Machine Learning I: Fractal 2

Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur
http://home.iitj.ac.in/~rn/

These slides are prepared from the following book:
Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Introduction to Probability and Random Variable

Random Variable

A random variable is a function $X : \mathcal{S} \to \mathcal{S}_X$. Here, $\mathcal{S}_X \subset \mathbb{R}$.

Toss a Coin

$\mathcal{S} = \{\mathsf{H}, \mathsf{T}\} = \{s_1, s_2\}$. $X(s_i) = \begin{cases} 0 & \text{If } s_i = \text{Head} \\ 1 & \text{If } s_i = \text{Tail} \end{cases}$

$\mathcal{S}_X = \{0, 1\}$.

Probability Mass Function

$$p_X[x_i] = P[X(x) = x_i].$$

Toss a Coin

$$\begin{aligned} p_X[0] &= P[X(s) = 0] = p \\ p_X[1] &= P[X(s) = 1] = 1 - p. \end{aligned}$$

Expectation

Let X be a discrete random variable with PMF p_X. Then, the expected value of X is defined as: $\mathbb{E}[X] = \sum_{x_i \in \mathcal{S}_X} x_i p_X[x_i]$.

Variance

Let X be a discrete random variable with the given PMF p_X. The variance of X is:

$$\text{var}(X) = \sum_i (x_i - \mathbb{E}[X])^2 p_X[x_i] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X].$$

Random Vector

A random vector is a function $\begin{bmatrix} X \\ Y \end{bmatrix} : \mathcal{S} \times \mathcal{S} \to \mathcal{S}_{X,Y}$. Here, $\mathcal{S}_{X,Y} \subset \mathbb{R}^2$.

Joint PMF

$p_{X,Y}[x_i, y_j] = P[X(s) = x_i, Y(s) = y_j], i = 1, 2, \ldots, N_X; j = 1, 2, \ldots, N_Y$

$p_{X,Y}[i,j]$	$j=0$	$j=1$
$i=0$	$\frac{1}{8}$	$\frac{1}{8}$
$i=1$	$\frac{1}{4}$	$\frac{1}{2}$

$p_{X,Y}[0,0] = \frac{1}{8},\ p_{X,Y}[0,1] = \frac{1}{8}$
$p_{X,Y}[1,0] = \frac{1}{4},\ p_{X,Y}[1,1] = \frac{1}{2}$

Marginal PMF

$$p_X[x_i] = \sum_{j=1}^{N_Y} p_{X,Y}[x_i, y_j] \text{ and } p_Y[y_j] = \sum_{i=1}^{N_X} p_{X,Y}[x_i, y_j]$$

Independence of two random variables

Two discrete random variables X and Y are independent, if $p_{X,Y}[x_i, y_j] = p_X[x_i] p_Y[y_j],\ \forall (x_i, y_j) \in \mathcal{S}_X \times \mathcal{S}_Y$.

Covariance and Correlation Coefficient

Consider two random variables X and Y with joint PMF $p_{X,Y}$.

$$\begin{aligned} \mathsf{var}(X+Y) &= \mathsf{var}(X) + \mathsf{var}(Y) + 2\mathsf{cov}(X,Y) \\ \mathsf{cov}(X,Y) &= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\ \rho_{X,Y} &= \frac{\mathsf{cov}(X,Y)}{\sqrt{\mathsf{var}(X)\mathsf{var}(Y)}} \end{aligned}$$

N-Dimensional Random Vector

A random vector is a function $\begin{bmatrix} X_1 & \cdots & X_N \end{bmatrix}^\top : \mathcal{S} \times \ldots \times \mathcal{S} \to \mathcal{S}_{X_1,\ldots,X_N} \subset \mathbb{R}^N$.

Joint PMF:

$$\begin{aligned} p_{X_1,X_2,\ldots,X_N}[x_1,x_2,\ldots,x_N] &= P[X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N] \\ p_{\mathbf{X}}[\mathbf{x}] &= P[\mathbf{X} = \mathbf{x}]. \end{aligned}$$

Expected Vector: $\mathbb{E}_{\mathbf{x}}[\mathbf{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[X_1] & \mathbb{E}_{X_2}[X_2] & \cdots & \mathbb{E}_{X_N}[X_N] \end{bmatrix}^\top$
Covariance Matrix: Let $\mathbf{X}$ be an N-dimensional random vector with the joint PMF $p_{\mathbf{X}}[\mathbf{x}]$.

$$\mathbf{C_X} = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top\right] = \begin{bmatrix} \mathsf{var}(X_1) & \mathsf{cov}(X_1,X_2) & \cdots & \mathsf{cov}(X_1,X_N) \\ \mathsf{cov}(X_2,X_1) & \mathsf{var}(X_2) & \cdots & \mathsf{cov}(X_2,X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{cov}(X_N,X_1) & \mathsf{cov}(X_N,X_2) & \cdots & \mathsf{var}(X_N) \end{bmatrix}$$

Covariance matrix of $\mathbf{Y} = \mathbf{AX}$ is equal to $\mathbf{C_Y} = \mathbf{AC_X A}^\top$.

Conditional Probability

The conditional probability of a random variable X taking on a value given that a second random variable Y has a specific value is defined as:

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

If the random variables X and Y are independent, then the conditional probability definition simplifies to $p(x|y) = p(x)$.

Law of Total Probability

For discrete random variables X and Y the law of total probability states that:

$$p(x) = \sum_y p(x,y) = \sum_y p(x|y)p(y)$$

Bayes' Rule

For discrete random variables X and Y, Bayes' rule states that:

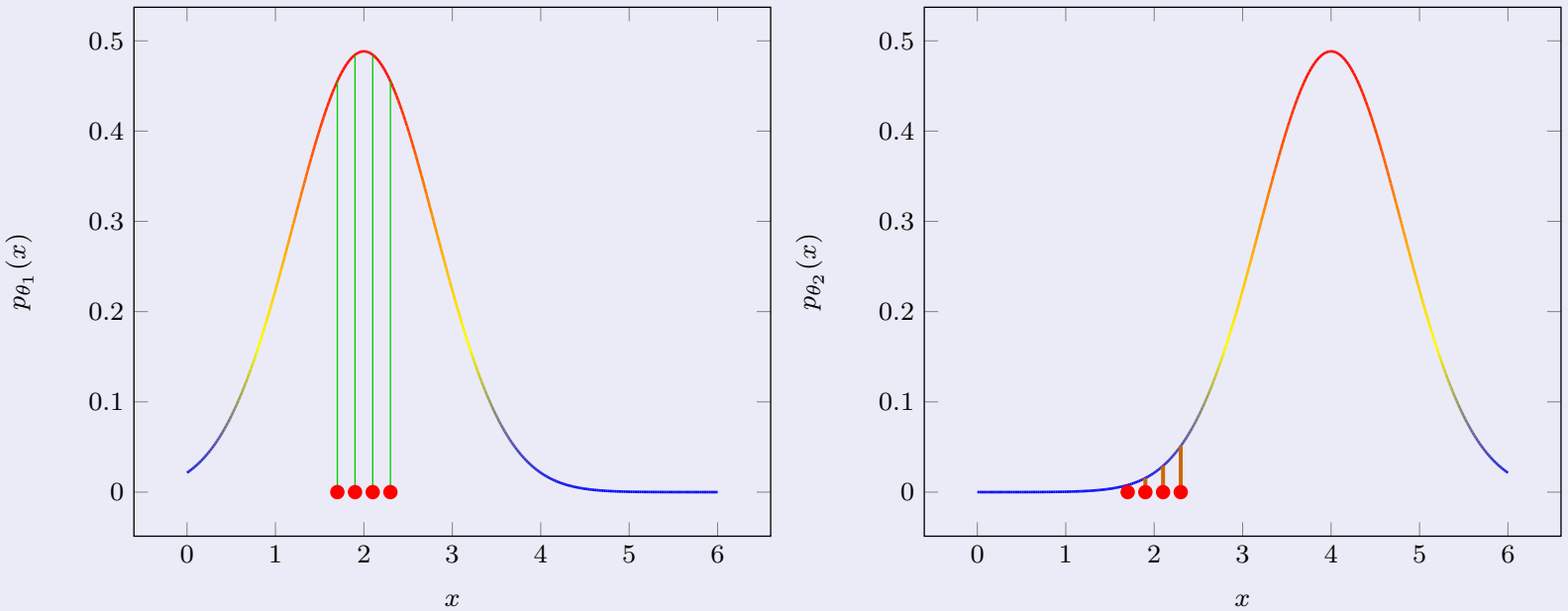
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Parameter Estimation

Given a dataset $\mathcal{S} = \{x_1, x_2, \ldots, x_m\}$ containing IID samples of an unknown distribution $p_{\text{data}}(x)$, learn a distribution $p_\theta(x)$ that is close as possible to $p_{\text{data}}(x)$. Assume that the $p_\theta(x)$ is parametrized by the parameter θ. Find θ such that $p_\theta(x)$ that is close as possible to $p_{\text{data}}(x)$.

Maximum Likelihood Estimator

Define the likelihood of the data $\mathcal{S}$ with respect to $p_\theta(x)$ as: $\ell(\mathcal{S}; \theta) = p_\theta(x_1, x_2, \ldots, x_m)$.



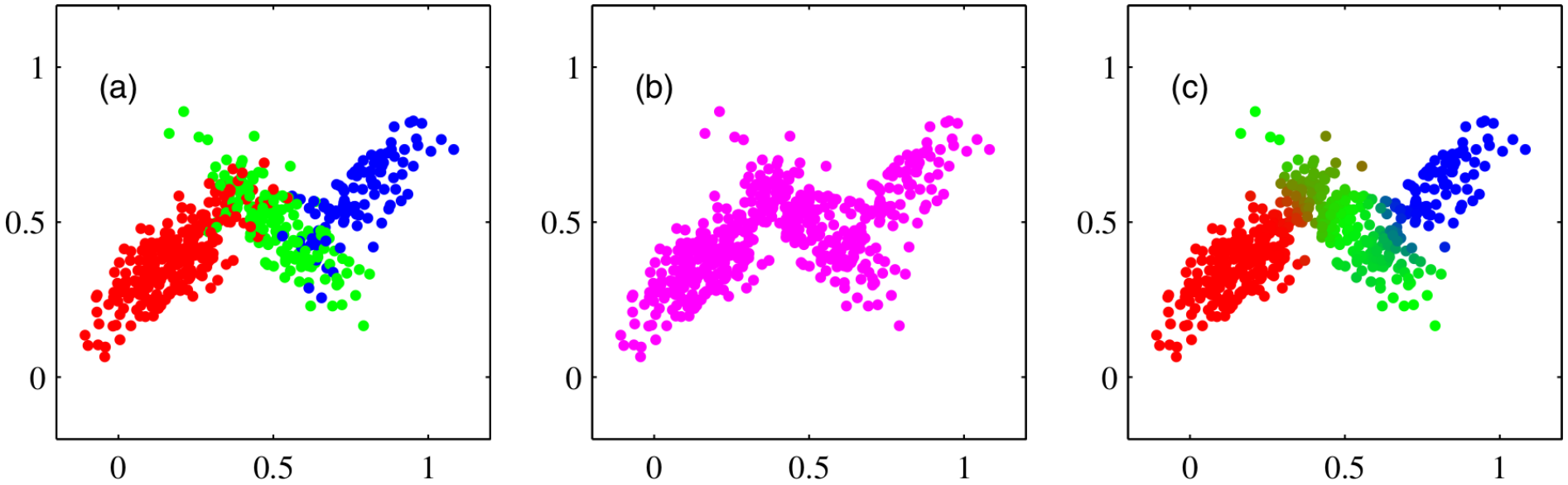
$$\hat{\theta} = \arg\max_\theta p_\theta(x_1, x_2, \ldots, x_m) = \arg\max_\theta \log(\ell(\mathcal{S}; \theta))$$

$$= \arg\max_\theta \sum_{i=1}^{m} \log p_\theta(x_i)$$

MLE: Example

- Assume that the samples of $\mathcal{S}$ are IID and drawn from a Gaussian distribution parameterized by $\theta = (\mu, \sigma)$.
- We can write the log-likelihood as:

$$\begin{aligned}
\ell(\mathcal{S};\theta) &= \sum_{i=1}^{m} \log p_\theta(x_i) = \sum_{i=1}^{m} \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu)^2 - m \times \log(\sigma\sqrt{2\pi}) \\
\frac{\partial}{\partial\mu}\ell(\mathcal{S};\theta) &= \frac{1}{\sigma^2} \sum_{i=1}^{m} (x_i - \mu) = 0 \\
\hat{\mu} &= \frac{1}{m} \sum_{i=1}^{m} x_i \\
\frac{\partial}{\partial\sigma}\ell(\mathcal{S};\theta) &= \frac{1}{\sigma^3} \sum_{i=1}^{m} (x_i - \mu)^2 - \frac{m}{\sigma} = 0 \\
\hat{\sigma} &= \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - \hat{\mu})^2}.
\end{aligned}$$



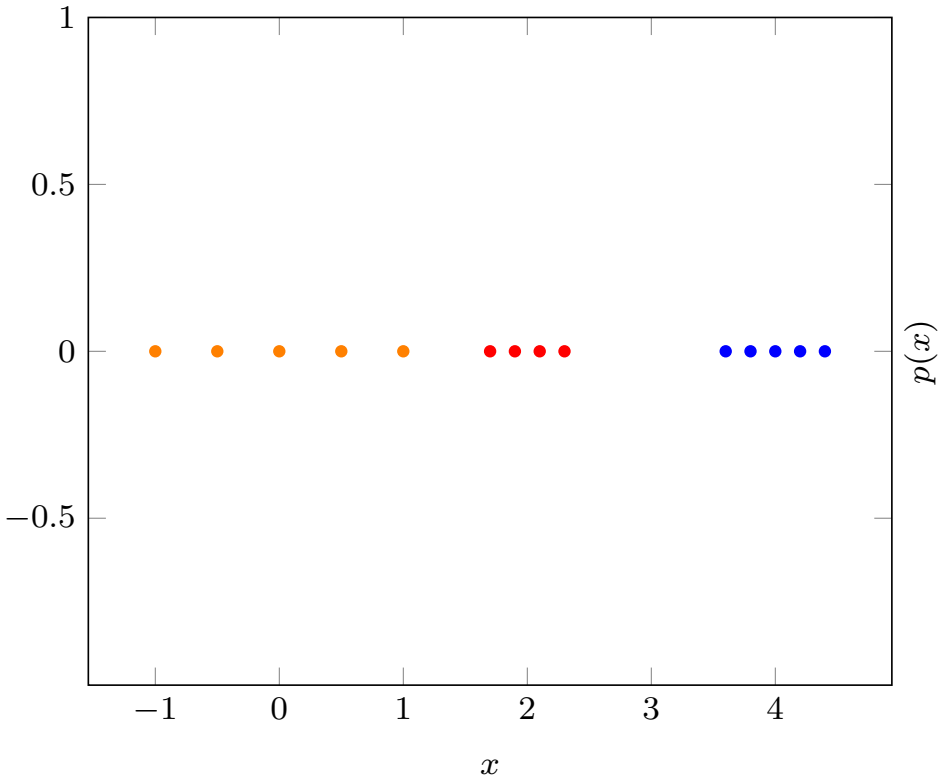
Gaussian Mixture Models: Problem Formulation

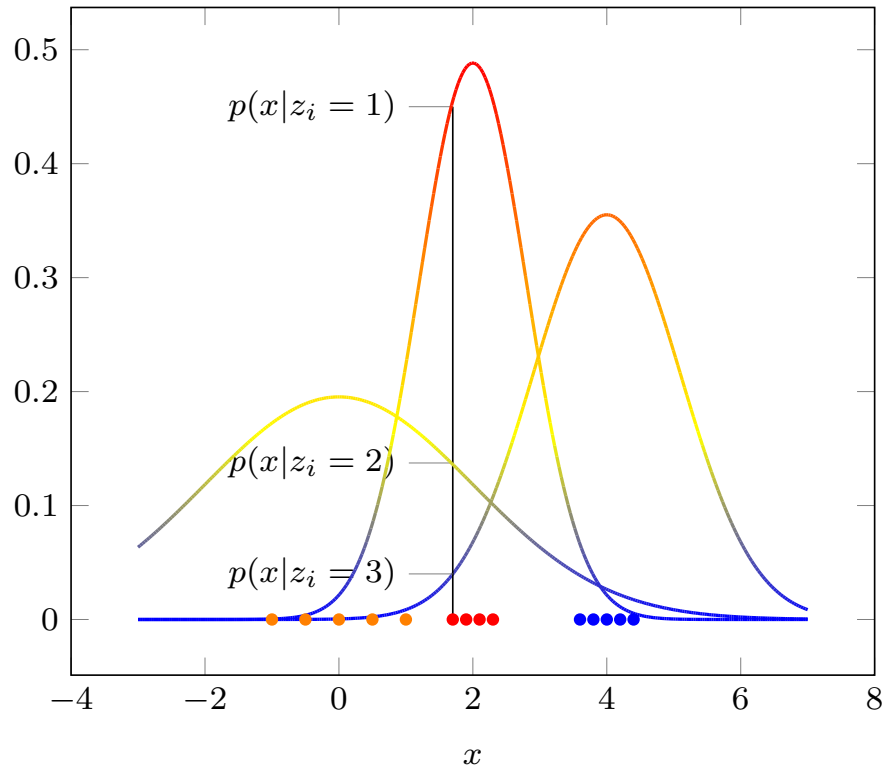
Given a set $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of n independent samples drawn from a mixture of k Gaussian distributions, find the following:

- Determine the probability of each point being sampled from a particular Gaussian distribution. (Cluster assignment probabilities, responsibilities)
- Estimate the parameters (mean vector and covariance matrix) for all k distributions. (Cluster representatives)

- We will use *latent random variables* for selecting one out of these k distributions.
- Assume that any $\mathbf{x}$ is sampled as follows. First, we choose a random number in $\{1, \ldots, k\}$. Let Z be the random variable corresponding to this choice, and denote $P[Z = z_i] = \pi_i$.
- Here, it is obvious to observe that $\pi_1 + \pi_2 + \cdots + \pi_k = 1$.
- Then, we sample $\mathbf{x}$ from the i-th Gaussian distribution:

$$p(\mathbf{x}|Z = z_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} = \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i).$$





Latent Variables and the EM Algorithm

- Therefore, the density of $\mathbf{x}$ can be written as:

$$
\begin{aligned}
p(\mathbf{x}) &= \sum_{i=1}^{k} p(\mathbf{x}, z_i) \\
&= \sum_{i=1}^{k} p(z_i) p(\mathbf{x}|z_i) \\
&= \sum_{i=1}^{k} \pi_i \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} \\
&= \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \\
\gamma(jt) &= p(z_t = 1|\mathbf{x}_j) \\
&= \frac{p(z_t = 1) p(\mathbf{x}|z_t = 1)}{p(\mathbf{x})} = \frac{\pi_t \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)}{\sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)}
\end{aligned}
$$

Maximum Likelihood Estimator

$$
\begin{aligned}
p(\mathbf{x}) &= \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\
p(\mathbf{X}) &= p(\mathbf{x}_1) \cdots p(\mathbf{x}_n) \\
\log(p(\mathbf{X})) &= \sum_{j=1}^{n} \log(p(\mathbf{x}_j)) \\
\ell(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) \\
\boldsymbol{\pi}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star &= \arg\max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \ell(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\nabla_{\boldsymbol{\mu}_t} \ell &= \nabla_{\boldsymbol{\mu}_t} \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) \\
&= \sum_{j=1}^{n} \frac{\pi_t \nabla_{\boldsymbol{\mu}_t} \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \\
&= \sum_{j=1}^{n} \frac{\pi_t \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \boldsymbol{\Sigma}_t^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_t) \\
&= \sum_{j=1}^{n} \gamma(jt) \boldsymbol{\Sigma}_t^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_t)
\end{aligned}
$$

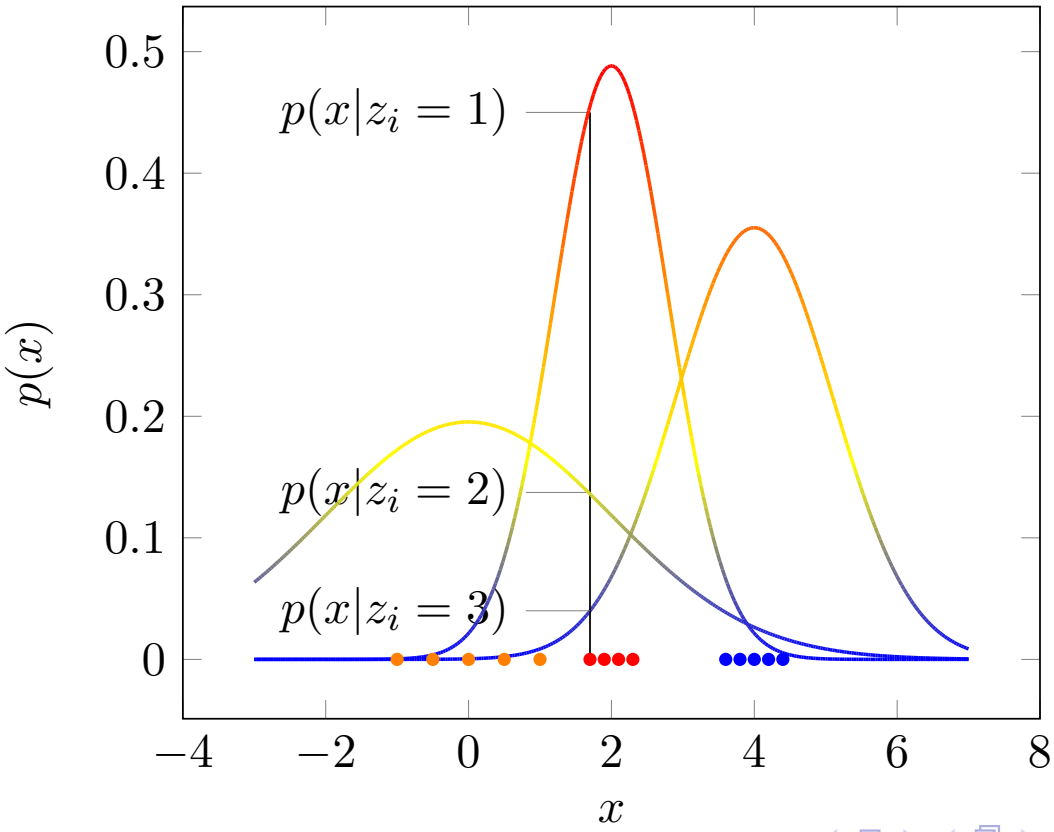
$$
\begin{aligned}
\nabla_{\boldsymbol{\mu}_t}\ell &= \sum_{j=1}^{n}\gamma(jt)\boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_t)=\mathbf{0}\Rightarrow\sum_{j=1}^{n}\gamma(jt)(\mathbf{x}_j-\boldsymbol{\mu}_t)=\mathbf{0}\\
\sum_{j=1}^{n}\gamma(jt)\mathbf{x}_j &= \boldsymbol{\mu}_t\sum_{j=1}^{n}\gamma(jt)\\
\boldsymbol{\mu}_t &= \frac{1}{n_t}\sum_{j=1}^{n}\gamma(jt)\mathbf{x}_j \quad \text{Here, } n_t=\sum_{j=1}^{n}\gamma(jt)\\
\nabla_{\boldsymbol{\Sigma}_t}\ell &= \sum_{j=1}^{n}\frac{\pi_t\nabla_{\boldsymbol{\Sigma}_t}\mathcal{N}(\mathbf{x}_j;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k}\pi_i\mathcal{N}(\mathbf{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}\\
&= \sum_{j=1}^{n}\frac{\pi_t\mathcal{N}(\mathbf{x}_j;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k}\pi_i\mathcal{N}(\mathbf{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}\left(-\boldsymbol{\Sigma}_t^{-1}+\boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_t)(\mathbf{x}_j-\boldsymbol{\mu}_t)^\top\boldsymbol{\Sigma}_t^{-1}\right)\\
&= \sum_{j=1}^{n}\gamma(jt)\left(-\boldsymbol{\Sigma}_t^{-1}+\boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_t)(\mathbf{x}_j-\boldsymbol{\mu}_t)^\top\boldsymbol{\Sigma}_t^{-1}\right)\\
\boldsymbol{\Sigma}_t^{\star} &= \frac{1}{n_t}\sum_{j=1}^{n}\gamma(jt)(\mathbf{x}_j-\boldsymbol{\mu}_t)(\mathbf{x}_j-\boldsymbol{\mu}_t)^\top
\end{aligned}
$$

$$
\begin{aligned}
\pi_t^\star &= \arg\max_{\boldsymbol{\pi}} \ell(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ subject to } \sum_{i=1}^{k} \pi_i = 1 \\
f &= \ell + \lambda \left(\sum_{i=1}^{k} \pi_i - 1 \right) \\
\nabla_{\pi_t} \ell + \lambda &= 0 \\
\sum_{t=1}^{k} \pi_t \nabla_{\pi_t} \ell + \lambda \sum_{t=1}^{k} \pi_t &= 0 \\
\sum_{t=1}^{k} \pi_t \nabla_{\pi_t} \ell &= -\lambda \\
\nabla_{\pi_t} \ell &= \sum_{j=1}^{n} \frac{\mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \\
\sum_{t=1}^{k} \sum_{j=1}^{n} \frac{\pi_t \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} &= \sum_{j=1}^{n} \frac{\sum_{t=1}^{k} \pi_t \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \sum_{j=1}^{n} 1 = n \Rightarrow n = -\lambda \\
\pi_t \nabla_{\pi_t} \ell &= -\lambda \pi_t \Rightarrow \pi_t \sum_{j=1}^{n} \frac{\mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} = -\lambda \pi_t \\
\sum_{j=1}^{n} \gamma(z_{jt}) = n \pi_t \Rightarrow \pi_t^\star &= \frac{1}{n} \sum_{j=1}^{n} \gamma(z_{it})
\end{aligned}
$$

$$\gamma(z_{jt}) \leftarrow \frac{\pi_t \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i\in[k]} \pi_i \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

$$\boldsymbol{\mu}_t \leftarrow \frac{1}{n_t}\sum_{j\in[n]} \gamma(z_{jt})\mathbf{x}_j \text{ and } \boldsymbol{\Sigma}_t \leftarrow \frac{1}{n_t}\sum_{j\in[n]} \gamma(z_{jt})(\mathbf{x}_j - \boldsymbol{\mu}_t)(\mathbf{x}_j - \boldsymbol{\mu}_t)^\top$$

$$\pi_t \leftarrow \frac{n_t}{n}. \text{ Here, } n_t = \sum_{j\in[n]} \gamma(z_{jt}).$$



Expectation Maximization

1: **Input**: $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{d \times n}$, where $p(\mathbf{x}) = \sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

2: **Maximize log-likelihood:** $\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}} \sum_{j \in [n]} \log \Big(\sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Big)$

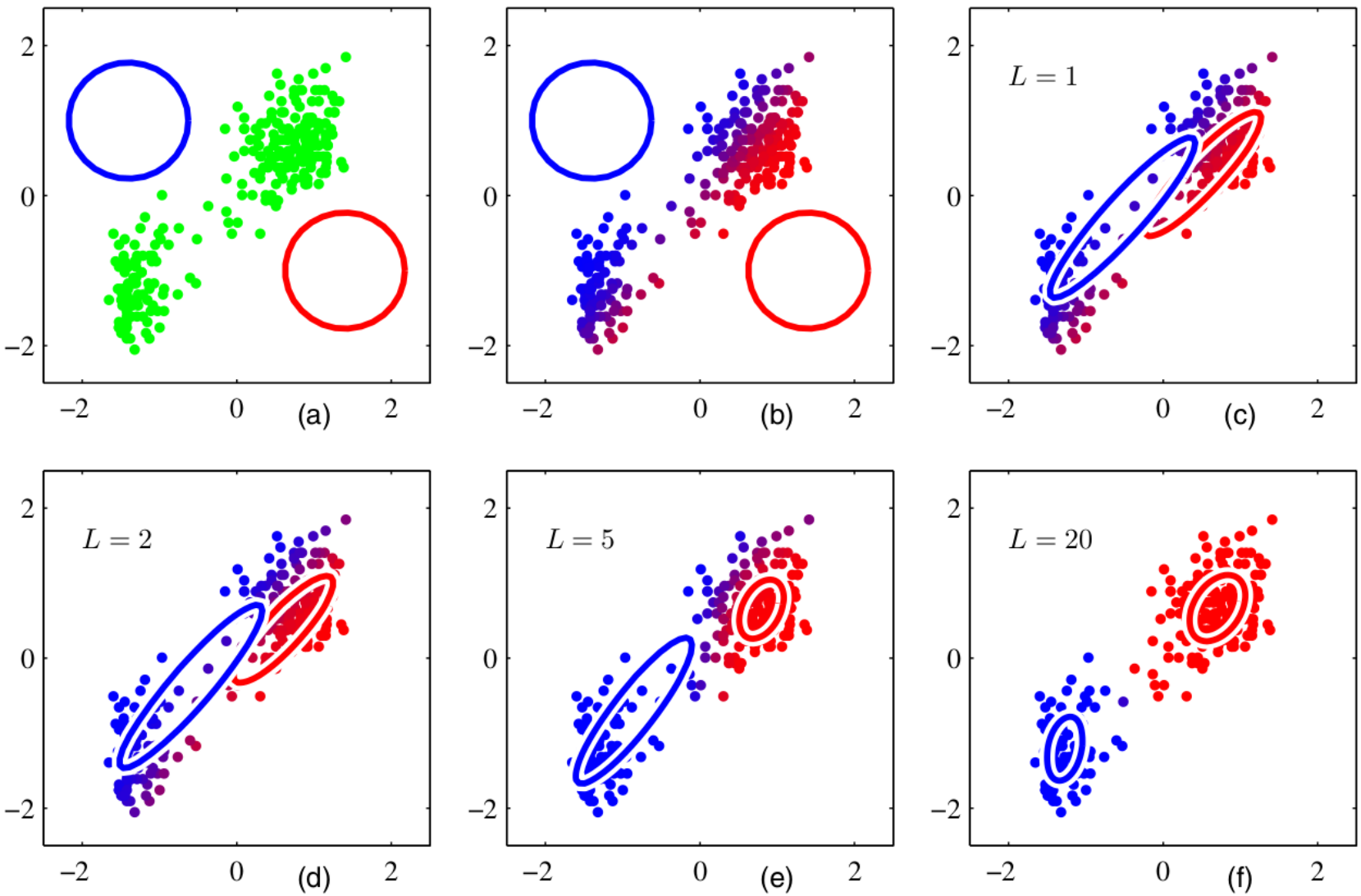
3: **Initialize:** $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$, and π_t, $\forall t \in [k]$.

4: **E** step. Evaluate the responsibilities using the current parameter values.

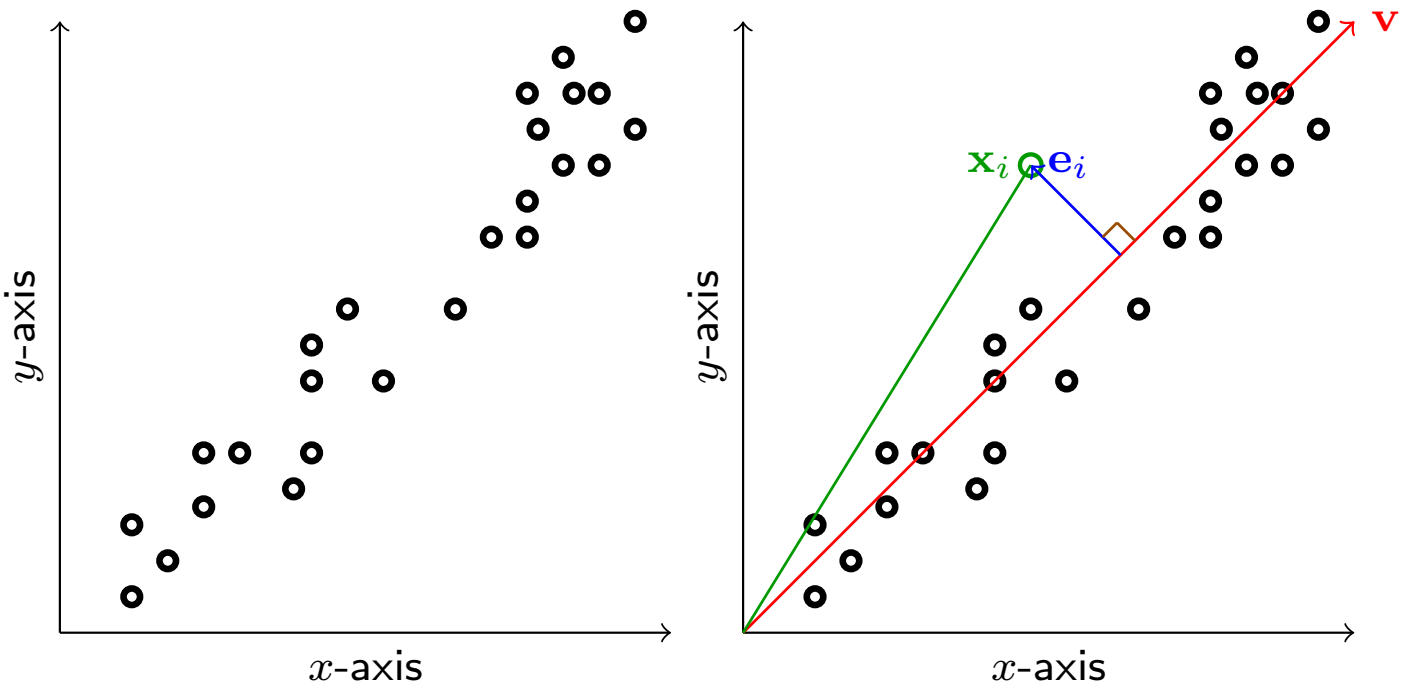
$$\gamma(z_{jt}) \leftarrow \frac{\pi_t \mathcal{N}(\mathbf{x}_j | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}.$$

5: **M** step. Re-estimate the parameters using the current responsibilities.

$$\begin{aligned} \boldsymbol{\mu}_t^{\text{new}} &\leftarrow \frac{1}{n_t} \sum_{j \in [n]} \gamma(z_{jt}) \mathbf{x}_j \\ \boldsymbol{\Sigma}_t^{\text{new}} &\leftarrow \frac{1}{n_t} \sum_{j \in [n]} \gamma(z_{jt}) (\mathbf{x}_j - \boldsymbol{\mu}_t^{\text{new}})(\mathbf{x}_j - \boldsymbol{\mu}_t^{\text{new}})^\top \\ \pi_t^{\text{new}} &\leftarrow \frac{n_t}{n}. \text{ Here, } n_t = \sum_{j \in [n]} \gamma(z_{jt}). \end{aligned}$$


2
0
−2
−2
0
2
(a)
(b)
$L = 1$
(c)
$L = 2$
(d)
$L = 5$
(e)
$L = 20$
(f)

The principal component analysis algorithm can be used to find the direction of maximum variance for given set of points. For example, consider the data points given in below figure. We can use the PCA algorithm to find the direction (shown as a red colored vector) of maximum variance. Now, let us formally consider a set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of n points in $\mathbb{R}^2$. Let $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix} \in \mathbb{R}^{2\times n}$ be a matrix containing these points. Show that the unit norm vector $\mathbf{v} \in \mathbb{R}^2$ along the maximum variance direction is the eigenvector corresponding to the largest eigenvalue of the matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$.



- Find an expression for the vector $\mathbf{e}_i$ in terms of $\mathbf{x}_i$ and $\mathbf{v}$. See the below diagram for the geometric relation between these quantities.
- Show that the unit norm vector $\mathbf{v}$ will be along the maximum variance direction if the error $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{e}_i\|_2^2$ is minimized.
- Show that $\underset{\mathbf{v}\in\mathbb{R}^2,\mathbf{v}^\top\mathbf{v}=1}{\text{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{e}_i\|_2^2 = \underset{\mathbf{v}\in\mathbb{R}^2,\mathbf{v}^\top\mathbf{v}=1}{\text{argmax}}\ \frac{1}{n}\mathbf{v}^\top\mathbf{X}\mathbf{X}^\top\mathbf{v}$.
- Show that the optimal $\mathbf{v}$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$.