Assignment 2A

Ludvig Johansson, Ossian Arn

December 2023

2A/1

2A.1

in PCA, one of the primary goals is to find the principal components. These principal components are linear combinations of the original variables that capture the most significant variation in the data. The first principal component represents the direction in the data space along which the variance of the data is maximized. The direction of maximum variance is important because it captures the dominant pattern or structure in the data. When PCA is applied to non-centered data (data with non-zero means), the first principal component might not necessarily correspond to the direction of maximum variance. Instead the first principal component might more or less align itself with the mean of the data.

2A.2

The argument regarding the use of Singular Value Decomposition (SVD) in Principal Component Analysis (PCA) suggests a potential misconception that a single SVD operation is adequate for conducting PCA on both the rows and the columns of a data matrix. At first glance, it appears feasible because one could extract the principal components for the rows of a matrix A by applying PCA to the columns of its transpose A^T, using the left singular vectors U as the principal components for A's rows. However, this approach is generally not valid. The crucial step in PCA involves centering the data, which means adjusting each dimension of the data so that its mean equals zero. This process ensures that each dimension (or feature) is centered around the origin. The challenge arises when applying this procedure to both the rows and the columns of A. When centering is done for a column-wise PCA (focusing on features), it involves subtracting the mean of each column from its values. This centering is specific to the features. However, if we then aim to perform PCA on the rows (observations) of A by working with A^T we encounter a problem. The initial centering adjusted the columns of A (now the rows of A^T) to have zero mean, but it did not center the columns of A^T which are the original rows of A.
For PCA on the observations, a separate centering process, followed by SVD on the appropriately centered matrix, is required to ensure accurate results. As a result, the initial belief that one SVD operation is sufficient for both types of PCA does not hold in general scenarios due to the distinct centering requirements for rows and columns.

2A.3

When you perform SVD on a centered data matrix Y, you get $Y = U\Sigma V^T$ where Σ contains the singular values. Given that our Y is zero centered, its variance is simply the sum of all squares of its elements.

$$Var(Y) = \sum_{i=0}^{d}\sum_{j=0}^{n} Y_{ij}$$

This is the square of the Frobenius norm of Y.

$$||Y||_F = \sqrt{\sum_{i=0}^{d}\sum_{j=0}^{n} Y_{ij}}$$

And since the Frobenius norm also can be expressed in terms of the matrices singular values σ_i^2

$$||Y||_F = \sqrt{\sum_{i=0}^{min(d,n)} \sigma_i^2}$$

Now using

$$Var(Y) = ||Y||_F^2$$

And the fact that $d < n$

$$Var(Y) = \sum_{i=0}^{d} \sigma_i^2$$

2A.4

Let's consider that we have already performed PCA on our dataset Y which is represented as a Y $d \times n$, and we have our projection matrix W which is a $d \times k$ matrix whose columns are the first k principal components of Y. The projected data points X are given by $X = W^T Y$, and this results in a $k \times n$ matrix X. The variance of the original data Y is given by the sum of the squares of all singular values σ_i^2 of Y.

$$Var(Y) = \sum_{i=0}^{d} \sigma_i^2$$

When we project Y onto the space spanned by the columns of W, which correspond to the first k principal components, we effectively reduce the dimensionality from d to k, keeping only the variance captured by these components. Since the singular values are ordered such that $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq ... \geq \sigma_d$ the variance of Y captured by the first k components is

$$Var(X) = \sum_{i=0}^{k} \sigma_i^2$$

This sum of squared singular values for the first k principal components represents the variance of the dataset Y that is preserved in the projected dataset X. It does not include the variance associated with the remaining $d - k$ components that were discarded during the projection.

2A.5

To show that the variance of the residual data Z is given by $Var(Z) = \sum_{i=k+1}^{d} \sigma_i^2$, where $Z = [z_1, ..., z_n]$ and each $z_i = y_i - WW^T y_i$, we use the properties of the Singular Value Decomposition (SVD) of Y and the properties of the variance.

Given that $Y = U\Sigma V^T$ and that W is composed of the first k columns of U, the projection of Y onto the space spanned by W is $WW^T Y$. The residual data points z_i represent the difference between the original data points y_i and their projection onto the space spanned by the first k principal components.

$$Z = Y - WW^T Y$$

Since W contains the first k columns of U, WW^T is equivalent to $U_k U_k^T$ where U_k is the matrix containing the first k columns of U. Thus, we have:

$$Z = Y - U_k U_k^T Y$$

$$\bar{z} = \sum_{i=1}^{n} \frac{1}{n} WW^T y_i = WW^T (\frac{1}{n} \sum_{i=1}^{n} y_i) = WW^T \cdot 0 = 0$$

Since Z should also be centered (mean 0) when Y has mean 0(seen above) the variance of Z is the sum of the squares of its elements, which is the Frobenius norm of Z:

$$\text{Var}(Z) = \|Z\|_F^2$$

Substituting Z into the equation for the variance, we get:

$$\text{Var}(Z) = \|Y - U_k U_k^T Y\|_F^2$$

Factorizing with the identity matrix

$$\text{Var}(Z) = \|(I - U_k U_k^T) Y\|_F^2$$

$(I - U_k U_k^T)$ essentially nullifies the components of Y in the subspace spanned by the first k principal components, leaving only the residual components.

The Frobenius norm of $(I - U_k U_k^T)Y$ is equal to the sum of the squares of the singular values not included in the first k components:

$$\text{Var}(Z) = \sum_{i=k+1}^{d} \sigma_i^2$$

This is because $(I - U_k U_k^T)Y$ effectively projects Y onto the orthogonal complement of the space spanned by the first k principal components, which corresponds to the remaining $d - k$ singular values.

$$\text{Var}(Y) = \text{Var}(WW^T Y) + \text{Var}(Z)$$

This means that the total variance of the original data Y is equal to the variance explained by the first k principal components (projection $WW^T Y$) plus the variance of the residual data Z, with:

$$\text{Var}(Y) = \sum_{i=1}^{d} \sigma_i^2$$

$$\text{Var}(WW^T Y) = \sum_{i=1}^{k} \sigma_i^2$$

$$\text{Var}(Z) = \text{Var}(Y) - \text{Var}(WW^T Y) = \sum_{i=1}^{d} \sigma_i^2 - \sum_{i=1}^{k} \sigma_i^2 = \sum_{i=k+1}^{d} \sigma_i^2$$

Hence, this proves that the variance of the residual data Z after projecting Y onto the first k principal components is given by the sum of the squared singular values from $k+1$ to d. And we can also conclude that Var(Original data) = Var(Variace explain by PCA) + Var(Residual data)

2A/2

2A.6

(i) **Projection error:** PCA and Johnson-Lindenstrauss are trying to preserver different things in the dataset. PCA minimizes the mean squared reconstruction error (given a linear transformation), which means that it preserves as much variance as possible given the number of allowed dimensions k. Johnson-Lindenstrauss on the other hand tries to preserve the pairwise distances between the points in the dataset while reducing the dimensionality. The error can be calculated according to the Johnson-Lindenstrauss lemma: $k > \frac{\ln(n)}{\epsilon^2}$ here k is the minimum number of dimensions, ϵ is the error and n is the number of datapoints.

(ii) **Computational efficiency:** PCA requires more complex calculations when compared to Johnson-Lindenstrauss. This is because PCA needs to find the eigenvalues of covariance matricies whereas Johnson-Lindenstrauss uses random projection which are computationally more efficient.

(iii) **Target usecases:** Using the previous answers one can deduct that PCA is useful when the variance and intrinsic structure of the data is important, Johnson-Lindenstrauss on the other hand is good when the pairwise distances are important, or when the dataset is large and on which PCA would require a lot of computation.

2A/3

2A.7