1. $\theta_1, \ldots \theta_K$ eigen vectors of kth largest eigen values of $C(\frac{1}{m}\phi\phi^T)$

θ_1 first principal eigen vector of C, $C\theta_1 = \lambda_1\theta_1$ and $\tilde{\phi}_i = (I - \theta_1\theta_1^T)\phi_i$

$$\tilde{\phi} = (I - \theta_1\theta_1^T)\phi \quad , \quad \theta_c^T\theta_d = \begin{cases} 0 & c \neq d \\ 1 & c = d \end{cases}$$

a) $\tilde{C} = \frac{1}{m}\tilde{\phi}\tilde{\phi}^T = \frac{1}{m}(I - \theta_1\theta_1^T)\phi\phi^T(I - \theta_1\theta_1^T)^T$

$= (I - \theta_1\theta_1^T) C (I - \theta_1\theta_1^T)$ [$(I - \theta_1\theta_1^T)$ is symmetric]

$= C - \theta_1\theta_1^T C - C\theta_1\theta_1^T + \theta_1\theta_1^T C\theta_1\theta_1^T$

Now, $C\theta_1 = \lambda_1\theta_1 \Rightarrow \theta_1^T C^T = \lambda_1\theta_1^T$ replacing in the eq^n

$\tilde{C} = C - \lambda_1\theta_1\theta_1^T - C\theta_1\theta_1^T + \lambda\theta_1\theta_1^T\theta_1\theta_1^T$ (now, $\theta_1^T\theta_1 = I$)

$$\boxed{\tilde{C} = C - \lambda_1\theta_1\theta_1^T}$$

b) We have, $\tilde{C}\theta_j = (C - \lambda_1\theta_1\theta_1^T)\theta_j = \lambda_j\theta_j - \lambda_1\theta_1\theta_1^T\theta_j$

Now if $j = 1$, $\tilde{C}\theta_1 = \lambda_1\theta_1 - \lambda_1\theta_1(\theta_1^T\theta_1) = (\lambda_1 - \lambda_1)\theta_1 = 0$

$j \neq 1$, $\tilde{C}\theta_j = \lambda_j\theta_j$ [$\theta_1^T\theta_j = 0$, $j \neq 1$]

Since $\theta_1, \theta_2, \ldots \theta_K$ are the first K eigen vectors with largest eigen values in C, therefore $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$.

Now for $\tilde{C}$, θ_j are the eigenvectors with eigen-values $(0, \lambda_2, \lambda_3, \ldots, \lambda_K)$

Therefore θ_2 is the eigen vector with the largest eigen value of $\tilde{C}$.

c)

```
findKEigenVectors(C, K, f}, lambda_list, eigen_list)
if lambda_list.length = K, return;  <---- base case      recursive
    lambda, u = f(C)
    C = C - lambda * u * u.Transpose
    lambda_list.append(lambda)
    eigen_list.append(u)
    findKEigenvectors(C, K, f, lambda_list, eigen_list)
```

2. Initialize Step :

Initialize μ_c, π_c and Σ matrix .

μ_c = mean feature vector for class c $\quad$ π_c = Probability of y taking cluster c

Σ = covariance matrix for all classes .

Repeat :

① Expectation Step :

The superscript (m) denotes the variable in m-th iteration .

We calculate the expectations, $\gamma_{ic}^{(m)}$ based on $\mu_c^{(m-1)}$, $\pi_c^{(m-1)}$ and $\Sigma^{(m-1)}$

$$\gamma_{ic} = E[y_i|x_i] = \frac{\pi_c^{(m-1)} \mathcal{N}(x_i | \mu_c^{(m-1)}, \Sigma^{(m-1)})}{\sum_{c=1}^{K} \pi_c^{(m-1)} \mathcal{N}(x_i | \mu_c^{(m-1)}, \Sigma^{(m-1)})}$$

② Maximization Step :

$$\mathcal{L} = E[\log P(x, y | \pi, \mu, \Sigma)] = \sum_{i=1}^{n} \sum_{c=1}^{k} \gamma_{ic} \{ \log \pi_c + \log \mathcal{N}(x_i | \mu_c, \Sigma) \}$$

Taking the derivative w.r.t π_c, μ_c and Σ^{-1} and setting those to 0, we can get the $\pi_c^{(m)}$, $\mu_c^{(m)}$ and $\Sigma^{(m)}$.

The second term in the $\mathcal{L}$ eduation could be,

$$\log \mathcal{N}(x_i | \mu_c, \Sigma) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_i - \mu_c)^T \Sigma^{-1} (x_i - \mu_c)$$

a) Taking derivative w.r.t to μ_c

$$\frac{\partial \mathcal{L}}{\partial \mu_c} = \sum_{i=1}^{n} \gamma_{ic} \frac{\partial (\log \mathcal{N}(x_i | \mu_c, \Sigma))}{\partial \mu_c}$$

$$0 = \sum_{i=1}^{n} \gamma_{ic} \frac{\partial (-\frac{1}{2} (x_i - \mu_c^{(m)})^T \Sigma^{-1} (x_i - \mu_c^{(m)}))}{\partial \mu_c}$$

$$\sum_{i=1}^{n} \gamma_{ic} \Sigma^{-1} (x_i - \mu_c^{(m)}) = 0 \Rightarrow \sum_{i=1}^{n} \gamma_{ic} x_i = \sum_{i=1}^{n} \gamma_{ic} \mu_c^{(m)}$$

$$\therefore \boxed{\mu_c^{(m)} = \frac{\sum_{i=1}^{n} \gamma_{ic} x_i}{\sum_{i=1}^{n} \gamma_{ic}}}$$

b) Taking derivative w.r.t Σ^{-1}

$$\frac{\partial \mathcal{L}}{\partial \Sigma^{-1}} = \sum_{i=1}^{n} \sum_{c=1}^{K} r_{ic} \left[\left(-\frac{1}{2} \frac{\partial}{\partial \Sigma^{-1}} \log |\Sigma| \right) - \frac{1}{2} \frac{\partial}{\partial \Sigma^{-1}} \left((x_i - \mu_c)^T \Sigma^{-1} (x_i - \mu_c) \right) \right]$$

(1) Since $x^T \Sigma^{-1} x$ is scalar, hence can be rewritten as $tr[x x^T \Sigma^{-1}]$

and $\frac{\partial}{\partial \Sigma^{-1}} tr[x x^T \Sigma^{-1}] = (x x^T)^T = x x^T$.

(2) $\frac{\partial}{\partial \Sigma^{-1}} (-\log |\Sigma|) = \frac{\partial}{\partial \Sigma^{-1}} \log |\Sigma^{-1}| = (\Sigma)^T = \Sigma$.

So, $$0 = \sum_{i=1}^{n} \sum_{c=1}^{K} \left[r_{ic} \left(\frac{1}{2} \Sigma^{(m)} - \frac{1}{2} (x_i - \mu_c)(x_i - \mu_c)^T \right) \right]$$

So, $$\Sigma^{(m)} = \frac{\sum_{i=1}^{n} \sum_{c=1}^{K} r_{ic} (x_i - \mu_c)(x_i - \mu_c)^T}{\sum_{i=1}^{n} \sum_{c=1}^{K} r_{ic}}$$

c) There's a constraint in maximizing π_c, $\sum_{c=1}^{K} \pi_c = 1$

Using Lagrange multiplier, $\mathcal{L}_\pi = \mathcal{L} + \lambda \left(1 - \sum_{c=1}^{K} \pi_c \right)$

$$\frac{\partial \mathcal{L}_\pi}{\partial \pi_c} = 0. \Rightarrow \sum_{i=1}^{n} r_{ic} / \pi_c^{(m)} - \lambda = 0$$

$$\pi_c^{(m)} = 1/\lambda \sum_{i=1}^{n} r_{ic}.$$

now $\sum_{c=1}^{K} 1/\lambda \sum_{i=1}^{n} r_{ic} = 1 \Rightarrow \lambda = \sum_{c=1}^{K} \sum_{i=1}^{n} r_{ic} = \sum_{i=1}^{n} 1 = n.$

[$\sum_{c=1}^{K} r_{ic} = 1$, sum of probabilities of a point being assigned to all the clusters is 1]

$$\therefore \pi_c^{(m)} = 1/n \sum_{i=1}^{n} r_{ic}$$

With $\mu_c^{(m)}$, $\Sigma^{(m)}$ and $\pi_c^{(m)}$ begin next iteration.