1 Principal Component Analysis

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 $\mathbf{e}_2$.

(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.

1.2 A Real Example

In a study a simple random sample of 100 bird species is collected. Three factors were measured: length (inches), wingspan (inches), and weight (ounces). Thus, $m = 3$ and $n = 100$. The covariance matrix S is calculated :

$$S = \begin{bmatrix} 91.43 & 171.92 & 297.99 \\ & 373.92 & 545.21 \\ & & 1297.26 \end{bmatrix}$$

As is customary, the entries below the diagonal were omitted, since the matrix is symmetric. Also, S was computed without dividing by $n-1$ (also a common practice).

(a) **(5 points)** Compute the eigenvalues and orthonormal eigenvectors.

(b) **(5 points)** Is there any of the orthonormal directions that can be omitted without loosing lot of information? If yes which one(s) and why?

(c) **(5 points)** How do you interpret the eigenvector(s) that contain(s) the most of information regarding this data?

2 Hidden Markov Model

In this problem, you will see how Hidden Markov Model generates sequences. 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.

(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 $o_{1:6}$? ($O_7 = \arg\max_O P(O | \boldsymbol{O}; \boldsymbol{\Theta})$).

3 Submission Instructions

Submission Instructions: You need to submit a soft copy and a hard copy of your solutions.

- All solutions must be typed into a `pdf` report (named `CSCI567_hw6_fall16.pdf`). If you choose handwriting instead of typing, you will get 40% points deducted.
- The soft copy should be a single `zip` file named `[lastname]_[firstname]_hw6_fall16.zip`. It should contain your `pdf` report (named `CSCI567_hw6_fall16.pdf`) having answers to all the problems, and the folder containing all your code. It must be submitted via Blackboard by **11:59pm** of the deadline date.
- The hard copy should be a printout of the report `CSCI567_hw6_fall16.pdf` and must be submitted to locker #19 at PHE building 1st floor by **5:00pm** of the deadline date.

Collaboration You may collaborate. However, collaboration has to be limited to discussion only and you need to write your own solutions and submit separately. You also need to list the names of people with whom you have discussed.