Mixture Models

Václav Šmídl

March 26, 2020

Mixture of Gaussians

2D data of 1000 points:

blue: 100 points

red: 200 points

green: 700 points

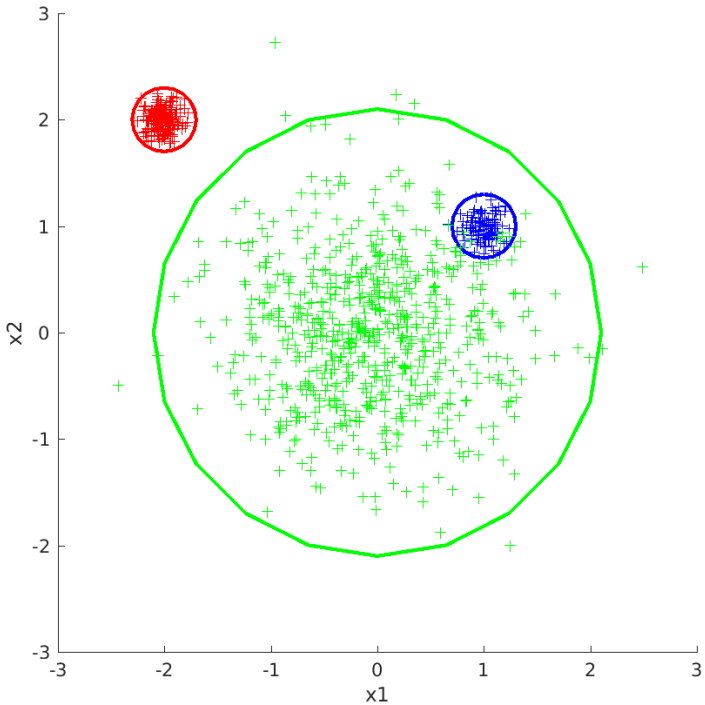
Each of them Gaussian distributed (component).

$$p_1(x) = \mathcal{N}([1;1], 0.1^2 I),$$
$$p_2(x) = \mathcal{N}([-2;2], 0.1^2 I),$$
$$p_3(x) = \mathcal{N}([0;0], 0.1^2 I),$$

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 $[0.1, 0.2, 0.7]$.

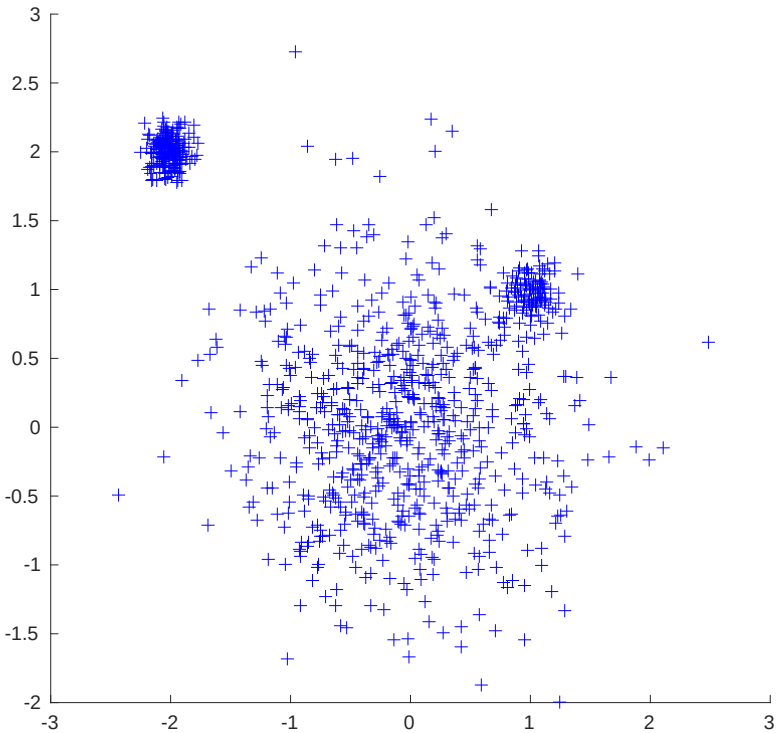


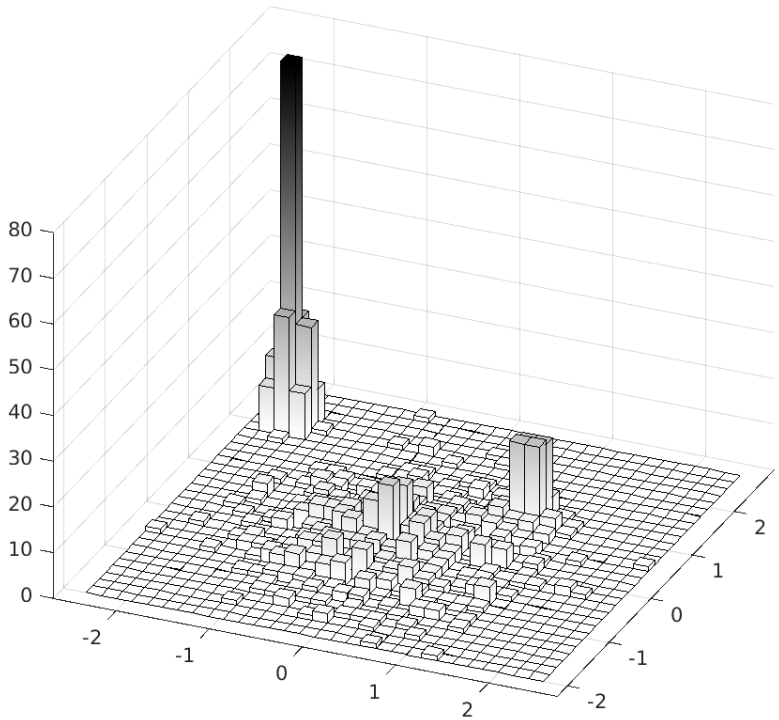
Mixture of Gaussians:

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

Estimation of mixture of Gaussians

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

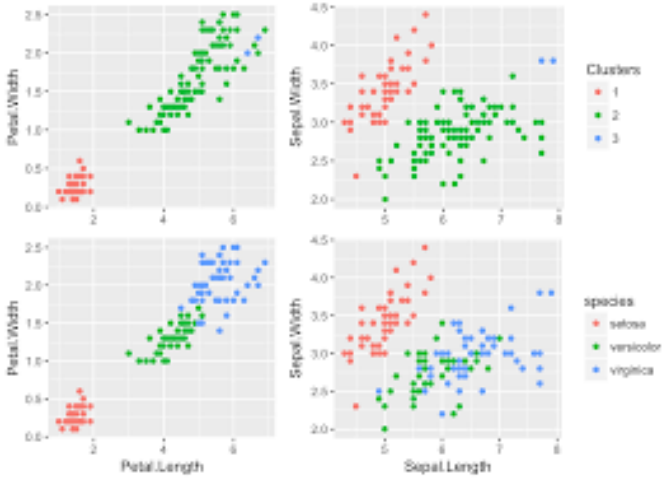


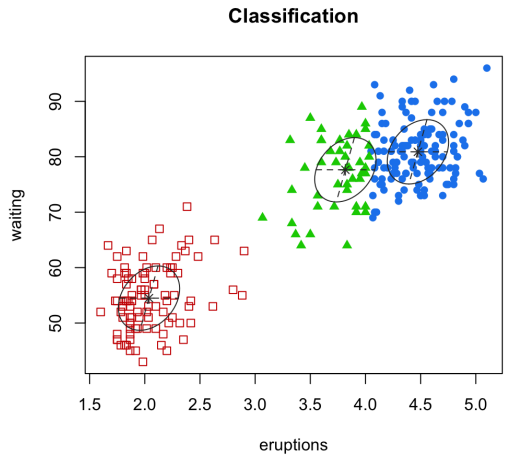


Task: find $\alpha_1 \ldots, \alpha_K$, $\mu_1, \ldots, \mu_K$ and $\Sigma_1, \ldots, \Sigma_K$ and possibly K.

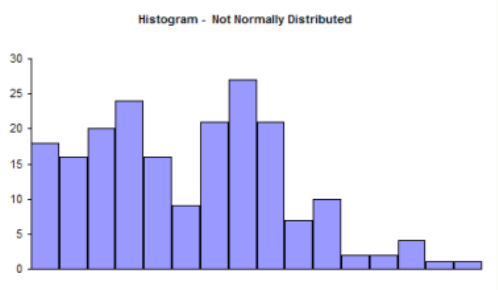
Uses of mixtures

Clustering we care about assignments:





Density representation: we do not care



Maximum likelihood estimation

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

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

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

Maximum likelihood estimation

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

$$
\begin{aligned}
p(x_i|\theta) &= \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k), \quad \forall i = 1, \ldots, n \\
p(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} p(x_i), \quad \log p(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i) \\
\log p(x_1, \ldots, x_n|\theta) &= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right) \\
\mathcal{N}(\mu_k, \Sigma_k) &= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right),
\end{aligned}
$$

Maximum likelihood estimation

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

$$
\begin{aligned}
p(x_i|\theta) &= \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k), \quad \forall i = 1, \ldots, n \\
p(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} p(x_i), \quad \log p(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i) \\
\log p(x_1, \ldots, x_n|\theta) &= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right) \\
\mathcal{N}(\mu_k, \Sigma_k) &= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right),
\end{aligned}
$$

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

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

and others.

Maximum likelihood continued

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

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

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

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** $\sum_k \alpha_k = 1$:

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

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

where λ is a Lagrange multiplier for $\sum_k \alpha_k = 1$.
Conditions of extrema (solved by alternating evaluation of):

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

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

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

Mixture estimation via latent variable

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

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

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

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

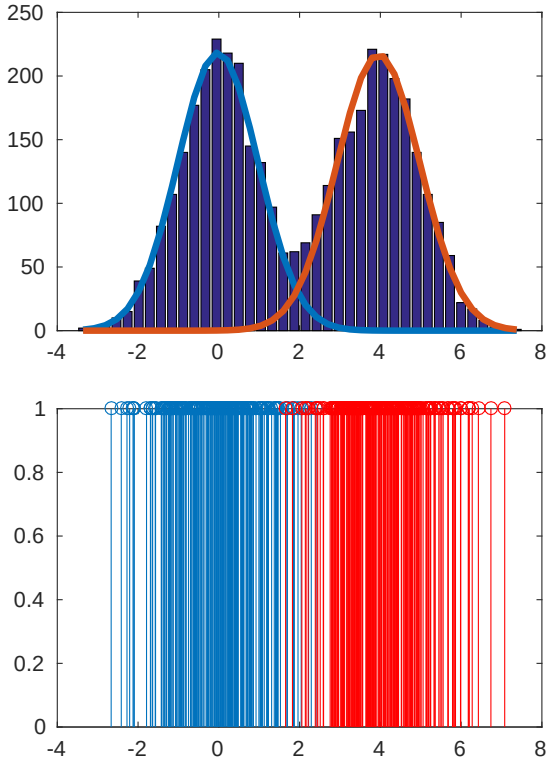
$$p(l_i) = \prod_{k=1}^{K} \alpha_k^{l_{i,k}}.$$

$$p(x_i) = \sum_k p(x_i, l_i)$$

$$= \sum_k p(x|l_i = \epsilon_k)p(l_i = \epsilon_k).$$

$$= \sum_k \mathcal{N}(\mu_k, \Sigma_k)\alpha_k,$$

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



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

Expectation maximization (EM) algorithm

Joint distribution:

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

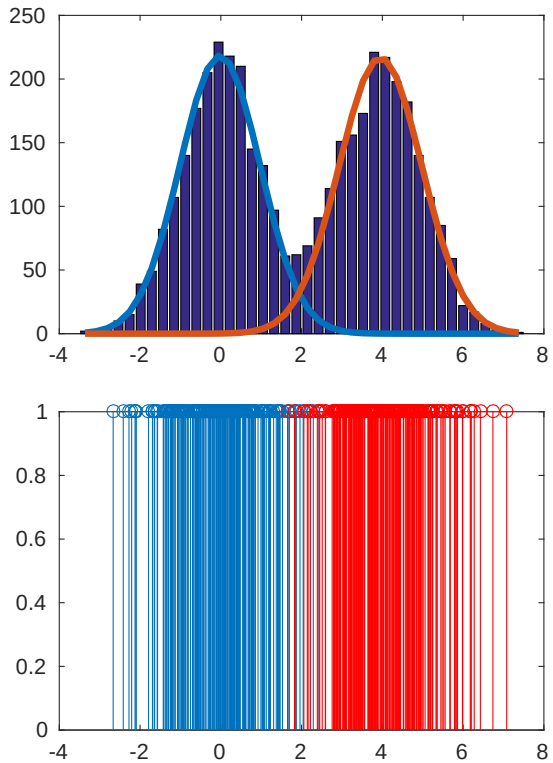
$$p(x_i|l_i) = \prod_k \mathcal{N}(x_i|\mu_k, \Sigma_k)^{l_{i,k}},$$

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

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 ith data point belongs to kth 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:

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

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

Maximizing log-likelihood

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

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

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

Maximum likelihood with latent variable

Maximizing log-likelihood

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

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

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

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

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

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

M-step:

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

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:

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

2. Recompute the component parameters

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

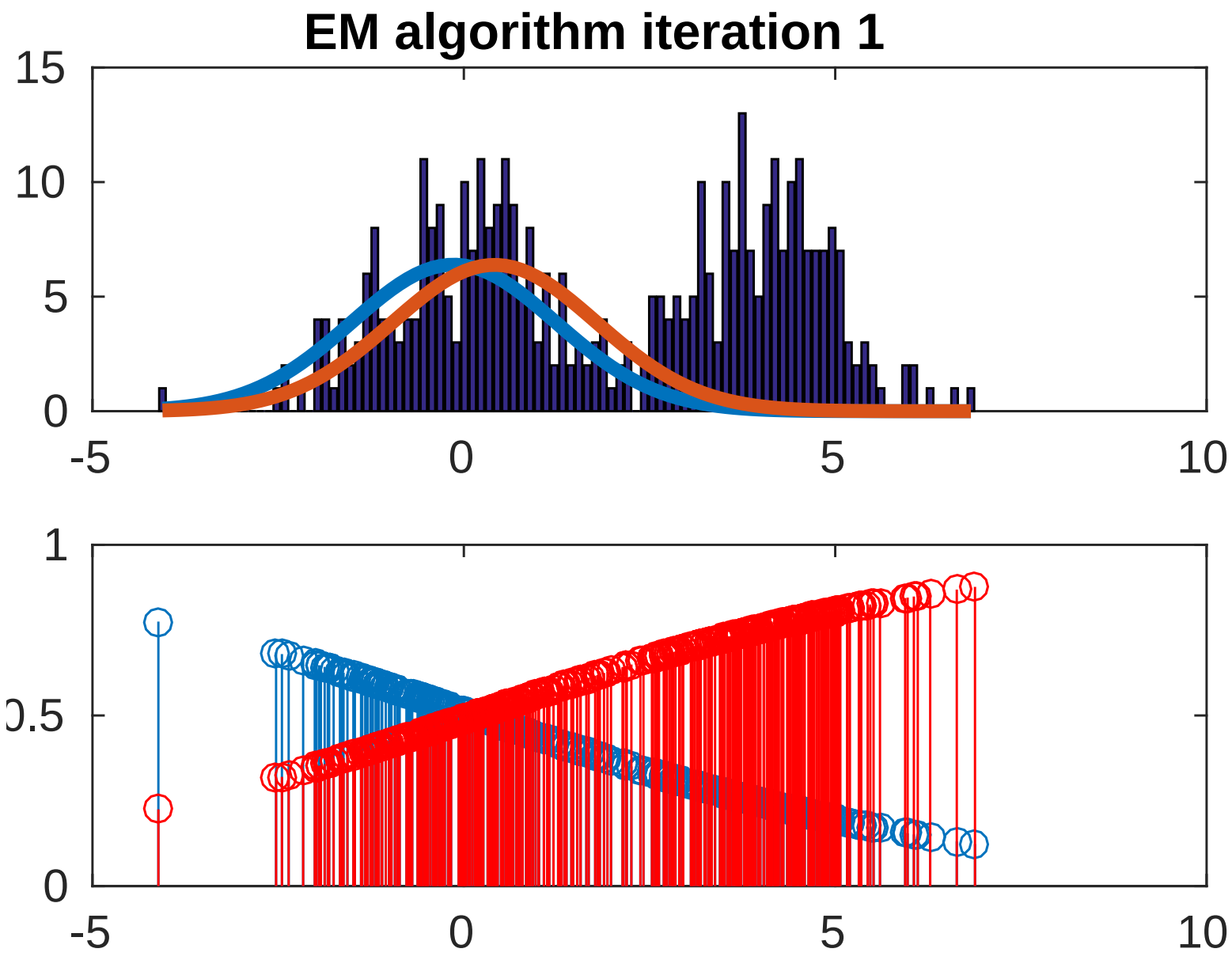
$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_i \hat{l}_{i,k} \left(x_i - \hat{\mu}_k\right) \left(x_i - \hat{\mu}_k\right)^T,$$

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

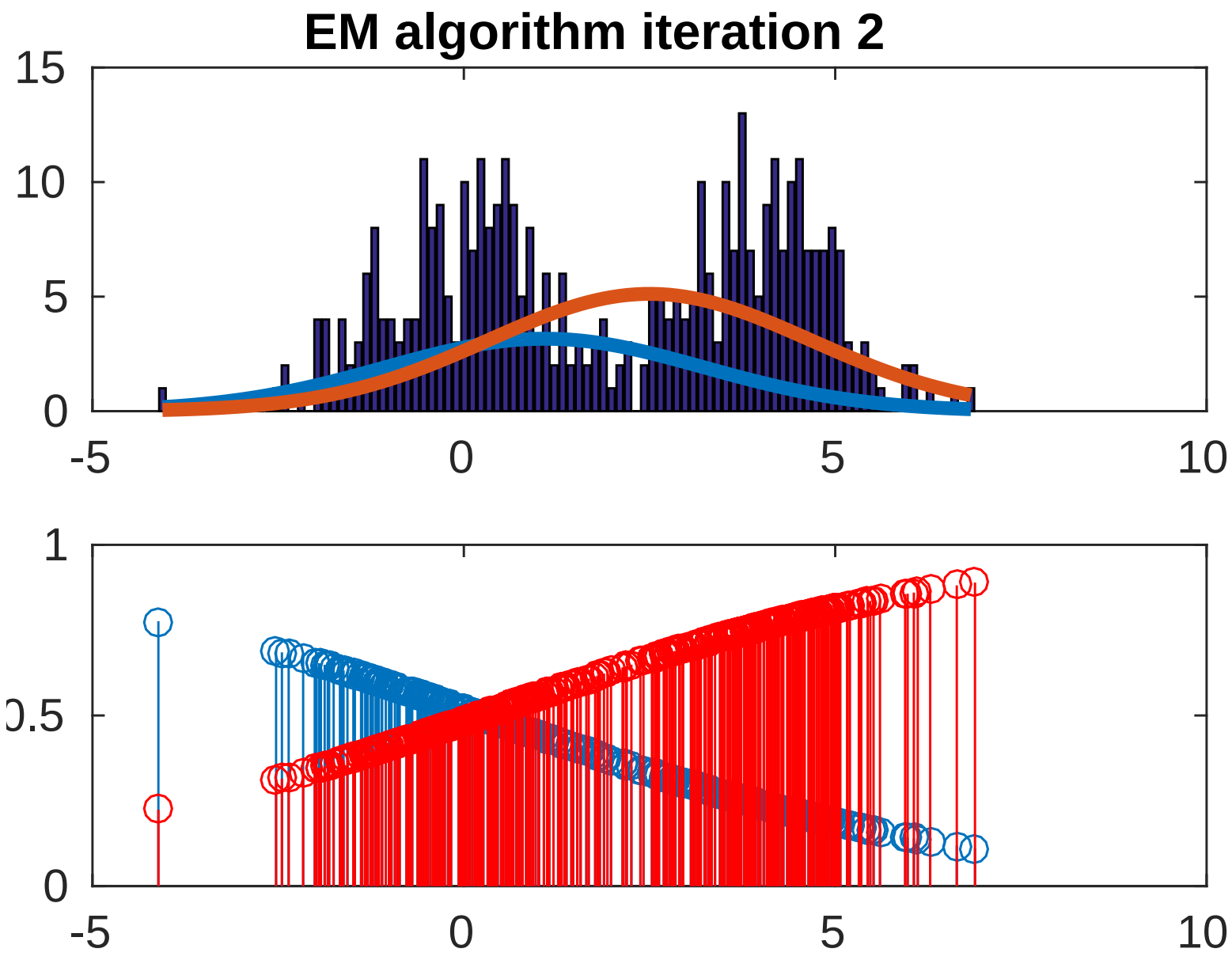
3. (Evaluate log-likelihood)

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

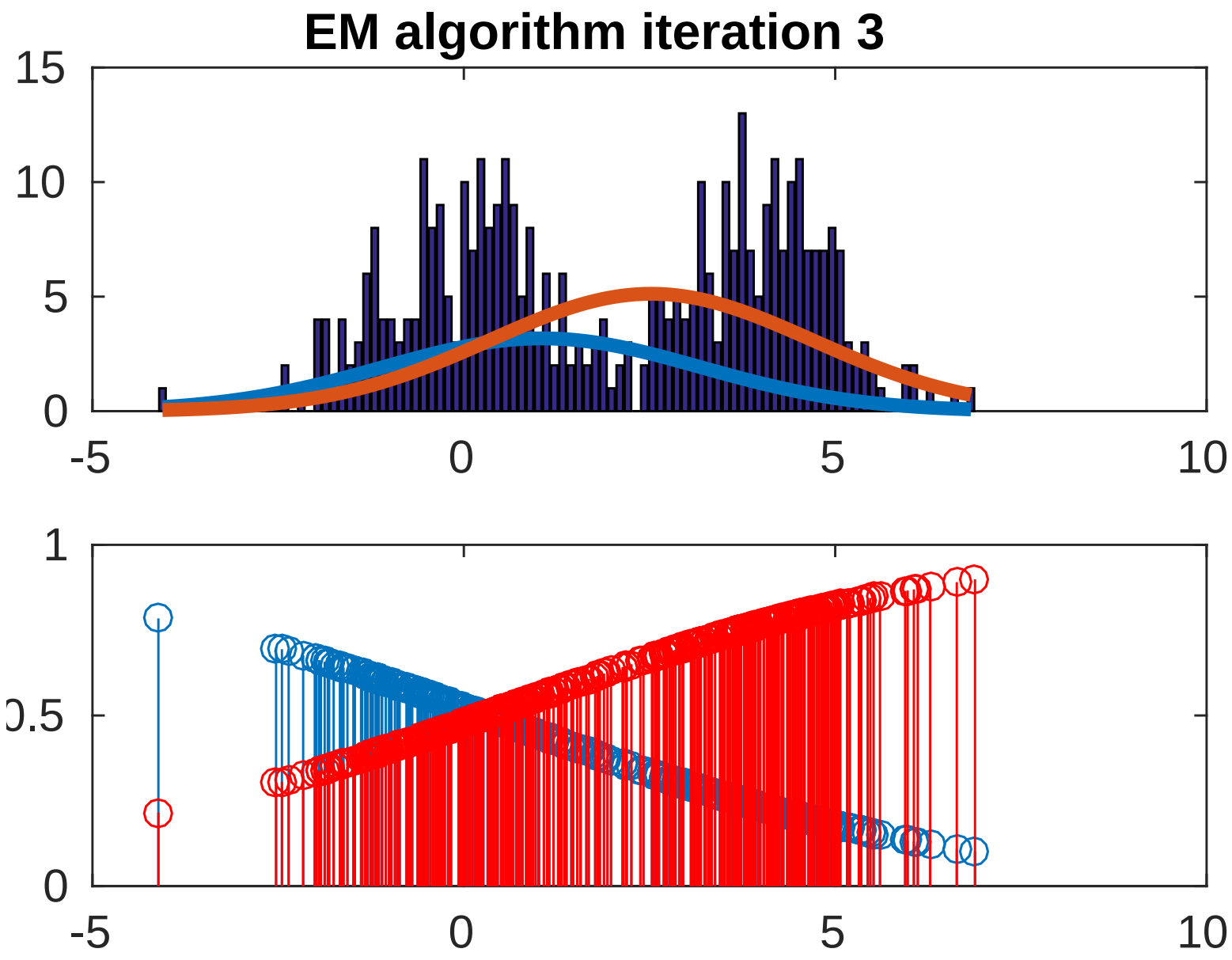
Expectation Maximization (EM) algorithm



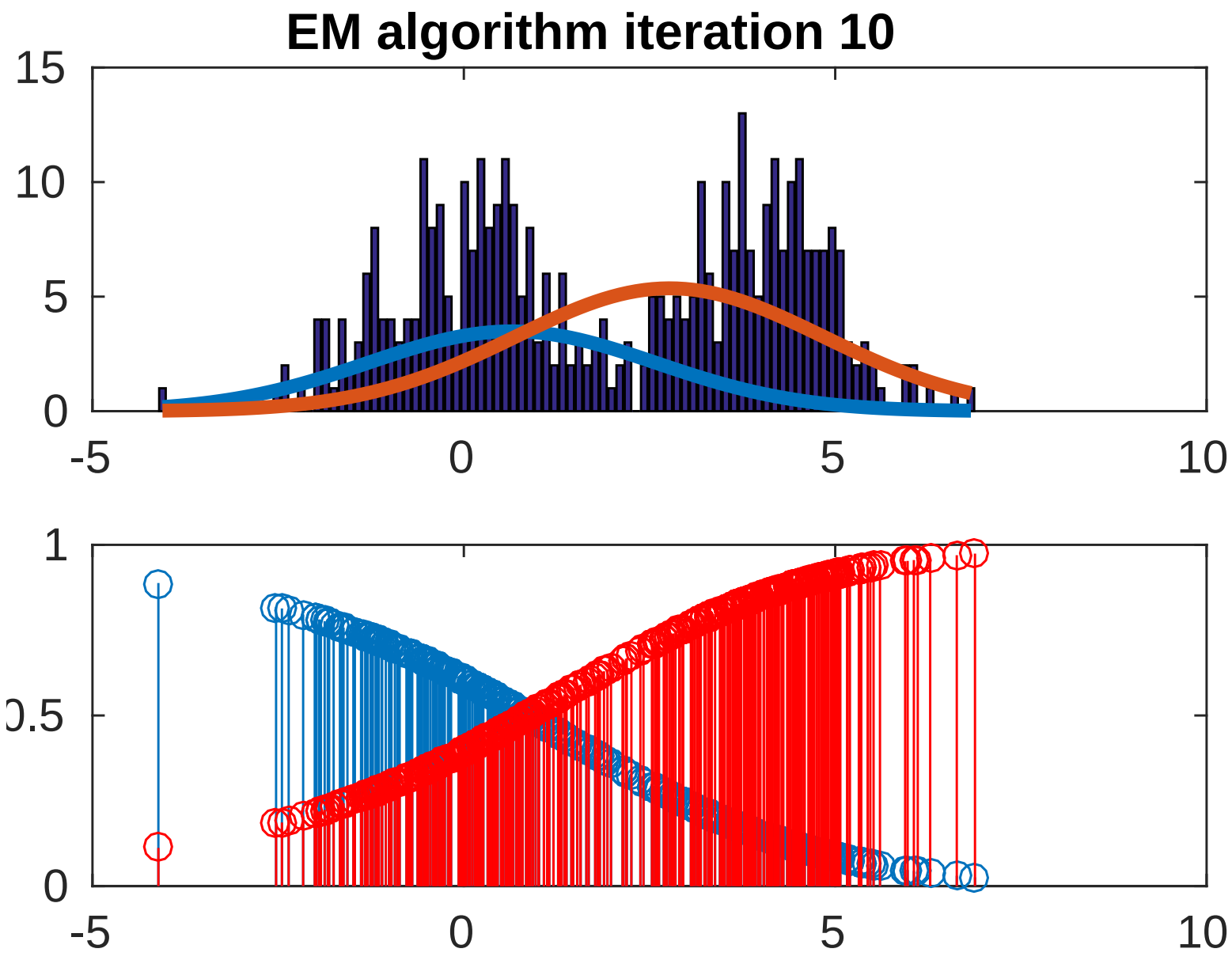
Expectation Maximization (EM) algorithm



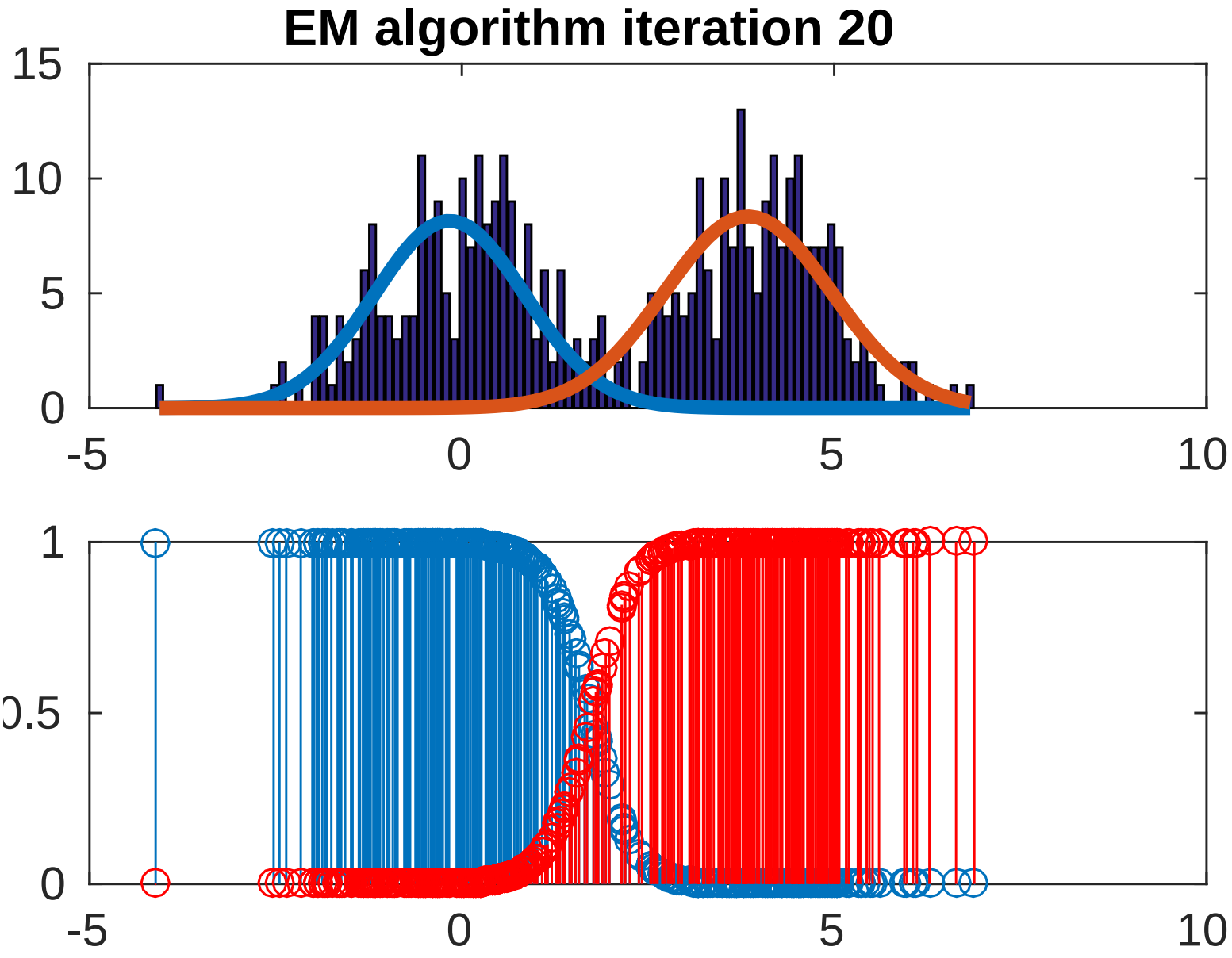
Expectation Maximization (EM) algorithm



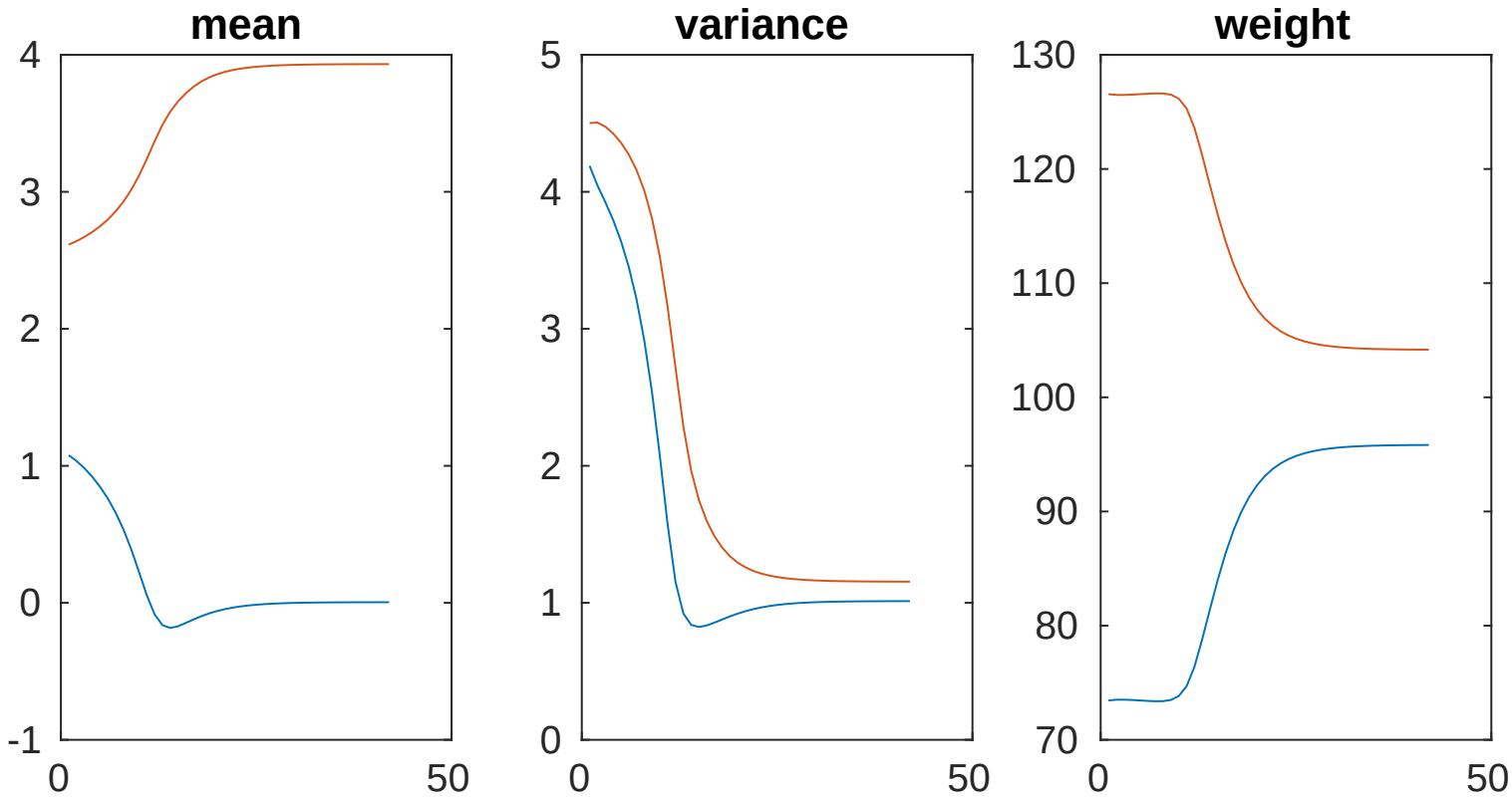
Expectation Maximization (EM) algorithm



Expectation Maximization (EM) algorithm



Expectation Maximization (EM) algorithm



Bayesian treatment

Joint distribution:

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

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

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

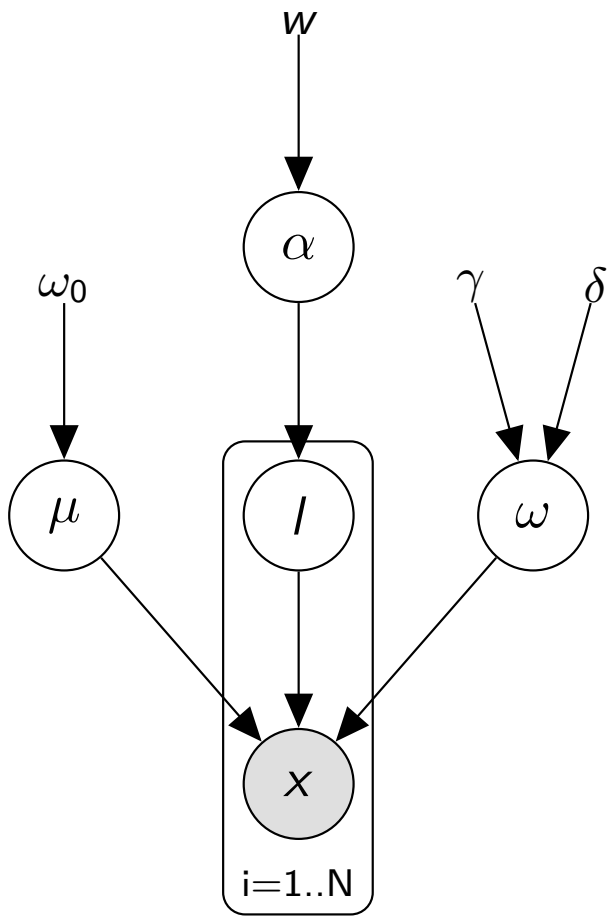
Priors for multivariate Normal (conjugate)

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

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

Priors for latent variable (conjugate)

$$p(\boldsymbol{\alpha}) = Di(w_0) \propto \prod_k \alpha_k^{w_{0,k}-1},$$



Variational Bayes for mixtures

Log-likelihood $\boldsymbol{x} = [x_1, \ldots, x_n], \boldsymbol{I} = [I_1, \ldots, I_n]$

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

Log-prior:

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

We will be looking for conditionally independent prior

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

With

$$\begin{aligned}q(\mu_k) &\propto \exp \mathrm{E}_{\Sigma_k,\alpha,\boldsymbol{I}}\left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right)\\ q(\Sigma_k) &\propto \exp \mathrm{E}_{\mu_k,\alpha,\boldsymbol{I}}\left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right)\\ q(\boldsymbol{\alpha}) &\propto \exp \mathrm{E}_{\mu_k,\Sigma_k,\boldsymbol{I}}\left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right),\\ q(I_i) &\propto \exp \mathrm{E}_{\mu_k,\Sigma_k,\boldsymbol{\alpha}}\left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right),\end{aligned}$$

Variational Bayes for mixtures

Log-likelihood $\boldsymbol{x} = [x_1, \ldots, x_n], \boldsymbol{I} = [I_1, \ldots, I_n]$

$$\log p(\boldsymbol{x}, \boldsymbol{I}|\theta) \propto \sum_{i=1}^{n} \sum_{k=1}^{K} I_{i,k} \left(\log(\mathcal{N}(x_i|\mu_k, \Sigma_k)) + \log \alpha_k\right)$$
$$\propto \sum_{i=1}^{n} \sum_{k=1}^{K} I_{i,k} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k) + \log \alpha_k\right)$$

Log-prior:

$$\log p(\theta) = \sum_{k=1}^{K} \log p(\alpha_k) + \log p(\mu_k) + \log p(\Sigma_k) + \log p(I_i)$$
$$= \sum_{k=1}^{K} (w_0 - 1) \log \alpha_k + 0 - \frac{\nu_0 + d + 1}{2} \log |\Sigma_k| - \mathrm{tr}(\Sigma_k^{-1} \Lambda_0) - \sum_i I_{i,k} \log K,$$

We will be looking for conditionally independent prior

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

With

$$q(\mu_k) \propto \exp \mathrm{E}_{\Sigma_k, \alpha, \boldsymbol{I}} \left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right)$$
$$q(\Sigma_k) \propto \exp \mathrm{E}_{\mu_k, \alpha, \boldsymbol{I}} \left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right)$$
$$q(\boldsymbol{\alpha}) \propto \exp \mathrm{E}_{\mu_k, \Sigma_k, \boldsymbol{I}} \left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right),$$
$$q(I_i) \propto \exp \mathrm{E}_{\mu_k, \Sigma_k, \boldsymbol{\alpha}} \left(\log(\boldsymbol{x}, \boldsymbol{I}|\theta) + \log p(\theta)\right),$$

Variational Bayes: factor $q(\mu_k)$

Joint likelihood,

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

$$\sum_{k=1}^{K}(w_0-1)\log\alpha_k + 0 - \frac{\nu_0+d+1}{2}\log|\Sigma_k| - \mathrm{tr}(\Sigma_k^{-1}\Lambda_0) - \sum_i I_{i,k}\log K$$

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

$$\log q(\mu_k) \propto \mathrm{E}\left\{\sum_{i=1}^{n} I_{i,k}\left(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)\right\}$$

$$\propto \mathrm{E}\left\{-\tfrac{1}{2}\sum_{i=1}^{n} I_{i,k}\left(x_i^T\Sigma_k^{-1}x_i - \mu_k^T\Sigma_k^{-1}x_i - x_i^T\Sigma_k^{-1}\mu_k - \mu_k^T\Sigma_k^{-1}\mu_k\right)\right\}$$

$$\text{expectation } \mathrm{E}(I_i) = \hat{I}_i, \mathrm{E}(\Sigma_k^{-1}) = \hat{\Sigma}_k^{-1})$$

$$\propto -\frac{1}{2}\left(-\mu_k^T\Sigma_k^{-1}\left(\sum_{i=1}^{n}\hat{I}_{i,k}x_i\right) - \left(\sum_{i=1}^{n}\hat{I}_{i,k}x_i^T\right)\Sigma_k^{-1}\mu_k - \mu_k^T\Sigma_k^{-1}\mu_k\right)$$

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

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

Variational Bayes: factor $q(\Sigma_k)$

Joint likelihood,

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

$$\sum_{k=1}^{K}(w_0-1)\log\alpha_k + 0 - \frac{\nu_0+d+1}{2}\log|\Sigma_k| - \mathrm{tr}(\Sigma_k^{-1}\Lambda_0) - \sum_i I_{i,k}\log K$$

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

$$\log q(\Sigma_k) \propto \mathrm{E}\left\{\sum_{i=1}^{n} I_{i,k}\left(-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)\right\} +$$

$$\mathrm{E}\left\{-\frac{\nu_0+d+1}{2}\log|\Sigma_k| - \frac{1}{2}\mathrm{tr}(\Sigma_k^{-1}\Lambda_0)\right\}$$

$$\propto -\frac{1}{2}(\nu_0+d+1+\sum_{i=1}^{n}\hat{I}_{i,k})\log|\Sigma_k|$$

$$\mathrm{E}\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma_k^{-1}\left(\sum_{i=1}^{n}\hat{I}_{i,k}(x_i-\mu_k)(x_i-\mu_k)^T + \Lambda_0\right)\right]\right\}$$

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

$$\nu_k = \nu_0 + \sum_{i=1}^{n}\hat{I}_{i,k}$$

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

Variational Bayes: factor $q(\alpha)$

Joint likelihood,

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

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

$$\log q(\Sigma_k) \propto \mathrm{E}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n} I_{i,k}\left(\log\alpha_k\right) + (w_0-1)\log\alpha_k\right\}$$
$$\propto \sum_{k=1}^{K}\sum_{i=1}^{n}\hat{I}_{i,k}\left(\log\alpha_k\right) + (w_0-1)\log\alpha_k$$

Which is recognized as Dirichlet

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

with assignment

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

Variational Bayes: factor $q(l_i)$

Joint likelihood,

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

$$\sum_{k=1}^{K}(w_0-1)\log\alpha_k + 0 - \frac{\nu_0+d+1}{2}\log|\Sigma_k| - \mathrm{tr}(\Sigma_k^{-1}\Lambda_0) - \sum_i l_{i,k}\log K$$

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

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

with assignment

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

Variational Bayes: moments

$$q(\mu_k) = \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_k) \qquad \mathrm{E}(\mu_k) = \hat{\mu}_k, \qquad \mathrm{E}(\mu_k^T Z \mu_k) = \hat{\mu}_k^T Z \hat{\mu}_k + \mathrm{tr}(Z\hat{\Sigma}_k),$$

$$q(\boldsymbol{\alpha}) = \mathcal{D}i(\boldsymbol{w}) \qquad \mathrm{E}(\alpha_k) = \frac{w_k}{\sum_k w_k}, \qquad \mathrm{E}(\log \alpha_k) = \psi(w_k) - \psi(\textstyle\sum_k w_k),$$

$$q(\Sigma_k) = i\mathcal{W}(\nu_k, \Lambda_k) \quad \mathrm{E}(\Sigma_k^{-1}) = \Lambda_k^{-1}\nu_k \qquad \mathrm{E}(\log|\Sigma_k|) = \log|\Lambda_k| + \psi_p(\nu_k/s) + d\log 2$$

$$q(l_i) = \mathcal{M}(\lambda_i) \qquad \mathrm{E}(l_{i,k}) = \frac{\lambda_{i,k}}{\sum \lambda_{i,k}}$$

Yielding:

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

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:

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

2. Recompute the component statistics

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

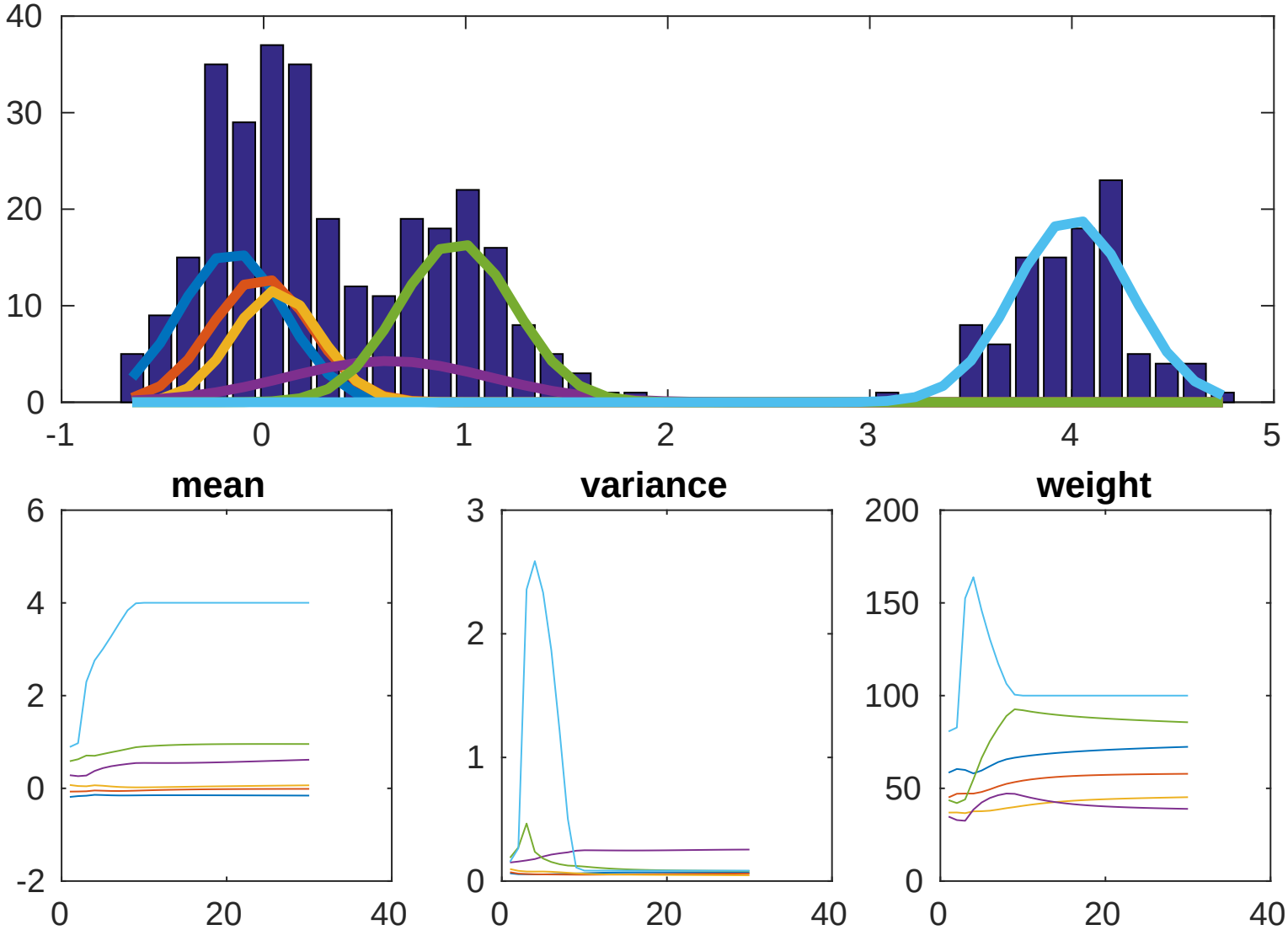
$$\Lambda_k = \frac{1}{w_k} \sum_i \hat{I}_{i,k} \left[(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T + \Sigma_{\mu,k} \right]$$

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

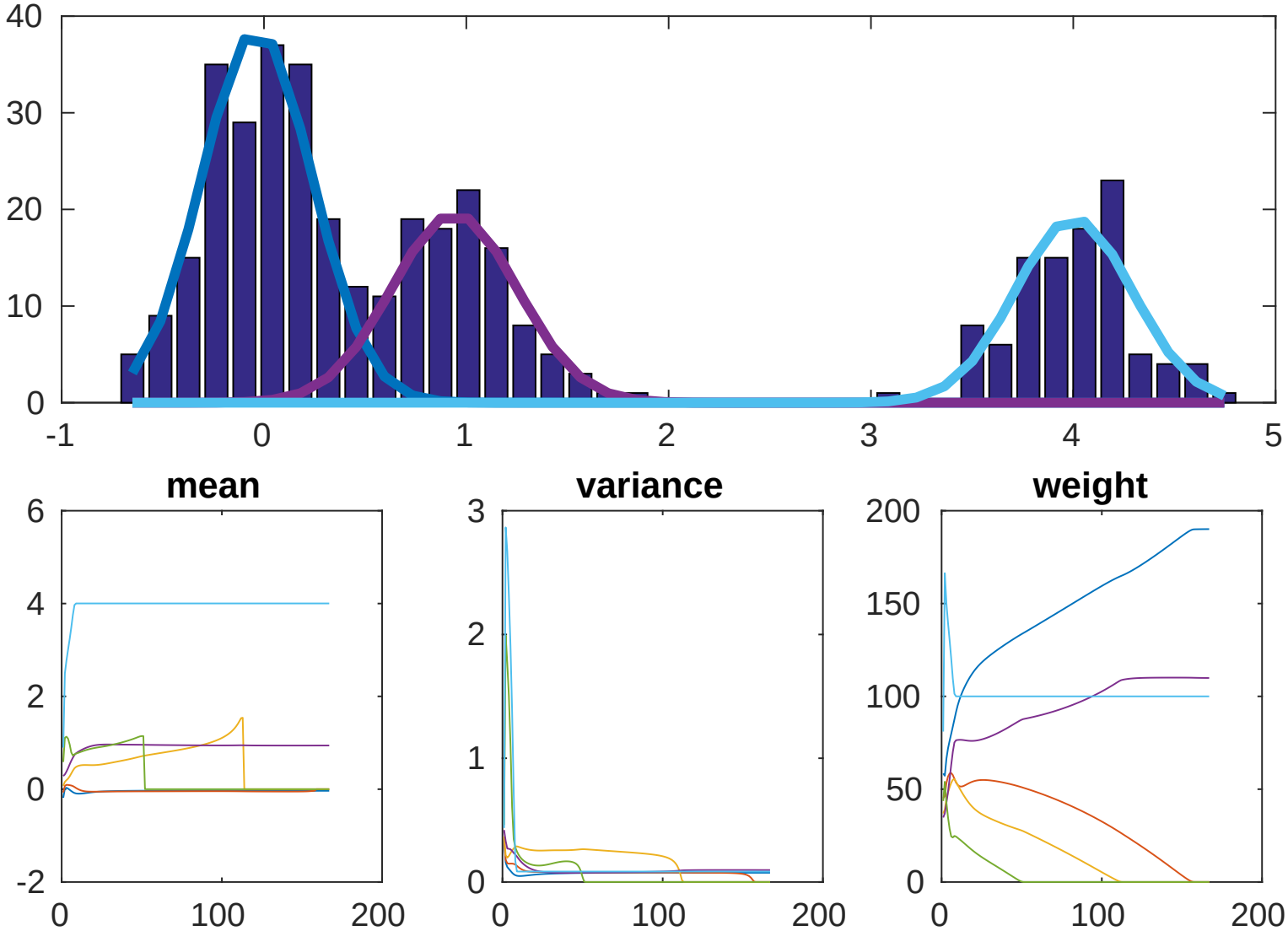
$$\hat{\alpha}_k = \frac{w_k}{\sum w_k}, \quad \nu_k = \nu_0 + \sum_{i=1}^n \hat{I}_{i,k},$$

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



Mixture of Gaussians in higher dimensions

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,

...

Mixture of Gaussians in higher dimensions

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.

Assignment II

Simulate 2d mixtures with components:

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

Estimation via	points
EM algorithm	20
VB algorithm	30