

Mixture Models

Václav Šmíd

April 18, 2022

Mixture of Gaussians

2D data of 1000 points:

blue: 100 points

red: 200 points

green: 700 points

Each of them Gaussian distributed (component).

$$p_1(x) = \mathcal{N}([1; 1], 0.1^2 I),$$

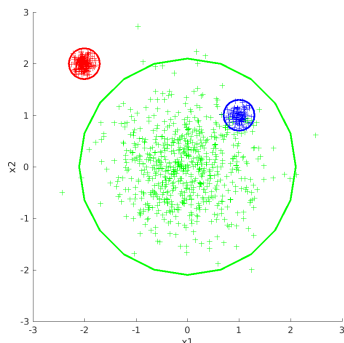
$$p_2(x) = \mathcal{N}([-2; 2], 0.1^2 I),$$

$$p_3(x) = \mathcal{N}([0; 0], 0.7^2 I),$$

Distribution of all data points:

$$p(x) = \frac{1}{10} p_1(x) + \frac{2}{10} p_2(x) + \frac{7}{10} p_3(x)$$

with weights [0.1, 0.2, 0.7].

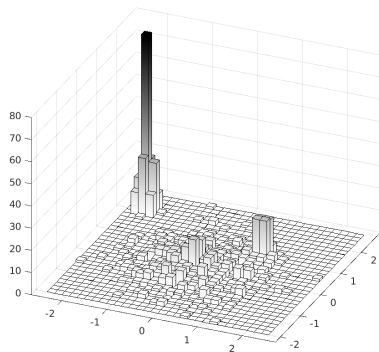
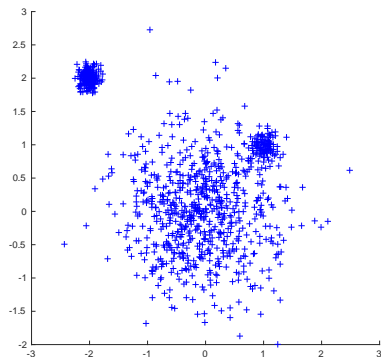


Mixture of Gaussians:

$$p(x) = \sum_{k=1}^3 \alpha_k \mathcal{N}(\mu_k, \Sigma_k)$$

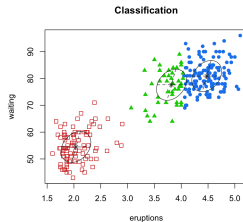
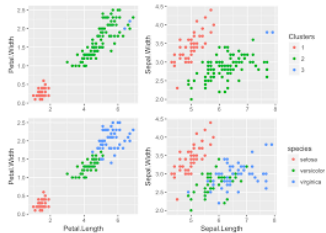
Estimation of mixture of Gaussians

Consider unsupervised scenario, where we do not know the assignment of each data point:

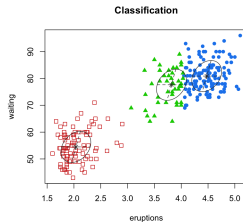
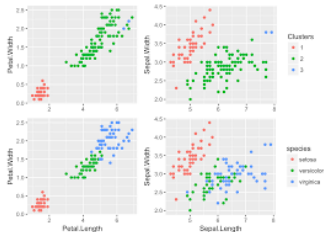


Task: find $\alpha_1, \dots, \alpha_K$, μ_1, \dots, μ_K and $\Sigma_1, \dots, \Sigma_K$ and possibly K .

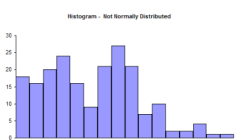
Clustering we care about assignments:



Clustering we care about assignments:



Density representation: we do not care (distribution of requests in network)



Maximum likelihood estimation

Probability distribution with parameters $\theta = \{\mu_k, \Sigma_k, \alpha_k\}_{k=1}^K$:

$$p(x_i|\theta) = \sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k), \quad \forall i = 1, \dots, n$$

$$p(x_1, \dots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta), \quad \log p(x_1, \dots, x_n|\theta) = \sum_{i=1}^n \log p(x_i|\theta)$$

Maximum likelihood estimation

Probability distribution with parameters $\theta = \{\mu_k, \Sigma_k, \alpha_k\}_{k=1}^K$:

$$p(x_i|\theta) = \sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k), \quad \forall i = 1, \dots, n$$

$$p(x_1, \dots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta), \quad \log p(x_1, \dots, x_n|\theta) = \sum_{i=1}^n \log p(x_i|\theta)$$

$$\log p(x_1, \dots, x_n|\theta) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right)$$

$$\mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left(-\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right),$$

Maximum likelihood estimation

Probability distribution with parameters $\theta = \{\mu_k, \Sigma_k, \alpha_k\}_{k=1}^K$:

$$p(x_i|\theta) = \sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k), \quad \forall i = 1, \dots, n$$

$$p(x_1, \dots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta), \quad \log p(x_1, \dots, x_n|\theta) = \sum_{i=1}^n \log p(x_i|\theta)$$

$$\log p(x_1, \dots, x_n|\theta) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right)$$

$$\mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left(-\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right),$$

Finding $\hat{\theta} = \arg \max_{\theta} \log p(x_1, \dots, x_n)$:

$$\frac{d}{d\mu_k} \log p(x_{1:n}) = \sum_{i=1}^n \frac{\alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i|\mu_k, \Sigma_k) \alpha_k} (\Sigma_k^{-1} (\mu_k - x_i)) \equiv 0,$$

and others.

Finding $\hat{\theta} = \arg \max_{\theta} \log p(x_1, \dots, x_n)$ **subject to** $\sum_k \alpha_k = 1$:

$$\frac{d}{d\mu_k} \log p(x_{1:n}) = \sum_{i=1}^n \frac{\alpha_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k} (\Sigma_k^{-1} (\mu_k - x_i)) \equiv 0,$$

$$\frac{d}{d\alpha_k} \log p(x_{1:n}) = \sum_{i=1}^n \frac{\mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k} + \lambda = 0,$$

where λ is a Lagrange multiplier for $\sum_k \alpha_k = 1$.

Maximum likelihood continued

Finding $\hat{\theta} = \arg \max_{\theta} \log p(x_1, \dots, x_n)$ **subject to** $\sum_k \alpha_k = 1$:

$$\frac{d}{d\mu_k} \log p(x_{1:n}) = \sum_{i=1}^n \frac{\alpha_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k} (\Sigma_k^{-1} (\mu_k - x_i)) \equiv 0,$$

$$\frac{d}{d\alpha_k} \log p(x_{1:n}) = \sum_{i=1}^n \frac{\mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k} + \lambda = 0,$$

where λ is a Lagrange multiplier for $\sum_k \alpha_k = 1$.

Conditions of extrema (solved by alternating evaluation of):

$$\hat{w}_{i,k} = \frac{\alpha_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k},$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_i w_{i,k} x_i, \quad n_k = \sum_i w_{i,k},$$

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_i w_{i,k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T, \quad \alpha_k = \frac{n_k}{n},$$

Mixture estimation via latent variable

Consider latent variable $l_i \in \{\epsilon_1, \dots, \epsilon_K\}$,
 $\epsilon_k = [0, 0, \dots, 1 \dots 0]$. (one-hot).

$$p(x_i, l_i) = p(x_i | l_i) p(l_i),$$

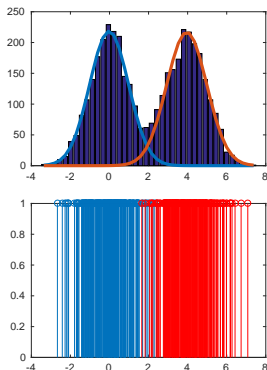
$$p(x_i | l_i) = \prod_k \mathcal{N}(\mu_k, \Sigma_k)^{l_{i,k}},$$

$$p(l_{i,k} = 1) = \alpha_k, \sum_k \alpha_k = 1$$

$$p(l_i) = \prod_{k=1}^K \alpha_k^{l_{i,k}}.$$

$$\begin{aligned} p(x_i) &= \sum_k p(x_i, l_i) \\ &= \sum_k p(x_i | l_i = \epsilon_k) p(l_i = \epsilon_k). \\ &= \sum_k \mathcal{N}(\mu_k, \Sigma_k) \alpha_k, \end{aligned}$$

Multinomial (Bernoulli) distribution $p(l_i)$.



- ▶ Each data point has a label from which component is generated.
- ▶ Estimation of the joint distribution $p(\theta, l_1 \dots l_n)$ is easier

Expectation maximization (EM) algorithm

Joint distribution:

$$p(x_i, l_i) = p(x_i | l_i) p(l_i), \quad \forall i$$

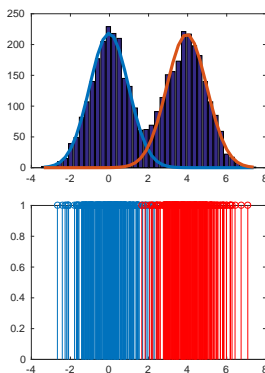
$$p(x_i | l_i) = \prod_k \mathcal{N}(x_i | \mu_k, \Sigma_k)^{l_{i,k}},$$

$$p(l_{i,k} = 1) = \alpha_k, \quad \sum_k \alpha_k = 1.$$

Conditional distribution

$$p(l_i = \epsilon_k | x_i) = \frac{p(x, l)}{p(x)} = \frac{\mathcal{N}(x | \mu_k, \Sigma_k) \alpha_k}{\sum_k \mathcal{N}(x | \mu_k, \Sigma_k) \alpha_k}$$

probability that i th data point belongs to k th component.



Maximum likelihood with latent variable

General EM algorithm [Dempster, Laird, Rubin, 1977]:

$$\hat{\theta} = \arg \max_{\theta} \int p(x|\theta, l)p(l)dl,$$

can be (approximately) found by alternating:

$$\text{E-step (over } l): q(\theta|\theta^{(j)}) = \int \log p(x, l|\theta)p(l|\theta^{(j)})dl$$

$$\text{M-step (of } \theta): \theta^{(j+1)} = \arg \max_{\theta} q(\theta|\theta^{(j)})$$

Maximizing log-likelihood

$$p(x_1, l_1, \dots, x_n, l_n|\theta) \propto \prod_{i=1}^n \prod_{k=1}^K \mathcal{N}(x_i|\mu_k, \Sigma_k)^{l_{i,k}} \alpha_k^{l_{i,k}}$$

$$\log p(x_1, l_1, \dots, x_n, l_n|\theta) \propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} (\log(\mathcal{N}(x_i|\mu_k, \Sigma_k)) + \log \alpha_k)$$

$$\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right)$$

Maximum likelihood with latent variable

Maximizing log-likelihood

$$\begin{aligned}\log p(x_1, l_1, \dots, x_n, l_n | \theta) &\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} (\log(\mathcal{N}(x_i | \mu_k, \Sigma_k)) + \log \alpha_k) \\ &\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right)\end{aligned}$$

E-step: $q(\theta | \theta^{(j)}) = \int \log p(\mathbf{x} | \theta, \mathbf{l}) p(\mathbf{l} | \theta^{(j)}) d\mathbf{l}$

$$\begin{aligned}q(\theta | \theta^{(j)}) &= \mathbb{E}_{\mathbf{l}} \left\{ \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T (\Sigma_k)^{-1} (x_i - \mu_k) + \log \alpha_k \right) \right\} \\ &= \sum_{i=1}^n \sum_{k=1}^K \hat{l}_{i,k}^{(j)} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T (\Sigma_k)^{-1} (x_i - \mu_k) + \log \alpha_k \right)\end{aligned}$$

$$\hat{l}_{i,k}^{(j)} = p(l_i = \epsilon_k | x_i, \theta^{(j)}) = \frac{\mathcal{N}(x_i | \mu_k^{(j)}, \Sigma_k^{(j)}) \alpha_k^{(j)}}{\sum_k \mathcal{N}(x_i | \mu_k^{(j)}, \Sigma_k^{(j)}) \alpha_k^{(j)}}$$

M-step:

$$\hat{\mu}_k^{(j+1)}, \hat{\Sigma}_k^{(j+1)} = \arg \min \left(-\frac{1}{2} n \log |\Sigma_k| - \frac{1}{2} \sum_{i=1}^n \hat{l}_{i,k} (x_i - \mu_k)^T (\Sigma_k)^{-1} (x_i - \mu_k) \right)$$

Expectation Maximization (EM) algorithm

[Dempster, Laird, Rubin, 1977]

Initialize: choose $\alpha_k^{(0)}, \mu_k^{(0)}, \Sigma_k^{(0)}, \forall k$

Iterate:

1. Compute expected labels:

$$p(l = \epsilon_k | x_i) = \hat{l}_{i,k} = \frac{\mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k}{\sum_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \alpha_k}$$

2. Recompute the component parameters

$$\hat{\mu}_k = \frac{1}{N_k} \sum_i \hat{l}_{i,k} x_i,$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_i \hat{l}_{i,k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T,$$

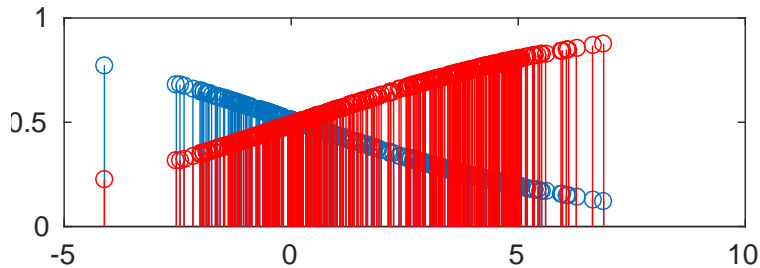
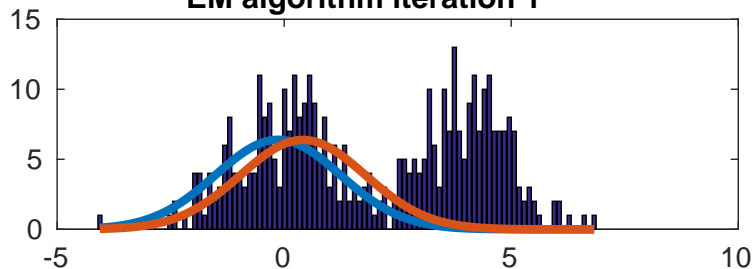
$$\hat{\alpha}_k = \frac{N_k}{N}, \quad N_k = \sum_i \hat{l}_{i,k}$$

3. (Evaluate log-likelihood)

$$\log p(x) = \sum_i \log \left(\sum_{k=1}^K \mathcal{N}(x_i | \hat{\mu}_k, \hat{\Sigma}_k) \hat{\alpha}_k \right)$$

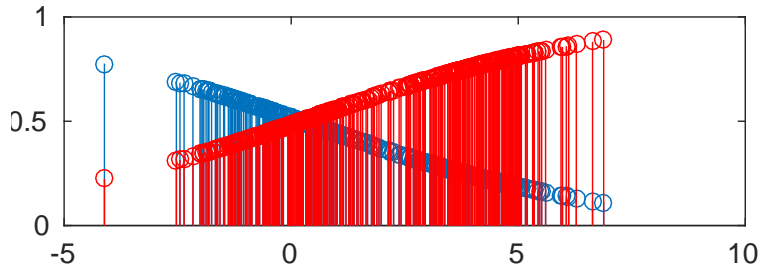
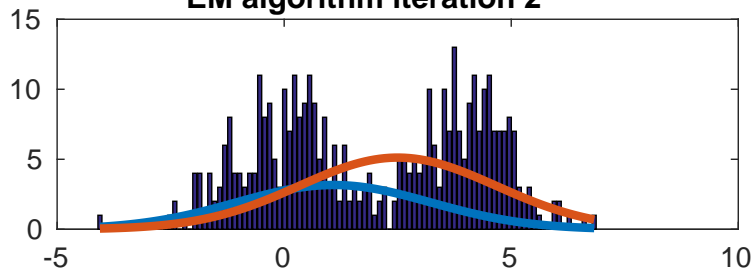
Expectation Maximization (EM) algorithm

EM algorithm iteration 1



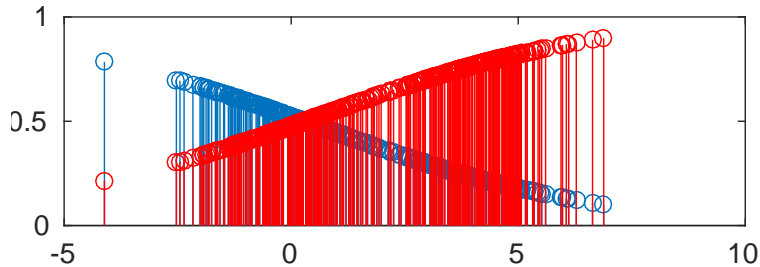
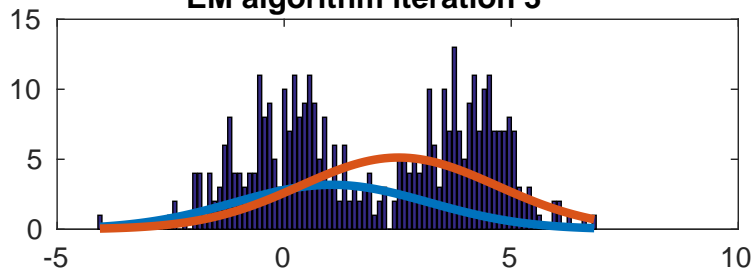
Expectation Maximization (EM) algorithm

EM algorithm iteration 2

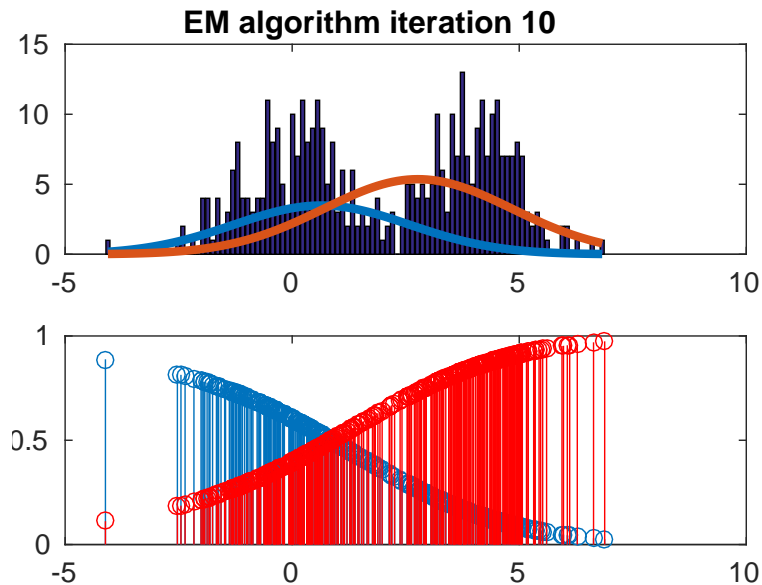


Expectation Maximization (EM) algorithm

EM algorithm iteration 3

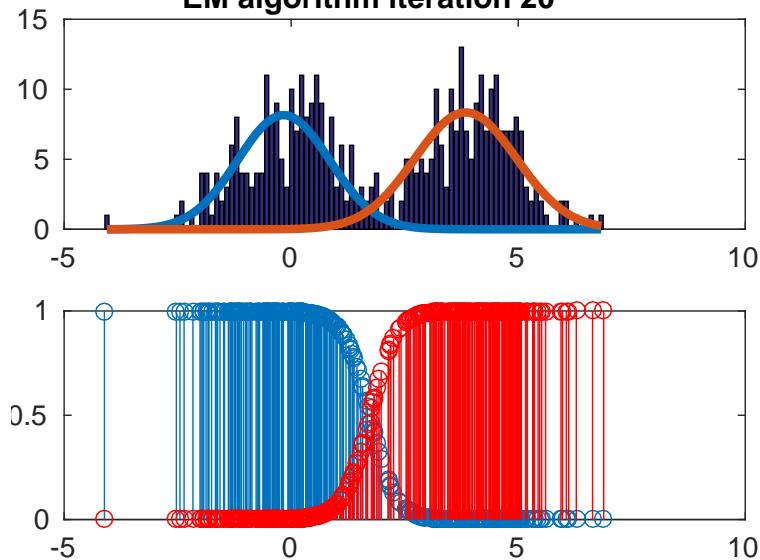


Expectation Maximization (EM) algorithm

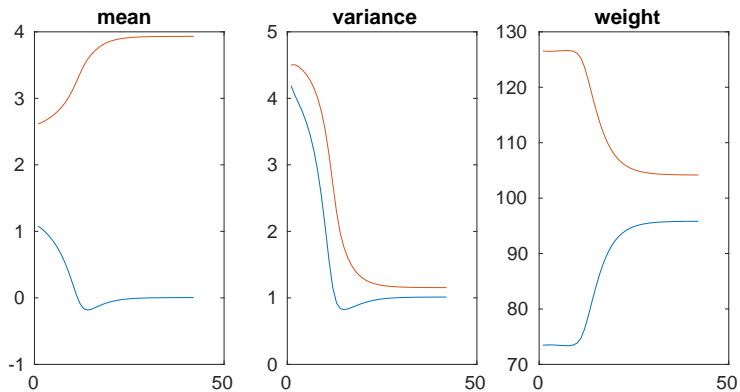


Expectation Maximization (EM) algorithm

EM algorithm iteration 20



Expectation Maximization (EM) algorithm



Bayesian treatment

Joint distribution:

$$p(x_i, l_i | \alpha) = p(x_i | l_i) p(l_i),$$

$$p(x_i | l_i) = \prod_k \mathcal{N}(\mu_k, \omega_k)^{l_{i,k}},$$

$$p(l_{i,k} = 1 | \alpha) = \alpha_k, \quad \sum_k \alpha_k = 1.$$

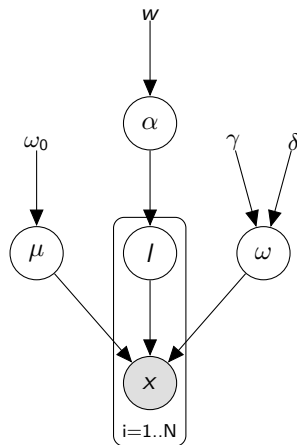
Priors for multivariate Normal (conjugate)

$$p(\mu_k) = \mathcal{N}(0, \infty),$$

$$p(\Sigma_k) = i\mathcal{W}(\nu_0, \Lambda_0) \propto |\Sigma_k|^{-\frac{\nu_0+d+1}{2}} e^{-\text{tr}(\Sigma_k^{-1}\Lambda_0)},$$

Priors for latent variable (conjugate)

$$p(\alpha) = \text{Di}(w_0) \propto \prod_k \alpha_k^{w_0, k-1},$$



Variational Bayes for mixtures

Log-likelihood $\mathbf{x} = [x_1, \dots, x_n]$, $\mathbf{l} = [l_1, \dots, l_n]$

$$\begin{aligned}\log p(\mathbf{x}, \mathbf{l} | \theta) &\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} (\log(\mathcal{N}(x_i | \mu_k, \Sigma_k)) + \log \alpha_k) \\ &\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right)\end{aligned}$$

Log-prior:

$$\begin{aligned}\log p(\theta) &= \sum_{k=1}^K \log p(\alpha_k) + \log p(\mu_k) + \log(\Sigma_k) \\ &= \sum_{k=1}^K (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \text{tr}(\Sigma_k^{-1} \Lambda_0),\end{aligned}$$

We will be looking for conditionally independent prior

$$p(\alpha, \{\mu_k, \Sigma_k\}_1^K, \mathbf{l} | \mathbf{x}) \approx q(\alpha) \prod_{k=1}^K q(\mu_k) q(\Sigma_k) \prod_{i=1}^n q(l_i)$$

With

$$q(\mu_k) \propto \exp E_{\Sigma_k, \alpha, \mathbf{l}} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta))$$

$$q(\Sigma_k) \propto \exp E_{\mu_k, \alpha, \mathbf{l}} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta))$$

$$q(\alpha) \propto \exp E_{\mu_k, \Sigma_k, \mathbf{l}} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta)),$$

$$q(l_i) \propto \exp E_{\mu_k, \Sigma_k, \alpha} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta)),$$

Variational Bayes for mixtures

Log-likelihood $\mathbf{x} = [x_1, \dots, x_n]$, $\mathbf{l} = [l_1, \dots, l_n]$

$$\begin{aligned}\log p(\mathbf{x}, \mathbf{l} | \theta) &\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} (\log(\mathcal{N}(x_i | \mu_k, \Sigma_k)) + \log \alpha_k) \\ &\propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right)\end{aligned}$$

Log-prior:

$$\begin{aligned}\log p(\theta) &= \sum_{k=1}^K \log p(\alpha_k) + \log p(\mu_k) + \log p(\Sigma_k) + \log p(l_i) \\ &= \sum_{k=1}^K (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \text{tr}(\Sigma_k^{-1} \Lambda_0) - \sum_i l_{i,k} \log K,\end{aligned}$$

We will be looking for conditionally independent prior

$$p(\alpha, \{\mu_k, \Sigma_k\}_1^K, \mathbf{l} | \mathbf{x}) \approx q(\alpha) \prod_{k=1}^K q(\mu_k) q(\Sigma_k) \prod_{i=1}^n q(l_i)$$

With

$$\begin{aligned}q(\mu_k) &\propto \exp E_{\Sigma_k, \alpha, \mathbf{l}} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta)) \\ q(\Sigma_k) &\propto \exp E_{\mu_k, \alpha, \mathbf{l}} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta)) \\ q(\alpha) &\propto \exp E_{\mu_k, \Sigma_k, \mathbf{l}} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta)), \\ q(l_i) &\propto \exp E_{\mu_k, \Sigma_k, \alpha} (\log(\mathbf{x}, \mathbf{l} | \theta) + \log p(\theta)),\end{aligned}$$

Variational Bayes: factor $q(\mu_k)$

Joint likelihood,

$$\log p(\mathbf{x}, \mathbf{I}, \theta) \propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right) + \sum_{k=1}^K (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \text{tr}(\Sigma_k^{-1} \Lambda_0) - \sum_i l_{i,k} \log K$$

For $q(\mu_k)$ we ignore all terms independent of μ_k :

$$\begin{aligned} \log q(\mu_k) &\propto \mathbb{E} \left\{ \sum_{i=1}^n l_{i,k} \left(-\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right) \right\} \\ &\propto \mathbb{E} \left\{ -\frac{1}{2} \sum_{i=1}^n l_{i,k} \left(x_i^T \Sigma_k^{-1} x_i - \mu_k^T \Sigma_k^{-1} x_i - x_i^T \Sigma_k^{-1} \mu_k - \mu_k^T \Sigma_k^{-1} \mu_k \right) \right\} \\ &\text{expectation } \mathbb{E}(l_i) = \hat{l}_i, \mathbb{E}(\Sigma_k^{-1}) = \hat{\Sigma}_k^{-1} \\ &\propto -\frac{1}{2} \left(-\mu_k^T \Sigma_k^{-1} \left(\sum_{i=1}^n \hat{l}_{i,k} x_i \right) - \left(\sum_{i=1}^n \hat{l}_{i,k} x_i^T \right) \Sigma_k^{-1} \mu_k - \mu_k^T \Sigma_k^{-1} \mu_k \right) \end{aligned}$$

which is recognized as Gaussian

$$q(\mu_k) = \mathcal{N}(\hat{\mu}_k, \Sigma_{\mu,k})$$

which is equal to the decomposition above with assignment

$$\hat{\mu}_k = \sum_{i=1}^n \hat{l}_{i,k} x_i, \quad \Sigma_{\mu,k} = \hat{\Sigma}_k / \sum_{i=1}^n \hat{l}_{i,k}$$

Variational Bayes: factor $q(\Sigma_k)$

Joint likelihood,

$$\log p(\mathbf{x}, \mathbf{I}, \theta) \propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right) + \\ \sum_{k=1}^K (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \text{tr}(\Sigma_k^{-1} \Lambda_0) - \sum_i l_{i,k} \log K$$

For $q(\Sigma_k)$ we ignore all terms independent of Σ_k :

$$\log q(\Sigma_k) \propto \mathbb{E} \left\{ \sum_{i=1}^n l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right) \right\} + \\ \mathbb{E} \left\{ -\frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \frac{1}{2} \text{tr}(\Sigma_k^{-1} \Lambda_0) \right\} \\ \propto -\frac{1}{2} (\nu_0 + d + 1 + \sum_{i=1}^n \hat{l}_{i,k}) \log |\Sigma_k| \\ \mathbb{E} \left\{ -\frac{1}{2} \text{tr} \left[\Sigma_k^{-1} \left(\sum_{i=1}^n \hat{l}_{i,k} (x_i - \mu_k)(x_i - \mu_k)^T + \Lambda_0 \right) \right] \right\}$$

Which is recognized as inverse Wishart

$$q(\Sigma_k) = i\mathcal{W}(\nu_k, \Lambda_k) \Leftrightarrow q(\Sigma_k^{-1}) = i\mathcal{W}(\nu_k, \Lambda_k^{-1})$$

with assignment

$$\nu_k = \nu_0 + \sum_{i=1}^n \hat{l}_{i,k} \\ \Lambda_k = \Lambda_0 + \sum_{i=1}^n \hat{l}_{i,k} \mathbb{E}_\mu \left((x_i - \mu_k)(x_i - \mu_k)^T \right)$$

Variational Bayes: factor $q(\alpha)$

Joint likelihood,

$$\log p(\mathbf{x}, \mathbf{I}, \theta) \propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right) + \sum_{k=1}^K (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \text{tr}(\Sigma_k^{-1} \Lambda_0) - \sum_i l_{i,k} \log K$$

For $q(\alpha)$ we ignore all terms independent of α_k :

$$\begin{aligned} \log q(\Sigma_k) &\propto \mathbb{E} \left\{ \sum_{k=1}^K \sum_{i=1}^n l_{i,k} (\log \alpha_k) + (w_0 - 1) \log \alpha_k \right\} \\ &\propto \sum_{k=1}^K \sum_{i=1}^n \hat{l}_{i,k} (\log \alpha_k) + (w_0 - 1) \log \alpha_k \end{aligned}$$

Which is recognized as Dirichlet

$$q(\alpha) = \mathcal{D}i(\mathbf{w})$$

with assignment

$$w_k = \sum_{i=1}^n \hat{l}_{i,k} + w_0$$

Variational Bayes: factor $q(l_i)$

Joint likelihood,

$$\log p(\mathbf{x}, \mathbf{l}, \theta) \propto \sum_{i=1}^n \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k \right) + \sum_{k=1}^K (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \text{tr}(\Sigma_k^{-1} \Lambda_0) - \sum_i l_{i,k} \log K$$

For $q(l_i)$ we ignore all terms independent of l_i :

$$\log q(l_i) \propto \sum_{k=1}^K l_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k - \log K \right)$$

Which is recognized as Multinomial

$$q(l_i) = \mathcal{M}(l_i)$$

with assignment

$$\lambda_{i,k} = \exp E_{\mu_k, \Sigma_k, \alpha_k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k - \log K \right)$$

Variational Bayes: moments

$$\begin{aligned}
 q(\mu_k) &= \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_k) & \mathbb{E}(\mu_k) &= \hat{\mu}_k, & \mathbb{E}(\mu_k^T Z \mu_k) &= \hat{\mu}_k^T Z \hat{\mu}_k + \text{tr}(Z \hat{\Sigma}_k), \\
 q(\alpha) &= \text{Di}(\mathbf{w}) & \mathbb{E}(\alpha_k) &= \frac{w_k}{\sum_k w_k}, & \mathbb{E}(\log \alpha_k) &= \psi(w_k) - \psi(\sum_k w_k), \\
 q(\Sigma_k) &= i\mathcal{W}(\nu_k, \Lambda_k) & \mathbb{E}(\Sigma_k^{-1}) &= \Lambda_k^{-1} \nu_k & \mathbb{E}(\log |\Sigma_k|) &= \log |\Lambda_k| + \psi_d\left(\frac{\nu_k}{2}\right) + d \log 2 \\
 q(l_i) &= \mathcal{M}(\lambda_i) & \mathbb{E}(l_{i,k}) &= \frac{\lambda_{i,k}}{\sum \lambda_{i,k}}
 \end{aligned}$$

with d being dimensionality of x . Yielding:

$$\begin{aligned}
 \lambda_{i,k} &= \exp \mathbb{E}_{\mu_k, \Sigma_k, \alpha_k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \alpha_k - \log K \right) \\
 &= \exp \left(-\frac{1}{2} (\log |\Lambda_k| - \psi_d(\frac{\nu_k}{2}) - d \log 2) - \frac{1}{2} (x_i - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x_i - \hat{\mu}_k) + \frac{\nu d}{\nu - d - 1} \right) \\
 &\quad \times \exp \left(\psi(w_k) - \psi(\sum_k w_k) - \log K \right) \\
 \Lambda_k &= \Lambda_0 + \frac{1}{2} \sum_{i=1}^n \hat{l}_{i,k} \left((x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T + \hat{\Sigma}_k \right)
 \end{aligned}$$

Variational Bayes (VB) algorithm

Initialize: choose $w_k^{(0)}, \hat{\mu}_k^{(0)}, \hat{\Sigma}_k^{(0)}, \Lambda_0, \nu_0, \forall k$

Iterate:

1. Compute $\lambda_{i,k}$ and expected labels:

$$p(l = \epsilon_k | x_i) = \hat{l}_{i,k} = \frac{\lambda_{i,k}}{\sum_k \lambda_{i,k}}, \quad w_k = \sum_{i=1}^n \hat{l}_{i,k} + w_0,$$

2. Recompute the component statistics

$$\hat{\mu}_k = \frac{1}{w_k} \sum_i \hat{l}_{i,k} x_i, \quad \Sigma_{\mu k} = \hat{\Sigma}_k / \sum_{i=1}^n \hat{l}_{i,k},$$

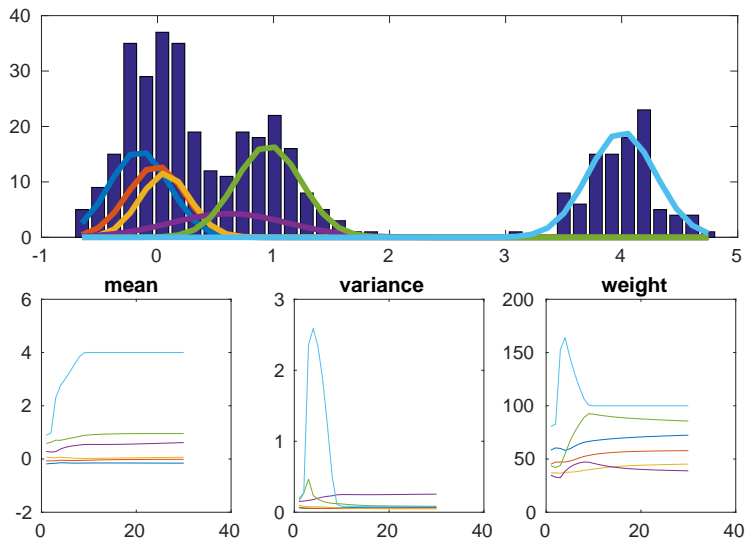
$$\Lambda_k = \frac{1}{w_k} \sum_i \hat{l}_{i,k} [(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T + \Sigma_{\mu,k}]$$

$$\hat{\Sigma}_k^{-1} = \Lambda_k \nu_k, \quad \hat{\Sigma}_k = \Lambda_k^{-1} (\nu_k - d - 1)^{-1}$$

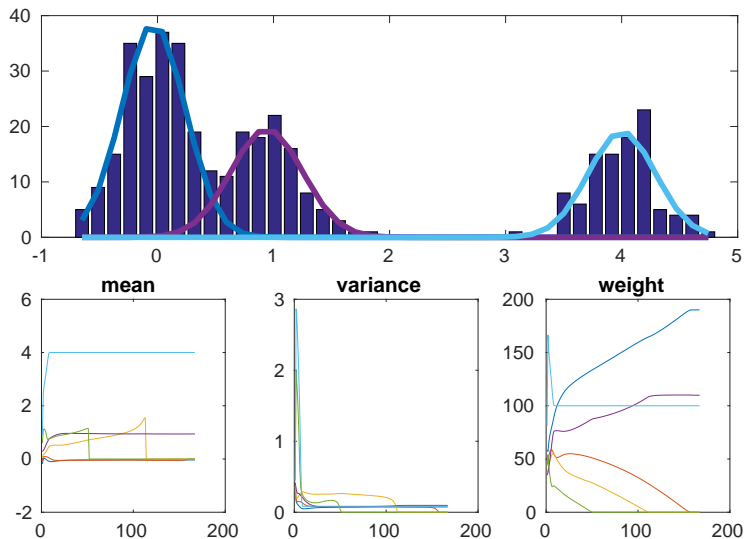
$$\hat{\alpha}_k = \frac{w_k}{\sum w_k}, \quad \nu_k = \nu_0 + \sum_{i=1}^n \hat{l}_{i,k},$$

3. Evaluate expected log-likelihood (if required)

EM: $\mu_{true} = \{0, 1, 4\}$, fit $K = 6$ components



VB: $\mu_{true} = \{0, 1, 4\}$ fit $K = 6$ components



Multivariate Gaussians in dimension d :

$$\begin{aligned}x &\sim \mathcal{N}(\mu, \Omega^{-1}), \\ \mu &\sim \mathcal{N}(\mu_0, (\tau\Omega)^{-1}) \\ \Omega &\sim \mathcal{W}(V, \nu),\end{aligned}$$

where \mathcal{W} is the Wishart distribution with ν degrees of freedom.

Covariance matrix:

full covariance: effective number of data $n_k > d$, $O(d^2)$,

scaled identity: homogenous noise σI , (k-means),

diagonal: ignoring rotation of ellipses,

low rank: only selected principal components,

...

Initialization:

random: over what space? cubic...

LHS: latin hypercube sampling

Number of component:

very many: slow convergence

birth and death: random generation

split and merge: evaluate which component to split and/or which two components join into one.

problematic.

Simulate 2d mixtures with components:

$$\begin{array}{lll} \mu_1 = [1; 1], & \Sigma_1 = \text{eye}(2), & \alpha_1 = 0.3, \\ \mu_2 = [-1; 1], & \Sigma_2 = \text{eye}(2), & \alpha_2 = 0.3, \\ \mu_3 = [0; -1], & \Sigma_3 = \text{diag}([2, 0.1]), & \alpha_3 = 0.4. \end{array}$$

Estimation via	points
EM algorithm	5
VB algorithm	5