Math 123 **Problem Set 3** **Due: Sunday 3/6 at 11:59 p.m. EST**

Instruction: Read the homework policy. For problems 5 and 6, include printed copies of your code with your final homework submission. You should submit a PDF copy of the homework and any associated codes on Gradescope. Your PDF must be a single file.

1. Given n data points $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ lying in $\mathcal{R}^d$, the covariance matrix is defined as follows:

$$\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$$

(a) Argue that $\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ where $\mathbf{X}$ is the n by d data matrix defined as

$$\mathbf{X} = \begin{pmatrix} \dots & \mathbf{x}_1^T & \dots \\ \dots & \mathbf{x}_2^T & \dots \\ & \vdots & \\ \dots & \mathbf{x}_n^T & \dots \end{pmatrix}$$

(b) Prove that the covariance matrix is positive semi-definite. [**Remark**: A matrix $\mathbf{A} \in \mathcal{R}^{n\times n}$ is positive semi-definite if $\mathbf{y}^T\mathbf{A}\mathbf{y} \geq 0$ for all $\mathbf{y} \in \mathcal{R}^n$.]

(c) Prove that all eigenvalues of the covariance matrix are non-negative. [**Hint**: Consider the quadratic form.]

(d) The covariance matrix is not necessarily positive definite. Let $d = 4$. Describe or give an example of data for which the covariance matrix is positive semi-definite but not positive definite.

(e) Let $d = 4$. Describe or give an example of data for which the covariance matrix is positive definite.

[**Remark**: For parts (d) and (e), data point must be not trivial e.g. $(0, 0, 0, 0)$].

2. Consider n data points $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ lying in $\mathcal{R}^d$. The mean of the data points is defined as $\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$. Let the points $\{\mathbf{x}_1 = \mathbf{y}_1 - \mu, \mathbf{x}_2 = \mathbf{y}_2 - \mu, ..., \mathbf{x}_n = \mathbf{y}_n - \mu\}$ denote the centered points i.e. the points have zero mean (centered at origin). Let $\mathbf{v}$ be a unit vector in $\mathcal{R}^d$ and let $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_n$ denote respectively the projection of the centered points onto the vector $\mathbf{v}$. The representation of the projections with respect to the coordinate $\mathbf{v}$ is $c_1 = (\mathbf{x}_1)^T\mathbf{v}, c_2 = (\mathbf{x}_2)^T\mathbf{v}, ..., c_n = (\mathbf{x}_n)^T\mathbf{v}$.

(a) Show that the variance of $\{c_1, c_2, ..., c_n\}$ is given by

$$V = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{v}^T\mathbf{x}_i\right)^2$$

(b) Show that the variance of $\{\mathbf{y}_1^T\mathbf{v}, \mathbf{y}_2^T\mathbf{v}, ..., \mathbf{y}_n^T\mathbf{v}\}$ is given by

$$V = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{v}^T\mathbf{x}_i\right)^2$$

(c) What is the implication of the results in (a) and (b) to finding the first principal component of the data? Briefly interpret your result.

3. Consider n data points $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ lying in $\mathcal{R}^d$. Let $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$ denote the first three principal components of the data.

(a) Let $\hat{\mathbf{x}}_i$ denote the projection of a data point $\mathbf{x}_i$ onto the subspace spanned by the first three principal components. What is the coordinate of $\hat{\mathbf{x}}_i$ with respect to the basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$?

(b) Let $\mathbf{y}$ be a point in $\mathcal{R}^d$ and let $\hat{\mathbf{y}}$ be the projection of $\mathbf{y}$ onto the subspace spanned by the first three principal components. Prove that $\hat{\mathbf{y}} = \mathbf{V}\mathbf{V}^T\mathbf{y}$ where

$$\mathbf{V} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

4. Let $\mathbf{A}$ be an $n \times n$ symmetric matrix.

(a) Prove that $\lambda_{\min}(\mathbf{A}) \leq \min_{1 \leq i \leq n} \mathbf{A}_{i,i}$ i.e. the minimum eigenvalue of $\mathbf{A}$ is upper bounded by the the minimum value in the diagonal.

(b) Prove that $\lambda_{\max}(\mathbf{A}) \geq \max_{1 \leq i \leq n} \mathbf{A}_{i,i}$ i.e. the maximum eigenvalue of $\mathbf{A}$ is lower bounded by the the maximum value in the diagonal.

[**Hint**: Use Rayleigh quotient.]

5. In this problem, we implement the principal component analysis algorithm and test it on 2-dimensional datasets.

(a) Given n points in $\mathcal{R}^2$, implement an algorithm that takes the data points as input and returns the principal components.
Remark: No credit will be given to using an inbuilt PCA function. However, you can use any inbuilt function to compute eigenvalues and eigenvectors. For example, in MATLAB, you can use the *eig* command.

(b) Load the data `gaussian_noisy` available in the HW3 folder on Canvas. Compute and display the first 2 principal components of the data. What is the percentage of variance for the first principal component? What is the percentage of variance for the second principal component? Summarize and explain your observations.

(c) Load the data `uniform_noisy` available in the HW3 folder on Canvas. Compute and display the first 2 principal components of the data. Compute the percentage of variance for the each of the first two principal components. Summarize and explain your observations.

6. In this problem, we consider the latent features obtained by the principal component analysis and explore the relationship between the number of principal components and average reconstruction error. The dataset we consider is the MNIST dataset which is a database of handwritten digits.

(a) Load the data `mnist_067` available in the HW3 folder on Canvas. The included data is a subset of the MNIST data consisting of the digits $0, 6$ and 7. There are 21072 digits and each digit is an image of size 28×28. With that, the data matrix is a matrix of size 21072×784. Note that each image is mapped into a 784 dimensional vector with each entry informing the gray pixel intensity. Project the data points onto the first two principal components. Show the representation of each data point in the coordinate of the principal components i.e. plot the latent representation of the data in $\mathcal{R}^2$. Use different colors to label the latent features according to the label of the digit. Is this a good low dimensional representation? Interpret your results.

(b) We now consider the digit 7 from the provided dataset. Define the average reconstruction error as follows

$$E = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2$$

where n is the number of digits of label 7, $\mathbf{x}_i$ is the input representation of the digit and $\hat{\mathbf{x}}_i$ is the reconstructed digit using K principal components of the data. Plot the average reconstruction error as a function of the number of principal components. Let the array of the number of principal components be $\{1, 10, 20, ..., 300\}$. Briefly discuss your results.

[**Remark**: For this problem, you can use an inbuilt PCA function from a solver of your choice.]