Day 19 - 15 November - Mixtures of Gaussian and EM Algorithms

Agenda:

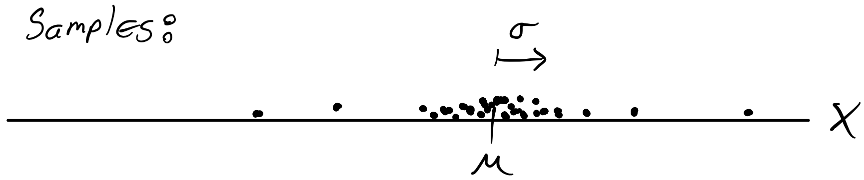
- Multivariate Gaussians
- Maximum Likelihood with Multivariate Gaussians
- Mixtures of Gaussians
- Expectation Maximization (EM) Algorithms

Multivariate Gaussians

A Gaussian in $\mathbb{R}$ follows the pdf

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

Samples:



Here $X \sim N(\mu, \sigma^2)$

$$\mathbb{E}(X) = \mu$$

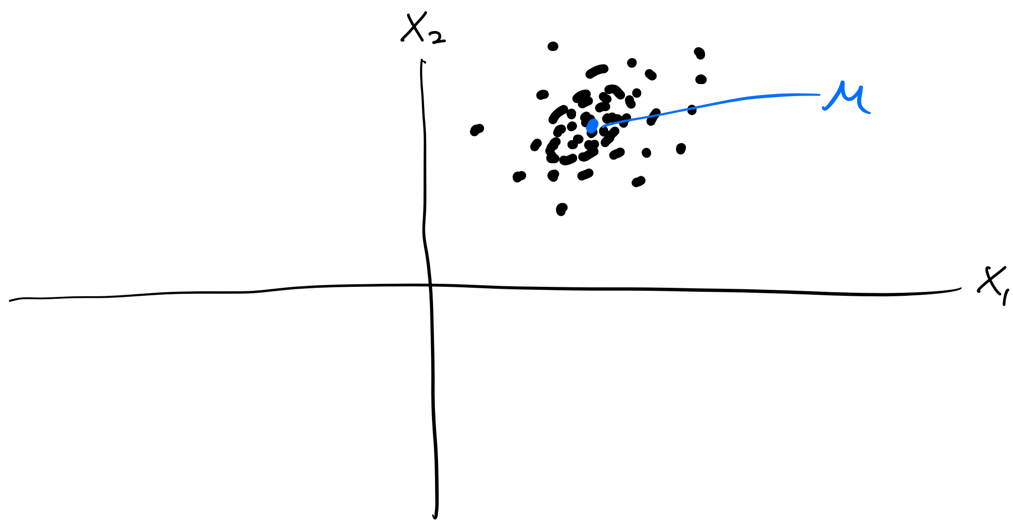
$$\mathbb{E}((X-\mu)^2) = \sigma^2$$

A multivariate Gaussian in $\mathbb{R}^d$ follows the pdf

$$f(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)\right)$$

$\mu \in \mathbb{R}^d$ $\quad$ $\Sigma \in \mathbb{R}^{d \times d}$

symmetric
positive definite



$X \sim \mathcal{N}(\mu, \Sigma)$

$\mathbb{E}(X) = \mu$ — mean

$\mathbb{E}((X-\mu)(X-\mu)^t) = \Sigma$ — covariance matrix

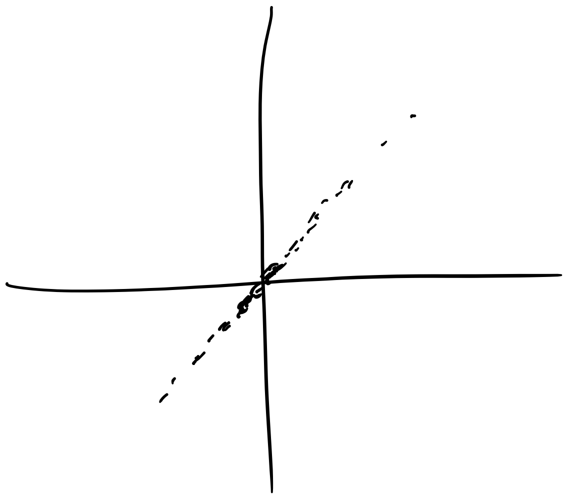
Note: Σ is positive semidefinite

why? $z^t \Sigma z = z^t \mathbb{E}(X-\mu)(X-\mu)^t z$

$= \mathbb{E}[z^t (X-\mu)(X-\mu)^t z]$

$= \mathbb{E}[((X-\mu)^t z)^2] \geq 0.$

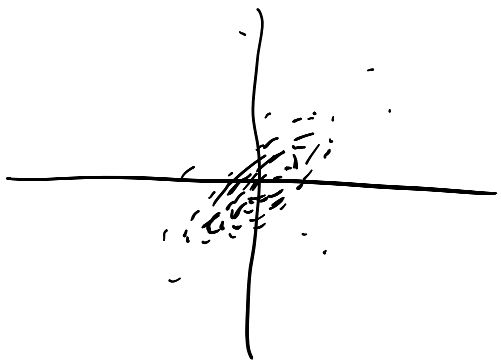
Eigenvectors of Σ with large eigenvalues provide directions w/ most variability



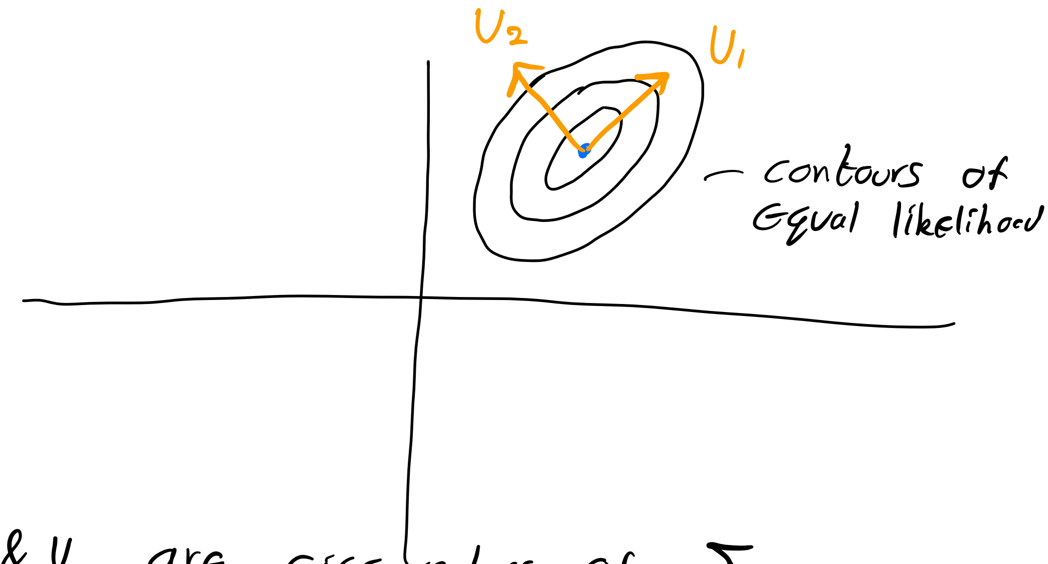
$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$\mathcal{N}(0, \Sigma)$$

$$\Sigma = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \cdots \end{pmatrix}^t$$



$$\Sigma = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cdots \end{pmatrix}^t$$



U_1 & U_2 are eigenvectors of Σ
w/ $\lambda_1 > \lambda_2$

Maximum Likelihood Estimation for Gaussians

Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$

Estimate μ & Σ.

How? Maximum Likelihood estimation

Likelihood of data $\underline{X}$:

$$P(\underline{X} \mid \mu, \Sigma) = \prod_{i=1}^{n} P(x_i \mid \mu, \Sigma)$$

$$\log P(\underline{X} \mid \mu, \Sigma) = \sum_{i=1}^{n} \log P(x_i \mid \mu, \Sigma)$$

$$= -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^t \Sigma^{-1} (X_i - \mu)$$

$$+ \text{ other terms}$$

MLE estimate of μ: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$$\nabla_\mu \log P(\underline{X} \mid \mu, \Sigma) = \sum_{i=1}^{n} \Sigma^{-1} (X_i - \mu) = 0$$

$$\Rightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} X_i$$

MLE Estimate of Σ:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^t$$

Issue: $\mathbb{E}[\hat{\Sigma}] = \frac{N-1}{N} \Sigma$ (estimator is biased)

Resolution:

$$\tilde{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^t$$

Mixtures of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \overbrace{\mathcal{N}(x \mid \mu_k, \Sigma_k)}^{\text{pdf eval. at } x}$$

where $\pi_k \geq 0$ & $\sum_{k=1}^{K} \pi_k = 1$.

To generate a sample:

Let $Z \sim \pi \qquad (P(Z=k) = \pi_k)$

$$X \sim \mathcal{N}(x \mid \mu_Z, \Sigma_Z)$$

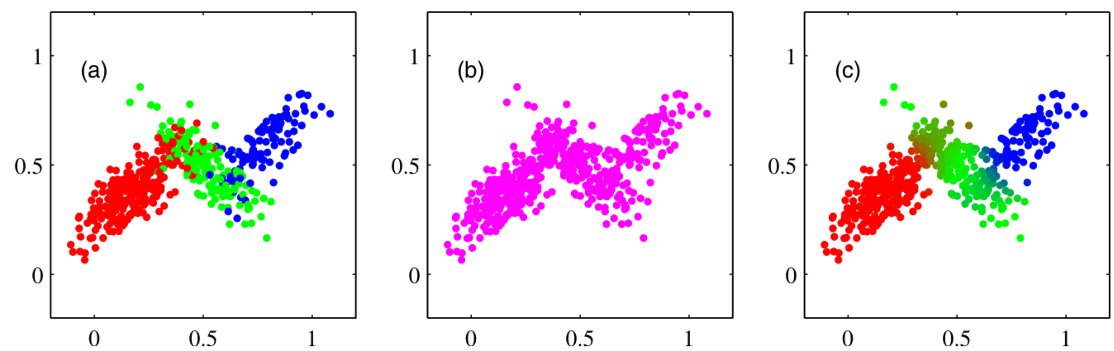


Figure 9.5 Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. (a) Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the three states of $\mathbf{z}$, corresponding to the three components of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution $p(\mathbf{x})$, which is obtained by simply ignoring the values of $\mathbf{z}$ and just plotting the $\mathbf{x}$ values. The data set in (a) is said to be *complete*, whereas that in (b) is *incomplete*. (c) The same samples in which the colours represent the value of the responsibilities $\gamma(z_{nk})$ associated with data point $\mathbf{x}_n$, obtained by plotting the corresponding point using proportions of red, blue, and green ink given by $\gamma(z_{nk})$ for $k = 1, 2, 3$, respectively

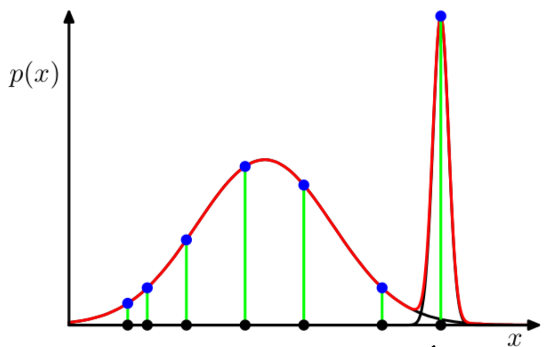
Goal: Use Maximum likelihood estimation to estimate the Gaussian mixture π, μ, Σ underlying a dataset $\{X_i\}_{i=1 \cdots n}$

Likelihood of data

$$\log p(X \mid \pi, \mu, \Sigma) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(X_i \mid \mu_k, \Sigma_k)$$

Issue: a singularity could arise

Figure 9.7 Illustration of how singularities in the likelihood function arise with mixtures of Gaussians. This should be compared with the case of a single Gaussian shown in Figure 1.14 for which no singularities arise.



can put arbitrarily narrow Gaussian around a single data point.

Expectation-Maximization for Gaussian Mixtures

What conditions should be satisfied at an optimum

$$\log p(X \mid \pi, \mu, \Sigma) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(X_i \mid \mu_k, \Sigma_k)$$

Set $\nabla_{\mu_k} \cdot = 0$,

$$0 = \nabla_{\mu_k} \log p(X \mid \pi, \mu, \Sigma) = \sum_{i=1}^{n} \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)} \Sigma_k^{-1}(x_i - \mu_k)$$

$$\Longrightarrow \quad \mu_k = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) x_i}{\sum_{i=1}^{n} \gamma(z_{ik})} \qquad \text{w/} \quad \gamma(z_{ik}) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)$$

Similarly, $\nabla_{\Sigma_k} \cdot = 0 \Longrightarrow \Sigma_k = \dfrac{\sum_{i=1}^{n} \gamma(z_{ik})(x_i - \mu_k)(x_i - \mu_k)^t}{\sum_{i=1}^{n} \gamma(z_{ik})}$

Finally, $\nabla_{\pi_k} \cdot = 0 \Longrightarrow \pi_k = \dfrac{\sum_{i=1}^{n} \gamma(z_{ik})}{n}$

Gives rise to EM algorithm:

1) Initialize μ_k, Σ_k, π_k

2) E step

update $\gamma(z_{ik}) = \dfrac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$

3) M step

update $\mu_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma(z_{ik}) x_i$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma(z_{ik})(x_i - \mu_k)(x_i - \mu_k)^t$$

$$\pi_k = N_k / n$$

w/ $N_k = \sum_{i=1}^{n} \gamma(z_{ik})$

4) Repeat 2&3 until stopping condition

Visualization

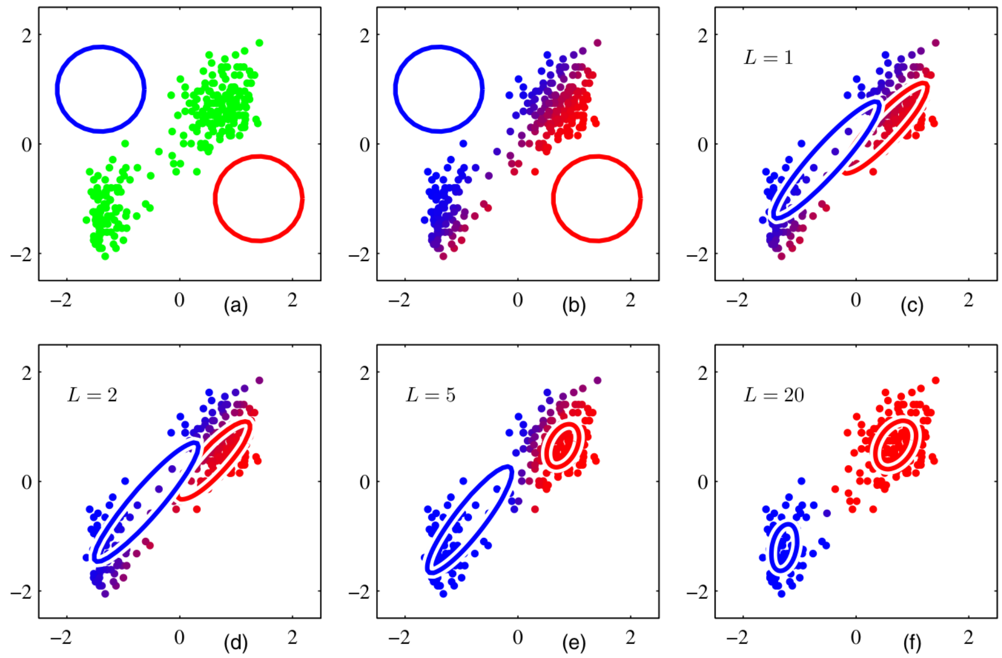
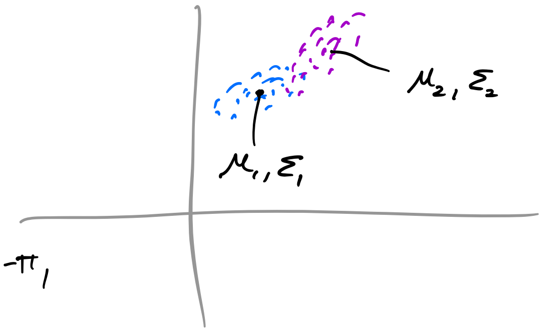


Figure 9.8 Illustration of the EM algorithm using the Old Faithful set as used for the illustration of the K-means algorithm in Figure 9.1. See the text for details.

EM and Gaussian Mixtures more abstractly

Data $\{X_i\}_{i=1\dots n}$



Model w/ 2 Gaussians

$$Z_i = \begin{cases} 1 & \text{w/ prob } \pi_1 \\ 2 & \text{w/ prob } \pi_2 = 1-\pi_1 \end{cases}$$

$$X_i \sim \mathcal{N}(X_i \mid \mu_{Z_i}, \Sigma_{Z_i})$$

Given $\{X_i\}$ estimate $\{\pi, \mu_1, \mu_2, \Sigma_1, \Sigma_2\} = \theta$

Likelihood of data

$$L(\theta; X, Z) = P(X, Z \mid \theta)$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{2} \Big(\underbrace{\mathcal{N}(X_i \mid \mu_j, \Sigma_j)}_{P(X \mid Z, \theta)} \underbrace{\pi_j}_{P(Z \mid \theta)} \Big)^{\mathbb{1}_{Z_i = j}}$$

Issue: dont know Z, so keep distribution over all values it could take and update that distribution

E Estimate dist over Z given $\theta = \hat{\theta}$

Compute $p(z_i = k \mid x_i, \hat{\theta})$

$$p(z_i = k \mid x_i, \hat{\theta}) = \frac{P(x_i \mid z_i = k, \hat{\theta})\, P(z_i = k \mid \hat{\theta})}{P(x_i \mid \hat{\theta})}$$

$$= \frac{\mathcal{N}(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)\, \hat{\pi}_k}{\sum_{j=1}^{2} \mathcal{N}(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)\, \hat{\pi}_j}$$

We can rewrite the likelihood function

$$Q(\theta, \hat{\theta}) = \mathbb{E}_{Z \mid X, \hat{\theta}} \log L(\theta \mid X, Z)$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z_i \mid x_i, \hat{\theta}} \log L(\theta \mid x_i, z_i)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \hat{\theta}) \log L(\theta \mid x_i, z_i)$$

M $\hat{\theta} \leftarrow \underset{\theta}{\arg\max}\ Q(\theta, \hat{\theta})$

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.

2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

   where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

   and return to step 2.