CSE – 426 Homework 7

Griffin Kent

1) Prove that if $\frac{p(x)}{q(x)} = c$ for any x and for two distributions p and q over the same domain for x, then $c = 1$.

(Proof): By the definition of a probability distribution, for both p and q, we know that $\sum_x p(x) = 1$ and $\sum_x q(x) = 1$; that is that $p(x_1) + p(x_2) + \cdots + p(x_n) = 1$ and $q(x_1) + q(x_2) + \cdots + q(x_n) = 1$ for all x_i in the domain of x. Given that $\frac{p(x)}{q(x)} = c$, then we can see that the following relationship must hold:

$$\frac{p(x_1)}{q(x_1)} = \frac{p(x_2)}{q(x_2)} = \cdots = \frac{p(x_n)}{q(x_n)} = c.$$

If we let $p(x_i) = q(x_i)c,\ \forall i = 1,2,\ldots,n$, then we can see that

$$\begin{aligned} &p(x_1) + p(x_2) + \cdots + p(x_n) \\ &= q(x_1)c + q(x_2)c + \cdots + q(x_n)c \\ &= c\big(q(x_1) + q(x_2) + \cdots + q(x_n)\big) \\ &= c \cdot 1 = c. \end{aligned}$$

Since we know that $\sum_x p(x) = 1$, we have shown $\sum_x p(x) = c = 1$. Therefore, if $\frac{p(x)}{q(x)} = c$ for any x, then $c = 1$.

■

2) For the Gaussian mixture model (GMM) with K Gaussian components, let the incomplete-data log-likelihood of an observed data point $\boldsymbol{x} \in \mathbb{R}^n$ be

$$\log \mathbb{P}(\boldsymbol{x}|\pi, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_k) = \log\left\{\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)\right\},$$

Where $\boldsymbol{\mu}_k, \Sigma_k$ are the mean and covariance of the k-th Gaussian, which has density $N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$. To estimate the mean vector during the EM algorithm of GMM, you will need to find the partial derivative of the log-likelihood with respect to $\boldsymbol{\mu}_k$. Prove that the partial derivative is

$$\frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k).$$

(Proof): To begin, we know that the Gaussian distribution above is defined as follows:

$$N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}.$$

Now, taking the partial derivative of the log-likelihood function with respect to $\boldsymbol{\mu}_k$, we have

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_k} = \frac{\partial}{\partial \boldsymbol{\mu}_k} \log\left\{\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)\right\}$$

$$= \frac{1}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k) \right\}$$

$$= \frac{1}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_k} \{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)\}$$

$$= \frac{1}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ \pi_k \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)} \right\}$$

$$= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \left(-\frac{1}{2}\right) \frac{\partial}{\partial \boldsymbol{\mu}_k} \{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)\}$$

Now, using the properties of Trace and Eq. (108) from the matrix cookbook, we have

$$= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \left(-\frac{1}{2}\right) \frac{\partial}{\partial \boldsymbol{\mu}_k} Tr\{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)\}$$

$$= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \left(-\frac{1}{2}\right) \left\{ \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k) \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}-\boldsymbol{\mu}_k) + \Sigma_k^{-T} (\boldsymbol{x}-\boldsymbol{\mu}_k) \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}-\boldsymbol{\mu}_k) \right\}$$

Since the covariance matrix Σ is symmetric, we have

$$= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \cdot \left(-\frac{1}{2}\right) \{-2\Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)\}$$

$$= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)} \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k).$$

Therefore, completing the proof.

■

3) Given data points $\boldsymbol{x}^{(1)} = [1,1]^T$, $\boldsymbol{x}^{(2)} = [1,-1]^T$, and $\boldsymbol{x}^{(3)} = [2,2]^T$, compute the covariance matrix $S = \frac{1}{3}\sum_{i=1}^{3} \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T}$. Then calculate all eigenvalues by writing down and solving the eigenvalue equation

$$\det(S - \lambda I_{2\times 2}) = 0.$$

(Solution): Using the given formula, we can calculate the covariance matrix of our data to be

$$S = \frac{1}{3}\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1,1] + \begin{bmatrix} 1 \\ -1 \end{bmatrix} [1,-1] + \begin{bmatrix} 2 \\ 2 \end{bmatrix} [2,2] \right\}$$

$$= \frac{1}{3}\left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \right\}$$

$$= \frac{1}{3}\begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix} = \begin{bmatrix} 2 & \frac{4}{3} \\ \frac{4}{3} & 2 \end{bmatrix}.$$

Now that we have S, we can find the eigenvalues:

$$\det(S-\lambda I_{2\times 2}) = \det\begin{bmatrix} 2-\lambda & \frac{4}{3} \\ \frac{4}{3} & 2-\lambda \end{bmatrix}$$

$$= (2-\lambda)(2-\lambda) - \left(\frac{4}{3}\right)^2 = 4 - 4\lambda + \lambda^2 - \frac{16}{9} = \lambda^2 - 4\lambda + \frac{20}{9}.$$

Using the quadratic equation, we have

$$\lambda = \frac{4 \pm \sqrt{16 - \frac{80}{9}}}{2} = \frac{10}{3}, \frac{2}{3}.$$

Therefore, the two eigenvalues are $\lambda_1 = \frac{10}{3}$, and $\lambda_2 = \frac{2}{3}$.

■

4) Let a basis be $B = [\boldsymbol{b}_1, \dots, \boldsymbol{b}_k] \in \mathbb{R}^{n\times k}$ and the projection matrix projecting any $\boldsymbol{x} \in \mathbb{R}^n$ to B be

$$P_B = B(B^T B)^{-1} B^T.$$

Prove that the product $P_B P_B$ is just P_B.

(Proof):

$$P_B P_B = [B(B^T B)^{-1} B^T][B(B^T B)^{-1} B^T]$$
$$= B(B^T B)^{-1} B^T B (B^T B)^{-1} B^T$$

Since a matrix multiplied by its inverse is the identity matrix, we have

$$= B(B^T B)^{-1} I B^T$$
$$= B(B^T B)^{-1} B^T = P_B.$$

Therefore, completing the proof.

■

5) (Graduate Only) Consider there are $K = 3$ topics that a piece of news may belong to. The news is a bag of words and can be represented by a vector $\boldsymbol{x}$ of n binary variables, where $x_j = 1$ if and only if the j-th word appears in the news, $j = 1, \dots, n$. The news word vector $\boldsymbol{x}$ is generated by first sampling one of the three topics. Denote the selected topic as k. Then sample each of the n words independently according to the Bernoulli distribution $x_j \sim \text{Bernoulli}(\mu_{kj})$ for the k-th topic. Note that μ_{kj} may be different for different k. Let the n-dimensional random column vector $\boldsymbol{x} \in \{0,1\}^n$ be generated from the mixture of K Bernoulli distributions as follows:

$$\mathbb{P}(\boldsymbol{z} = \boldsymbol{e}_k | \boldsymbol{\pi}) = \pi_k, \qquad (5)$$

$$\mathbb{P}(\boldsymbol{x} | \boldsymbol{z} = \boldsymbol{e}_k; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = \prod_{j=1}^{n} \mu_{kj}^{x_j} (1-\mu_{kj})^{(1-x_j)}, \qquad (6)$$

Where the mixture parameters $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$ and the K Bernoulli vectors $\boldsymbol{\mu}_k = [\mu_{k1}, \dots, \mu_{kn}]^T \in [0,1]^n$ are given. In other words, conditioned on k being sampled using Eq (5), x_j is sampled from a Bernoulli distribution with mean μ_{kj} (μ_{ki} is not necessarily equal to μ_{kj}).

$\boldsymbol{e}_k \in \{0,1\}^K$ is a binary one-hot vector with the k-th entry being 1 and the remaining entries being 0. Prove that the mean and covariance matrix of the random vector $\boldsymbol{x}$ are:

$$\mathbb{E}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k,$$

$$\mathrm{cov}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \{\Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\} - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T.$$

Where Σ_k is a diagonal matrix of size $n \times n$, with the i-th diagonal element of Σ_k being $\mu_{ki}(1-\mu_{ki})$.

[*Hints: By definition, the mean of a random vector is the vector with each element being the mean of the corresponding random variable in the random vector:* $\mathbb{E}[\boldsymbol{x}]_j = \mathbb{E}[x_j]$ *for* $j = 1, \dots, n$. *The* (i,j)*-th entry of the covariance matrix* $\mathrm{cov}[\boldsymbol{x}]$ *is the covariance of the two random variables* x_i *and* x_j. *Please find the mean vector first and you will need to use the total probability* $\mathbb{P}(x_j) = \sum_{\boldsymbol{z}} \mathbb{P}(x_j, \boldsymbol{z})$ *for any value of* x_j. *The covariance matrix is more difficult and another total probability equation, possibly involving* x_i *and* x_j, *is needed.*]

(Proof): We have the following

$$\mathbb{P}(x_j) = \sum_{\boldsymbol{z}} \mathbb{P}(x_j, \boldsymbol{z}) = \sum_{\boldsymbol{z}} \mathbb{P}(\boldsymbol{z} = \boldsymbol{e}_k | \boldsymbol{\pi}) \mathbb{P}(x_j | \boldsymbol{z} = \boldsymbol{e}_k; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$$
$$= \sum_{k=1}^{K} \pi_k \mu_{kj}^{x_j} (1-\mu_{kj})^{(1-x_j)}.$$

Taking the expectation of $\boldsymbol{x}$, we have

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}[x_1] \\ \mathbb{E}[x_2] \\ \vdots \\ \mathbb{E}[x_n] \end{bmatrix} = \begin{bmatrix} \sum_{x_1} x_1 \mathbb{P}(x_1) \\ \sum_{x_2} x_2 \mathbb{P}(x_2) \\ \vdots \\ \sum_{x_n} x_n \mathbb{P}(x_n) \end{bmatrix} = \begin{bmatrix} \sum_{x_1} x_1 \sum_{k=1}^{K} \pi_k \mu_{k1}^{x_1} (1-\mu_{k1})^{(1-x_1)} \\ \sum_{x_2} x_2 \sum_{k=1}^{K} \pi_k \mu_{k2}^{x_2} (1-\mu_{k2})^{(1-x_2)} \\ \vdots \\ \sum_{x_n} x_n \sum_{k=1}^{K} \pi_k \mu_{kn}^{x_n} (1-\mu_{kn})^{(1-x_n)} \end{bmatrix}$$

$$= \begin{bmatrix} 0 \sum_{k=1}^{K} \pi_k \mu_{k1}^{x_1=0} (1-\mu_{k1})^{(1-(x_1=0))} + 1 \sum_{k=1}^{K} \pi_k \mu_{k1}^{x_1=1} (1-\mu_{k1})^{(1-(x_1=1))} \\ 0 \sum_{k=1}^{K} \pi_k \mu_{k2}^{x_2=0} (1-\mu_{k2})^{(1-(x_2=0))} + 1 \sum_{k=1}^{K} \pi_k \mu_{k2}^{x_2=1} (1-\mu_{k2})^{(1-(x_2=1))} \\ \vdots \\ 0 \sum_{k=1}^{K} \pi_k \mu_{k3}^{x_3=0} (1-\mu_{k3})^{(1-(x_3=0))} + 1 \sum_{k=1}^{K} \pi_k \mu_{k3}^{x_3=1} (1-\mu_{k3})^{(1-(x_3=1))} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} \pi_k \mu_{k1} \\ \sum_{k=1}^{K} \pi_k \mu_{k2} \\ \vdots \\ \sum_{k=1}^{K} \pi_k \mu_{k3} \end{bmatrix}$$

$$= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k.$$

■

Now, for the covariance:

We know that the covariance matrix of a random vector $x \in \mathbb{R}^n$ is the symmetric matrix $C \in \mathbb{R}^{n\times n}$ where the (i,j)-th entry is defined as

$$C_{ij} \triangleq cov(x_i, x_j) = \mathbb{E}_{x_i x_j}[x_i x_j] - \mathbb{E}_{x_i}[x_i]\mathbb{E}_{x_j}[x_j].$$

In matrix-vector notation, this is equivalent to

$$\text{cov}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T.$$

The expected value of the product of two discrete random variables $\mathbb{E}_{x_i x_j}[x_i x_j]$, for $i \neq j$, is defined as

$$\mathbb{E}_{x_i x_j}[x_i x_j] \triangleq \sum_{x_j}\sum_{x_i} x_i x_j \mathbb{P}(x_i, x_j)$$

We can further derive the joint probability of x_i and x_j as

$$\mathbb{P}(x_i, x_j) = \sum_{\boldsymbol{z}} \mathbb{P}(x_i, x_j, \boldsymbol{z}) = \sum_{\boldsymbol{z}} \mathbb{P}(\boldsymbol{z} = \boldsymbol{e}_k|\boldsymbol{\pi})\mathbb{P}(x_i, x_j|\boldsymbol{z} = \boldsymbol{e}_k; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$$

$$= \sum_{k=1}^{K} \pi_k \left(\mu_{kj}^{x_j}(1-\mu_{kj})^{(1-x_j)}\right)\left(\mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}\right).$$

Since x_i and x_j are Bernoulli distributed random variables, substituting these expressions for $\mathbb{P}(x_i, x_j)$ and $\mathbb{P}(x_i, x_i)$ back into the joint expectations; for $i \neq j$, we have

$$\mathbb{E}_{x_i x_j}[x_i x_j] = \sum_{x_j}\sum_{x_i} x_i x_j \mathbb{P}(x_i, x_j)$$

$$= \sum_{x_j}\sum_{x_i} x_i x_j \left[\sum_{k=1}^{K} \pi_k \left(\mu_{kj}^{x_j}(1-\mu_{kj})^{(1-x_j)}\right)\left(\mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}\right)\right]$$

$$= \sum_{k=1}^{K} \pi_k \mu_{kj}\mu_{ki}.$$

And for $i = j$, we have

$$\mathbb{E}_{x_i x_i}[x_i x_i] = \sum_{x_i}\sum_{x_i} x_i x_i \mathbb{P}(x_i, x_i)$$

$$= \sum_{x_i}\sum_{x_i} x_i x_i \left[\sum_{k=1}^{K} \pi_k \left(\mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}\right)\left(\mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}\right)\right]$$

$$= \sum_{k=1}^{K} \pi_k \mu_{ki}\mu_{ki} = \sum_{k=1}^{K} \pi_k \mu_{ki}^2.$$

Looking back at the covariance matrix, we can expand out the $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}]$ term, which yields

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$$

$$= \begin{bmatrix} \mathbb{E}[x_1x_1] & \mathbb{E}[x_1x_2] & \cdots & \mathbb{E}[x_1x_n] \\ \mathbb{E}[x_2x_1] & \mathbb{E}[x_2x_2] & \cdots & \mathbb{E}[x_2x_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[x_nx_1] & \mathbb{E}[x_nx_2] & \cdots & \mathbb{E}[x_nx_n] \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} \pi_k \mu_{k1}^2 & \sum_{k=1}^{K} \pi_k \mu_{k1}\mu_{k2} & \cdots & \sum_{k=1}^{K} \pi_k \mu_{k1}\mu_{kn} \\ \sum_{k=1}^{K} \pi_k \mu_{k2}\mu_{k1} & \sum_{k=1}^{K} \pi_k \mu_{k2}^2 & \cdots & \sum_{k=1}^{K} \pi_k \mu_{k2}\mu_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{K} \pi_k \mu_{kn}\mu_{k1} & \sum_{k=1}^{K} \pi_k \mu_{kn}\mu_{k2} & \cdots & \sum_{k=1}^{K} \pi_k \mu_{kn}^2 \end{bmatrix}$$

$$= \sum_{k=1}^{K} \pi_k \begin{bmatrix} \mu_{k1}^2 & \mu_{k1}\mu_{k2} & \cdots & \mu_{k1}\mu_{kn} \\ \mu_{k2}\mu_{k1} & \mu_{k2}^2 & \cdots & \mu_{k2}\mu_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{kn}\mu_{k1} & \mu_{kn}\mu_{k2} & \cdots & \mu_{kn}^2 \end{bmatrix}$$

$$= \sum_{k=1}^{K} \pi_k \left\{ \begin{bmatrix} \mu_{k1}(1-\mu_{k1}) & 0 & \cdots & 0 \\ 0 & \mu_{k1}(1-\mu_{k1}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{kn}(1-\mu_{kn}) \end{bmatrix} + \begin{bmatrix} \mu_{k1}\mu_{k1} & \mu_{k1}\mu_{k2} & \cdots & \mu_{k1}\mu_{kn} \\ \mu_{k2}\mu_{k1} & \mu_{k2}\mu_{k2} & \cdots & \mu_{k2}\mu_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{kn}\mu_{k1} & \mu_{kn}\mu_{k2} & \cdots & \mu_{kn}\mu_{kn} \end{bmatrix} \right\}$$

$$= \sum_{k=1}^{K} \pi_k \{\Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\}$$

Where Σ_k is a diagonal matrix with the i-th diagonal element being $\mu_{ki}(1-\mu_{ki})$. Substituting this back into our covariance matrix, we have the desired form

$$\mathrm{cov}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \{\Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\} - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T.$$

■