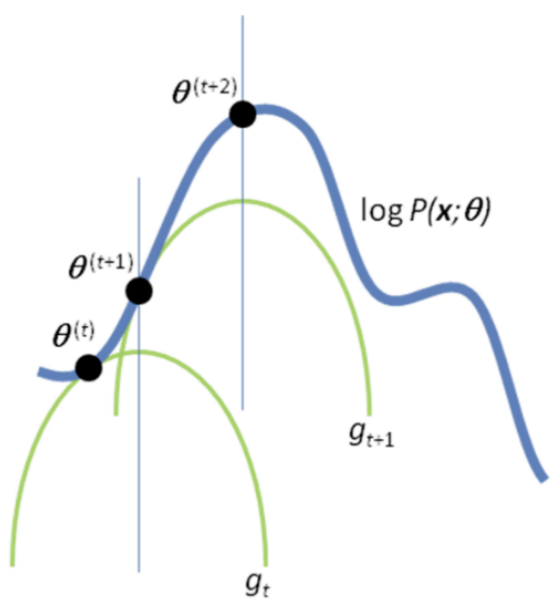
🍏

15강: EM Algorithm & Factor Analysis

EM Algorithm



다시 한 번 EM 알고리즘을 통해 gaussian mixture 모형의 파라미터를 구해보자 .

1. E-step

: $w_j^{(i)}$를 계산한다.

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$$

$Q_i(z^{(i)}=j)$는 Q분포 하에서 $z^{(i)}$=j일 확률을 나타낸다.

2. M-step

: ϕ, μ, Σ에 대해서 다음을 maximize해야한다.

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{k} Q_i(z^{(i)}=j) \log \frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{Q_i(z^{(i)}=j)} \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot \phi_j}{w_j^{(i)}}
\end{aligned}$$

다음과 같은 식을 각각의 파라미터에 대해 미분하고 0이 되는 값을 찾으면 된다.

- μ_1에 대해서 최대화

$$\begin{aligned}
&\nabla_{\mu_l} \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot \phi_j}{w_j^{(i)}} \\
&= \quad -\nabla_{\mu_l} \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j) \\
&= \quad \frac{1}{2}\sum_{i=1}^{n} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T\Sigma_l^{-1}x^{(i)} - \mu_l^T\Sigma_l^{-1}\mu_l \\
&= \quad \sum_{i=1}^{n} w_l^{(i)} \left(\Sigma_l^{-1}x^{(i)} - \Sigma_l^{-1}\mu_l\right)
\end{aligned}$$

위 식을 0으로 두고 μ_1에 대해서 풀면 다음과 같다.

$$\mu_l := \frac{\sum_{i=1}^{n} w_l^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_l^{(i)}},$$

- ϕ에 대해서 최대화 - ϕ에 의존하는 항만 남겨둔다.

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j.$$

이 때, ϕ_j는 제약조건 $\sum_{j=1}^{k}\phi_j = 1$이 있기 때문에 라그랑지안으로 풀면 다음과 같다.

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta(\sum_{j=1}^{k} \phi_j - 1),$$

$$\frac{\partial}{\partial \phi_j} \mathcal{L}(\phi) = \sum_{i=1}^{n} \frac{w_j^{(i)}}{\phi_j} + \beta \qquad \phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)}}{-\beta}$$

Factor Analysis

$x^{(i)} \in \mathbb{R}^d$ 가 가우시안 혼합 구조를 띌 때, 주로 EM 알고리즘을 통해 혼합 모형을 적합할 수 있다.
단, 이는 우리의 training set의 크기
n이 데이터 셋의 차원 d보다 큰 경우이다.

그렇다면 $d \gg n$의 경우를 고려해보자.

이런 경우는 혼합 가우시안은 고사하고 단일 가우시안 모델을 적합하기도 힘들다.

이는 n개의 데이터들이 $\mathbb{R}^d$의 subspace를 span하기 때문이다.

만약 가우시안 모델을 적합한다고 할 때, MLE을 이용한 추정값은 다음과 같다.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$$
$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

다만, Σ 가 singular, 즉 , Σ^{-1} 값이 존재하지 않기 때문에, $1/|\Sigma|^{1/2} = 1/0$이 되고,

multivariate Gaussian distribution을 계산할 수 없다.

이를 일반화하자면, 어느정도로 $n \gg d$가 성립하지 않는다면, mle를 통한 추정값은 영… 좋지 않을 것이다.

이렇게 적은 데이터셋으로 인해 발생하는 문제를 해결하는 방법으로는

첫째, **covariance term Σ에 제약 조건을 주는 방법** (후술하겠지만 영 시원찮다)

둘째, **EM 알고리즘을 통해 Factor Analysis를 진행하는 방법**이 있다.

Restrictions of Σ

1. Diagonal covariance matrix

$$\Sigma_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)^2$$

첫번째 옵션은 시그마를 대각행렬로 제한하는 것이다.

2. Diagonal covariance matrix & diagonal entries must all be equal

$$\sigma^2 = \frac{1}{nd}\sum_{j=1}^{d}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)^2$$

즉, $\Sigma = \sigma^2 I$ 인 경우 이다. 이 경우 gaussian의 contours는 동그라미 모양이다.

위의 두 방법을 이용해 제약규칙을 준다면, 우리는 $n \geq 2$일 때 non-singular한 Σ를 얻는다.
하지만,
Σ에 제약조건을 주는 것은 서로 다른 변량x_i, x_j 간 **상관관계가 없고 독립**인 것으로 간주하고 모델링하는 것이다. 즉, 데이터의 정보를 어느 정도 잃는것이라고 볼 수 있다.

만약 우리가 두 변수간 흥미로운 상관관계를 알아내고 싶다면, 이렇게 Σ에 제약조건을 주는 방법은 썩 좋은 방법이 아니다. 그렇다면, Σ에 제약조건을 주지 않고, **변수 간 상관관계를 알 수 있는 요인 분석**에 대해 알아보도록 하자.

Marginals and conditionals of Gaussians

요인분석에 대해 본격적으로 알아보기에 앞서 marginals and conditionals of Gaussians에 대해 알아보자.

다음과 같은 multivariate gaussian distribution이 있다고 하자.

$$x \sim \mathcal{N}(\mu, \Sigma)$$
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- $x_1 \in \mathbb{R}^r, x_2 \in \mathbb{R}^sandx \in \mathbb{R}^{r+s}$
- $\mu_1 \in \mathbb{R}^r, \mu_2 \in \mathbb{R}^s, \Sigma_{11} \in R^{r\times r}, \Sigma_{12} \in R^{r \times s}$
- covariance matrices are symmetric → $\Sigma_{12} = \Sigma_{21}^T$

이 때,

$$\begin{aligned} Cov(x) &= \Sigma \\ &= E[(x-\mu)(x-\mu)^T] \\ &= E\Big[\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T\Big] \\ &= E\begin{bmatrix} (x_1-\mu_1)(x_1-\mu_1)^T & (x_1-\mu_1)(x_2-\mu_2)^T \\ (x_2-\mu_2)(x_1-\mu_1)^T & (x_2-\mu_2)(x_2-\mu_2)^T \end{bmatrix} \end{aligned}$$

1. marginal distribution of x_1 : $x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$

2. conditional distribution : $x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$

- $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$
- $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

The Factor analysis model

파라미터 설명

우리가 관심있는 것은 joint distribution $p(x,z) = p(x|z)p(z)$이다.

이 때 , $z \in \mathbb{R}^K$는 latent random variable이다.

$$z \sim \mathcal{N}(0, I)$$
$$x|z \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

- $\mu \in \mathbb{R}^{d\times d}, \Lambda \in \mathbb{R}^{d\times k}, \Psi \in \mathbb{R}^{d\times k}$
- $k < d$ (k는 데이터 셋의 개수)
  - 데이터 셋 $x^{(i)}$가 k 차원의 다변량 가우시안 분포 $z^{(i)}$에서 샘플링되어서 생성됨
  - $\mu + \Lambda z^{(i)}$ 연산 → $\mathbb{R}^d$차원의 아핀 공간으로 매핑
    covariance
    Ψ noise를 추가

이를 바탕으로 fatctor analysis 모델을 재정의하면 다음과 같다.

$$z \sim \mathcal{N}(0, I)$$
$$\epsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda z + \epsilon$$

이 때, ϵ과 z는 독립이다.

Joint Gaussian distribution of z and x

이제 이 요인분석 모델이 정확히 어떤 분포를 띄는지 정의하자.

확률변수 z와 x는 joint Gaussian distribution을 띈다.

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$$
$$\Sigma = \begin{bmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{bmatrix}$$

이제 μ_{zx}와 Σ를 구해보자.

1. μ_{zx}

$$\begin{aligned} E[x] &= E[\mu + \Lambda z + \epsilon] \\ &= \mu + \Lambda E[z] + E[\epsilon] \\ &= \mu \end{aligned}$$

이기 때문에, μ_{zx}는 다음과 같다. ($\because E[z] = 0$)

$$\mu_{zx} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}$$

2. Σ : 각각 $\Sigma_{zz}, \Sigma_{zx}, \Sigma_{xx}$를 구해보자

- $\Sigma_{zz} = E[(z - E[z])(z - E[z])^T] = I$ ($\because z \sim \mathcal{N}(0, I)$)

- $\Sigma_{zx} = E[(z-E[z])(x-E[x])^T]$

$$\begin{aligned} E[(z-E[z])(x-E[x])^T] &= E[z(\mu+\Lambda z+\epsilon-\mu)^T] \\ &= E[zz^T]\Lambda^T + E[z\epsilon^T] \\ &= \Lambda^T \end{aligned}$$

- $\Sigma_{xx} = E[(x-E[x])(x-E[x])^T]$

$$\begin{aligned} E[(x-E[x])(x-E[x])^T] &= E[(\mu+\Lambda z+\epsilon-\mu)(\mu+\Lambda z+\epsilon-\mu)^T] \\ &= E[\Lambda zz^T\Lambda^T + \epsilon z^T\Lambda^T + \Lambda z\epsilon^T + \epsilon\epsilon^T] \\ &= \Lambda E[zz^T]\Lambda + E[\epsilon\epsilon^T] \\ &= \Lambda\Lambda^T + \Psi \end{aligned}$$

3. 이제 다시 합쳐보자!

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix})$$

이를 통해, 우리는 x의 marginal 분포가 $\sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$를 따른다는 것을 알 수 있다.

주어진 파라미터로 log-likelihood를 적으면 다음과 같다.

$$\ell(\mu, \Lambda, \Psi) = log\prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\Lambda\Lambda^T + \Psi|^{1/2}} exp(-\frac{1}{2}(x^{(i)}-\mu)^T(\Lambda\Lambda^T+\Psi)^{-1}(x^{(i)}-\mu))$$