CS182: Introduction to Machine Learning –Expectation-Maximization (EM) algorithm and Gaussian Mixture Models (GMM)

Yujiao Shi
SIST, ShanghaiTech

Spring, 2025

Parameter Estimation with Latent Variables

- Consider a generative model with joint distr. $p(\mathbf{X}, \mathbf{Z}|\Theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n)$
  - Observed data: $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$
  - Latent variables: $\mathbf{Z} = \{\boldsymbol{z}_n\}_{n=1}^{N}$. All the model parameters: Θ

Parameter Estimation with Latent Variables

上海科技大学 ShanghaiTech University

- Consider a generative model with joint distr. $p(\mathbf{X}, \mathbf{Z}|\Theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n)$
  - Observed data: $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$
  - Latent variables: $\mathbf{Z} = \{\boldsymbol{z}_n\}_{n=1}^{N}$. All the model parameters: Θ
- Goal: Estimate the model parameters Θ via MLE (or MAP)

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{(when } \mathbf{Z} \text{ is discrete)}$$

$$= \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z} \quad \text{(when } \mathbf{Z} \text{ is continuous)}$$

Parameter Estimation with Latent Variables

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta)$$

Parameter Estimation with Latent Variables

上海科技大学

iversity

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)}$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad (\text{where } q(\mathbf{Z}) \text{ is some dist.}) \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})}
\end{aligned}
$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i))$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta}$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \textstyle\sum \lambda_i f(x_i)) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i))$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality

Parameter Estimation with Latent Variables

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i))$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})}$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i))$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, the above inequality becomes equality

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}}$$

Parameter Estimation with Latent Variables

- Consider the 'incomplete" data log likelihood

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.}
\end{aligned}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, the above inequality becomes equality

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log p(\mathbf{X}|\Theta)$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality

$$\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta)
\end{aligned}$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\begin{aligned}\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)} \\ &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)) \\ \log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}\end{aligned}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality

$$\begin{aligned}\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\ &= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)\end{aligned}$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad (\text{where } q(\mathbf{Z}) \text{ is some dist.})$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad (\text{concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i))$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, the above inequality becomes equality

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log p(\mathbf{X}|\Theta)$$

$$= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) = \log p(\mathbf{X}|\Theta)$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, we have

Parameter Estimation with Latent Variables

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality

$$
\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}
$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, we have

$$
\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
$$

Parameter Estimation with Latent Variables

上海科技大学

- Consider the ‘incomplete” data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, the above inequality becomes equality

$$
\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}
$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, we have

$$
\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.} = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] + \text{const.}
$$

Parameter Estimation with Latent Variables

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some dist.)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i))$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, the above inequality becomes equality

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X}, \Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log p(\mathbf{X}|\Theta)$$

$$= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) = \log p(\mathbf{X}|\Theta)$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \Theta)$, we have

$$\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta) \log p(\mathbf{X}, \mathbf{Z}|\Theta) + \text{const.} = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] + \text{const.}$$

- Thus $\log p(\mathbf{X}|\Theta)$ is tightly lower-bounded by $\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$ which EM maximizes

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**

The Expectation Maximization (EM) Algorithm

上海科技大学

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}

The Expectation Maximization (EM) Algorithm

上海科技大学

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

- **M (Maximization) step:**

The Expectation Maximization (EM) Algorithm

上海科技大学

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

$$\begin{aligned}\Theta^{new} &= \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{old}) \quad \text{(if doing MLE)} \\ \Theta^{new} &= \arg\max_{\Theta} \{\mathcal{Q}(\Theta, \Theta^{old}) + \log p(\Theta)\} \quad \text{(if doing MAP)}\end{aligned}$$

The Expectation Maximization (EM) Algorithm

上海科技大学

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X},\Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior

$$\mathcal{Q}(\Theta,\Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \log p(\mathbf{X},\mathbf{Z}|\Theta)$$

- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

$$\begin{aligned} \Theta^{new} &= \arg\max_{\Theta} \mathcal{Q}(\Theta,\Theta^{old}) \quad \text{(if doing MLE)} \\ \Theta^{new} &= \arg\max_{\Theta} \{\mathcal{Q}(\Theta,\Theta^{old}) + \log p(\Theta)\} \quad \text{(if doing MAP)} \end{aligned}$$

- If the log-likelihood or the parameter values not converged then set $\Theta^{old} = \Theta^{new}$ and go to the E step.

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

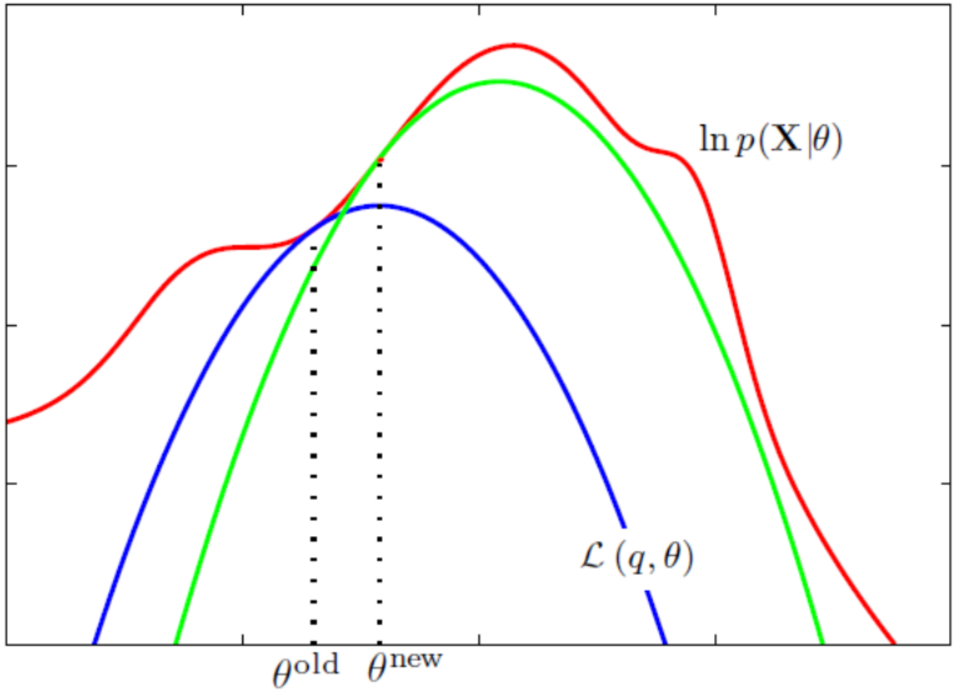
$$\begin{aligned} \Theta^{new} &= \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{old}) \quad \text{(if doing MLE)} \\ \Theta^{new} &= \arg\max_{\Theta} \{\mathcal{Q}(\Theta, \Theta^{old}) + \log p(\Theta)\} \quad \text{(if doing MAP)} \end{aligned}$$

- If the log-likelihood or the parameter values not converged then set $\Theta^{old} = \Theta^{new}$ and go to the E step.

The algorithm converges to a local maxima of $p(\mathbf{X}|\Theta)$ (as we saw)

EM: A View in the Parameter Space

- E-step: Update of q makes the $\mathcal{L}(q, \Theta)$ curve touch the $\log p(\mathbf{X}|\Theta)$ curve
- M-step gives the maxima Θ^{new} of $\mathcal{L}(q, \Theta)$
- Next E-step readjusts $\mathcal{L}(q, \Theta)$ curve (green) to meet $\log p(\mathbf{X}|\Theta)$ curve again
- This continues until a local maxima of $\log p(\mathbf{X}|\Theta)$ is reached



EM: Some Comments

- A general framework for parameter estimation in latent variable models
- Very widely used in problems with "missing data", e.g., missing features, or missing labels (semi-supervised learning)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated using EM
- More advanced probabilistic inference algorithms are based on similar ideas
  - E.g., variational Bayesian inference
- Very easy to extend to online learning setting and gives SGD like algorithms (will post a reading on "Online EM" on the class webpage)
- Note: The E and M steps may not always be possible to perform exactly (approximate inference methods may be needed in such cases)

Recap: GMM

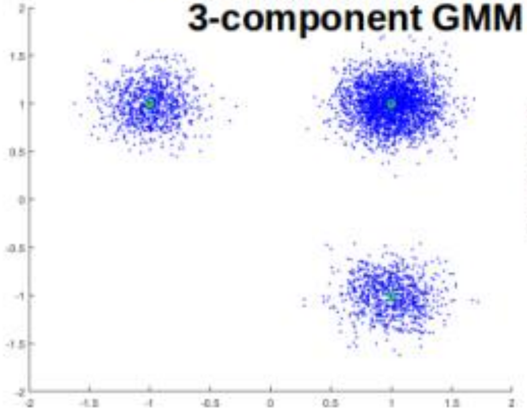
- The generative story for each $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$
  - First choose one of the K mixture components as

$$\boldsymbol{z}_n \sim \text{Multinomial}(\boldsymbol{z}_n|\pi) \qquad \text{(from the prior } p(\boldsymbol{z}) \text{ over } \boldsymbol{z}\text{)}$$

  - Suppose $\boldsymbol{z}_n = k$. Now generate $\boldsymbol{x}_n$ from the k-th Gaussian as

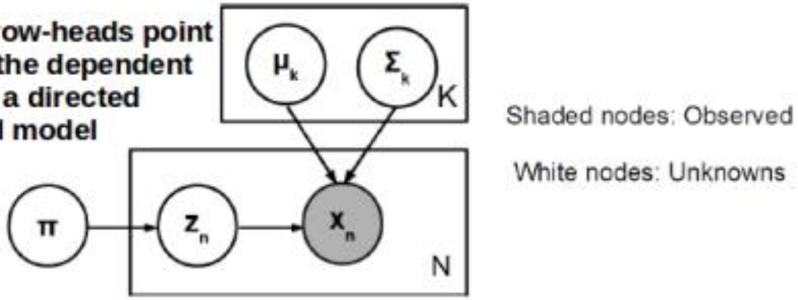
$$\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \text{(from the data distr. } p(\boldsymbol{x}|\boldsymbol{z})\text{)}$$

Some simulated data from a 3-component GMM



Directed Graphical Model for a K-component GMM

Note: Arrow-heads point towards the dependent nodes in a directed graphical model



Recap: Learning GMM

上海科技大学 ShanghaiTech University

We derive the Expectation-Maximization (EM) algorithm for GMM with K components:

- Observed data $X = \{x_1, \ldots, x_N\}$
- Latent variables $Z = \{z_1, \ldots, z_N\}$, where $z_i \in \{1, \ldots, K\}$.
- The parameters are $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$.

1. Complete-Data Likelihood

The joint distribution of observed data and latent data is:

$$p(X, Z|\Theta) = \prod_{i=1}^{N} p(x_i, z_i)\, p(z_i|\Theta) = \prod_{i=1}^{N} \pi_{z_i} \mathcal{N}\left(x_i \middle| \mu_{z_i}, \Sigma_{z_i}\right)$$

The complete-data log-likelihood is:

$$\log p(X, Z|\Theta) = \sum_{i=1}^{N} \log \pi_{z_i} + \log \mathcal{N}\left(x_i \middle| \mu_{z_i}, \Sigma_{z_i}\right)$$

Learning GMM

2. E-Step

- Compute the posterior responsibility $\gamma_{ik} = p\left(z_i = k \middle| x_i, \Theta^{old}\right)$:

$$\gamma_{ik} = p\left(z_i = k \middle| x_i, \Theta^{old}\right)$$

$$= \frac{p\left(x_i \middle| z_i = k, \Theta^{\text{old}}\right) p\left(z_i = k \middle| \Theta^{\text{old}}\right)}{p\left(x_i \middle| \Theta^{\text{old}}\right)}$$

$$= \frac{\pi_k^{old} \mathcal{N}\left(x_i \middle| \mu_k^{old}, \Sigma_k^{old}\right)}{\sum_{j=1}^{K} \pi_j^{old} \mathcal{N}\left(x_i \middle| \mu_j^{old}, \Sigma_j^{old}\right)}$$

- This is a soft assignment of x_i to cluster k.

Learning GMM

上海科技大学 ShanghaiTech University

3. **M-step: Maximize $Q(\Theta, \Theta^{old})$**
   - The Q-function is the expected complete-data log-likeligood

$$Q(\Theta, \Theta^{old}) = \mathrm{E}_{Z|X,\Theta^{old}}[\log p(X, Z|\Theta)] = \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}[\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)]$$

   - Update Mixing coefficients π_k:
     - Maximize Q w.r.t. π_k under the constraint $\sum_k \pi_k = 1$:

$$\mathcal{L} = Q(\Theta, \Theta^{old}) + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{i=1}^{N}\frac{\gamma_{ik}}{\pi_k} - \lambda = 0 \quad \Rightarrow \quad \pi_k \propto \sum_{i=1}^{N}\gamma_{ik}$$

Learning GMM

上海科技大学 ShanghaiTech University

3. **M-step: Maximize $\boldsymbol{Q(\Theta, \Theta^{old})}$**

$$Q(\Theta, \Theta^{old}) = \mathrm{E}_{Z|X,\Theta^{old}}[\log p(X, Z|\Theta)] = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}[\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)]$$

- Update Mixing coefficients π_k:

$$\mathcal{L} = Q(\Theta, \Theta^{old}) + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{i=1}^{N} \frac{\gamma_{ik}}{\pi_k} - \lambda = 0 \quad \Rightarrow \quad \pi_k \propto \sum_{i=1}^{N} \gamma_{ik}$$

  - Enforce constraint $\sum_k \pi_k = 1$:

$$\lambda = \sum_{k=1}^{K}\sum_{i=1}^{N} \gamma_{ik} = N \quad \Rightarrow \quad \pi_k^{new} = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ik}$$

Learning GMM

3. **M-step: Maximize $Q(\Theta, \Theta^{old})$**

$$Q(\Theta, \Theta^{old}) = \mathrm{E}_{Z|X,\Theta^{old}}[\log p(X, Z|\Theta)] = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}[\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)]$$

- The probability density function for a D -dimensional Gaussian:

$$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

- Update Means μ_k:

$$\frac{\partial Q}{\partial \mu_k} = \sum_{i=1}^{N} \gamma_{ik} \Sigma_k^{-1}(x_i - \mu_k) = 0 \quad \Longrightarrow \mu_k^{new} = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}$$

Learning GMM

3. M-step: Maximize $Q(\Theta, \Theta^{old})$

$$Q(\Theta, \Theta^{old}) = \mathrm{E}_{Z|X,\Theta^{old}}[\log p(X, Z|\Theta)] = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}[\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)]$$

- The probability density function for a D-dimensional Gaussian:

$$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

- Update Covariances Σ_k: $\boxed{\frac{\partial \log|\Sigma_k|}{\partial \Sigma_k^{-1}} = -\Sigma_k \qquad \frac{\partial \mathrm{tr}(A\Sigma_k^{-1})}{\partial \Sigma_k^{-1}} = A^T}$

$$\frac{\partial Q}{\partial \Sigma_k^{-1}} = \frac{1}{2}\sum_{i=1}^{N} \gamma_{ik}[\Sigma_k - (x_i - \mu_k)(x_i - \mu_k)^{\mathsf{T}}] = 0$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k^{new})(x_i - \mu_k^{new})^{\mathsf{T}}}{\sum_{i=1}^{N} \gamma_{ik}}$$

Learning GMM: Summary

上海科技大学 ShanghaiTech University

1. Initialize the Parameters $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ randomly, or using K-means

2. Iterate until convergence (e.g., when $\log p(X|\Theta)$ ceases to increase

   a) E-step:

$$\gamma_{ik} = p\left(z_i = k \middle| x_i, \Theta^{old}\right) = \frac{\pi_k^{old}\mathcal{N}\left(x_i \middle| \mu_k^{old}, \Sigma_k^{old}\right)}{\sum_{j=1}^{K}\pi_j^{old}\mathcal{N}\left(x_i \middle| \mu_j^{old}, \Sigma_j^{old}\right)}$$

   b) M-step:

$$\pi_k^{new} = \frac{1}{N}\sum_{i=1}^{N}\gamma_{ik}$$

$$\mu_k^{new} = \frac{\sum_{i=1}^{N}\gamma_{ik}x_i}{\sum_{i=1}^{N}\gamma_{ik}}$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N}\gamma_{ik}(x_i - \mu_k^{new})(x_i - \mu_k^{new})^{\top}}{\sum_{i=1}^{N}\gamma_{ik}}$$