Principal Component Analysis and Connection to Singular Value Decomposition

Kian Kermani

December 10, 2023

Abstract

In statistical machine learning, techniques such as Principal Component Analysis (PCA) are used to find an orthogonal projection of high dimensional data to a low dimensional subspace. We want the low dimensional representation to be a good approximation of the original data in order to preserve the most important information and make downstream analysis easier. In this paper, we construct the PCA algorithm by seeking to minimize the reconstruction error and later on demonstrate the relationship between Singular Value Decomposition (SVD) and PCA.

1 Introduction

The modern landscape of data science often entails the analysis of high-dimensional data sets. The use of traditional statistical methods to analyze such data sets can be difficult due to the large number of variables. Thus, we employ dimension reduction techniques that allow us to focus on the important features of a data set. Principal Component Analysis is one such technique that will aid in our goal of being able to construct meaningful analyses on these large sets.

We can employ the thinking that if a variable has a large variance, then it is probably important in terms of informing us about the nature of the data. Suppose we have our data matrix X and we look at the column vectors x_i contained within. If we take those column vectors and transform them such that the transformed variables have maximal variance and are uncorrelated, we get the principal components of the data. Those transformed variables will give us the most information about the data.

To measure the relationship between each pairing of variables in the data, we can take the covariance of these pairings and store them in a matrix. Since this matrix would be a symmetric matrix due to the properties of the covariance function, eigenvectors corresponding to distinct eigenvalues will be orthogonal, which sounds quite similar to how we described principal components. In fact, we can use certain matrix factorizations that make analysis simpler and obtaining the eigenvectors/eigenvalues with less work. Eigen-decomposition and Singular value decomposition (SVD) provide ways to decompose that matrix into smaller parts.

Throughout the rest of the paper, we will take this intuitive understanding we built of the tools needed for PCA and expand on them with formal definitions. In section 2, we establish a few preliminary tools that will be utilized throughout the paper. In section 3, we introduce the algorithm for PCA by optimizing something called the reconstruction error, which builds off of the information gained from the preliminary definitions. Lastly, in Section 4, we show an alternative way to derive PCA using truncated Singular Value Decomposition.

2 Background

We need a key result from linear algebra to begin laying down the foundation for PCA, Theorem 2.2, which is known as the Spectral Theorem. The proof of which, will be aided by Lemma 2.1.

Lemma 2.1. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix ($A^T = A$) and V be an arbitrary subspace of $\mathbb{R}^{n\times n}$ that is A-invariant. Then $V^\perp$ (the orthogonal complement of V), is invariant under A, and A restricted to V is symmetric.

Proof. Let $\hat{v} \in V^\perp$ and $v \in V$. Then

$$A\hat{v}v = \hat{v}Av = 0$$

since A is symmetric, V is invariant under A (so $Av \in V$), and because $\hat{v} \in V^\perp$. This implies that $A\hat{v} \in V^\perp$. Thus $V^\perp$ is invariant under A. □

Theorem 2.2. Let $A \in \mathbb{R}^{n\times n}$ where A is a symmetric matrix, then A has exactly n (possibly not all distinct) eigenvalues that are all real. Associated eigenvectors can be chosen to form an orthonormal basis.

Proof. We proceed by induction on n. Since the $n = 1$ case is trivial, suppose the result holds from $1 \leq k - 1 < n$ for $k \in \mathbb{N}$. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and q be an eigenvector of A such that $||q|| = 1$, since a symmetric matrix in a non-trivial vector space has a guaranteed eigenvalue and thus an associated eigenvector. Let $\text{span}(q) = V$. Therefore, V is a 1-dimensional subspace of $R^{n\times n}$. Take the orthogonal complement of V, and notice that $\dim(V^\perp) < \dim(\mathbb{R}^{n\times n})$ If we take A and restrict it to $V^\perp$, we see that $A|_{V^\perp}$ is also symmetric by 2.1 simply by replacing V with $V_\perp$. By our initial assumption, $V^\perp$ will have an orthonormal basis comprised of eigenvectors of $A|_{V^\perp}$. Simply adjoin q to this new orthonormal basis of $V^\perp$ and we get an orthonormal basis of $\mathbb{R}^{n\times n}$ consisting of eigenvectors of A, completing the proof. □

As a direct consequence of the Spectral Theorem, we are able to derive the following matrix decomposition.

Corollary 2.3. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix, then the *eigen-decompostion* of A can be written as the following:

$$A = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$ is an $n \times n$ diagonal matrix made up of the eigenvalues of A, and Q is an orthogonal matrix (that is, $QQ^T = Q^TQ = I$) with columns that are unit eigenvectors $q_1, ..., q_n$ of A.

Proof. Let A be a symmetric $n \times n$ matrix, then by 2.2 we know that A has n real eigenvalues and corresponding eigenvectors that form an orthonormal basis. Let $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$, where each λ_i is an eigenvalue of A. Let $q_1, ..., q_n$ be the eigenvectors of A that comprise the columns of matrix Q. Since each q_i is orthonormal, we have $Q^TQ = I$. This implies Q is invertible where $Q^{-1} = Q^T$. Let e_i be an element of the standard basis, so for any i we have

$$\begin{aligned} Q^TAQe_i &= Q^TAq_i \\ &= Q^T\lambda_i q_i \\ &= \lambda_i Q^T q_i \\ &= \lambda_i e_i \\ &= \Lambda e_i \end{aligned}$$

Therefore $\Lambda = Q^TAQ$ which means we can conclude $A = Q\Lambda Q^T$. Furthermore, for any j we have

$$(\sum_{i=1}^{n} \lambda_i q_i q_i^T)q_j = \lambda_j q_j = Aq_j$$

Thus, $A = \sum_{i=1}^{n} \lambda_i q_i q_i^T$ and we have completed the proof. □

We need a way of measuring the distance between two matrices, as reducing the dimension of a data set requires finding a good approximation to it. So we use Definition 2.4 to aid us in this manner.

Definition 2.4. Let $A \in \mathbb{R}^{n\times p}$. Define the *Frobenius norm* of A as

$$||A||_F = (\sum_{i=1}^{n}\sum_{j=1}^{p}|a_{ij}|^2)^{\frac{1}{2}} = \text{tr}(A^TA)^{\frac{1}{2}}$$

where a_{ij} are the entries of A and $\text{tr}(\cdot)$ represents the trace of a matrix.

The Frobenius Norm will be able to measure the size of a matrix A, and for our purposes, the distance between A and some lower dimensional approximation $\hat{A}$. When we later try to minimize the distance between a given data set and some reduced form of it, we will want to use the covariance matrix as a way of representing our data.

Definition 2.5. Recall that the covariance between two random variables X and Y measures the degree to which they are related. Covariance is defined as follows

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

Where $E[X]$ is defined as the expectation of X. Now let $x \in \mathbb{R}^n$, we define its *covariance matrix* as

$$\Sigma = \text{Cov}(x) = \begin{bmatrix} \text{Cov}(x_1,x_1) & \text{Cov}(x_1,x_2) & \dots & \text{Cov}(x_1,x_n) \\ \text{Cov}(x_2,x_1) & \text{Cov}(x_2,x_2 & \dots & \text{Cov}(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(x_n,x_1) & \text{Cov}(x_n,x_2) & \dots & \text{Cov}(x_n,x_n) \end{bmatrix}$$

By using the property of covariance where $\text{Cov}(x_i,x_j) = \text{Cov}(x_j,x_i)$, we can see that this matrix is also symmetric.

3 The PCA Algorithm

Throughout this section, we will try to find a good approximation to our data matrix, $X \in \mathbb{R}^{n\times p}$. Additionally, we will assume that this matrix is centered. That is, if μ represents the mean of each column vector x_i in X, then we would have

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = 0$$

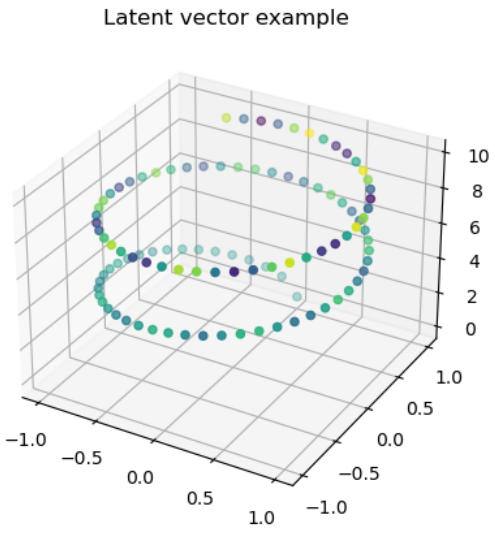
We can begin by finding a low-dimensional representation of each column vector x_i.

Definition 3.1. A *latent vector* $z_i \in \mathbb{R}^\ell$ is a low-dimensional variable that is not directly observable, but can be used to describe the original data.

Example 3.2. In Figure 1, we define a variable alpha which is not actually observable in our output. However, alpha is used to describe our other 4 observable variables, the x, y, z coordinates, and the color of each point. Hence, alpha is a latent vector/variable.

```python
alpha = np.array(range(100)) # latent variable
x_coord = np.sin(25 * alpha) # 1st observable variable
y_coord = np.cos(25 * alpha) # 2nd observable variable
z_coord = np.sqrt(alpha)     # 3rd observable variable
color = x+2                  # 4th observable variable
```

(a) Python code



(b) Output of code using Matplotlib

Figure 1: Using python to demonstrate latent vectors/variables

We will assume that x_i can be approximated by a linear combination of latent vectors $z_i \in \mathbb{R}^\ell$ and basis vectors $w_1, ..., w_\ell$ with each $w_i \in \mathbb{R}^p$.

$$x_i \approx \sum_{k=1}^{\ell} z_{ik} w_k$$

These basis vectors will later be shown to be none other than the eigenvectors of the covariance matrix, which can be recovered through eigen-decomposition. In order to measure the error that results from our approximation, we define what is known as the reconstruction error.

Definition 3.3. Let Z be a matrix containing the low-dimensional vectors z_i of x_i. Let W be a matrix comprised of basis vectors $w_1, ..., w_\ell$ such that W is an orthogonal matrix. We create the (average) *reconstruction error* as follows:

$$\begin{aligned} L(W, Z) &= \frac{1}{n} ||X - ZW^T||_F^2 \\ &= \frac{1}{n} ||X^T - WZ^T||_F^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} ||x_i - Wz_i||^2 \end{aligned}$$

We want to minimize the reconstruction error so that we may find the best possible approximation to X. The solution to this optimization problem is found in Theorem 3.4, which is the main result of our paper.

Theorem 3.4. The reconstruction error is minimized by setting $\hat{W} = U_\ell$ where U_ℓ contains the ℓ eigenvectors with the largest eigenvalues of the covariance matrix.

Proof. To begin, we should try to find the best 1-dimensional solution and then proceed by induc-

tion. Start by letting $z_1 = [z_{11} \ ... \ z_{n1}] \in \mathbb{R}^n$. The reconstruction error is then given by

$$\begin{aligned} L(w_1, z_1) &= \frac{1}{n}\sum_{k=1}^{n} ||x_k - z_{k1}w_1||^2 \\ &= \frac{1}{n}\sum_{k=1}^{n} (x_k - z_{k1}w_1)^T(x_k - z_{k1}w_1) \\ &= \frac{1}{n}\sum_{k=1}^{n} (x_k^T x_k - 2z_{k1}w_1^T x_k + z_{k1}^2 w_1^T w_1) \\ &= \frac{1}{n}\sum_{k=1}^{n} (x_k^T x_k - 2z_{k1}w_1^T x_k + z_{k1}^2) \end{aligned}$$

Where $w_1^T w_1 = 1$ since we are assuming orthonormality. Now we will take a derivative with respect to z_{k1} and set the equation to zero.

$$\begin{aligned} \frac{\partial}{\partial z_{k1}} L(w_1, z_1) &= \frac{1}{n}(-2w_1^T x_k + 2z_{k1}) \\ 0 &= \frac{1}{n}(-2w_1^T x_k + 2z_{k1}) \\ z_{k1} &= w_1^T x_k \end{aligned} \tag{1}$$

If we plug this back in, we get the loss for the weights.

$$\begin{aligned} L(w_1) &= L(w_1, z_1^*(w_1)) \\ &= \frac{1}{n}\sum_{k=1}^{n} (x_k^T x_k - z_{k1}^2) \\ &= C - \frac{1}{n}\sum_{k=1}^{n} z_{k1}^2 \end{aligned}$$

With C being some constant. Using the result given by (1), where $z_{k1} = w_1^T x_k$, we can continue to solve for w_1.

$$\begin{aligned} L(w_1) &= -\frac{1}{n}\sum_{k=1}^{n} z_{k1}^2 \\ &= -\frac{1}{n}\sum_{k=1}^{n} w_1^T x_k x_k^T w_1 \\ &= -w_1^T \Sigma w_1 \end{aligned}$$

where Σ is the covariance matrix obtained from X. We now take w_1, such that $||w_1|| = 1$, and continue to optimize.

$$L(w_1) = w_1^T \Sigma w_1 - \lambda_1(w_1^T w_1 - 1)$$

With λ_1 being a Lagrange multiplier. Once again, taking derivatives and setting equal to zero, we have

$$\begin{aligned} \frac{\partial}{\partial w_1} L(w_1) &= 2\Sigma w_1 - 2\lambda_1 w_1 = 0 \\ \Sigma w_1 &= \lambda_1 w_1 \end{aligned}$$

This implies that λ_1 is an eigenvalue of the covariance matrix Σ and w_1 is an eigenvector. If we left multiply by w_1^T we see

$$w_1^T \Sigma w_1 = \lambda_1$$

Simply picking the eigenvector that corresponds to the largest eigenvalue allows us to maximize this quantity and therefore minimize the loss. We could continue to proceed by induction in the same manner as we did above, and this would lead us to the realization that what we have done is essentially performed eigen-decomposition on the covariance matrix. That is,

$$\Sigma = W \Lambda W^T$$

where W is comprised of eigenvectors of Σ and Λ is the diagonal matrix with the corresponding eigenvalues. Thus, to minimize the reconstruction error, we simply create a matrix U_ℓ which contains the ℓ eigenvectors (our principal components) with the largest eigenvalues of the covariance matrix and set that equal to W.

$$\hat{W} = U_\ell$$

□

4 An Alternative Approach to PCA Using SVD

For the previous section, we computed PCA using eigenvector methods. However, we will find an alternative method that is equivalent and less computationally intensive. We start off by introducing Singular Value Decomposition.

Theorem 4.1. A non-square matrix A has *singular vectors* and *singular values* instead of eigenvectors and eigenvalues. Let A be an $n \times p$ matrix with rank r $(1 \leq r \leq \min(n, p))$. Then s is a singular value with left and right singular vectors u and v if

$$Av = su \text{ and } A^T u = sv$$

Also, there exists an $n \times r$ matrix $U = [u_1 \ ... \ u_r]$ a $p \times r$ matrix $V = [v_1 \ ... \ v_r]$ and a $r \times r$ diagonal matrix $S = \text{diag}(s_1, ..., s_r)$ such that

$$A = USV^T = \sum_{i=1}^{r} s_i u_i v_i^T$$

where $U^T U = I_r = V^T V$ and $s_1 \geq ... \geq s_r \geq 0$. Each u_i and v_i are unit vectors.

Proof. If we let A be an $n \times p$ matrix, we can take $A^T A$ which would then be a $p \times p$ symmetric matrix. Using Theorem 2.2 and Corollary 2.3 we are able to write $A^T A$ as

$$A^T A = V \Lambda V^T$$

where V is a $p \times p$ orthogonal matrix comprised of the orthonormal eigenvectors of $A^T A$, and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_r)$ is a diagonal matrix made of eigenvalues where $\lambda_1 \geq ... \geq \lambda_r > 0$. Let $i = 1, ..., r$, $s_i = \sqrt{\lambda_i}$, and $u_i = \frac{1}{s_i} A v_i$. Vectors u_i are orthonormal:

$$\begin{aligned} u_i^T u_j &= \frac{1}{s_i s_j} v_i^T A^T A v_j \\ &= \frac{s_j^2}{s_i s_j} v_i^T v_j \quad \text{since } v_j \text{ is an eigenvector of } A^T A. \\ &\text{Also, since the } v_i \text{ are orthonormal vectors we have} \\ &= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \end{aligned}$$

Now we show that u_i and v_i are left and right singular vectors,

$$A^T u_i = \frac{1}{s_i} A^T A v_i = \frac{s_i^2}{s_i} v_i = s_i v_i$$

Given $u_1, ..., u_r$, choose vectors $u_{r+1}, ..., u_n$ to complete the orthonormal basis for $\mathbb{R}^n$ and let $U = [u_1\ u_2\ ...\ u_r\ ...\ u_n]$. Then let S be the $n \times p$ diagonal matrix given by

$$\begin{bmatrix} s_1 & 0 & \dots & & & 0 \\ 0 & s_2 & 0 & \dots & & 0 \\ \vdots & & & & & \\ 0 & 0 & & \dots & s_r & \\ 0 & 0 & & \dots & 0 & \dots \\ \vdots & & & & & \\ 0 & 0 & & \dots & & 0 \end{bmatrix}$$

Since the columns of U and V form an orthonormal basis for $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively, if we only use the first r columns of U and V in addition to truncating S to a $r \times r$ matrix we can begin to recover the form for SVD. Thus we now have

$$\begin{aligned} U &= AVS^{-1} \\ US &= AV \\ USV^T &= A \end{aligned}$$

□

Using singular value decomposition, the rank of A simply becomes the number of non-zero singular values. The left and right singular vectors also form an orthonormal basis for the column space of A and A^T respectively. Additionally, SVD allows us to end up with an alternative expression for the Frobenius norm that will help us prove Theorem 4.3.

Proposition 4.2. Let σ_i be the singular values of A and $r = \text{rank}(A)$, then

$$||A||_F = (\sum_{i=1}^{r} s_i^2)^{\frac{1}{2}}$$

Proof. We begin by decomposing A using SVD such that $A = USV^T$. Now observe

$$||A||_F = ||U^T A||_F = ||U^T AV||_F = ||S||_F = \text{tr}(S^T S)^{\frac{1}{2}} = (\sum_{i=1}^{r} s_i^2)^{\frac{1}{2}}$$

□

Now we will consider the following situation. If we were to take the SVD of A, which we will call $\hat{A}$, we would naturally see that

$$||A - \hat{A}||_F^2 = 0$$

since A and $\hat{A}$ are equivalent by 4.1. However, consider another scenario where we take the first k columns of $\hat{A}$, that is, $\hat{A}_k = U_k S_k V_k^T$. With this choice of $\hat{A}_k$, we arrive at what is called a truncated SVD.

Theorem 4.3. Let $A = USV^T$ be the SVD of A, and then choose $A_k = U_k S_k V_k^T$ such that we use the first k columns of U and V. Then the error from this rank-k approximation is given by

$$||A - A_k||_F = \sum_{i=k+1}^{r} s_i$$

If $k = r = rank(A)$ there will be no error resulting from this decomposition, however, if $k < r$ we find some error, which we denote as a *truncated SVD*.

Proof. The result follows by simple application of Proposition 4.2. □

With Theorem 4.3 we will find that we can recover the expression for PCA that we arrived at in the previous section.

Theorem 4.4. Performing PCA using eigen-decomposition of the covariance matrix is equivalent to using the SVD of a data matrix X.

Proof. To begin, let $U_\Sigma \Lambda_\Sigma U_\Sigma^T$ be the top ℓ eigen-decomposition of the covariance matrix Σ. Using Theorem 3.4 from the last section we have

$$W = U_\Sigma$$

Now we create the ℓ-truncated SVD approximation to data matrix X with

$$X \approx U_X S_X V_X^T$$

Since the right singular vectors of X are also the eigenvectors of $X^T X$, we see that

$$V_X = U_\Sigma = W$$

To find the principal components let

$$Z = XW = W_X S_X V_X^T = U_X S_X$$

Upon approximately reconstructing the data we get,

$$\hat{X} = ZW^T = U_X S_X V_X^T$$

Which is exactly the same as a truncated SVD approximation. Clearly, this will be able to minimize the reconstruction error given by

$$L(W, Z) = \frac{1}{n} ||X - ZW^T||_F^2$$

Thus, PCA can be performed via eigen-decomposition of the covariance matrix Σ or an SVD decomposition of X. □

References

[1] Sheldon Jay Axler. *Linear Algebra Done Right*. Third. Undergraduate Texts in Mathematics. Springer International Publishing, 2015. ISBN: 9783319110806 3319110802. DOI: 10.1007/978-3-319-11080-6. URL: http://linear.axler.net/.

[2] Zenon Gniazdowski. "On the Correlation between Random Variables and their Principal Components". In: *Zeszyty Naukowe WWSI* 17.28 (Sept. 2023), p. 41. ISSN: 1896-396X, 2082-8349. DOI: 10.26348/znwwsi.28.41. URL: https://doi.org/10.26348/znwwsi.28.41.

[3] Ian T. Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2015.0202. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2015.0202.

[4] Xiaocan Li, Shuo Wang, and Yinghao Cai. *Tutorial: Complexity analysis of Singular Value Decomposition and its variants*. 2019. arXiv: 1906.12085 [math.NA].

[5] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[6] Henri De Plaen and Johan A. K. Suykens. *A Dual Formulation for Probabilistic Principal Component Analysis*. 2023. arXiv: 2307.10078 [cs.LG].