CS316: Introduction to Data Science
Major Exam Practice

Anis Koubaa
akoubaa@psu.edu.sa
Prince Sultan University,
College of Computer and Information Sciences

Date: Tuesday, Oct 23, 2024
Duration: 60 Minutes

Instructions

- This exam is a closed book and closed notes.
- Electronic devices are not allowed except for a simple calculator.
- Read each question carefully and answer to the best of your ability.
- Write your name and student ID in the space provided below.

Student Information

Student ID:	Student Name:

1 CS316: Mathematical Problems on PCA and SVD

Time: 45 minutes **Score:** 20 points **Mapped CLOs:** CLO 2, CLO 3

Objective:

Apply Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) on a dataset to understand dimensionality reduction and data compression techniques. This exercise will enhance your understanding of linear algebra concepts applied to real-world data scenarios.

Dataset:

For the exercises, use the following hypothetical dataset representing scores in Math, Physics, and English of 3 students:

Student	Math	Physics
A	85	78
B	72	65
C	90	88

Questions on PCA:

Answer:

Step 1: Compute the Mean and Standard Deviation of Each Variable

Calculate the mean and standard deviation of each variable:

$$\text{Math Mean} = \frac{85 + 72 + 90}{3} = \frac{247}{3} = 82.33$$

$$\text{Physics Mean} = \frac{78 + 65 + 88}{3} = \frac{231}{3} = 77.00$$

Standard Deviation Calculation:

The standard deviation (σ) for each variable is calculated using the formula:

$$\text{Std} = \sqrt{\frac{\sum(X_i - \mu)^2}{n-1}}$$

For Math:

$$\text{Math Std} = \sqrt{\frac{(85 - 82.33)^2 + (72 - 82.33)^2 + (90 - 82.33)^2}{3-1}}$$

Expanding the terms:

$$(85 - 82.33)^2 = (2.67)^2 = 7.13$$

$$(72 - 82.33)^2 = (-10.33)^2 = 106.78$$

$$(90 - 82.33)^2 = (7.67)^2 = 58.85$$

Summing them:

$$\text{Math Variance} = \frac{7.13 + 106.78 + 58.85}{2} = \frac{172.76}{2} = 86.38$$

Taking the square root gives:

$$\text{Math Std} = \sqrt{86.38} = 9.29$$

For Physics:

$$\text{Physics Std} = \sqrt{\frac{(78 - 77)^2 + (65 - 77)^2 + (88 - 77)^2}{3 - 1}}$$

Expanding the terms:

$$(78 - 77)^2 = (1)^2 = 1$$
$$(65 - 77)^2 = (-12)^2 = 144$$
$$(88 - 77)^2 = (11)^2 = 121$$

Summing them:

$$\text{Physics Variance} = \frac{1 + 144 + 121}{2} = \frac{266}{2} = 133$$

Taking the square root gives:

$$\text{Physics Std} = \sqrt{133} = 11.53$$

Step 2: Standardize the Data
Subtract the mean and divide by the standard deviation to standardize the data:

Student	**Math (Standardized)**	**Physics (Standardized)**
A	$\frac{85-82.33}{9.29} = 0.29$	$\frac{78-77}{11.53} = 0.09$
B	$\frac{72-82.33}{9.29} = -1.11$	$\frac{65-77}{11.53} = -1.04$
C	$\frac{90-82.33}{9.29} = 0.83$	$\frac{88-77}{11.53} = 0.95$

Step 3: Compute the Covariance Matrix
Calculate the covariance between each pair of variables using the formula:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum (X_i - \mu_X)(Y_i - \mu_Y)$$

Since the data is standardized, the mean of each standardized variable is 0, simplifying the calculations:

$$\text{Cov}(\text{Math}, \text{Math}) = 1$$
$$\text{Cov}(\text{Math}, \text{Physics}) = 0.98$$
$$\text{Cov}(\text{Physics}, \text{Physics}) = 1$$

Thus, the covariance matrix is:

$$\begin{bmatrix} 1 & 0.98 \\ 0.98 & 1 \end{bmatrix}$$

Importance of the Covariance Matrix in PCA

The covariance matrix is essential in PCA as it provides information about the variance of each variable and the relationships (correlations) between them. By using standardized data, PCA ensures that each variable contributes equally to the analysis. The principal components, derived from the eigenvalues and eigenvectors of this matrix, represent the directions of maximum variance in the data. By projecting the data onto these components, PCA allows for dimensionality reduction while retaining the most significant features, aiding in better analysis and visualization.

Q2: Eigenvalue Decomposition

Time: 5 minutes **Score:** 3 points **Mapped CLOs:** CLO 2, CLO 3

Question: Perform eigenvalue decomposition on the covariance matrix. List the eigenvalues and eigenvectors.

Answer:

Step 1: Eigenvalue Decomposition of the Covariance Matrix

Given the covariance matrix:

$$\begin{bmatrix} 1 & 0.98 \\ 0.98 & 1 \end{bmatrix}$$

The characteristic equation is obtained by solving:

$$\det(\mathbf{C} - \lambda \mathbf{I}) = 0$$

$$\det\left(\begin{bmatrix} 1-\lambda & 0.98 \\ 0.98 & 1-\lambda \end{bmatrix}\right) = 0$$

Expanding the determinant:

$$(1-\lambda)^2 - (0.98)^2 = 0$$

$$\lambda^2 - 2\lambda + 0.0396 = 0$$

Using the quadratic formula:

$$\lambda = \frac{2 \pm \sqrt{3.8416}}{2}$$

$$\lambda_1 = 1.9845, \quad \lambda_2 = 0.0155$$

Step 2: Find Eigenvectors

For $\lambda_1 = 1.9845$:

$$\begin{bmatrix} -0.9845 & 0.98 \\ 0.98 & -0.9845 \end{bmatrix} \mathbf{v}_1 = 0$$

Solving gives:

$$\mathbf{v}_1 = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$$

For $\lambda_2 = 0.0155$:

$$\begin{bmatrix} 0.9845 & 0.98 \\ 0.98 & 0.9845 \end{bmatrix} \mathbf{v}_2 = 0$$

Solving gives:

$$\mathbf{v}_2 = \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$$

Eigenvalues:

$$\lambda_1 = 1.9845, \quad \lambda_2 = 0.0155$$

Eigenvectors:

$$\mathbf{v}_1 = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$$

The eigenvalues indicate the amount of variance explained by each corresponding eigenvector direction. The eigenvector $\mathbf{v}_1$ corresponds to the direction of maximum variance (the first principal component), while $\mathbf{v}_2$ corresponds to the direction of the second principal component.

Q3: Variance Explained

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 2, CLO 3

Question: Determine the percentage of variance explained by each principal component. Which principal component would you choose to retain for a reduced dimensionality and why?

Answer:

Step 1: Calculate Total Variance

The total variance is the sum of the eigenvalues:

$$\text{Total Variance} = \lambda_1 + \lambda_2 = 1.9845 + 0.0155 = 2$$

Step 2: Calculate Explained Variance Ratios

The percentage of variance explained by each principal component is calculated as:

$$\text{Explained Variance Ratio for } \mathbf{v}_1 = \frac{\lambda_1}{\text{Total Variance}} \times 100 = \frac{1.9845}{2} \times 100 = 99.23\%$$

$$\text{Explained Variance Ratio for } \mathbf{v}_2 = \frac{\lambda_2}{\text{Total Variance}} \times 100 = \frac{0.0155}{2} \times 100 = 0.77\%$$

Step 3: Principal Component Selection
Since $\mathbf{v}_1$ explains 99.23% of the variance, while $\mathbf{v}_2$ explains only 0.77%, it is more effective to retain $\mathbf{v}_1$ for dimensionality reduction. This will capture most of the data's variance while reducing the dataset to a single dimension, thus simplifying the analysis without losing significant information.

Q4: Data Projection

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 2, CLO 3

Question: Project the original data onto a 2D plane using the first two principal components. Provide the coordinates of the projected data.

Answer:
Step 1: Projection Matrix
The projection matrix is formed using the eigenvectors:

$$\mathbf{P} = \begin{bmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}^T$$

Step 2: Project the Standardized Data
The standardized data is projected onto the 2D space defined by the principal components using:

$$\mathbf{Z} = \mathbf{X}_{\text{standardized}} \cdot \mathbf{P}$$

The projected coordinates for each student are:

$$\text{Student A: } [0.1416, 0.2643]$$

$$\text{Student B: } [-0.0506, -1.5222]$$

$$\text{Student C: } [-0.0910, 1.2579]$$

These coordinates represent the positions of the original data points in the new space defined by the two principal components, preserving the directions of maximum variance.

Question: Project the original data onto a 2D plane using the highest principal components. Provide the coordinates of the projected data.

Answer:
Step 1: Select the Highest Principal Component
The highest principal component ($\mathbf{v}_1$) is:

$$\mathbf{v}_1 = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$$

Step 2: Project the Standardized Data onto $\mathbf{v}_1$

The projection of the standardized data onto the 1D space defined by $\mathbf{v}_1$ is calculated using:

$$\mathbf{Z} = \mathbf{X}_{\text{standardized}} \cdot \mathbf{v}_1$$

Step 3: Projected Coordinates
The resulting coordinates of the data points in the 1D space are:

Student A: 0.2643

Student B: -1.5222

Student C: 1.2579

These coordinates represent the data points projected onto the direction of maximum variance, captured by $\mathbf{v}_1$. This 1D projection retains most of the variability in the data, making it suitable for dimensionality reduction while preserving key information.

Questions on SVD:

Q5: Singular Value Decomposition

Time: 5 minutes **Score:** 3 points **Mapped CLOs:** CLO 2, CLO 3

Question: Apply Singular Value Decomposition on the standardized data matrix. List the singular values and the corresponding left and right singular vectors.

Answer:

Step 1: Given the Standardized Data Matrix

$$\mathbf{X}_{\text{standardized}} = \begin{bmatrix} 0.29 & 0.09 \\ -1.11 & -1.04 \\ 0.83 & 0.95 \end{bmatrix}$$

Step 2: Compute the Covariance Matrix

$$\mathbf{C} = \mathbf{X}_{\text{standardized}}^T \mathbf{X}_{\text{standardized}} = \begin{bmatrix} 2.00 & 0.98 \\ 0.98 & 2.00 \end{bmatrix}$$

Step 3: Eigenvalues and Singular Values
The eigenvalues of $\mathbf{C}$ are:

$$\lambda_1 = 1.9923, \quad \lambda_2 = 0.1758$$

The singular values are the square roots of the eigenvalues:

$$\mathbf{\Sigma} = \begin{bmatrix} 1.4129 & 0 \\ 0 & 0.4191 \end{bmatrix}$$

Step 4: Right Singular Vectors (V)

The right singular vectors ($\mathbf{V}$) are the eigenvectors of $\mathbf{C}$:

$$\mathbf{V} = \begin{bmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{bmatrix}$$

Step 5: Left Singular Vectors (U)
$\mathbf{U}$ is computed using:

$$\mathbf{U} = \mathbf{X}_{\text{standardized}} \mathbf{V} \mathbf{\Sigma}^{-1}$$

$$\mathbf{U} = \begin{bmatrix} -0.1326 & -0.8057 \\ 0.7640 & 0.2880 \\ -0.6314 & 0.5177 \end{bmatrix}$$

Final Results:
Singular Values: $[1.9923, 0.1758]$
Left Singular Vectors (U):

$$\begin{bmatrix} -0.1326 & -0.8057 \\ 0.7640 & 0.2880 \\ -0.6314 & 0.5177 \end{bmatrix}$$

Right Singular Vectors ($\mathbf{V}^T$):

$$\begin{bmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{bmatrix}$$

The singular values indicate the magnitude of the principal components in the transformed space, while the left singular vectors ($\mathbf{U}$) represent the transformed coordinates of the data points. The right singular vectors ($\mathbf{V}$) define the directions of the principal components in the original space.

Q6: Relationship Between Singular Values and Eigenvalues

Time: 5 minutes **Score:** 2 points **Mapped CLOs:** CLO 2, CLO 3

Question: Explain how the singular values obtained in Q5 relate to the eigenvalues obtained from the covariance matrix in PCA.

Answer:
The relationship between the singular values from Singular Value Decomposition (SVD) and the eigenvalues from Principal Component Analysis (PCA) is fundamental to understanding how these two methods analyze data:

1. Singular Values and Eigenvalues:
- When we perform SVD on a standardized data matrix $\mathbf{X}_{\text{standardized}}$, we obtain singular values ($\sigma_1, \sigma_2, \ldots$) in the diagonal matrix $\mathbf{\Sigma}$. - The square of each singular value corresponds to an eigenvalue of the covariance matrix $\mathbf{X}^T_{\text{standardized}} \mathbf{X}_{\text{standardized}}$. - Mathematically, if σ_1 is a singular value, then σ_1^2 is an eigenvalue of the covariance matrix.

2. Example from Q5:
- In Q5, the singular values were $\sigma_1 = 1.9923$ and $\sigma_2 = 0.1758$. - Squaring these gives us:

$$\sigma_1^2 = (1.9923)^2 = 3.97 \approx \lambda_1$$

$$\sigma_2^2 = (0.1758)^2 = 0.03 \approx \lambda_2$$

where λ_1 and λ_2 are the eigenvalues obtained from the covariance matrix in PCA.

3. Intuition:
- Singular values measure the "stretch" applied to the data along each principal component direction during the transformation. - Eigenvalues represent the amount of variance in the data along each principal component direction. - This means that larger singular values correspond to principal components that explain more variance in the data, which is also reflected in larger eigenvalues.

In summary, the singular values from SVD, when squared, directly relate to the eigenvalues from PCA. This relationship helps us understand how much variance each principal component captures and how the data is transformed through SVD.

Q7: Data Reconstruction

Time: 5 minutes **Score:** 3 points **Mapped CLOs:** CLO 2, CLO 3

Question: Using the results from SVD, reconstruct the original data matrix using only the first two singular values and corresponding vectors. What does this reconstruction represent in terms of data compression?

Answer:

Step 1: Understand SVD Decomposition
Singular Value Decomposition (SVD) decomposes the standardized data matrix $\mathbf{X}_{\text{standardized}}$ into three matrices:

$$\mathbf{X}_{\text{standardized}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Where: - $\mathbf{U}$: Left singular vectors (matrix of size $m \times n$) - $\mathbf{\Sigma}$: Diagonal matrix of singular values (matrix of size $n \times n$) - $\mathbf{V}^T$: Transpose of the right singular vectors (matrix of size $n \times n$)

Step 2: Truncated SVD for Reconstruction
To reconstruct the data using only the first singular value and its corresponding vectors, we use a truncated version of $\mathbf{\Sigma}$, $\mathbf{U}$, and $\mathbf{V}$:

$$\mathbf{\Sigma}_k = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad \mathbf{U}_k = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix}, \quad \mathbf{V}_k^T = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

Step 3: Perform the Multiplications
The reconstructed matrix using the top two components is computed as:

$$\mathbf{X}_{\text{reconstructed}} = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$$

Expanding this, the multiplication proceeds as:

$$\mathbf{U}_k\mathbf{\Sigma}_k = \begin{bmatrix} -0.1326 & -0.8057 \\ 0.7640 & 0.2880 \\ -0.6314 & 0.5177 \end{bmatrix} \begin{bmatrix} 1.9923 & 0 \\ 0 & 0.1758 \end{bmatrix} = \begin{bmatrix} -0.2643 & -0.1415 \\ 1.5222 & 0.0506 \\ -1.2579 & 0.0910 \end{bmatrix}$$

Then, multiply with $\mathbf{V}_k^T$:

$$\mathbf{X}_{\text{reconstructed}} = \begin{bmatrix} -0.2643 & -0.1415 \\ 1.5222 & 0.0506 \\ -1.2579 & 0.0910 \end{bmatrix} \begin{bmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} = \begin{bmatrix} 0.29 & 0.09 \\ -1.11 & -1.04 \\ 0.83 & 0.95 \end{bmatrix}$$

Step 4: Interpret the Reconstruction

The reconstructed matrix $\mathbf{X}_{\text{reconstructed}}$ approximates the original standardized data matrix using only the first two singular values and their corresponding vectors.

Step 5: What Does This Mean for Data Compression?

- By using only the top two singular values, we retain the majority of the information (variance) in the original data while discarding less significant components. - This is an example of data compression because we reduce the dimensionality of the data from potentially higher dimensions down to the two most significant directions (principal components). - This approach captures most of the important features of the data with fewer dimensions, making computations faster while still maintaining the essential structure of the data.

Summary: The reconstructed matrix is a compressed version of the original data, retaining the most important features. It is especially useful when we want to reduce the complexity of the data without losing too much information.