Principal Component Analysis (PCA)

Hyun Jae Stephen Chu*
Korea University, Department of Economics

August 17, 2024

1 Introduction

Recently, the availability of big data sets, combined with the advances in econometrics, statistics and machine learning has generated interest in prediction under high-dimensional predictor space where a researcher faces a large number of predictor variables. Under such environment, standard regression methods lose efficient due to the inclusion of irrelevant variables and face multicollinerity. In order to resolve the so-called curse of dimensionality problem, various approaches have been proposed.

One way to improve prediction accuracy under the high-dimensional predictor space is to reduce the dimensionality by selecting a few statistically significant predictor variables. However, a crucial limitation of such method is that the statistical significance depends on a certain selection of significance level, i.e. p-value, by the researcher. More importantly, the significance level implies the probability of false rejection. Thus, there is always a possibility of selecting the wrong predictor variable. In this sense, reducing the dimension of the predictor space by selecting the statistically significant variables cannot be the antidote to the curse of dimensionality.

A popular alternative to handle problems under the high-dimensional predictor space is the **Principal Component Analysis (PCA)**, which is a type of a factor model. The idea is to reduce the dimensionality of a data set consisting of a large number of predictor variables while retaining as much as possible of the variation present in the data set. This can be done by constructing certain components (factors) using all available predictor variables in the data set. For example, suppose the data generating process is

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_{20} x_{2t} + e_t$$

Using the data of 20 predictor variables, one can construct a certain number (say two) of principal components, denoted as z_{1t} and z_{2t}. Thus, the dimension of the predictor variables can be successfully reduced. Moreover, since PCs have the characteristic that each are uncorrelated to one another, there are no multicollinearities between them. Then the next question will be how to apply PCA and construct such principal components.

*E-mail: stephenchu@korea.ac.kr

2 Eigenvalue, Eigenvector and Eigen Decomposition

Before going into the details of the PCA method, certain necessary mathematical concepts will be introduced.

Definition 1. *Suppose a square matrix A with $dim(A) = K \times K$. Then such square matrix has K **eigenvalues**, each denoted as λ_i, such that*

$$|A - \lambda_i \cdot I_K| = 0$$

*holds. Using each eigenvalue, the corresponding **eigenvectors**, denoted as $v_i = (e_1, \cdots, e_K)'$, is the vector that satisfies the following relation.*

$$A \cdot v_i = \lambda_i \cdot v_i$$

Example 1. *Suppose*

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Then

$$|A - \lambda \cdot I_2| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1 = 0$$

Thus,

$$\lambda^2 - 4 \cdot \lambda + 3 = 0$$

leading to $\lambda = 1, 3$ or $\lambda_1 = 1$ and $\lambda_2 = 3$. Now, in order to derive the eigenvector, one can solve the following relation.

$$A \cdot v_i = \lambda_i \cdot v_i$$

Note that sometimes, the row of the coefficient matrix is linearly dependent, i.e. the solution is undetermined. In order to derive a unique solution, we use the method of "normalization". That is, we add an equation that satisfies $\sum_{i=1}^{K} a_i^2 = 1$.

- *if $\lambda_1 = 1$*

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 1 \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \Rightarrow \begin{cases} 2a_1 + a_2 = a_1 & \Rightarrow a_1 + a_2 = 0 \\ a_1 + 2a_2 = a_2 & \Rightarrow a_1 + a_2 = 0 \end{cases}$$

  Here, one can observe that the coefficient matrix is linearly dependent. Thus, by applying the normalization technique stated above, one can solve the following system of equations.

$$\begin{cases} a_1 + a_2 = 0 \\ (a_1)^2 + (a_2)^2 = 1 \end{cases} \Rightarrow \begin{cases} a_2 = -a_1 \\ 2(a_1)^2 = 1 \end{cases}$$

  Then one gets $a_1 = \pm\frac{1}{\sqrt{2}}$ and $a_2 = \mp\frac{1}{\sqrt{2}}$, leading to

$$v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$$

- *if* $\lambda_2 = 3$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 3 \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \Rightarrow \begin{cases} -a_1 + a_2 = 0 \\ a_1 - a_2 = 0 \end{cases}$$

Again, the coefficient matrix is linearly dependent. Thus, one can solve the following system of equations with an additional condition regarding the normalization method.

$$\begin{cases} a_1 - a_2 = 0 \\ (a_1)^2 + (a_2)^2 = 1 \end{cases} \Rightarrow \begin{cases} a_1 = a_2 \\ 2(a_1)^2 = 1 \end{cases}$$

Then one gets $a_1 = \pm\frac{1}{\sqrt{2}}$ *and* $a_2 = \pm\frac{1}{\sqrt{2}}$*, leading to*

$$v_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Definition 2. *By stacking the eigenvectors side-by-side, one can construct the* ***eigenvector matrix****, denoted as* C*, such that*

$$C = \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix}$$

Theorem 1. *A square matrix* A *can be* ***eigen decomposed*** *as*

$$A = C \cdot \Lambda \cdot C^{-1}$$

where C *is its eigenvector matrix and* Λ *is a diagonal matrix with the diagonal elements as the eigenvalues of* A*. That is,*

$$A = \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_K \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_K \end{bmatrix}$$

In this sense, one can ***diagonalize*** *the initial square matrix* A *as follows.*

$$\Lambda = C^{-1} \cdot A \cdot C$$

Corollary 1. *If the square matrix* A *is*

1. *positive definite, i.e. all* $\lambda_i > 0$
2. *symmetric*

then the eigenvector matrix C *is orthogonal. More precisely,*

$$C \cdot C' = I \iff C^{-1} = C'$$

Thus, the theorem above can be rewritten as follows.

$$A = C \cdot \Lambda \cdot C'$$

3 Principal Component Analysis

3.1 Model Framework

Now let us go into the details of **principal component analysis (PCA)**, i.e. constructing the principal components of a large number of predictor variables. Consider the standard regression model such that

$$y = X \cdot \beta + \varepsilon,\ \varepsilon \sim iid(0, \sigma^2)$$

where y is a vector of the dependent variable with T observations, X is the $T \times K$ matrix of predictor variables and ε is the vector of error terms. Precisely, suppose that there are K predictor variables with T observations. That is,

$$X = (X_1, \cdots, X_K)$$

where $dim(X_i) = T \times 1$ so that $dim(X) = T \times K$ and $Var(X) = \Sigma$. Note that this variance-covariance matrix of X is the key term of PCA. One remark is that since the variance-covariance matrix Σ depends on the scale of the predictor variables, it is general to convert it to the correlation matrix, denoted as Γ, by standardizing each predictor variable such that the mean is 0 and variance is 1. Precisely,

$$\bar{X_i} = 0 \text{ and } Var(X_i) = 1$$

Then the sample correlation matrix estimator is

$$\Gamma \equiv \frac{1}{T-1} \cdot X'X$$

3.2 Definition of Principal Component

Note that the number of principal components are the same as the number of predictor variables used, i.e. there are K principal components. Then each principal component, denoted as Z_i, is defined to be constructed by a linear combination using the predictor variables as follows.

$$\begin{aligned} Z_1 &= a_{11}X_1 + \cdots + a_{1K}X_K \equiv X \cdot a_1 \\ Z_2 &= a_{21}X_1 + \cdots + a_{2K}X_K \equiv X \cdot a_2 \\ &\vdots \\ Z_K &= a_{K1}X_1 + \cdots + a_{KK}X_K \equiv X \cdot a_K \end{aligned}$$

where $(a_{i1})^2 + \cdots + (a_{iK})^2 = 1$ and a_i is the vector of coefficients such that

$$a_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{iK} \end{pmatrix}$$

Here, the condition $\sum_{j=1}^{K}(a_{ij})^2 = 1$ is the most widely used *normalization contraint* of weighting the predictor variables, while other methods such as $max_j|a_{ij}| = 1$ for all $i = 1, \cdots, K$ can be alternatively used. By combining the principal components and coefficient vectors, we can obtain a vector of principal components and matrix of coefficents such that

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_K \end{pmatrix} \text{ and } A = \begin{bmatrix} a_1 & \cdots & a_K \end{bmatrix}$$

Therefore, the vector of principal components can be expressed as a matrix form such as

$$Z = X \cdot A \text{ where } X'X = \Gamma$$

3.3 Derivation of Principal Components

Now that we have defined the principal components, we will focus on how to derive them, i.e. how to derive the coefficient vector a_i or more generally the coefficent matrix A. Precisely, the objective is to maximize the variance of Z given the condition that $\sum_{j=1}^{K}(a_{ij})^2 = 1$. Thus, since $Z'Z = A'X'XA = A'\Gamma A$, the maximization problem is

$$\max_{\{a_1,\cdots,a_K\}} Z'Z \text{ s.t. } (a_{i1})^2 + \cdots + (a_{iK})^2 = 1 \ \forall \ i = 1, \cdots, K$$

The reason of maximizing the variance of Z is to find the coefficients that best explain the variance of the predictor variables X_i. Moreover, it turns out that the coefficient a_i^* obtained as the solution of the maximization problem that constructs the $i-$th principal component, is in fact the eigenvector corresponding to the $i-$th largest eigenvalue of Γ.

Theorem 2. *When solving the following maximization problem,*

$$\max_{\{a_1,\cdots,a_K\}} Z'Z \ \textit{s.t.} \ (a_{i1})^2 + \cdots + (a_{iK})^2 = 1 \ \forall \ i = 1, \cdots, K$$

the $i-$th solution a_i^* is the eigenvector corresponding to the $i-$th largest eigenvalue of Γ.

Proof. First, suppose $K = 1$, i.e. let us focus on a_1. Using the relation such that

$$Z_1'Z_1 = a_1'X'Xa_1 = a_1'\Gamma a_1$$

the maximization problem with the normalization contraint can be expressed as follows.

$$\max_{a_1} a_1'\Gamma a_1 \text{ s.t. } a_1'a_1 = 1$$

Then the equivalent lagrangian function is

$$\begin{aligned} &\Rightarrow \mathcal{L} = a_1'\Gamma a_1 + \lambda(1 - a_1'a_1) \\ &\partial a_1\text{: } \Gamma a_1 - \lambda a_1 = 0 \end{aligned} \tag{1}$$

Then equation (1) can be expressed as

$$\Rightarrow (\Gamma - \lambda \cdot I_K) = 0$$

Therefore, λ is the eigenvalue of Γ and a_1 is the corresponding eigenvector.

As mentioned above, an eigenvector and eigenvalue of a certain matrix is not unique. In fact, if $dim(\Gamma) = K \times K$, then one can obtain at most K eigenvalues. Then which of the K eigenvectors gives $X \cdot a_1 = Z_1$ the maximum variance? Note that the maximand of the maximization problem can be further expressed as follows.

$$a_1'\Gamma a_1 = a_1'\lambda a_1 = \lambda a_1' a_1 = \lambda$$

Thus, one can notice that the objective of the maximization problem is equivalent to maximizing λ. In this sense, a_1 is the eigenvector corresponding to the largest eigenvalue of Γ and $Var(Z_1) = a_1'\Gamma a_1 = \lambda_1$ which is the largest eigenvalue.

Now suppose $K = 2$, leading to the existance of two principal components. Since the derivation of the first PC is the same as the $K = 1$ case, we focus on the second PC. One must note that the second PC, denoted as $Z_2 = X \cdot a_2$, is assumed to be uncorrelated with Z_1. That is,

$$Cov(Z_1, Z_2) = Cov(X \cdot a_1, X \cdot a_2) = 0$$

Since the covariance of the two PCs can be further expressed as follows

$$\begin{aligned} &Cov(X \cdot a_1, X \cdot a_2) && = a_1' X' X a_2 \\ &= a_1'\Gamma a_2 = a_2'\Gamma a_1 \cdots \text{ since } \Gamma \text{ is symmetric} \\ &= a_2'\lambda_1 a_1 \\ &= \lambda_1 a_2' a_1 = \lambda_1 a_1' a_2 \end{aligned}$$

we can use any of the four equations stated below to specify the zero correlation restriction between Z_1 and Z_2.

$$\begin{aligned} a_1'\Gamma a_2 = 0 \qquad & a_2'\Gamma a_1 = 0 \\ a_1' a_2 = 0 \qquad & a_2' a_1 = 0 \end{aligned} \tag{2}$$

Without loss of generality, let us choose the last expression. Also, note that the normalization constraint is again necessary. Then the maximization problem for the second PC is

$$\begin{aligned} \max_{a_2} a_2'\Gamma a_2 \text{ s.t. } & a_2' a_2 = 1 \\ & a_2' a_1 = 0 \end{aligned}$$

Then the equivalent lagrangian function is

$$\begin{aligned} &\Rightarrow \mathcal{L} = a_2'\Gamma a_2 + \lambda(1 - a_2' a_2) - \phi \cdot a_2' a_1 \\ &\partial a_2\text{: } \Gamma a_2 - \lambda a_2 - \phi a_1 = 0 \end{aligned} \tag{3}$$

$$\Rightarrow a_1'\Gamma a_2 - \lambda a_1' a_2 - \phi \cdot a_1' a_1 = 0 \tag{4}$$

Since the first two terms are zero by equation (2) and $a_1' a_1 = 1$ by the normalization constraint, equation (4) reduces to $\phi = 0$. Therefore, equation (3) can be expressed as

$$\Rightarrow (\Gamma - \lambda \cdot I_K) = 0$$

so λ is once again an eigenvalue of Γ and a_2 is the corresponding eigenvector.

Again, $\lambda = a_2'\Gamma a_2$ so the goal of our optimization problem is to maximize λ as large as possible. Assuming that Γ does not have any repeated eignevalues, λ cannot equal λ_1. If it did, it follows that $a_2 = a_1$, violating the constraint that $a_1' a_2 = 0$. Therefore, $\lambda \equiv \lambda_2$ must be the second largest eigenvalue of Γ, and a_2 is the corresponding eigenvector.

Now by a similar process, one can expand to a finite K case, i.e. $K = 3, 4, \cdots$. It can be shown that for the third, fourth,$\cdots$,Kth PCs, the vectors of coefficients $a_3, a_4, \cdots, a_K$ are the eigenvectors of Γ corresponding to $\lambda_3, \lambda_4, \cdots, \lambda_K$ that are the third, fourth largest,$\cdots$, and the smallest eigenvalue, respectively. □

3.4 Interpretation of Principal Components

Until now, we have seen that the principal component matrix Z can be derived by solving the maximization problem

$$\max_{\{a_1, \cdots, a_K\}} A'\Gamma A$$

Moreover, by definition, Γ is positive definite and symmetric. Thus, one can diagolalize Γ by applying the previous Corollary 1 as follows.

$$\begin{aligned} &\Gamma = X' \cdot X = A \cdot \Lambda \cdot A' \\ &\Rightarrow Z' \cdot Z = A' X' X A = \Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_K \end{pmatrix} \end{aligned} \tag{5}$$

Precisely, equation (5) can be written as

$$\Rightarrow \begin{pmatrix} Z_1' Z_1 & \cdots & Z_1' Z_K \\ \vdots & \ddots & \vdots \\ Z_K' Z_1 & \cdots & Z_K' Z_K \end{pmatrix} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_K \end{pmatrix} \tag{6}$$

Then one can interpret equation (6) as follows.

1. λ_i implies the variation of the $i-$th PC, Z_i. That is

$$Var(Z_i) = \lambda_i$$

2. Since $\lambda_1 > \cdots > \lambda_K$, Z_1 has the largest variation while Z_K has the smallest variation among all PCs.

3. $Z_i'Z_j = 0$ for all $i \neq j$ implies that all PCs are uncorrelated, i.e. orthogonal.

3.5 Application to Linear Regression Models

4 Existence of Principal Components

For simplicity, let us suppose that a researcher is constructing a principal component using two predictor variables X_1 and X_2. That is,

$$Z_1 = a_{11}X_1 + a_{12}X_2$$
$$Z_2 = a_{21}X_1 + a_{22}X_2$$

Further, suppose that $Cor(X_1, X_2) \approx 0$. Then the correlation matrix of $X = (X_1, X_2)$, denoted as Γ, is

$$\Gamma = \begin{pmatrix} 1 & 0.01 \\ 0.01 & 1 \end{pmatrix}$$

leading to the fact that

$$\Rightarrow |\Gamma - \lambda \cdot I_2| = \begin{vmatrix} 1-\lambda & 0.01 \\ 0.01 & 1-\lambda \end{vmatrix} = 0$$

Therefore, the two eigenvalues become $\lambda_1 \approx 1.01$ and $\lambda_2 \approx 0.99$. Then one can interpret that the variation of the first PC Z_1 is not significantly larger than that of Z_2. In other words, there is no principal component between X_1 and X_2. Loosely spoken, even if there is a PC between the predictor variables, it cannot represent them. Conversely, if the difference between eigenvalues are large, then we say that the principal component explains most of the variation of the predictor variables. Precisely, this can be mathematically shown as follows.

Remark 1. *Suppose a square matrix A such that*

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Then the ***trace*** *of A, denoted as $tr(A)$ is the sum of the diagonal terms of A, i.e.*

$$tr(A) = a + d$$

Moreover,

$$tr(ABC) = tr(BCA) = tr(CAB)$$

Applying remark 1 to the correlation matrix,

$$tr(\Gamma) = tr\left(\frac{X'X}{T-1}\right)$$
$$= tr\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

$$
\begin{aligned}
&= 1 + \cdots + 1 = K \\
&= tr(A\Lambda A') \\
&= tr(\Lambda A' A) \\
&= tr(\Lambda) \cdots \text{ since } A'A = I_K \\
&= \lambda_1 + \cdots + \lambda_K
\end{aligned}
$$

Therefore,

$$\sum_{i=1}^{K} \lambda_i = K = \sum_{i=1}^{K} Var(X_i) \tag{7}$$

Using the relation in equation (7), one can observe that if $\lambda_1 >> \lambda_2, \cdots, \lambda_K$, then Z_1 explains a large proportion of the variation of predictor variables where the fraction of Z_1 is

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_K}$$

5 Singular Value Decomposition

An alternative way of applying PCA is to use the **Singular Value Decomposition (SVD)** method. Before going into the details, we first introduce the **spectral decomposition** which is used to prove the SVD.

Theorem 3. *(Spectral Decomposition)*

*Suppose the variance-covariance matrix Σ of X. Then one can decompose Σ using the **Spectral Decomposition** method as follows.*

$$\Sigma = \lambda_1 a_1 a_1' + \lambda_2 a_2 a_2' + \cdots + \lambda_K a_K a_K' = \sum_{j=1}^{K} \lambda_j a_j a_j'$$

Proof. Recall that by eigen decomposition stated in Corollary 1, one gets

$$\Sigma = \underbrace{A}_{K \times K} \cdot \underbrace{\Lambda}_{K \times K} \cdot \underbrace{A'}_{K \times K}$$

Precisely,

$$
\begin{aligned}
A \cdot \Lambda \cdot A' &= \begin{bmatrix} a_1 & a_2 & \cdots & a_K \end{bmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_K \end{pmatrix} \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_K' \end{bmatrix} \\
&= \lambda_1 a_1 a_1' + \lambda_2 a_2 a_2' + \cdots + \lambda_K a_K a_K' \\
&= \sum_{j=1}^{K} \lambda_j a_j a_j'
\end{aligned}
$$

□

Theorem 4. *(Singular Value Decomposition)*

Suppose an arbitrary matrix X of $dim(X) = T \times K$. Then X can be written as

$$X = \underbrace{U}_{T\times r} \cdot \underbrace{L}_{r\times r} \cdot \underbrace{A'}_{r\times K}$$

where U and A are orthogonal, i.e. $U'U = UU' = I_r$ and $A'A = AA' = I_r$, L is a diagonal matrix, and r is the rank of X. In our case, L is the diagonal matrix of eigenvalues as the diagonal terms and A is the eigenvector matrix of $\Gamma = \frac{1}{T-1}X'X$ respectively.

Proof. Recall that $\Gamma = \frac{1}{T-1}X'X$ if X is standardized. Denote λ_i as the $i-$th eigenvalue of $X'X$. Consider the spectral decomposition of $X'X$, i.e.

$$\begin{aligned}(T-1)\Gamma \equiv X'X &= l_1 a_1 a_1' + l_2 a_2 a_2' + \cdots + l_r a_r a_r' + \cdots + l_K a_K a_K' \\ &= l_1 a_1 a_1' + l_2 a_2 a_2' + \cdots + l_r a_r a_r'\end{aligned}$$

where l_i and a_i is each the eigenvalue and corresponding eigenvector of $X'X$. More importantly, the last equality is due to the fact that $rank(X) = r$, hence $rank(X'X) = r$, so that the last $K - r$ eigenvalues of X are zero.

Define

$$A \equiv \underbrace{\begin{bmatrix} \underbrace{a_1}_{K\times 1} & \cdots & a_r \end{bmatrix}}_{K\times r} \text{ and } L \equiv \begin{pmatrix} l_1^{\frac{1}{2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_K^{\frac{1}{2}} \end{pmatrix}$$

and

$$U \equiv \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} = \begin{bmatrix} l_1^{-\frac{1}{2}} \cdot \underbrace{X}_{T\times K} \cdot \underbrace{a_1}_{K\times 1} & \cdots & l_r^{-\frac{1}{2}} \cdot X \cdot a_r \end{bmatrix}$$

Thus, L is diagonal, and A and U are orthogonal due to the normalization and zero correlation constraint explained in section 3 above.

Now let us show that $X = ULA'$.

$$\begin{aligned} ULA' &= U \cdot \begin{bmatrix} l_1^{\frac{1}{2}} a_1' \\ \vdots \\ l_r^{\frac{1}{2}} \cdot a_r' \end{bmatrix} \\ &= \sum_{j=1}^{r} u_j \cdot l_j^{\frac{1}{2}} \cdot a_j' \\ &= \sum_{j=1}^{r} l_j^{-\frac{1}{2}} \cdot X \cdot a_j \cdot l_j^{\frac{1}{2}} \cdot a_j' \\ &= \sum_{j=1}^{r} X \cdot a_j a_j' \\ &= \sum_{j=1}^{r} X \cdot a_j a_j' \end{aligned}$$

where the last equality is due to the fact that a_j for $j = (r+1), (r+2), \cdots, K$ are the eigenvectors

of $X'X$ corresponding to zero eigenvalues. Thus,

$$\sum_{j=1}^{K} a_j a_j' = 1$$

since the $K \times K$ matrix with its k =th column denoted as a_k is orthogonal, i.e. has orthonormal rows which are unit vectors. Therefore,

$$ULA' = X \sum_{j=1}^{K} a_j a_j' = X$$

as desired. □

Then how can one apply this Singular Value Decomposition to PCA? It turns out that by the following relationship between the SVD and Eigen Decomposition, one can gain the same eigenvalue and corresponding eigenvector.

Corollary 2. *SVD and the Eigen Decomposition have the following relationship. If X has SVD $X = ULA'$, then*

$$X'X = AL'U'ULA = A(L'L)A'$$
$$XX' = ULA'AL'U' = U(LL')U'$$

since U and A are orthogonal.

Remark 2. *As shown by Corollary 2, the SVD provides a computationally efficient method of actually finding the PCs. That is, if one finds U, L, A that satisfies SVD,*

- *the columns of A (referred to as the **right-singular vectors**) are eigenvectors of $X'X$.*
- *the columns of U (referred to as the **left-singular vecotrs**) are the eigenvectors of XX'.*
- *the non-zero elements of L are the square roots of the non-zero eigenvalues of $X'X$ or XX'.*

One bonus of using SVD is that one can also get an additional term U, which is the scaled version of the so-called PC scores. This can be shown as follows.

$$X = ULA'$$
$$\Rightarrow XA = ULA'A = UL$$

since A is orthogonal. Note that XA is defined to be the matrix of PC scores whose $k-$th column consists of the PC scores for the $k-$th PC. The PC scores z_{ij} are therefore

$$z_{ij} = u_{ij} \cdot l_j^{\frac{1}{2}}$$

for all $i = 1, \cdots, T$ and $j = 1, \cdots, r$. This can be expressed in matrix form as

$$Z = UL \text{ or } U = Z \cdot L^{-1}$$