2.5 Exercises

Exercise 2.1 (Some Properties of PCA). Let $x$ be a random vector with covariance matrix $\Sigma_x$. Consider a linear transformation of $x$:

$$y = W^T x,$$  \hspace{1cm} (2.34)

where $y \in \mathbb{R}^d$ and $W$ is a $D \times d$ orthogonal matrix. Let $\Sigma_y = W^T \Sigma_x W$ be the covariance matrix for $y$. Show that

1. The trace of $\Sigma_y$ is maximized by $W = U_d$, where $U_d$ consists of the first $d$ (normalized) eigenvectors of $\Sigma_x$.
2. The trace of $\Sigma_y$ is minimized by $W = \tilde{U}_d$, where $\tilde{U}_d$ consists of the last $d$ (normalized) eigenvectors of $\Sigma_x$.

Exercise 2.2 (Subspace Angles). Given two $d$-dimensional subspaces $S_1$ and $S_2$ in $\mathbb{R}^d$, define the largest subspace angle $\theta_i$ between $S_1$ and $S_2$ to be the largest possible angle between any two vectors in $S_1$ and $S_2$ respectively. Let $U_1 \in \mathbb{R}^{D \times d}$ be an orthogonal matrix whose columns form a basis for $S_1$ and similarly $U_2$ for $S_2$. Then show that if $\sigma_1$ is the largest singular value of the matrix $W = U_1^T U_2$, then we have

$$\cos(\theta_i) = \sigma_1.$$  \hspace{1cm} (2.35)

Similarly, one can define the rest of the subspace angles as $\cos(\theta_i) = \sigma_i, i = 2, \ldots, d$ from the rest of the singular values of $W$.

Exercise 2.3 (Fixed-Rank Approximation of a Matrix). Given an arbitrary full-rank matrix $A \in \mathbb{R}^{m \times n}$, find the matrix $B \in \mathbb{R}^{m \times n}$ with a fixed rank $r < \min\{m, n\}$ such that the Frobenius norm $\|A - B\|_F$ is minimized. The Frobenius norm of a matrix $M$ is defined to be $\|M\|_F^2 = \text{trace}(M^T M)$. (Hint: Use the SVD of $A$ to guess the matrix $B$ and then prove its optimality.)

Exercise 2.4 (Identification of Auto-Regressive Exogenous (ARX) Systems). A popular model that people use to analyze a time series $\{y_t\}_{t \in \mathbb{Z}}$ is the linear auto-regressive model:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_n y_{t-n} + \epsilon_t, \quad \forall t, y_t \in \mathbb{R},$$  \hspace{1cm} (2.36)

where $\epsilon_t \in \mathbb{R}$ models the modeling error or noise and it is often assumed to be a white-noise random process. Now suppose that you are given the values of $y_t$ for a sufficiently long period of time.

1. Show that in the noise free case, i.e. $\epsilon_t \equiv 0$, regardless of the initial conditions, the vectors $x_t = [y_t, y_{t-1}, \ldots, y_{t-n}]^T$ for all $t$ lie on an $n$-dimensional hyperplane in $\mathbb{R}^{n+1}$. What is the normal vector to this hyperplane?
2. Now consider the case with noise. Describe how you may use PCA to identify the unknown model parameters $(a_1, a_2, \ldots, a_n)$?

Exercise 2.5 (Basis for an Image). Given a gray-level image $I$, consider all of its $b \times b$ blocks, denoted as $\{B_i \in \mathbb{R}^{b \times b}\}$. We would like to approximate each block as a superposition of $d$ base blocks, say $\{\tilde{B}_j \in \mathbb{R}^{b \times b}\}_{j=1}^d$. That is,

$$B_i = \sum_{j=1}^d a_{ij} \tilde{B}_j + E_i,$$  \hspace{1cm} (2.37)
where $E_i \in \mathbb{R}^{b \times b}$ is the possible residual from the approximation. Describe how you can use PCA to identify an optimal set of $d$ base blocks so that the residual is minimized.

In Section 1.2.1, we have seen an example in which a similar process can be applied to an ensemble of face images, where the first $d = 3$ principal components are computed for further classification. In the computer vision literature, the corresponding base images are called “eigen faces.”

**Exercise 2.6 (Probability of Selecting a Subset of Inliers).** Imagine we have 80 samples from a four-dimensional subspace in $\mathbb{R}^5$. However, the samples are contaminated with another 20 samples that are far from the subspace. We want to estimate the subspace from randomly drawn subsets of four samples. In order to be of probability 0.95 that one of the subsets contains only inliers, what is the smallest number of subsets that we need to draw?

**Exercise 2.7 (Ranking of Webpages).** PCA is actually used to rank webpages on the Internet by many popular search engines. One way to see this is to view the Internet as a directed graph $G = (V, E)$, where every webpage, denoted as $p_i$, is a node in $V$, and every hyperlink from $p_i$ to $p_j$, denoted as $e_{ij}$, is a directed edge in $E$. We can assign each webpage $p_i$ an “authority” score $x_i$ that indicates how many other webpages point to it and a “hub” score $y_i$ that indicates how many other webpages it points out to. Then, the authority score $x_i$ depends on how many hubs point to $p_i$ and the hub score $y_i$ depends on how many authorities $p_i$ points to. Let $L$ be the adjacent matrix of the graph $G$ (i.e. $L_{ij} = 1$ if $e_{ij} = E$), $x$ the vector of the authority scores and $y$ of the hub scores.

1. Justify that the following relationships hold:
   \[ y' = Lx, \quad x' = L^T y, \quad x = x'/\|x'\|, \quad y = y'/\|y'\|. \]  
   (2.38)

2. Show that $x$ is the eigenvector of $L^T L$ and $y$ is the eigenvector of $L L^T$ associated with the largest eigenvalue (why not the others). Explain how $x$ and $y$ can be computed from the singular value decomposition of $L$.

In the literature, this is known as the **HyperText Induced Topic Selection (HITS)** algorithm [Kleinberg, 1999, Ding et al., 2004]. In fact, the same algorithm can also be used to rank any competitive sports such as football teams and chess players.

**Exercise 2.8 (Karhunen-Loève Transform).** The Karhunen-Loève transform (KLT) can be thought as a generalization of PCA from a random vector $x \in \mathbb{R}^D$ to a random process $x(t), t \in \mathbb{R}$. When $x(t)$ is a (zero-mean) second-order stationary random process, its auto correlation function is defined to be $K(t, \tau) = E[x(t)x(\tau)]$ for all $t, \tau \in \mathbb{R}$.

1. Show that $K(t, \tau)$ has a family of orthonormal eigen-functions \( \{ \phi_i(t) \}_{i=1}^{\infty} \) that are defined as
   \[ \int K(t, \tau)\phi_i(\tau) \, d\tau = \lambda_i \phi_i(t), \quad i = 1, 2, \ldots \]  
   (2.39)
   (Hint: First show that $K(t, \tau)$ is a positive definite function and then use Mercer’s Theorem.)

2. Show that with respect to the eigen-functions, we original random process can be decomposed as
   \[ x(t) = \sum_{i=1}^{\infty} x_i \phi_i(t), \]  
   (2.40)
   where \( \{ x_i \}_{i=1}^{\infty} \) are a set of uncorrelated random variables.