An Alternative Exposition of PCA

Ronak Mehta

1 Introduction

The purpose of this note is to take a slightly different route in motivating and understanding principle components analysis (PCA). The algorithm involves taking the singular value decomposition (SVD) of the data matrix, but often, it is less clear why exactly the left singular vectors associated with large singular values are directions oh high variance. Students have sometimes never seen the singular value decomposition in other contexts before encountering PCA in a statistics course, and the details are swept under the rug. Here, we construct the singular value decomposition by relating it to the spectral decomposition. We then analyze the population covariance and sample covariance matrix of a random vector, interpret its eigenvectors and eigenvalues, and relate them to uncorrelated bases for the random variable. Finally, we show that the SVD provides an efficient way to compute the orthogonalized data matrix, which is the goal of PCA. This note is meant to be supplementary, as we do not cover important connections such as finding the subspace with minimum reconstruction error.

2 Linear Algebra Review

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 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.

- A set of vectors $\mathbf{v}_1, ..., \mathbf{v}_d \in \mathbb{R}^n$ are 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

  $$\mathbf{X}_{ij}^\top = \mathbf{X}_{ji}.$$

  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 (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-multiplyin scales the columns.

- A real number $\lambda \in \mathbb{R}$ and 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

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* **U** *and* **V** *are called the* **left** *and* **right singular vectors***, respectively, while the diagonal elements of* **S** *are called the* **singular values** *of* **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

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

where $[\mathrm{GS}]_{(d-n)\times d}$ 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

Let $\mathbf{x}$ be a random vector that realizes in $\mathbb{R}^d$. The mean vector $\mu = \mathbb{E}\left[\mathbf{x}\right]$ is the expected value of $\mathbf{x}$ taken element wise. The covariance matrix

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

3 is analogous to the variance formula for univariate random variable x, i.e. $\operatorname{Var}\left(x\right) = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2$. Each entry of the covariance matrix represents the covariance between components of $\mathbf{x}$, that is, $\boldsymbol{\Sigma}_{ij} = \operatorname{Cov}\left(x_i, x_j\right)$. The covariance matrix $\boldsymbol{\Sigma}$ is positive semi-definite. Give any $\mathbf{z} \neq 0 \in \mathbb{R}^d$, we have:

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

This value is nonnegative for any $\mathbf{z}$, completing the proof. 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. $\operatorname{Var}\left(\mathbf{z}^\top\mathbf{x}\right) = 0$. For $\boldsymbol{\Sigma}$ to be positive definite, it would then mean that $\operatorname{Var}\left(\mathbf{z}^\top\mathbf{x}\right) > 0$ for every non-zero $\mathbf{z}$, meaning that no linear combination of the coordinates of $\mathbf{x}$ can result in a constant.

In statistical applications, we are usually given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, where every row of X is an independent observation of a random variable $\mathbf{x}$. The j-th column contains the values of the j-th dimension of each of the n observations of $\mathbf{x}$. If $\boldsymbol{\Sigma}$ is P.D., then no column of X can be represented as a linear combination of the others. This reveals a goal - if the data matrix did have columns that were linearly dependent, we might want to drop those columns, reducing the dimension of our dataset, which might have benefits for statistical inference. This can be done losslessly if $\boldsymbol{\Sigma}$ is not P.D. and has some of its eigenvalues (hence singular values) set to zero. What if there is an eigenvalue λ_j of $\boldsymbol{\Sigma}$ such that

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

for some small $\epsilon > 0$? This means that there is a dimension of $\mathbf{x}$ (hence a column of $\mathbf{X}$) that is **approximately** a linear combination of the others. We still might be able to benefit statistically by dropping such a column from our dataset. If two columns are (or nearly are) linear scalings of one another, which one should we drop? As it turns out, in any dataset, we can have columns highly correlated with one another (nearly linear combinations of one another). This means that the non-diagonal entries of $\mathbf{\Sigma}$ can be far from zero. Is there a way to represent the data such that each column affects the data independently? In other words, does there exist a basis to represent the dimensions of the data are uncorrelated.?

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

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

What is the interpretation of $\mathbf{\Lambda}$?

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

The random vector $\mathbf{V}^\top \mathbf{x}$ is just $\mathbf{x}$ in a rotated basis, but 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 $\mathbf{v}_{\cdot 1}, ..., \mathbf{v}_{\cdot 1}$, 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. Letting λ_j be the j-th diagonal element of $\mathbf{\Lambda}$ is , we have

$$\lambda_j = \mathrm{Var}\left(z_j\right) = \mathrm{Var}\left(\mathbf{v}_{\cdot j}^\top \mathbf{x}\right)$$

because

$$\mathbf{z} = \mathbf{V}^\top \mathbf{x} = \begin{bmatrix} \mathbf{v}_{\cdot 1}^\top \mathbf{x} \\ \vdots \\ \mathbf{v}_{\cdot d}^\top \mathbf{x} \end{bmatrix}$$

We can also write

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

as the orthogonal projection of $\mathbf{x}$ onto the orthonormal basis $\{\mathbf{v}_{\cdot 1}, ..., \mathbf{v}_{\cdot d}\}$. In Principle Components Analysis (PCA), we are interested in recovering the 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.

5 Estimating Principle Components

Let's assume that $\mathbb{E}[\mathbf{x}] = \mathbf{0}_d$. (We can achieve this in practice by subtracting the sample mean from each point.) Then, a natural estimate of the covariance matrix is

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

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

and we can estimate the loading matrix

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

From here, columns can be dropped by any method. While this checks out theoretically, from a numerical viewpoint, there are some shortcomings. Both the computation of $\hat{\mathbf{\Sigma}}$ as well as its eigendecomposition are notoriously expensive operations. Additionally, eigendecomposition is less numerically stable than singular value decomposition. (You should believe that taking the eigendecomposition of $\mathbf{X}\mathbf{X}^\top$ is not how the SVD of $\mathbf{X}$ is typically computed by linear algebra libraries.) 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 $\mathbf{\Sigma}$, as

$$\tfrac{1}{\sqrt{n}}\mathbf{X}^\top \left(\tfrac{1}{\sqrt{n}}\mathbf{X}^\top\right)^\top = \tfrac{1}{n}\mathbf{X}^T\mathbf{X} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^\top$$

and this the $\hat{\mathbf{V}}$ used in the SVD of $\tfrac{1}{\sqrt{n}}\mathbf{X}^\top$. 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 $\hat{V}$ in the SVD of 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.

Summary In this section, we set understand our data better by introducing a new coordinate system in which the individual dimensions of our data points were uncorrelated. After confirming that such a basis exists, we identified the (new) directions that have the largest variance. Because these directions depend on unknown population parameters, we discussed how to estimate them from data. In the process, we discovered why the SVD of the data matrix achieves this goal, and how to interpret both the left singular vectors and the singular values, which are both used in PCA. Of note is that other motivations of PCA typically aim to find $k < d$ uncorrelated directions that have the maximum variance, as opposed to here, where we picked one particular uncorrelated basis and studied which among these directions have maximum variance.