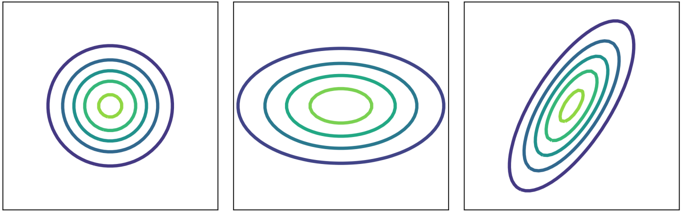
Gaussian Mixture Models

Horst Petschenig, Institute of Theoretical Computer Science

The Gaussian distribution re-visited

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

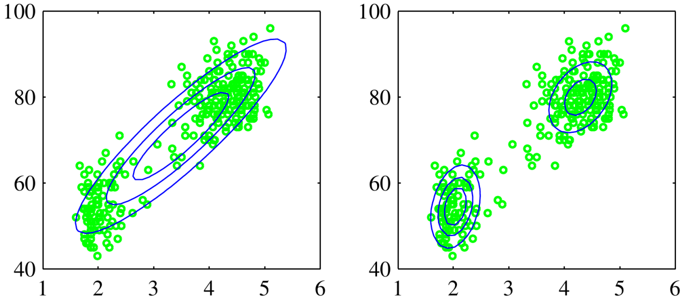
$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}$$

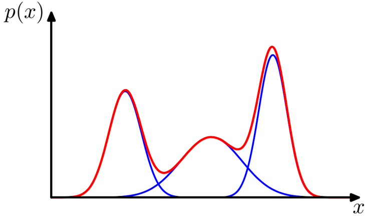
“Downside” of Gaussians

- How to model real-world data?



- What if we could have more than one Gaussian?

Gaussian mixtures



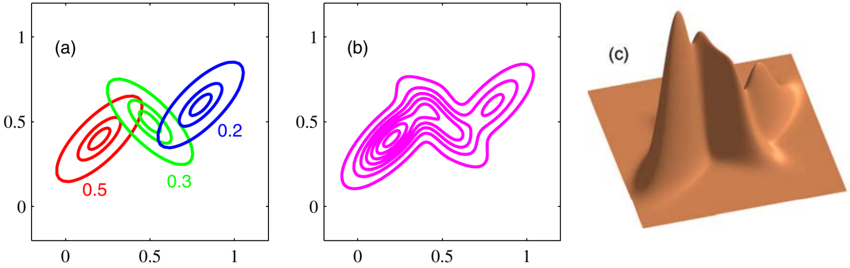
- **Idea:** Create mixture distributions
- Superposition of K Gaussians

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

- Mixing coefficients $0 \leq \pi_k \leq 1$

$$\sum_{k=1}^{K} \pi_k = 1$$

Example: Mixture of 3 Gaussians



Properties of GMMs

- Marginal density

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x}|k)$$

- responsibilities (posterior prob.) $\gamma_k(\boldsymbol{x}) = p(k|\boldsymbol{x})$

$$\gamma_k(\boldsymbol{x}) = \frac{p(k)p(\boldsymbol{x}|k)}{\sum_m p(m)p(\boldsymbol{x}|m)}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_m \pi_m \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m, \Sigma_m)}$$

GMM Maximum Likelihood Estimation

- Recall the likelihood function

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

- How to find parameters $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ and $\boldsymbol{\Sigma} = \{\Sigma_1, \ldots, \Sigma_K\}$?

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

Latent variable models

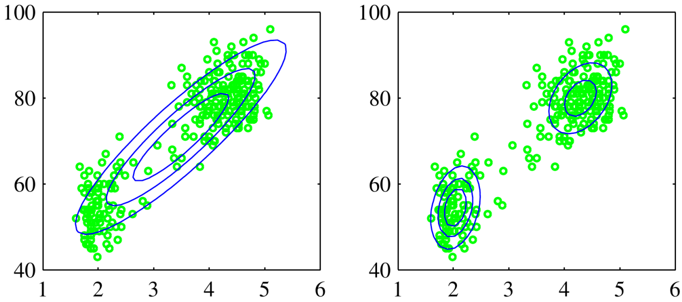
- Some models have latent (hidden) variables never observed in training data
- Assume that observed variables arise from a hidden common "cause"

+ fewer parameters
+ compute a compressed representation of the data (bottleneck)

− harder to fit

Latent variable: cluster label

- How to model real-world data?



EM for Gaussian mixtures

Recall the log-likelihood function

$$\mathcal{L} = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{1}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m) \right)}{\partial \boldsymbol{\mu}_m}$$

Trick: $\partial \ln x = \frac{1}{x} \Rightarrow 1 = x \partial \ln x$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\ln(\pi_k) + \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right) \right)}{\partial \boldsymbol{\mu}_k}$$

EM for Gaussian mixtures

Recall the log-likelihood function

$$\mathcal{L} = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{1}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m) \right)}{\partial \boldsymbol{\mu}_m}$$

Trick: $\partial \ln x = \frac{1}{x} \Rightarrow 1 = x \partial \ln x$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\ln(\pi_k) + \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right) \right)}{\partial \boldsymbol{\mu}_k}$$

EM for Gaussian mixtures

Recall the log-likelihood function

$$\mathcal{L} = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{1}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m) \right)}{\partial \boldsymbol{\mu}_m}$$

Trick: $\partial \ln x = \frac{1}{x} \Rightarrow 1 = x \partial \ln x$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\ln(\pi_k) + \ln\left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\right)\right)}{\partial \boldsymbol{\mu}_k}$$

EM for Gaussian mixtures

Recall the log-likelihood function

$$\mathcal{L} = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{1}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m) \right)}{\partial \boldsymbol{\mu}_m}$$

Trick: $\partial \ln x = \frac{1}{x} \Rightarrow 1 = x \partial \ln x$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)} \frac{\partial \left(\ln(\pi_k) + \ln \left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right) \right)}{\partial \boldsymbol{\mu}_k}$$

Derivation for mean μ_k

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \underbrace{\boxed{\frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{m=1}^{K} \pi_m \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_m, \Sigma_m)}}}_{\text{posterior prob. } \gamma_k(\boldsymbol{x})} \frac{\partial \left(\ln(\pi_k) + \ln\left(\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\right)\right)}{\partial \boldsymbol{\mu}_k}$$

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

$$\frac{\partial \ln\left(\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)\right)}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\left(\Sigma^{-1} + \left(\Sigma^{-1}\right)^T\right)(\boldsymbol{x} - \boldsymbol{\mu}) = -\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = -\sum_{n=1}^{N} \gamma_k(\boldsymbol{x}) \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Derivation for mean $\boldsymbol{\mu}_k$

$$\frac{\mathcal{L}}{\partial \boldsymbol{\mu}_k} = -\sum_{n=1}^{N} \gamma_k(\boldsymbol{x})\Sigma_k^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)$$

$$0 \stackrel{!}{=} -\sum_{n=1}^{N} \gamma_k(\boldsymbol{x})\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)$$

$$\boldsymbol{\mu}_k \sum_{n=1}^{N} \gamma_k(\boldsymbol{x}) = \sum_{n=1}^{N} \gamma_k(\boldsymbol{x})\boldsymbol{x}_n$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_k(\boldsymbol{x})\boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_k(\boldsymbol{x})}$$