1 PCA

For the following problems, we have N *zero-mean* data points $\mathbf{x}_i \in \mathbb{R}^{D\times 1}$ and $\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \in \mathbb{R}^{D\times D}$ is the sample covariance matrix of the dataset.

1.1 Derivation of Second Principal Component

(a) (**5 points**) Let cost function

$$J = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i - p_{i1}\mathbf{e}_1 - p_{i2}\mathbf{e}_2\right)^{\mathrm{T}}\left(\mathbf{x}_i - p_{i1}\mathbf{e}_1 - p_{i2}\mathbf{e}_2\right)$$

with $\mathbf{e}_1$ and $\mathbf{e}_2$ are the orthonormal vector basis for the dimensionality reduction, i.e. $\|\mathbf{e}_1\|_2 = 1$, $\|\mathbf{e}_2\|_2 = 1$, and $\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_2 = 0$, and some coefficients p_{i1} and p_{i2}.
Show that $\frac{\partial J}{\partial p_{i2}} = 0$ yields $p_{i2} = \mathbf{e}_2^{\mathrm{T}}\mathbf{x}_i$, i.e. the projection length of data point $\mathbf{x}_i$ along vector $\mathbf{e}_2$.
Answer:

$$\begin{aligned}
J &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i - p_{i1}\mathbf{e}_1 - p_{i2}\mathbf{e}_2\right)^{\mathrm{T}}\left(\mathbf{x}_i - p_{i1}\mathbf{e}_1 - p_{i2}\mathbf{e}_2\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{\mathrm{T}} - p_{i1}\mathbf{e}_1^{\mathrm{T}} - p_{i2}\mathbf{e}_2^{\mathrm{T}}\right)\left(\mathbf{x}_i - p_{i1}\mathbf{e}_1 - p_{i2}\mathbf{e}_2\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i + p_{i1}^2\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_1 + p_{i2}^2\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_2 - 2p_{i1}\mathbf{e}_1^{\mathrm{T}}\mathbf{x}_i + 2p_{i1}p_{i2}\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_2 - 2p_{i2}\mathbf{e}_2^{\mathrm{T}}\mathbf{x}_i\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i + p_{i1}^2(1) + p_{i2}^2(1) - 2p_{i1}\mathbf{e}_1^{\mathrm{T}}\mathbf{x}_i + 2p_{i1}p_{i2}(0) - 2p_{i2}\mathbf{e}_2^{\mathrm{T}}\mathbf{x}_i\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i + p_{i1}^2 + p_{i2}^2 - 2p_{i1}\mathbf{e}_1^{\mathrm{T}}\mathbf{x}_i - 2p_{i2}\mathbf{e}_2^{\mathrm{T}}\mathbf{x}_i\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J}{\partial p_{i2}} &= 2p_{i2} - 2\mathbf{e}_2^{\mathrm{T}}\mathbf{x}_i = 0\\
p_{i2} &= \mathbf{e}_2^{\mathrm{T}}\mathbf{x}_i
\end{aligned}$$

(b) (**5 points**) Show that the value of $\mathbf{e}_2$ that minimizes cost function

$$\tilde{J} = -\mathbf{e}_2^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2 + \lambda_2\left(\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_2 - 1\right) + \lambda_{12}\left(\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_1 - 0\right)$$

is given by the eigenvector associated with the second largest eigenvalue of $\boldsymbol{S}$.
λ_2 is the Lagrange Multiplier for equality constraint $\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_2 = 1$ and λ_{12} is the Lagrange Multiplier for equality constraint $\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_1 = 0$.
Hint: Recall that $\boldsymbol{S}\mathbf{e}_1 = \lambda_1\mathbf{e}_1$ ($\mathbf{e}_1$ is the normalized eigenvector associated with the largest eigenvalue λ_1 of $\boldsymbol{S}$) and $\frac{\partial \mathbf{y}^{\mathrm{T}}\boldsymbol{A}\mathbf{y}}{\partial \mathbf{y}} = \left(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right)\mathbf{y}$. Also notice that $\boldsymbol{S}$ is a symmetric matrix.
Answer:
Taking partial derivative of $\tilde{J}$ with respect to Lagrange Multiplier λ_2 yields:

$$\begin{aligned}\frac{\partial \tilde{J}}{\partial \lambda_2} &= \mathbf{e}_2^{\mathrm{T}}\mathbf{e}_2 - 1 = 0\\ \mathbf{e}_2^{\mathrm{T}}\mathbf{e}_2 &= 1\end{aligned}$$

Taking partial derivative of $\tilde{J}$ with respect to Lagrange Multiplier λ_{12} yields:

$$\frac{\partial \tilde{J}}{\partial \lambda_{12}} = \mathbf{e}_2^{\mathrm{T}}\mathbf{e}_1 = 0$$

Taking partial derivative of $\tilde{J}$ with respect to Lagrange Multiplier $\mathbf{e}_2$ yields:

$$\begin{aligned}\frac{\partial \tilde{J}}{\partial \mathbf{e}_2} &= -\left(\boldsymbol{S} + \boldsymbol{S}^{\mathrm{T}}\right)\mathbf{e}_2 + 2\lambda_2\mathbf{e}_2 + \lambda_{12}\mathbf{e}_1 = 0\\ &-2\boldsymbol{S}\mathbf{e}_2 + 2\lambda_2\mathbf{e}_2 + \lambda_{12}\mathbf{e}_1 = 0 \qquad (1)\end{aligned}$$

Pre-multiply (or left-multiply) the equation 1 with $\mathbf{e}_2^{\mathrm{T}}$ yields:

$$\begin{aligned}-2\mathbf{e}_2^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2 + 2\lambda_2\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_2 + \lambda_{12}\mathbf{e}_2^{\mathrm{T}}\mathbf{e}_1 &= 0\\ -2\mathbf{e}_2^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2 + 2\lambda_2(1) + \lambda_{12}(0) &= 0\\ \lambda_2 &= \mathbf{e}_2^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2\end{aligned}$$

Pre-multiply (or left-multiply) the equation 1 with $\mathbf{e}_1^{\mathrm{T}}$ yields:

$$\begin{aligned}-2\mathbf{e}_1^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2 + 2\lambda_2\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_2 + \lambda_{12}\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_1 &= 0\\ -2\mathbf{e}_1^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2 + 2\lambda_2(0) + \lambda_{12}(1) &= 0\\ \lambda_{12} &= 2\mathbf{e}_1^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_2\\ \lambda_{12} &= 2\mathbf{e}_1^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}}\mathbf{e}_2\\ \lambda_{12} &= 2\left(\boldsymbol{S}\mathbf{e}_1\right)^{\mathrm{T}}\mathbf{e}_2\\ \lambda_{12} &= 2\left(\lambda_1\mathbf{e}_1\right)^{\mathrm{T}}\mathbf{e}_2\\ \lambda_{12} &= 2\lambda_1\mathbf{e}_1^{\mathrm{T}}\mathbf{e}_2\\ \lambda_{12} &= 2\lambda_1(0)\\ \lambda_{12} &= 0\end{aligned}$$

Substituting $\lambda_{12} = 0$ into equation 1:

$$\begin{aligned}-2\boldsymbol{S}\mathbf{e}_2 + 2\lambda_2\mathbf{e}_2 + (0)\,\mathbf{e}_1 &= 0\\ \boldsymbol{S}\mathbf{e}_2 &= \lambda_2\mathbf{e}_2\end{aligned}$$

Thus, $\mathbf{e}_2$ is an eigenvector associated with eigenvalue λ_2 of $\boldsymbol{S}$.
Substituting $\lambda_2 = \mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2$ and $\lambda_{12} = 0$ into the definition of cost function $\tilde{J}$ yields:

$$\begin{aligned}
\tilde{J} &= -\mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 + \lambda_2 \left(\mathbf{e}_2^{\mathrm{T}} \mathbf{e}_2 - 1\right) + \lambda_{12} \left(\mathbf{e}_2^{\mathrm{T}} \mathbf{e}_1 - 0\right) \\
&= -\mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 + \mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 \left(\mathbf{e}_2^{\mathrm{T}} \mathbf{e}_2 - 1\right) + (0) \left(\mathbf{e}_2^{\mathrm{T}} \mathbf{e}_1 - 0\right) \\
&= -\mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 + \mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 \mathbf{e}_2^{\mathrm{T}} \mathbf{e}_2 - \mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 \\
&= -\mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 + \mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 (1) - \mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 \\
&= -\mathbf{e}_2^{\mathrm{T}} \boldsymbol{S} \mathbf{e}_2 \\
&= -\lambda_2
\end{aligned}$$

Thus, to minimize $\tilde{J}$, we should pick the maximum possible value for λ_2. Since λ_1 is the largest eigenvalue of $\boldsymbol{S}$, λ_2 should be the second largest eigenvalue of $\boldsymbol{S}$, and $\mathbf{e}_2$ is the eigenvector associated with eigenvalue λ_2.

1.2 Derivation of PCA Residual Error

(a) (**5 points**) Prove that for a data point $\mathbf{x}_i$:

$$\|\mathbf{x}_i - \sum_{j=1}^{K} p_{ij} \mathbf{e}_j\|_2^2 = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - \sum_{j=1}^{K} \mathbf{e}_j^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_j$$

Hint: The most common method to proof a mathematical equation of this flavor is by using mathematical induction. To perform a proof by mathematical induction in this case, first show that the equation above holds for the base case $K = 1$, and then using the assumption that the equation holds for $K = k - 1$, show that the equation also holds for $K = k$, for any $1 \le k \le D$.
Use the fact that $\mathbf{e}_j^{\mathrm{T}} \mathbf{e}_j = 1$ (length of eigenvector $\mathbf{e}_j$ is 1) and $\mathbf{e}_j^{\mathrm{T}} \mathbf{e}_m = 0$ for $j \neq m$ (eigenvectors are perpendicular each other for square symmetric matrix $\boldsymbol{S}$). Also, use definition $p_{ij} = \mathbf{e}_j^{\mathrm{T}} \mathbf{x}_i$.
Answer:
Proof by mathematical induction:

- Base case $K = 1$:

$$\begin{aligned}
\|\mathbf{x}_i - p_{i1} \mathbf{e}_1\|_2^2 &= \left(\mathbf{x}_i - p_{i1} \mathbf{e}_1\right)^{\mathrm{T}} \left(\mathbf{x}_i - p_{i1} \mathbf{e}_1\right) \\
&= \left(\mathbf{x}_i^{\mathrm{T}} - p_{i1} \mathbf{e}_1^{\mathrm{T}}\right) \left(\mathbf{x}_i - p_{i1} \mathbf{e}_1\right) \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - p_{i1} \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_1 - p_{i1} \mathbf{e}_1^{\mathrm{T}} \mathbf{x}_i + p_{i1}^2 \mathbf{e}_1^{\mathrm{T}} \mathbf{e}_1 \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - 2 p_{i1} \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_1 + p_{i1}^2 (1) \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - 2 \mathbf{e}_1^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_1 + \mathbf{e}_1^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_1 \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - \mathbf{e}_1^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_1
\end{aligned}$$

  Thus the equation holds for base case $K = 1$

- Now, assuming that the equation holds for $K = k - 1$, i.e.:

$$\|\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij} \mathbf{e}_j\|_2^2 = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - \sum_{j=1}^{k-1} \mathbf{e}_j^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{e}_j$$

we will show that the equation also holds for $K = k$, as follows:

$$
\begin{aligned}
\|\mathbf{x}_i - \sum_{j=1}^{k} p_{ij}\mathbf{e}_j\|_2^2 &= \left(\mathbf{x}_i - \sum_{j=1}^{k} p_{ij}\mathbf{e}_j\right)^{\mathrm{T}} \left(\mathbf{x}_i - \sum_{j=1}^{k} p_{ij}\mathbf{e}_j\right) \\
&= \left(\left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right) - p_{ik}\mathbf{e}_k\right)^{\mathrm{T}} \left(\left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right) - p_{ik}\mathbf{e}_k\right) \\
&= \left(\left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right)^{\mathrm{T}} - p_{ik}\mathbf{e}_k^{\mathrm{T}}\right) \left(\left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right) - p_{ik}\mathbf{e}_k\right) \\
&= \left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right)^{\mathrm{T}} \left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right) - \left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right)^{\mathrm{T}} p_{ik}\mathbf{e}_k \\
&- p_{ik}\mathbf{e}_k^{\mathrm{T}} \left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right) + p_{ik}^2\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_k \\
&= \|\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\|_2^2 - \left(\mathbf{x}_i^{\mathrm{T}} - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j^{\mathrm{T}}\right) p_{ik}\mathbf{e}_k - p_{ik}\mathbf{e}_k^{\mathrm{T}} \left(\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\right) + p_{ik}^2\,(1) \\
&= \|\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\|_2^2 - p_{ik}\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_k + \sum_{j=1}^{k-1} p_{ij}p_{ik}\mathbf{e}_j^{\mathrm{T}}\mathbf{e}_k - p_{ik}\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i + \sum_{j=1}^{k-1} p_{ij}p_{ik}\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_j + p_{ik}^2 \\
&= \|\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\|_2^2 - 2p_{ik}\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_k + \sum_{j=1}^{k-1} p_{ij}p_{ik}\,(0) + \sum_{j=1}^{k-1} p_{ij}p_{ik}\,(0) + p_{ik}^2 \\
&= \|\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\|_2^2 - 2p_{ik}\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_k + p_{ik}^2 \\
&= \|\mathbf{x}_i - \sum_{j=1}^{k-1} p_{ij}\mathbf{e}_j\|_2^2 - 2\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_k + \mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_k \\
&= \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{k-1} \mathbf{e}_j^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_j - \mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_k \\
&= \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{e}_j^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_j
\end{aligned}
$$

Thus the equation holds for any $1 \le K \le D$

(b) (**5 points**) Now show that

$$
J_K \triangleq \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\mathbf{e}_j^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_j\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\lambda_j
$$

Hint: recall that $\mathbf{e}_j^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_j = \lambda_j\mathbf{e}_j^{\mathrm{T}}\mathbf{e}_j = \lambda_j$

Answer:

$$
\begin{aligned}
J_K &\triangleq \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\mathbf{e}_j^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_j\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{e}_j^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_j \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\mathbf{e}_j^{\mathrm{T}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{e}_j \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\mathbf{e}_j^{\mathrm{T}}\boldsymbol{S}\mathbf{e}_j \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\lambda_j\mathbf{e}_j^{\mathrm{T}}\mathbf{e}_j \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\lambda_j
\end{aligned}
$$

(c) (**5 points**) If $K = D$ principal components are used, there is no truncation, so $J_D = 0$. Use this to show that the error from only using $K < D$ principal components is given by

$$
J_K = \sum_{j=K+1}^{D}\lambda_j
$$

Answer:
When $K = D$:

$$
\begin{aligned}
J_D = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{D}\lambda_j &= 0 \\
\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i &= \sum_{j=1}^{D}\lambda_j
\end{aligned}
$$

Thus for $K < D$:

$$
\begin{aligned}
J_K &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i - \sum_{j=1}^{K}\lambda_j \\
&= \left(\sum_{j=1}^{D}\lambda_j\right) - \left(\sum_{j=1}^{K}\lambda_j\right) \\
&= \sum_{j=K+1}^{D}\lambda_j
\end{aligned}
$$

1.3 A Real Example

(a) The eigenvectors and values are as follows:

$$u_1 = \begin{bmatrix} 0.22 \\ 0.41 \\ 0.88 \end{bmatrix}$$

$$u_2 = \begin{bmatrix} 0.25 \\ 0.85 \\ -0.46 \end{bmatrix}$$

$$u_3 = \begin{bmatrix} 0.94 \\ -0.32 \\ -0.08 \end{bmatrix}$$

$$\lambda_1 = 1626.52 \quad \lambda_2 = 128.99 \quad \lambda_3 = 7.10$$

(b) u_2 and u_3 can be omitted because the correspoding eigenvalues for these two directions are contributing a small amount to the total variation of the data. In fact u_1 accounts for $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = 92.8\%$ of the data vairation and u_2 acounts for 7.32% of variation in data.The remaining principal component, explaining only 0.40% of the data, is negligible compared to the first two.

(c) We might think of u_1 as giving a generalized notion of "size" that incorporates length, wingspan, and weight. Indeed, all three entries of u_1 have the same sign, indicating that birds with larger "size" tend to have larger length, wingspan, and weight.

2 Hidden Markov Model (25 Points)

In this problem, you will implement Hidden Markov Model. First, please read forward, backward, and Viterbi algorithm in the lecture note.

A simple DNA sequence is $\boldsymbol{O} = \overline{O_1 O_2 \cdots O_T}$, with each component O_i takes from $\{A, C, G, T\}$. Assume it is generated from a Hidden Markov Model controlled by a hidden variable X, which takes two possible states S_1, S_2.

This HMM has the following parameters $\boldsymbol{\Theta} = \{\pi_i, a_{ij}, b_{ik}\}$ for $i, j = 1, 2$ and $k \in \{A, C, G, T\}$:

- Initial state distribution π_i for $i = 1, 2$:

$$\pi_1 = P(X_1 = S_1) = 0.6; \pi_2 = P(X_1 = S_2) = 0.4.$$

- Transition probabilities $a_{ij} = P(X_{t+1} = S_j | X_t = S_i)$ for any $t \in \mathbb{N}^+$, $i = 1, 2$, and $j = 1, 2$:

$$a_{11} = 0.7, a_{12} = 0.3; a_{21} = 0.4, a_{22} = 0.6.$$

- Emission probabilities $b_{ik} = P(O_t = k | X_t = S_i)$ for any $t \in \mathbb{N}^+$, $i = 1, 2$, and $k \in \{A, C, G, T\}$:

$$b_{1A} = 0.4, b_{1C} = 0.2, b_{1G} = 0.3, b_{1T} = 0.1;$$

$$b_{2A} = 0.2, b_{2C} = 0.4, b_{2G} = 0.1, b_{2T} = 0.3;$$

Assume we have an observed sequence $\boldsymbol{O} = \overline{O_1 O_2 \cdots O_6} = ACCGTA$, please answer the following questions with step-by-step computations and explanation for full credits. Your code should return all following answers when we run it.

(a) (**5 points**) *Probability of an observed sequence.* Calculate $P(\boldsymbol{O}; \boldsymbol{\Theta})$.

(b) (**5 points**) *Filtering.* Calculate $P(X_6 = S_i | \boldsymbol{O}; \boldsymbol{\Theta})$ for $i = 1, 2$.

(c) (**5 points**) *Smoothing.* Calculate $P(X_4 = S_i | \boldsymbol{O}; \boldsymbol{\Theta})$ for $i = 1, 2$.

(d) (**5 points**) *Most likely explanation.* Compute $\boldsymbol{X} = \overline{X_1 X_2 \cdots X_6} = \arg\max_{\boldsymbol{X}} P(\boldsymbol{X} | \boldsymbol{O}; \boldsymbol{\Theta})$.

(e) (**5 points**) *Prediction.* Compute $P(O_7 | \boldsymbol{O}; \boldsymbol{\Theta})$. Then, which observation is most likely after $o_{1:6}$? ($O_7 = \arg\max_O P(O | \boldsymbol{O}; \boldsymbol{\Theta})$).

Answer:

(a) (**5 points**) $P(\boldsymbol{O}; \boldsymbol{\Theta}) = 0.0002738928(\text{log}: -8.20277376901)$.

$$\begin{aligned}
&\alpha_1(j) = P(X_1 = S_j, o_1) = P(o_1 | X_1 = S_j) P(X_1 = S_j) \\
&\alpha_t(j) = P(X_t = S_j, o_{1:t}) = P(o_t | X_t = S_j) \sum_i a_{ij} \alpha_{t-1}(i) \\
&P(o_{1:T}) = \sum_j \alpha_T(j)
\end{aligned}$$

(b) (**5 points**) $P(X_6 = S_1 | \boldsymbol{O}; \boldsymbol{\Theta}) = 0.67355987452$, $P(X_6 = S_2 | \boldsymbol{O}; \boldsymbol{\Theta}) = 0.32644012548$.

$$\begin{aligned}
&\beta_T(j) = 1 \text{ for any } j \\
&\beta_{t-1}(i) = P(o_{t:T} | X_{t-1} = S_i) = \sum_j \beta_t(j) a_{ij} P(o_t | X_t = S_j) \\
&\gamma_t(j) = P(X_t = S_j | o_{1:T}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_j' \alpha_t(j')\beta_t(j')}
\end{aligned}$$

(c) (**5 points**) $P(X_4 = S_1 | \boldsymbol{O}; \boldsymbol{\Theta}) = 0.705017437479$, $P(X_4 = S_2 | \boldsymbol{O}; \boldsymbol{\Theta}) = 0.294982562521$.

$$\gamma_t(j) = P(X_t = S_j | o_{1:T}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_j' \alpha_t(j')\beta_t(j')}$$

(d) (**5 points**) $\boldsymbol{X} = \overline{X_1 X_2 \cdots X_6} = \arg\max_{\boldsymbol{X}} P(\boldsymbol{X} | \boldsymbol{O}; \boldsymbol{\Theta}) = S_1 S_1 S_2 S_1 S_2 S_1$. $\delta_t(j)$ is the probability of the most likely path ending with j at time t.

$$\begin{aligned}
\delta_t(j) &= \max_{x_1, x_2, \cdots, x_{t-1}} P(X_1 = x_1, X_2 = x_2, \cdots, X_{t-1} = x_{t-1}, X_t = S_j, o_{1:t} | \Theta) \\
&= \max_i \delta_{t-1}(i) a_{ij} P(o_t | X_t = S_j)
\end{aligned}$$

Thus, $\arg\max_j \delta_t(j)$ tells which state is most likely at time t given $o_{1:t}$.

(e) (**5 points**) $O_7 = \arg\max_O P(O | \boldsymbol{O}; \boldsymbol{\Theta}) = A$.
$P(O_7 = A) = 0.3204$, $P(O_7 = C) = 0.2796$, $P(O_7 = G) = 0.2204$, $P(O_7 = T) = 0.1796$. $P(O_7 = A)$ is the highest probability

```
alpha:
[[  2.40000000e-01,   8.00000000e-02],
 [  4.00000000e-02,   4.80000000e-02],
```

```
 [  9.44000000e-03,   1.63200000e-02],
 [  3.94080000e-03,   1.26240000e-03],
 [  3.26352000e-04,   5.81904000e-04],
 [  1.84483200e-04,   8.94096000e-05]]
beta:
[[  7.99764000e-04,   1.02436800e-03],
       [  2.87580000e-03,   3.30960000e-03],
       [  1.22100000e-02,   9.72000000e-03],
       [  4.90000000e-02,   6.40000000e-02],
       [  3.40000000e-01,   2.80000000e-01],
       [  1.00000000e+00,   1.00000000e+00]]
gamma:
[[ 0.70079739,  0.29920261],
       [ 0.41998913,  0.58001087],
       [ 0.42083034,  0.57916966],
       [ 0.70501744,  0.29498256],
       [ 0.40512084,  0.59487916],
       [ 0.67355987,  0.32644013]]
delta:
[[  2.40000000e-01,   8.00000000e-02],
       [  3.36000000e-02,   2.88000000e-02],
       [  4.70400000e-03,   6.91200000e-03],
       [  9.87840000e-04,   4.14720000e-04],
       [  6.91488000e-05,   8.89056000e-05],
       [  1.93616640e-05,   1.06686720e-05]]
Prob (d):  [0 0 1 0 1 0]  1.93616640e-05
Prob (e):
Prob of next state: [0.6021, 0.3979]
Prob of next observatoin: [0.3204, 0.2796, 0.2204, 0.1796]
```