|\boldsymbol{\mu_j^{(\tau)}}, \Sigma_j^{(\tau)})}{p(\boldsymbol{x_i}|\boldsymbol{\theta^{(\tau)}})} = \frac{\alpha_j^{(\tau)} p(\boldsymbol{x_i}|\boldsymbol{\mu_j^{(\tau)}}, \Sigma_j^{(\tau)})}{\sum_{k=1}^{M} \alpha_k^{(\tau)} p(\boldsymbol{x_i}|\boldsymbol{\mu_k^{(\tau)}}, \Sigma_k^{(\tau)})}$$

④ **[M-step]** E-step에서 얻어진 기대치를 이용하여 $\tau+1$번째 파라미터 $\boldsymbol{\theta^{(\tau+1)}}$를 계산한다.

$$\boldsymbol{\mu_j^{(\tau)}} = \frac{\sum_{i=1}^{N} r_{ij}^{(\tau)} \boldsymbol{x_i}}{\sum_{i=1}^{N} r_{ij}^{(\tau)}}, \quad \Sigma_j^{(\tau)} = \frac{\sum_{i=1}^{N} r_{ij}^{(\tau)} (\boldsymbol{x_i} - \boldsymbol{\mu_j^{(\tau)}})(\boldsymbol{x_i} - \boldsymbol{\mu_j^{(\tau)}})^T}{\sum_{i=1}^{N} r_{ij}^{(\tau)}}, \quad \alpha_j^{(\tau)} = \frac{1}{N} \sum_{i=1}^{N} r_{ij}^{(\tau)}$$

⑤ 파라미터가 수렴할 때까지 혹은 원하는 로그우도값이 얻어질 때까지 E-step(③)과 M-step(④)를 반복한다.

10.4 일반화된 EM 알고리즘

EM 알고리즘은 가우시안 혼합 모델에 국한된 학습 알고리즘이 아니라, 다양한 확률 모델에서 파라미터 추정을 위해 사용되는 일반적인 학습법이다. 특히 은닉변수를 가진 확률 모델에서는 은닉변수와 파라미터를 함께 추정하는데 사용될 수 있어서, 다른 최적화 알고리즘과는 차별되는 특성을 가지고 있다. 또한 가우시안 혼합 모델의 경우와 같이 원래 주어진 모델에는 은닉변수가 존재하지 않는다 하더라도, 이것을 은닉변수를 가진 모델로 변형하여 표현함으로써 EM 알고리즘을 적용할 수 있다. 이 절에서는 EM 알고리즘을 적용할 수 있는 은닉변수를 가진 확률 모델을 일반적인 식으로 정의하고, 이에 대한 EM 알고리즘에 대해 알아보겠다.

은닉변수란, 일반변수와는 달리 외부에서는 관찰될 수 없는 값에 대한 변수로, 데이터가 생성되는 시스템을 보다 잘 설명하기 위해 도입되는 변수들이다. 앞서 살펴본 가우시안 혼합 모델의 경우, 처음 모델을 정의할 때에는 외부에서 관찰되는 변수 x만을 사용하였으나, 데이터가 생성되는 과정을 좀 더 자세히 고려한다면, 관찰된 데이터 x는 혼합 모델을 이루는

여러 가우시안 성분 중 하나로부터 생성된다고 볼 수 있으므로 이를 나타내는 확률변수를 추가적으로 도입할 수 있다. 그러나 이는 일반적으로 관찰되지 않는 값이므로 은닉변수가 된다. 이와 같이 은닉변수를 가지는 다양한 확률 모델이 존재하는데, 시계열 분석에서 널리 사용되는 은닉마르코프 모델(HMM, Hidden Markov Model)이나 베이지안망 (Bayesian Networks)과 같은 그래피컬 모델(Gramphical Model) 등이 대표적인 예가 된다. 이 절에서는 특정 모델에 국한시키지 않고 은닉변수를 가지는 일반적인 확률 모델에 대하여 EM 알고리즘을 설명한다.

관찰된 데이터에 대한 확률변수를 $\boldsymbol{x}$, 은닉 확률변수를 $\boldsymbol{z}$, 그리고 이 모델을 설명하기 위한 파라미터를 $\boldsymbol{\theta}$라고 하면, 전체 시스템에 대한 확률밀도함수는 두 변수의 결합 확률밀도함수 $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$로 나타낼 수 있다. 우리는 확률변수 $\boldsymbol{x}$에 대해 관찰된 데이터 $X=\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}\}$를 가지고 있으므로, 이에 대한 로그우도 $\ln p(X|\boldsymbol{\theta})$를 최대로 하는 파라미터를 추정해야 한다. 이 때 $p(\boldsymbol{x}|\boldsymbol{\theta})$는 결합확률분포 $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$를 $\boldsymbol{z}$에 대해 적분(이산변수의 경우 합산)한 주변 확률분포로 계산될 수 있음을 이용하면 관찰된 데이터 X에 대한 로그우도는 다음 식과 같이 정의할 수 있다.

$$\ln p(X|\boldsymbol{\theta}) = \ln \sum_{\boldsymbol{z}} p(X, \boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{\theta}) \qquad \text{[식 10-28]}$$

이 로그우도함수는 확률변수 $\boldsymbol{x}$에 대한 데이터만으로 정의되는 것으로 <불완전 데이터 로그우도 (incomplete-data log-likelihood)>라고 부른다. 그런데 이 로그우도는 데이터 집합 X의 각각의 데이터들이 서로 독립이라고 하더라도 이것을 로그를 취한 합산 형태로 간소화할 수 없다. 결과적으로 [식 10-28]을 최대화하는 파라미터를 찾는 최적화 문제는 풀기 어려운 형태가 된다.

이러한 문제를 해결하기 위하여, 먼저 은닉 확률변수를 $\boldsymbol{z}$에 대해서도 관찰된 데이터 $Z=\{\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_N}\}$ 가 주어진다고 가정해 보자. 그러면 두 데이터 집합 X와 Z에 대한 우도함수를 다음과 같이 정의할 수 있을 것이다.

$$\ln p(X, Z|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(\boldsymbol{x_i}, \boldsymbol{z_i}|\boldsymbol{\theta}) \qquad \text{[식 10-29]}$$

이 우도함수는 확률 모델을 정의하는 모든 확률변수에 대하여 데이터가 관찰되어 정의된 우도함수로서, 이를 <완전 데이터 로그우도(complete-data log-likelihood)>라고 부른다. 이 로그우도는 [식 10-29]에서 보이는 것처럼 각각의 데이터의 독립성을 이용하여 각 데이터의 로그우도를 합산한 형태로 표현가능하고, 가우시안 분포처럼 p가 지수함수로 주어지는 경우 최적화 과정이 매우 간단해 진다.

그런데 은닉변수를 가진 문제에서는 확률변수 $\boldsymbol{z}$에 대한 데이터 $Z=\{\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_N}\}$ 는 주어지지 않다. 따라서 구체적인 데이터를 사용하는 대신 $\boldsymbol{z}$에 대한 기대치를 사용해야 한다. 즉 완전 데이터 로그우도 함수를 $\boldsymbol{z}$의 확률분포함수를 이용하여 적분 (이산변수인 경우 합산)을 취한 값을 다음 [식 10-30]과 같이 정의한다.

$$\mathrm{E}_{\boldsymbol{z}}[\ln p(X,\ \boldsymbol{z}|\boldsymbol{\theta})]=\sum_{z} p(\boldsymbol{z}|X,\boldsymbol{\theta}^{(\tau)})\ln p(X,\ \boldsymbol{z}|\boldsymbol{\theta}) \qquad \text{[식 10-30]}$$

앞 절에서는 데이터 x_i와 파라미터 $\boldsymbol{\theta}^{(\tau)}$가 주어졌을 때 $\boldsymbol{z}$의 기대치 E[$\boldsymbol{z}|x_i,\boldsymbol{\theta}^{(\tau)}$]를 각각 계산하였다. 그런데 일반적인 확률 모델에서는 이 값을 계산하는 것이 어려운 경우가 존재한다. 또한 실제로 M-step에서 사용되는 것은 $\boldsymbol{z}$의 기대치 자체가 아니라, $\boldsymbol{z}$의 기대치를 이용한 [식 10-30]과 같이 정의되는 완전 데이터 로그우도의 기대치이다.

이렇게 정의된 완전 데이터 로그우도의 기대치를 계산한 후, 이를 최대화하는 파라미터를 찾음으로써 원하는 추정치를 얻게 된다. [식 10-30]을 자세히 들여다보면, 현재의 파라미터 $\boldsymbol{\theta}^{(\tau)}$와 관찰된 데이터 $X=\{\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_N}\}$ 를 가지면 $\boldsymbol{z}$에 대한 조건부 확률 $p(\boldsymbol{z}|X,\boldsymbol{\theta}^{(\tau)})$ 을 얻을 수 있고, 이것을 이용하여 로그우도 $\ln p(X,\ \boldsymbol{z}|\boldsymbol{\theta})$ 의 기대치를 계산하게 된다. 그런데 로그우도 $\ln p(X,\ \boldsymbol{z}|\boldsymbol{\theta})$ 는 $\boldsymbol{\theta}$의 함수이므로, 결국 [식 10-30]은 파라미터 $\boldsymbol{\theta}$의 함수가 되며, 또한 $p(\boldsymbol{z}|X,\boldsymbol{\theta}^{(\tau)})$ 가 $\boldsymbol{\theta}^{(\tau)}$에 의존하여 결정되므로 현재 주어진 상수값 $\boldsymbol{\theta}^{(\tau)}$에도 의존하여 그 형태가 변하게 되는 성질을 가진다. 이에 주목하여 [식 10-30]은 일반적으로 함수 $\mathrm{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(\tau)})$ 라고 나타내고 <Q함수 (Q-learning)>라고 부른다.

E-step에서는 현재 주어진 상수값 $\boldsymbol{\theta}^{(\tau)}$과 데이터 $X=\{\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_N}\}$ 를 이용하여 함수 $\mathrm{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(\tau)})$ 를 계산하고, 이어서 M-step에서는 얻어진 $\mathrm{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(\tau)})$ 를 최대화하는 새로운 파라미터 $\boldsymbol{\theta}^{(\tau+1)}$를 찾게 된다. 이 과정을 반복함으로써 결국 불완전 데이터 $X=\{\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_N}\}$ 에 대한 로그우도를 극대화하는 파라미터를 얻게 된다. 지금까지 기술한 일반화된 EM 알고리즘을 정리하면 다음과 같다.

[일반적인 EM 알고리즘의 수행 단계]

① 주어진 데이터 집합 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\}$ 을 설명할 수 있는 은닉변수를 가진 확률 모델 $p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})$ 을 정의한다.

② 각 파라미터의 초기치 $\boldsymbol{\theta}^{(0)}$를 임의로 설정한다.

③ **[E-step]** τ번째 반복 단계에서 주어진 파라미터 $\boldsymbol{\theta}^{(\tau)}$를 이용하여 Q함수를 얻는다.

$$\mathrm{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(\tau)})=\mathrm{E}_{\boldsymbol{z}}[\ln p(X,\ \boldsymbol{z}|\boldsymbol{\theta})]=\sum_{z} p(\boldsymbol{z}|X,\boldsymbol{\theta}^{(\tau)})\ln p(X,\ \boldsymbol{z}|\boldsymbol{\theta})$$

④ **[M-step]** E-step에서 얻어진 Q함수를 최대화 하는 파라미터$\boldsymbol{\theta}^{(\tau+1)}$를 계산한다.

$$\boldsymbol{\theta}^{(\tau+1)} = \mathrm{argmax}_{\boldsymbol{\theta}}\{Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(\tau)})\}$$

⑤ 파라미터가 수렴할 때까지 혹은 원하는 Q값이 얻어질 때까지 E-step(③)과 M-step(④)를 반복한다.

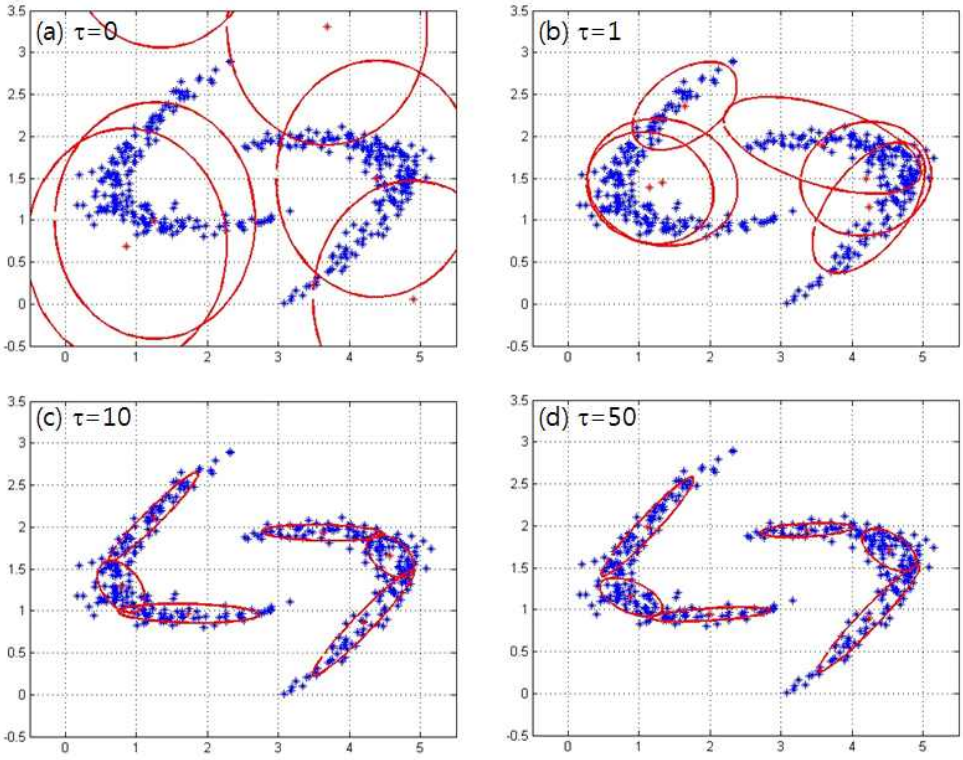
지금까지 이 절에서 살펴본 내용의 정리하면, EM 알고리즘이란 결국 관찰된 데이터로부터

정의되는 불완전 데이터 로그우도를 직접적으로 최대화하는 것이 불가능한 경우에 그것을 해결하는 방법으로 볼 수 있다. 이를 위해 은닉변수를 도입하여 새로운 완전 데이터 로그우도를 정의하고, 이것의 기대치로 정의되는 새로운 함수 $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\tau)})$ 를 이용하여 파라미터를 최대화하는 전략을 취한다. 앞서 살펴본 바와 같이 $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\tau)})$ 함수, 혹은 완전 데이터에 대한 로그우도는 지수함수 형태의 확률 밀도함수의 경우에 쉽게 계산될 수 있는 특성을 가지고 있다. 만약 그렇지 못한 확률 밀도함수를 사용하면 [식 10-30]에 정의된 함수도 최적화하기 힘든 경우도 존재한다. 이러한 경우에는 $\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\tau)})$ 를 최대화하는 파라미터 $\boldsymbol{\theta}^{(\tau+1)}$ 를 찾는 대신 어느 정도 큰 값을 줄 수 있는 근사된 최적화 파라미터를 찾는 전략을 취하기도 한다. 이와 같이 주어진 문제에 따라 그에 적합한 형태로 변형된 다양한 EM 알고리즘의 변형들이 존재한다. 자세한 내용은 이 장의 마지막에 소개하는 참고자료를 참조하기 바란다.

10.5 매트랩을 이용한 실험

[그림 10-2]에 주어진 데이터 분포를 가우시안 혼합 모델을 이용하여 추정하는 매트랩 프로그램을 구현해 보겠다. [프로그램 10-1]에 이를 위한 EM 알고리즘 코드를 제시하였다. 이와 함께 [프로그램 10-1-1]과 [프로그램 10-1-2]에는 부가적으로 사용되는 함수를 정의해 두었다. [프로그램 10-1-1]은 입력 데이터 집합의 가우시안 밀도 함수를 계산하는 것으로 [프로그램 10-1]에서 각 성분의 확률값을 계산하기 위하여 호출하여 사용한다. [프로그램 10-1-2]는 학습 도중에 파라미터 값들이 어떻게 변하는지를 그림으로 표현하기 위하여 데이터 집합과 함께 평균 파라미터의 위치와 타원으로 표현한 공분산 파라미터의 크기를 2차원 공간상에 표시해 주는 함수이다. [프로그램 10-1]에서 학습을 수행하면서 정기적으로 이 함수를 호출하여 파라미터의 변화를 표시하였다.

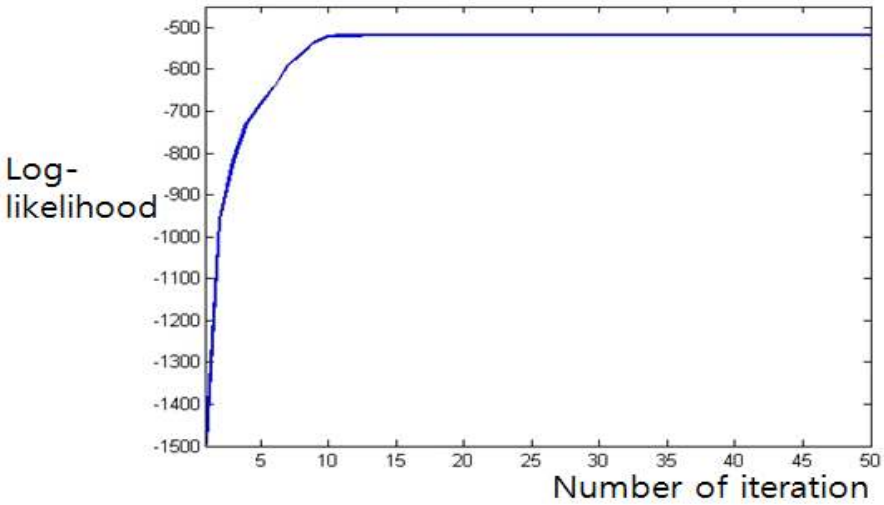
[프로그램 10-1]에서는 먼저 데이터를 불러온 후 관련된 변수값(데이터 크기, 성분개수 등)을 초기화하고 파라미터도 초기화한다. 여기서는 성분의 수 M을 6으로 두었으나 이는 자유롭게 설정가능하다. 평균 파라미터는 입력 공간 내에서 랜덤하게 설정하였으며, 공분산 파라미터는 모든 성분이 동일하게 단위행렬이 되도록 하였다. 혼합계수 파라미터도 모든 성분이 동일한 값을 가지도록 설정하였다. 이어서 전체 데이터 집합에 대하여 E-step(r_{ij}값의 계산)과 M-step(파라미터의 수정)을 반복하여 수행한다. 최대 반복횟수를 100으로 두었으나 이 역시 자유롭게 설정할 수 있다. 또한 E-step으로 r_{ij}의 값이 계산되면 이와 함께 로그우도도 계산하여 학습이 진행됨에 따른 변화를 저장할 수 있도록 하였다. 저장된 로그우도 값은 프로그램의 마지막에 그래프로 그리도록 하였다.



[그림 10-4] 학습 횟수에 따른 파라미터의 변화

[그림 10-4]에는 [프로그램 10-1-2]에서 정의된 함수 drawgraph에 의해 그려진 파라미터의 변화를 보이고 있다. [그림 10-4a]에 보이는 것이 초기화된 파라미터로, 랜덤하게 설정된 평균과 단위행렬로 설정된 공분산이 나타나 있다. [그림 10-4b]가 한 번 학습을 수행한 결과이다. 한 번의 학습에 의해서도 어느 정도 평균의 위치를 찾아감을 볼 수 있다. 이어서 [그림 10-4c]에서는 10번 학습 후의 파라미터 값을, [그림 10-4d]에서는 50번 학습한 후의 파라미터의 값을 나타내었다. 그림에서 50번 학습 후에는 적절한 평균과 공분산을 추정할 수 있음을 볼 수 있다.

[그림 10-5]에는 학습이 진행됨에 따른 학습 데이터의 로그우도의 변화를 나타내었다. 학습의 초기에는 로그우도에 급격한 변화가 있으며, 10번 정도 학습 후에는 어느 정도 수렴값에 도달하여 그 이후에 천천히 변화하여 수렴된 값을 찾아감을 볼 수 있다. 이는 [그림 10-4]에서 살펴본 결과와 잘 부합된다고 볼 수 있겠다. 이 때 로그우도의 값은 1보다 작은 확률값에 대해 로그를 취한 것이므로 음수가 됨에 유의하자.



[그림 10-5] 학습 횟수에 따른 로그우도의 변화

프로그램 10-1 EM Algorithm for Gaussian Mixtures

2차원 데이터를 위한 가우시안 혼합 모델의 EM 알고리즘

```
load data10_2                    % 데이터 불러오기
X=data';                         % X: 학습 데이터
N=size(X,2);                     % N: 데이터의 수
M=6;                             % M: 가우시안 성분의 수
Mu=rand(M,2)*5;                  % 파라미터의 초기화 (평균)
for i=1:M                        % 파라미터의 초기화 (분산)
    Sigma(i,1:2,1:2) = [1 0; 0 1];
end
alpha=zeros(6,1)+1/6;            % 파라미터의 초기화 (혼합계수)
drawgraph(X, Mu, Sigma, 1);      % 그래프그리기 함수 호출

Maxtau=100;                      % 최대 반복횟수 설정
for tau=1: Maxtau
   %%% E-step %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
   for j=1:M                     % $p(x_i|\mu_j,\sigma_j^2)$ 계산
      px(j,:) = gausspdf(X, Mu(j,:), reshape(Sigma(j,:,:),2,2));
   end
   sump=px'*alpha                % $\alpha_j p(x_i|\mu_j,\sigma_j^2)$ 계산
   for j=1:M                     % $r_{ij}$ 계산
      r(:,j) = (alpha(j)*px(j,:))'./sump;
   end
   L(tau)=sum(log(sump));        %현재 파라미터의 로그우도 계산

   %%% M-step %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
   for j=1:M
        sumr=sum(r(:,j))               % $r_{ij}$의 성분별 합산
        Rj=repmat(r(:,j),1,2)';        % 행렬 계산을 위한 준비
        Mu(j,:) = sum(Rj.*X,2)/sumr    % 새로운 평균
        % 새로운 공분산 계산
        rxmu= (X-repmat(Mu(j,:),N,1)').*Rj;
        Sigma(j,1:2,1:2)= rxmu*(X-repmat(Mu(j,:),N,1)')'/sumr;
        alpha(j)=sumr/N;               % 새로운 혼합계수
   end
   if (mod(tau,10)==1)                 % 그래프그리기 함수 호출
        drawgraph(X, Mu, Sigma, ceil(tau/10)+1);
   end
end

drawgraph(X, Mu, Sigma, tau);          % 그래프그리기 함수 호출
figure(tau+1); plot(L);                % 로그우도의 변화 그래프
```

프로그램 10-1-1 가우시안 확률밀도값 계산 함수 gausspdf

입력과 파라미터(μ_j, σ_j^2)를 받아서 가우시안 확률밀도값 $p(x_i|\mu_j, \sigma_j^2)$를 반환

```
001 function [out]=gausspdf(X, mu, sigma)      % 함수 정의
002 n=size(X,1);                               % 입력 벡터의 차원
003 N=size(X,2);                               % 데이터의 수
004 Mu=repmat(mu',1,N);                        % 행렬 연산을 위한 준비
005 % 확률밀도값 계산
006 out =
007 (1/((sqrt(2*pi))^n*sqrt(det(sigma))))*exp(-diag((X-Mu)'*inv(sigma)*(X-Mu))/2);
```

프로그램 10-1-2 데이터와 파라미터를 그래프로 표현하는 함수 drawgraph

데이터와 파라미터(μ_j, σ_j^2)를 받아서 데이터와 각 성분별 평균의 위치를 2차원 공간에 표시하고 공분산에 따른 확률밀도의 등고선을 타원으로 표현

```
001 function drawgraph(X, Mu, Sigma, cnt)      % 함수 정의
002 M=size(Mu,1);                              % 성분의 수
003 figure(cnt);                               % 데이터 그리기
004 plot(X(1,:), X(2,:), '*'); hold on
005 axis([-0.5 5.5 -0.5 3.5]); grid on
006 plot(Mu(:,1), Mu(:,2), 'r*');              % 평균 파라미터 그리기
007 for j=1:M
008    sigma=reshape(Sigma(j,:,:),2,2);        % 공분산에 따른 타원 그리기
009    t=[-pi:0.1:pi]';
010    A=sqrt(2)*[cos(t) sin(t)]*sqrtm(sigma)+repmat(Mu(j,:), size(t),1);
011    plot(A(:,1), A(:,2), 'r-', 'linewidth' ,2 );
012 end
```

연습문제

1. [그림 10-1]에 나타난 데이터 집합에 대하여 EM 알고리즘으로 가우시안 혼합 모델을 추정하는 프로그램을 직접 매트랩으로 구현해 본다. 이를 위해 다음과 같은 절차를 따르시오.
   (1) 평균이 2.0이고 표준편차가 0.5인 가우시안 분포를 따르는 데이터 350개를 생성하시오.
   (2) 평균이 0.5이고 표준편차가 0.5인 가우시안 분포를 따르는 데이터 400개를 생성하시오.
   (3) 평균이 -2.0이고 표준편차가 0.5인 가우시안 분포를 따르는 데이터 250개를 생성하시오.
   (4) (1)~(3)에서 생성된 데이터를 합쳐서 모두 1000개의 데이터를 가지는 집합 X를 만들고, 이 분포를 histogram으로 그려서 나타내 보시오.
   (5) 1000개의 데이터 각각이 세 그룹 중 어디에 속하는지 알고 있다는 전제하에서, 각 그룹의 표본평균과 표본분산을 계산하고, 각 그룹에 전체 데이터에 대해서 차지하는 비율을 계산하여 3개의 성분을 가지는 가우시안 혼합 모델을 찾아보시오.
   (6) (5)에서 찾아진 가우시안 혼합 모델 함수를 그래프로 그려서 (4)에서 그린 히스토그램과 그 형태가 일치하는지 확인하시오.
   (7) 데이터들이 어떤 그룹에서 생성되었는지를 모른다는 전제하에, EM 알고리즘을 이용하여 적절한 가우시안 혼합 모델을 찾아보시오. 이때, 성분의 수는 3개로 하고, 파라미터의 초기값은 임의로 정한다.
   (8) (7)에서 찾아진 가우시안 혼합 모델 함수를 그래프로 그려서 (4)의 히스토그램과 (5)에서 찾은 함수와 비교해 보시오.

2. 3개의 그룹을 가지는 2차원 데이터를 적절히 생성하여 문제1번과 같은 과정을 수행해 보시오. 단, 히스토그램을 그리는 대신 2차원 공간상에 데이터들을 점으로 표시하고, 혼합 모델 함수를 그리는 대신 찾아진 각 가우시안 성분의 contour를 2차원 공간에 그려서 비교하시오.

3. 10.5절에서 사용한 데이터에 대하여 성분의 수(M의 값)와 초기치를 변화시켜 가면서 그 결과를 비교해 보시오.

4. [그림 10-2]와 같이 크게 두 클래스로 나눌 수 있는 데이터 집합을 생성하고, 각 클래스에 대하여 가우시안 혼합 모델로 확률분포를 추정하시오. 그리고 이를 이용하여 베이지안 분류기로 분류를 수행해 보시오.

5. 7장에서 살펴본 K-means 알고리즘은 EM 알고리즘과 같이 반복 알고리즘의 형태를 가진다. K-means 알고리즘과 가우시안 혼합 모델을 위한 EM 알고리즘을 비교해 보고 그 관계를 생각해 보시오.

참고자료

이 장에서 소개한 EM알고리즘은 본문에서 기술한 바와 같이 대표적인 반복적 파라미터 추정법의 하나로, 가우시안 혼합 모델 뿐 아니라 은닉변수를 다양한 문제들에 대해 연구되어 왔다. 관련된 다양한 연구들을 소개하고 있는 대표적인 문헌으로 [McLachlan & Krishnan 08]이 있으며, 보다 간단하고 쉬운 입문 자료로 [Borman 04]와 [Bilmes 98]도 활용될 수 있겠다.

[McLachlan & Krishnan 08] G. J. McLachlan and T. Krishnan. The EM Algorithm and Extensions (2nd ed.). Wiley-Interscience. 2008.

[Borman 04] S. Borman, The expectation maximization algorithm – a short tutorial, (http://www.seanborman.com/publications/EM_algorithm.pdf), 2004.

[Bilmes 98] J. Bilmes, " A gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," IEEE trans. on Systems, Man, and Cybernetics (Part B: Cybernetics), 28(3):301-315, 1998.