An Alternate Exposition of PCA

Ronak Mehta

1 Introduction

The purpose of this note is to present the principle components analysis (PCA) method alongside two important background topics: the construction of the singular value decomposition (SVD) and the interpretation of the covariance matrix of a random variable. Surprisingly, students have typically not seen these topics before encountering PCA in a statistics or machine learning course. Here, our focus is to understand *why* the SVD of the data matrix solves the PCA problem, and to interpret each of its components exactly. That being said, this note is not meant to be a self-contained introduction to the topic. Important motivations, such the maximum total variance or minimum reconstruction error optimizations over subspaces are not covered. Instead, this can be considered a supplementary background review to read before jumping into other expositions of PCA.

2 Linear Algebra Review

2.1 Notation

$\mathbb{R}^d$ is the set of real d-dimensional vectors, and $\mathbb{R}^{n\times d}$ is the set of real n-by-d matrices. A vector $\mathbf{x} \in \mathbb{R}^d$, is denoted as a bold lower case symbol, with its j-th element as x_j. A matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is denoted as a bold upper case symbol, with the element at row i and column j as X_{ij}, and $\mathbf{x}_{i\cdot}$ and $\mathbf{x}_{\cdot j}$ denoting the i-th row and j-th column respectively (the center dot might be dropped if it is clear in context). The identity matrix $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the matrix of ones on the diagonal and zeros elsewhere. The vector of all zeros is denoted $\mathbf{0}_d \in \mathbb{R}^d$, where as the matrix of all zeros is denoted $\mathbf{0}_{n\times d} \in \mathbb{R}^{n\times d}$. $\mathbf{e}_d^{(i)} \in \mathbb{R}^d$ is the d-dimensional vector with 1 in position i and 0 elsewhere.

2.2 Definitions and Results

- A set of vectors $\mathbf{v}_1, ..., \mathbf{v}_d \in \mathbb{R}^n$ is called **linearly independent** if for any real numbers $\alpha_1, ..., \alpha_d \in \mathbb{R}$,

  $$\alpha_1\mathbf{v}_1 + ... + \alpha_d\mathbf{v}_d = \mathbf{0}_n \implies \alpha_1 = ... = \alpha_d = 0.$$

  A set of vectors are called **linearly dependent** if they are not linearly independent.

- The Euclidean norm $||\mathbf{x}||_2$ of a vector $\mathbf{x} \in \mathbb{R}^d$ is given by

  $$||\mathbf{x}||_2 = \sqrt{\mathbf{x}^\top\mathbf{x}}.$$

- A square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is called **invertible** or **nonsingular** if that exists a matrix $\mathbf{A}^{-1}$ such that

  $$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_d.$$

  $\mathbf{A}^{-1}$ is called the **inverse** of $\mathbf{A}$, and is unique. A square matrix is invertible if and only if its columns are linearly independent.

- The **transpose** $\mathbf{X}^\top \in \mathbb{R}^{d\times n}$ of a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the matrix given by

$$X^\top_{ij} = X_{ji}.$$

For any two matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$.

- A square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is called **symmetric** if

$$\mathbf{A} = \mathbf{A}^\top.$$

- A set of vectors $\mathbf{v}_1, ..., \mathbf{v}_n \in \mathbb{R}^d$ are called **orthonormal** if

$$\mathbf{v}_{\cdot i}^\top \mathbf{v}_{\cdot j} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

If $n \leq d$, then the **Gram-Schmidt** algorithm can be used to produce $d - n$ more vectors $\mathbf{v}_{n+1}, ..., \mathbf{v}_d$ such that the set of $\mathbf{v}_1, ..., \mathbf{v}_d$ are orthonormal.

- A square matrix $\mathbf{V} \in \mathbb{R}^{d\times d}$ is called **orthogonal** if

$$\mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_d$$

For orthogonal matrices, the inverse $\mathbf{V}^{-1} = \mathbf{V}^\top$, by above. The columns of an orthogonal matrix are orthonormal. Note that if $\mathbf{V}$ is orthogonal, then $\mathbf{V}^\top$ is also orthogonal, meaning the rows of $\mathbf{V}$ are also orthognormal. These are also called rotation matrices, as applying them to a vector does not change the norm or distance between vectors (try it out!), thus only rotating the vector in space.

- A square matrix $\mathbf{D}$ is called **diagonal** if $D_{ij} = 0$ for all $i \neq j$. Pre-multiplying by a diagonal matrix scales the rows of a matrix, while post-multiplying scales the columns.

- A real number $\lambda \in \mathbb{R}$ and a non-zero vector $\mathbf{v} \in \mathbb{R}^d$ are called an **eigenvalue** and **eigenvector**, respectively, of a square matrix $\mathbf{A}$ if

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}. \tag{1}$$

A square matrix is invertible if and only if all of its eigenvalues are nonzero.

- A square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is called **diagonalizable** if there exists an invertible matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{A} = \mathbf{W}\mathbf{D}\mathbf{W}^{-1} \tag{2}$$

We say that $\mathbf{A}$ is *diagonalized* by $\mathbf{W}$. A matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is diagonalizable if and only if it has d linearly independent eigenvectors. To see this, we can write 2 as

$$\mathbf{A}\mathbf{W} = \mathbf{W}\mathbf{D},$$

and notice that every column satisfies 1. We are assured that these eigenvectors are linearly independent because $\mathbf{W}$ is invertible, therefore having linearly independent columns. The same steps can be used in reverse to achieve the "only if" direction. This means that when diagonalizing a matrix, the columns of $\mathbf{W}$ *are* the eigenvectors, and the diagonal entries of $\mathbf{D}$ are the eigenvalues. The decomposition 2 is called the **spectral decomposition** or **eigendecomposition**.

- A symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is called **positive definite** (P.D.) if for all $\mathbf{z} \in \mathbb{R}^d$ such that $\mathbf{z} \neq \mathbf{0}_d$,

$$\mathbf{z}^\top \mathbf{A}\mathbf{z} > 0.$$

  A symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is called **positive semi-definite** (P.S.D.) if for all $\mathbf{z} \in \mathbb{R}^d$,

$$\mathbf{z}^\top \mathbf{A}\mathbf{z} \geq 0.$$

  A symmetric matrix is positive definite if and only if all of its eigenvalues are positive. A symmetric matrix is positive semi-definite if and only if all of its eigenvalues are non-negative. As a result, a positive definite matrix is always invertible.

- The spectral theorem for real matrices states that any symmetric matrix can be diagonalized by an orthogonal matrix.

  Theorem 2.1 (Spectral Theorem)**.** *Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be symmetric. Then there exists an orthogonal $\mathbf{V} \in \mathbb{R}^{d \times d}$, and (real) diagonal $\mathbf{D} \in \mathbb{R}^{d \times d}$, such that:*

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$$

3 Construction of the SVD

Understanding the proof that the SVD exists will clarify its relationship to the eigendecomposition, which will make its use in PCA clear. We will first show an intermediate result, which will get us most of the way to the SVD.

Theorem 3.1. *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $n \leq d$. There exists an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$, a diagonal matrix $\mathbf{S}' \in \mathbb{R}^{n \times n}$, and a matrix $\mathbf{V}' \in \mathbb{R}^{n \times d}$ with orthonormal rows, such that*

$$\mathbf{X} = \mathbf{U}\mathbf{S}'\mathbf{V}'$$

Proof. Consider the matrix $\mathbf{A} = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n \times n}$. $\mathbf{A}$ is symmetric because

$$\mathbf{A}^\top = (\mathbf{X}\mathbf{X}^\top)^\top = (\mathbf{X}^\top)^\top \mathbf{X}^\top = \mathbf{X}\mathbf{X}^\top = \mathbf{A}$$

$\mathbf{A}$ is also positive semi-definite because for any $\mathbf{z} \in \mathbb{R}^n$,

$$\mathbf{z}^\top \mathbf{A}\mathbf{z} = \mathbf{z}^\top \mathbf{X}\mathbf{X}^\top \mathbf{z} = ||\mathbf{X}^\top \mathbf{z}||_2 \geq 0.$$

Thus, $\mathbf{A}$ admits an eigendecomposition

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

with $\mathbf{U}$ orthogonal and $D_{ii} \geq 0$ for $i = 1, ..., n$ from positive semi-definiteness. Let $\sigma_i = \sqrt{D_{ii}}$, and construct

$$\mathbf{S}' = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix},$$

with off-diagonal entries set to 0. Let $\mathbf{u}_{\cdot 1}, ..., \mathbf{u}_{\cdot n}$ be the columns and $\mathbf{U}$ and $\mathbf{v}_{1\cdot}, ..., \mathbf{v}_{n\cdot}$ be the (to be defined) rows of $\mathbf{V}'$. Arbitrarily, let $i = 1, ..., k$ index the rows such that $\sigma_i > 0$ and $i = k+1, ..., n$ index the rows with $\sigma_i = 0$. Then, for $i = 1, ..., k$, let

$$\mathbf{v}_{i\cdot} = \frac{1}{\sigma_i} \mathbf{u}_{\cdot i}^\top \mathbf{X}.$$

We first check that these rows are orthonormal.

$$\begin{aligned}
\mathbf{v}_{i\cdot} \mathbf{v}_{j\cdot}^\top &= \frac{1}{\sigma_i} \mathbf{u}_{\cdot i}^\top \mathbf{X} \left(\frac{1}{\sigma_j} \mathbf{u}_{\cdot j}^\top \mathbf{X} \right)^\top \\
&= \frac{1}{\sigma_i \sigma_j} \mathbf{u}_{\cdot i}^\top \mathbf{X} \mathbf{X}^\top \mathbf{u}_{\cdot j} \\
&= \frac{1}{\sigma_i \sigma_j} \mathbf{u}_{\cdot i}^\top \mathbf{A} \mathbf{u}_{\cdot j} \\
&= \frac{1}{\sigma_i \sigma_j} \mathbf{u}_{\cdot i}^\top \mathbf{U} \mathbf{D} \mathbf{U}^\top \mathbf{u}_{\cdot j} \\
&= \frac{1}{\sigma_i \sigma_j} (\mathbf{e}_n^{(i)})^\top \mathbf{D} (\mathbf{e}_n^{(j)}) \\
&= \begin{cases} \frac{\sigma_i \sigma_j}{\sigma_i \sigma_j} = 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

For the remaining rows of $\mathbf{V}'$, that is, $i = k+1, ..., n$, we can use the Gram-Schmidt algorithm to produce any set of vectors such that the rows remain orthonormal. Finally, we check that the proposed $\mathbf{U}$, $\mathbf{S}'$, and $\mathbf{V}'$ actually satisfy $\mathbf{X} = \mathbf{U}\mathbf{S}'\mathbf{V}'$. This is the same as claiming that $\mathbf{U}^\top \mathbf{X} = \mathbf{S}'\mathbf{V}'$, as $\mathbf{U}$ is orthogonal (its transpose is its inverse).

$$\begin{aligned}
\mathbf{S}'\mathbf{V}' &= \begin{bmatrix} \sigma_1 & & & & & & \\ & \ddots & & & & & \\ & & \sigma_k & & & & \\ & & & \sigma_{k+1} & & & \\ & & & & \ddots & & \\ & & & & & \sigma_n & \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1} \mathbf{u}_{\cdot 1}^\top \mathbf{X} \\ \vdots \\ \frac{1}{\sigma_k} \mathbf{u}_{\cdot k}^\top \mathbf{X} \\ \mathbf{v}_{k+1\cdot} \\ \vdots \\ \mathbf{v}_{n\cdot} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{u}_{\cdot 1}^\top \mathbf{X} \\ \vdots \\ \mathbf{u}_{\cdot k}^\top \mathbf{X} \\ \mathbf{0}_{(n-k) \times d} \end{bmatrix}
\end{aligned}$$

On the other hand,

$$\mathbf{U}^\top \mathbf{X} = \begin{bmatrix} \mathbf{u}_{\cdot 1}^\top \\ \vdots \\ \mathbf{u}_{\cdot k}^\top \\ \mathbf{u}_{\cdot k+1}^\top \\ \vdots \\ \mathbf{u}_{\cdot n}^\top \end{bmatrix} \mathbf{X}$$

$$= \begin{bmatrix} \mathbf{u}_{\cdot 1}^\top \mathbf{X} \\ \vdots \\ \mathbf{u}_{\cdot k}^\top \mathbf{X} \\ \mathbf{u}_{\cdot k+1}^\top \mathbf{X} \\ \vdots \\ \mathbf{u}_{\cdot n}^\top \mathbf{X} \end{bmatrix}$$

Thus, the first k rows of $\mathbf{S}'\mathbf{V}'$ equal the first k rows of $\mathbf{U}^\top \mathbf{X}$. We now much show that the last $n-k$ rows of $\mathbf{U}^\top \mathbf{X}$ are zero. If $\sigma_i = 0$, then $D_{ii} = \sigma_i^2 = 0$. We also have that

$$||\mathbf{u}_{\cdot i}^\top \mathbf{X}||_2^2 = \mathbf{u}_{\cdot i}^\top \mathbf{X}\mathbf{X}^\top \mathbf{u}_{\cdot i} = \mathbf{u}_{\cdot i}^\top \mathbf{A}\mathbf{u}_{\cdot i} = \mathbf{u}_{\cdot i}^\top \mathbf{U}\mathbf{D}\mathbf{U}^\top \mathbf{u}_{\cdot i} = (\mathbf{e}_n^{(i)})^\top \mathbf{D}(\mathbf{e}_n^{(i)}) = D_{ii} = 0$$

With norm zero, this means that $\mathbf{u}_{\cdot i}^\top \mathbf{X} = \mathbf{0}_d^\top$. Finally, we have $\mathbf{X} = \mathbf{U}\mathbf{S}'\mathbf{V}'$ with $\mathbf{U}$ orthogonal, $\mathbf{S}'$ diagonal, and $\mathbf{V}'$ with orthonormal rows. □

Using the above fact, we can produce the SVD.

Theorem 3.2 (Singular Value Decomposition)**.** *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$. There exists an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$, a matrix $\mathbf{S} \in \mathbb{R}^{n \times d}$ with $S_{ij} = 0$, and an orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$, such that*

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$$

This decomposition is called the **singular value decomposition** *(SVD). The columns of $\mathbf{U}$ and $\mathbf{V}$ are called the* **left** *and* **right singular vectors**, *respectively, while the diagonal elements of $\mathbf{S}$ are called the* **singular values** *of $\mathbf{X}$.*

Proof. If $n \leq d$, we can apply Theorem 3.1 to get $\mathbf{X} = \mathbf{U}\mathbf{S}'\mathbf{V}'$. Let

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}' & \mathbf{0}_{n \times (d-n)} \end{bmatrix}$$

and

$$\mathbf{V}^\top = \begin{bmatrix} \mathbf{V}' \\ [\text{GS}]_{(d-n) \times d} \end{bmatrix},$$

where $[\text{GS}]_{(d-n) \times d}$ denotes we fill in the bottom $d-n$ rows via Gram-Schmidt. The conditions of the theorem are satisfied. If $n > d$, then we can take the SVD of $\mathbf{X}^\top$, and transpose the resulting matrices to achieve the same result. □

In the proofs above, not that when we take the SVD of $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, $\mathbf{U}$ (the matrix of left singular vectors) contains the eigenvectors of $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$, while the singular values are the square roots of the singular values of $\mathbf{A}$. Finally, note that for a positive semi-definite matrix, the spectral decomposition and the singular value decomposition are the same (try going through the above steps, swapping the original matrix with its spectral decomposition). This means that the singular values and eigenvalues of a P.S.D. matrix are the same as well.

4 Covariance Matrix of a Random Variable

In this section, we discuss the covariance matrix, which is another major component of PCA. Understanding what this matrix and its eigenpairs represent is key to understanding why the algorithm gets us what we want. Let $\mathbf{x}$ be a random vector that realizes in $\mathbb{R}^d$. The mean vector $\mu = \mathbb{E}[\mathbf{x}]$ is the expected value of $\mathbf{x}$ taken element wise. The covariance matrix

$$\mathbf{\Sigma} = \text{Cov}(\mathbf{x}) = \mathbb{E}\left[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top\right] = \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] - \mu\mu^\top. \tag{3}$$

3 is analogous to the variance formula for univariate random variable x, i.e. $\text{Var}(x) = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$. Each entry of the covariance matrix represents the covariance between components of $\mathbf{x}$, that is, $\mathbf{\Sigma}_{ij} = \text{Cov}(x_i, x_j)$. The covariance matrix $\mathbf{\Sigma}$ is positive semi-definite, as given any $\mathbf{z} \in \mathbb{R}^d$, we have:

$$\begin{aligned}
\mathbf{z}^T\mathbf{\Sigma}\mathbf{z} &= \mathbf{z}^\top\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] - \mu\mu^\top\right)\mathbf{z} \\
&= \mathbb{E}\left[\mathbf{z}^\top\mathbf{x}\mathbf{x}^\top\mathbf{z}\right] - \mathbf{z}^\top\mu\mu^\top\mathbf{z}] \\
&= \mathbb{E}\left[(\mathbf{z}^\top\mathbf{x})^2\right] - (\mathbf{z}^\top\mu)^2 \\
&= \text{Var}\left(\mathbf{z}^\top\mathbf{x}\right)
\end{aligned}$$

$\text{Var}\left(\mathbf{z}^\top\mathbf{x}\right)$ is nonnegative for any $\mathbf{z}$, completing the proof. Even when $\mathbf{z}$ is non-zero, we cannot say more than that, because $\mathbf{z}^\top\mathbf{x}$ can be constant even when each coordinate of $\mathbf{x}$ has variance. Take for example

$$\mathbf{x} = \begin{bmatrix} y \\ 1 - y \end{bmatrix}$$

where $y \sim \text{Unif}(0, 1)$. Letting $\mathbf{z} = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, we have that $\mathbf{z}^\top\mathbf{x} = 1$ with probability 1, i.e. $\text{Var}\left(\mathbf{z}^\top\mathbf{x}\right) = 0$. For $\mathbf{\Sigma}$ to be positive definite, it would then mean that $\text{Var}\left(\mathbf{z}^\top\mathbf{x}\right) > 0$ for every non-zero $\mathbf{z}$, meaning that no linear combination of the coordinates of $\mathbf{x}$ can result in a random variable that is almost surely a constant.

Generalizing this, let $\mathbf{v}$ be an eigenvector of $\mathbf{\Sigma}$ associated with eigenvalue 0, if one exists (i.e. $\mathbf{\Sigma}$ is P.S.D. but not P.D.). We have that $\mathbb{E}\left[\mathbf{v}^\top\mathbf{x}\right] = \mathbf{v}^\top\mu$. Then,

$$\text{Var}\left(\mathbf{v}^\top\mathbf{x}\right) = \mathbf{v}^\top\mathbf{\Sigma}\mathbf{v} = 0 \implies \mathbb{P}\left[\mathbf{v}^\top\mathbf{x} = \mathbf{v}^\top\mu\right] = 1$$

Let's assume that $\mu = \mathbf{0}_d$, which would be true if we centered our data. Then $\mathbb{P}\left[\mathbf{v}^\top\mathbf{x} = 0\right] = 1$. Now, let k be the rank of $\mathbf{\Sigma}$. Because $\mathbf{\Sigma}$ is P.S.D., its rank is equal to the number of nonzero

eigenvalues. (To see this, note that P.S.D. matrices have their spectral decomposition and SVD equal, so the number of nonzero singular values is the number of nonzero eigenvalues.) Let $\mathbf{v}_1$, ..., $\mathbf{v}_k$ be the eigenvectors associated with the k non-zero eigenvalues, and $\mathbf{v}_{k+1}$, ..., $\mathbf{v}_d$ be the ones associated with zero. Take any $\mathbf{v} \in \text{span}\{\mathbf{v}_{k+1}, ..., \mathbf{v}_n\}$, i.e. $\mathbf{v} = \sum_{j=k+1}^{d} \alpha_j \mathbf{v}_j$.

$$\begin{aligned}
\mathbb{P}\left[\mathbf{v}^\top \mathbf{x} \neq 0\right] &= \mathbb{P}\left[\sum_{j=k+1}^{d} \alpha_j \mathbf{v}_j^\top \mathbf{x} \neq 0\right] \\
&\leq \mathbb{P}\left[\bigcup_{j=k+1}^{d} \alpha_j \mathbf{v}_j^\top \mathbf{x} \neq 0\right] \\
&\leq \sum_{j=k+1}^{d} \mathbb{P}\left[\alpha_j \mathbf{v}_j^\top \mathbf{x} \neq 0\right] \\
&= 0
\end{aligned}$$

Thus, $\mathbf{x}$ is almost surely in the orthogonal space of $\text{span}\{\mathbf{v}_{k+1}, ..., \mathbf{v}_d\}$, and because $\mathbf{v}_1, ..., \mathbf{v}_d$ form an orthonormal basis for $\mathbb{R}^d$, we have

$$\mathbb{P}\left[\mathbf{x} \in \text{span}\{\mathbf{v}_1, ..., \mathbf{v}_k\}\right] = 1$$

Thus, $\mathbf{x}$, while realized in $\mathbb{R}^d$, essentially lives on a subspace of dimension k. While we know that this is true if $\mathbf{\Sigma}$ has 0 as an eigenvalue, what do the eigenvalues of $\mathbf{\Sigma}$ actually represent?

We know that $\mathbf{\Sigma}$ is P.S.D., so we can write

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

Let's interpret of $\mathbf{\Lambda}$ and $\mathbf{V}$. We know that from $\mathbf{V}$, we can get the subspace that $\mathbf{x}$ actually lives in, as it contains the eigenvectors of $\mathbf{\Sigma}$ as its columns. Additionally, $\mathbf{V}$ is orthogonal, so $\mathbf{V}^\top = \mathbf{V}^{-1}$.

$$\begin{aligned}
\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \implies \mathbf{\Lambda} &= \mathbf{V}^\top \mathbf{\Sigma} \mathbf{V} \\
&= \mathbf{V}^\top \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] - \mathbb{E}\left[\mathbf{x}\right]\mathbb{E}\left[\mathbf{x}\right]^\top\right)\mathbf{V} \\
&= \mathbb{E}\left[\mathbf{V}^\top \mathbf{x}\left(\mathbf{V}^\top \mathbf{x}\right)^\top\right] - \mathbb{E}\left[\mathbf{V}^\top \mathbf{x}\right]\mathbb{E}\left[\mathbf{V}^\top \mathbf{x}\right]^\top \\
&= \text{Cov}\left(\mathbf{V}^\top \mathbf{x}\right)
\end{aligned}$$

The random vector $\mathbf{V}^\top \mathbf{x}$ is just $\mathbf{x}$ in a rotated basis. However, in this basis, all of the dimensions of $\mathbf{V}^\top \mathbf{x}$ are uncorrelated! ($\mathbf{V}^\top \mathbf{x}$ is a transformation of $\mathbf{x}$ into the basis of $\mathbf{v}_1, ..., \mathbf{v}_d$, the columns of $\mathbf{V}$, because $\mathbf{V}$ is orthogonal.) Now, we can consider the random vector $\mathbf{z} = \mathbf{V}^\top \mathbf{x}$ with linearly independent dimensions. By the above argument, letting λ_j be the j-th diagonal element of $\mathbf{\Lambda}$ is, we have

$$\lambda_j = \text{Var}\left(z_j\right) = \text{Var}\left(\mathbf{v}_j^\top \mathbf{x}\right)$$

because

$$\mathbf{z} = \mathbf{V}^\top \mathbf{x} = \begin{bmatrix} \mathbf{v}_1^\top \mathbf{x} \\ \vdots \\ \mathbf{v}_d^\top \mathbf{x} \end{bmatrix}$$

We can also write

$$\mathbf{x} = \mathbf{V}\mathbf{z} = z_1\mathbf{v}_1 + ... + z_d\mathbf{v}_d = (\mathbf{v}_1^\top \mathbf{x})\mathbf{v}_1 + ... + (\mathbf{v}_d^\top \mathbf{x})\mathbf{v}_d = \sum_{j=1}^{d} (\mathbf{v}_j^\top \mathbf{x})\mathbf{v}_j$$

as the orthogonal projection of $\mathbf{x}$ onto the orthonormal basis $\{\mathbf{v}_{\cdot 1}, ..., \mathbf{v}_{\cdot d}\}$. This lets us interpret the eigenvectors associated with 0 as directions with no variance, restricting $\mathbf{x}$ to the other directions. We will keep this interpretation in mind in Section 5.

We know that $\text{Cov}(\mathbf{z}) = \mathbf{\Lambda}$ is diagonal, so the dimensions of $\mathbf{z}$ are uncorrelated. How might this help us? Consider some $\beta \in \mathbb{R}^d$, and the random variable $\beta^\top \mathbf{z}$ (such as the prediction in linear regression).

$$\text{Var}\left(\beta^\top \mathbf{z}\right) = \text{Var}\left(\sum_{j=1}^{d} \beta_j z_j\right) = \sum_{j=1}^{d} \beta_j^2 \text{Var}(z_j) + \sum_{i \neq j} \beta_i \beta_j \text{Cov}(z_i, z_j)$$

If the dimensions of $\mathbf{z}$ are uncorrelated, then $\text{Cov}(z_i, z_j) = 0$ for $i \neq j$, so it's clear that dropping a dimension will necessarily decrease the variance of $\beta^\top \mathbf{z}$. If the covariance terms are non-zero, then it is unclear how variance is affected when dropping dimensions. So, it is a problem of general interest to be able to represent an arbitrary random variable $\mathbf{x}$ as a rotation $\mathbf{z}$ that has this property of interpretability. Of course, the matrix $\mathbf{V}$ depends on $\mathbf{\Sigma}$, which we do not have access to as it depends on the true distribution of $\mathbf{x}$. In the next section, we see how to estimate the quantities of interest from data, putting together the ideas from Section 3 and Section 4.

5 Principle Components Analysis

In Section 3, we constructed the SVD of $\mathbf{X}$ using the spectral decomposition of $\mathbf{X}\mathbf{X}^\top$, and saw that the eigenvectors of $\mathbf{X}\mathbf{X}^\top$ became the left singular vectors of $\mathbf{X}$. Given that $\mathbf{X}\mathbf{X}^\top$ is P.S.D., we also saw that the number of nonzero singular values of $\mathbf{X}$ was equal to the rank of $\mathbf{X}\mathbf{X}^\top$. In 4, we observed properties of the covariance matrix $\mathbf{\Sigma}$ of random variable $\mathbf{x}$. Specifically, letting $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the spectral decomposition of $\mathbf{\Sigma}$, the variable $\mathbf{z} = \mathbf{V}^\top \mathbf{x}$ has uncorrelated dimensions, each with variance equal to the corresponding eigenvalue of Σ. We saw that if any of these eigenvalues are zero, $\mathbf{x}$ (hence $\mathbf{z}$) lies on a lower dimensional subspace in $\mathbb{R}^d$. Thus, representing $\mathbf{x}$ is this lower-dimensional representation will be a more faithful view of the data and might come with its own statistical benefits. In this section, we will answer two questions, namely how to estimate this representation from sample data and how to interpret eigenvalues that are very close to zero. The PCA algorithm will fall out of these answers.

Let's first see how we might infer this orthogonalized representation from a dataset $\mathbf{x}_1, ..., \mathbf{x}_n$ of independent observations of $\mathbf{x}$. We will stack these observations as a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where i-th row of $\mathbf{X}$ is $\mathbf{x}_i$, and the j-th column contains the values of the j-th dimension for each of the

n observations. If $\mathbf{\Sigma}$ is P.D., then no column of $\mathbf{X}$ can be represented as a linear combination of the others.

If letting $\mathbf{z}_i = \mathbf{V}^\top \mathbf{x}_i$ (hence $\mathbf{z}_i^\top = \mathbf{x}_i^\top \mathbf{V}$), consider the matrix $\mathbf{Z}$, containing the $\mathbf{z}_i$'s as rows. We see that

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_n^\top \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{V} \\ \vdots \\ \mathbf{x}_n^\top \mathbf{V} \end{bmatrix} = \mathbf{X}\mathbf{V}$$

In Principle Components Analysis (PCA), we are interested in recovering this matrix $\mathbf{Z} = \mathbf{X}\mathbf{V} \in \mathbb{R}^{n \times d}$, the orthogonalized representation of $\mathbf{X}$. $\mathbf{Z}$ is called the **loading matrix**, while the columns of $\mathbf{V}$ are called the **principle components**. Other than the benefit of being able to observe the components of $\mathbf{x}$ that are uncorrelated, which is interesting in its own right, we can drop columns that have low values of λ_j, as they may not describe the principle patterns in the data. Of course, because $\mathbf{V}$ relies on the population parameter $\mathbf{\Sigma}$, we do not have access to it in general. How can we estimate $\mathbf{Z}$?

Let's assume that $\mathbb{E}[\mathbf{x}] = \mu = \mathbf{0}_d$, as before. Then, a natural estimate of the covariance matrix and its spectral decomposition is is

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top \mathbf{X} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^\top$$

and we can estimate the loading matrix

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{V}}$$

While this checks out mathematically, from a numerical viewpoint, there are some shortcomings. Both the computation of $\hat{\mathbf{\Sigma}}$ as well as its eigendecomposition are expensive operations. Additionally, eigendecomposition is less numerically stable than singular value decomposition. Is there a way to compute $\hat{\mathbf{Z}}$ without either of the above steps? The answer is in the SVD! Consider the SVD of the matrix below.

$$\tfrac{1}{\sqrt{n}}\mathbf{X} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^\top$$

This matrix is chosen so that "$\hat{\mathbf{V}}$" is the same "$\hat{\mathbf{V}}$" that we discussed before - an estimate of the eigenvectors of $\hat{\mathbf{\Sigma}}$, as

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top \mathbf{X} = \tfrac{1}{\sqrt{n}}\mathbf{X}^\top \left(\tfrac{1}{\sqrt{n}}\mathbf{X}^\top\right)^\top = \hat{\mathbf{V}}\hat{\mathbf{S}}^\top \hat{\mathbf{U}}^\top \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^\top = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^\top.$$

This is because $\hat{\mathbf{U}}^\top \hat{\mathbf{U}} = \mathbf{I}_n$ and $\hat{\mathbf{S}}^\top \hat{\mathbf{S}} = \hat{\mathbf{\Lambda}}$, as the singular values of $\frac{1}{\sqrt{n}}\mathbf{X}^\top$ are the square roots of the eigenvalues of $\frac{1}{n}\mathbf{X}^\top \mathbf{X} = \hat{\mathbf{\Sigma}}$. To compute $\hat{\mathbf{Z}}$, we can apply

$$\begin{aligned}
\hat{\mathbf{Z}} &= \mathbf{X}\hat{\mathbf{V}} \\
&= \sqrt{n}\left(\tfrac{1}{\sqrt{n}}\mathbf{X}\right)\hat{\mathbf{V}} \\
&= \sqrt{n}\left(\hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^\top\right)\hat{\mathbf{V}} \\
&= \sqrt{n}\hat{\mathbf{U}}\hat{\mathbf{S}}\left(\hat{\mathbf{V}}^\top \hat{\mathbf{V}}\right) \\
&= \sqrt{n}\hat{\mathbf{U}}\hat{\mathbf{S}}
\end{aligned}$$

The orthogonalized $\hat{\mathbf{Z}}$ can be computed fully by the SVD of $\mathbf{X}$! You may have not seen the $\sqrt{n}$ factor before, but this is only written so that the d-by-d matrix $\hat{\mathbf{V}}$ in the SVD of $\mathbf{X}$ is the same as the eigenvector matrix of $\hat{\mathbf{\Sigma}}$. Multiplying the entire dataset by a number will not affect the result of statistical inference.

It is now clear how to compute the representation $\mathbf{Z}$ with uncorrelated dimensions. If any of the dimensions had zero variance, we would see a column of zeros in this matrix. In reality, this is very unlikely to happen, and will occur with 0 probability for any non-degenerate random variable. What is much more likely, however, is that there is an eigenvalue λ_j of $\mathbf{\Sigma}$ such that

$$\lambda_j = \epsilon \approx 0$$

for some small $\epsilon > 0$. This means that along direction $\mathbf{v}_j$, $\mathbf{x}$ has very little variance, so $\mathbf{x}$ approximately lies on the subspace $\mathbf{v}_j^\perp$, that is the space orthogonal to $\mathbf{v}_j$. This also means that a column of $\mathbf{X}$ is **approximately** a linear combination of the others. We thus accomplish a similar goal by dropping columns from $\mathbf{Z}$ associated with small singular values of $\mathbf{X}$, which are also square roots of small eigenvalues of $\mathbf{X}\mathbf{X}^\top$. Most real data will be of this "approximately low rank" form. There are various heuristics for determining whether a column has enough variance to remain in the reduced representation, which may differ by application.

In summary, the SVD of $\mathbf{X}$ gives us the ability to efficiently and stably represent the data in an interpretable form, after which we can reduce it and perform downstream inference. The use of the SVD in PCA can fall out of other motivations, such as finding the subspace of minimum reconstruction error, but this exposition might highlight the meaning behind each of the elements in the final result.

6 Summary

In this note, we reviewed linear algebra fundamentals and constructed the singular value decomposition mathematically to better understand its elements and properties. We then analyzed the covariance matrix of a random variable, discovering that its eigenvectors are basis vectors for which coordinates of the random variable are uncorrelated, and its eigenvalues are exactly the variance in those dimensions. We then explored principle components analysis as a means to estimate the representation of data in this basis, as well as the motivation for doing so. PCA is a well-studied topic, and this is one of many developments of the material. Feedback is appreciated!