Singular Value Decomposition

Aiden Kenny
Franklin & Marshall College
Summer 2019

1 Introduction

The purpose of this paper is to provide a comprehensive overview of the *singular value decomposition* (SVD) of a matrix. Any $n \times p$ matrix $\mathbf{A}$ can be decomposed into three separate matrices, given as

$$\mathbf{A} = \mathbf{USV}^T, \tag{1}$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix, $\mathbf{S}$ is an $n \times p$ diagonal matrix, and $\mathbf{V}$ is a $n \times n$ orthogonal matrix.

We could continue to discuss the properties of the three matrices in (1), but I believe it is important to ensure that the reader has a solid foundation in the necessary theory and terminology needed. For this reason, section 2 provides various linear algebra pre-requisites needed to understand the SVD. It is not until section 3 that (1) is explored further from a theoretical perspective. Section 4 describes a systematic approach to determining the SVD of a matrix, as well as various computational methods.

2 Linear Algebra: Notation and Methods

In this section we have a brief review of the theory of several linear algebra topics relative to SVD. Note that this review is in no way comprehensive or flushed-out, as that would lead us too far astray from the goal of this paper. For example, topics such as elementary operations or inverse matrices, which are extremely important for linear algebra as a whole, or the idea of treating a matrix as a linear transformation, are completely ignored.

2.1 Vectors and Matrices

For a positive integer n, a *vector* (or *n-vector*) is an ordered list (known as an *array*) of n numbers, called *elements* or *scalars*. The *order* of a vector is the number of elements it contains, so an n-vector has an order of n. We will generally denote a vector using a lowercase boldface letter , e.g. $\mathbf{v}$. The i^{th} element in $\mathbf{v}$ is denoted using a lowercase letter (the same letter that denotes the vector), e.g. v_i is the i^{th} element of $\mathbf{v}$). For our purposes, we will assume that the elements of a vector are all real numbers; that is, $v_i \in \mathbb{R}$ for all v_i. The notation $\mathbb{R}^n$ is used to denote all n-vectors with real elements.

When we want to express the specific elements of a vector, it is written as either

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \tag{2}$$

or

$$\mathbf{v} = (v_1, \ldots, v_n). \tag{3}$$

Vectors are, by default, thought of as "columns," so (2) seems more natural in that sense. However, for notational purposes, it is often easier to express vectors in the form of (3). Geometrically, an n-vector may be thought of as a point in n-dimensional space, where each element of the matrix is the coordinate along one of the axis.

An *$n \times p$ matrix* is a 2-dimensional array of numbers, with n *rows* and p *columns*. The numbers n and p denote the *shape* of the matrix; if a matrix has the same number of rows and columns (i.e. $n = p$),

then the matrix is said to be *square*. We will denote a matrix using a capital boldface letter, e.g. $\mathbf{A}$. We sometimes also wish to denote a matrix as $\mathbf{A}_{n\times p}$, to indicate the shape of the matrix at a glace. To denote an element in the i^{th} row and j^{th} column of a matrix, we use a lowercase letter (the same letter that denotes the matrix) combined with two subscripts i and j, e.g. $a_{i,j}$ is the element in the i^{th} row and j^{th} column of the matrix $\mathbf{A}$. We again assume that the elements of a matrix are all real numbers. The notation $\mathbb{R}^{n\times p}$ is used to denote all $n \times p$ matrices with real elements.

When we want to express the specific elements of a matrix, it is written as either

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \cdots & a_{n,p} \end{pmatrix} \tag{4}$$

or

$$\mathbf{A} = (a_{i,j}), \tag{5}$$

for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$.

Two matrices are equal only if they have the same shape and the corresponding elements of each matrix are equal. So, for two matrices $\mathbf{A}_{n\times p}$ and $\mathbf{B}_{m\times q}$ to be equal, it must be true that $n = m$, $p = q$, and $a_{i,j} = b_{i,j}$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$.

2.1.1 Relationship Between Vectors and Matrices

There is a close relationship between vectors and matrices. A *column vector* is an $n \times 1$ matrix, and a *row vector* is a $1 \times n$ matrix. Although it is important to understand that vectors and matrices are fundamentally different objects, for our purposes we will treat them as one in the same. So from now on, (2) and (3) will be columns vectors. To explicitly examine the elements, we denote a row vector as

$$\mathbf{v} = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \tag{6}$$

or

$$\mathbf{v} = (v_1, \ldots, v_n)^T. \tag{7}$$

Again, (6) explains the row vector in a more "natural" sense, while (7) is used for notational purposes.

Conversely, we can think of an $n \times p$ matrix $\mathbf{A}$ as two different groups of vectors. First, we can think of the *rows* of $\mathbf{A}$ as a collection of row vectors, that is, the matrix is a collection of n vectors all with order p. We can also think of the *columns* of $\mathbf{A}$ as a collection of column vectors; that is, the matrix is a collection of p different vectors, all with order n.

$$\mathbf{A} = \underbrace{\begin{pmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,p} \end{pmatrix}}_{\text{A collection of row vectors}} = \underbrace{\begin{pmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,p} \end{pmatrix}}_{\text{A collection of column vectors}} \tag{8}$$

As a result, one additional way to notate a matrix is by writing a matrix as a row vector, where each entry in the vector is a column vector (i.e. a vector of vectors). This is given by

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{pmatrix}, \tag{9}$$

where $\mathbf{a}_i$ is the i^{th} column vector of $\mathbf{A}$.

2.2 Matrix (and Vector) Addition and Scalar Multiplication

Given two matrices $\mathbf{A}_{n\times p}$ and $\mathbf{B}_{n\times p}$, we can perform *matrix addition* on them to get a new matrix, generally denoted as $\mathbf{A} + \mathbf{B}$. This operation is defined by its elements and is given by

$$(a + b)_{i,j} = a_{i,j} + b_{i,j}, \tag{10}$$

for all $i \in \{1, \ldots, n\}$ and $j \in \{i, \ldots, p\}$. In other words, to get the $(i,j)^{\text{th}}$ element of $\mathbf{A} + \mathbf{B}$, just add together the $(i,j)^{\text{th}}$ elements of $\mathbf{A}$ and $\mathbf{B}$.

Another basic operation is *scalar multiplication*, where a matrix is multiplied by some real number c. Denoted as $c\mathbf{A}$, this operation is again defined by its elements and is given by

$$(ca)_{i,j} = c \cdot a_{i,j}. \tag{11}$$

In other words, to get the $(i,j)^{\text{th}}$ element of $c\mathbf{A}$, just multiply the $(i,j)^{\text{th}}$ element of $\mathbf{A}$ by c.

These operations are the same for vectors (remember that we are treating them as essentially the same type of object). That is, given two vectors $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$, we have $\mathbf{v} + \mathbf{w} = (v_1 + w_1, \ldots, v_n + w_n)$ and $c\mathbf{v} = (cv_1, \ldots, cv_n)$.

2.3 Matrix (and Vector) Multiplication

Keeping in mind that we will be treating vectors as $n \times 1$ matrices, we will now introduce the idea of multiplying two matrices together. *Matrix Multiplication* between two matrices $\mathbf{A}_{n\times p}$ and $\mathbf{B}_{p\times q}$ gives us a third matrix, denoted as $\mathbf{AB}$, and is defined by its elements as

$$(ab)_{i,j} = \sum_{k=1}^{p} a_{i,k} b_{k,j}. \tag{12}$$

This formula can be somewhat daunting at first glance, and there are several subtleties that need to be understood. First, we can see that in order for this operation to be defined, the number of *columns* of $\mathbf{A}$ must equal the number of *rows* of $\mathbf{B}$. As a result, the new matrix $\mathbf{AB}$ has n rows and q columns; that is, it has the same number of rows as $\mathbf{A}$ and the same number of columns as $\mathbf{B}$. (12) gives us a systematic way to compute each element of $\mathbf{AB}$ by using the rows of $\mathbf{A}$ and the columns of $\mathbf{B}$, but this process is rather tedious, and computer software such as `R` or `MATLAB` can be used to easy multiply matrices.

2.4 Some Special Vectors and Matrices

We have already spend a good deal of time talking about the transpose. However, there are several other special vectors and matrices that we should become familiar with before going forward.

2.4.1 The Null Vector and the Ones Vector

The *null vector*, denoted as $\mathbf{0}$, is simply an n-vector where *every* element is equal to 0. In other words, $\mathbf{0} = (0, \ldots, 0)$. Similarly, the *ones vector*, denoted as $\mathbf{1}$, is an n-vector where every element is equal to 1. In other words, $\mathbf{1} = (1, \ldots, 1)$.

2.4.2 The Transpose

For an $n \times p$ matrix $\mathbf{A}$, its *transpose*, denoted as $\mathbf{A}^T$, is defined as

$$\mathbf{A}^T = (a^T_{i,j}) = (a_{j,i}), \tag{13}$$

for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. In other words, we take the rows of $\mathbf{A}$ and re-write them as the columns of $\mathbf{A}^T$. Note that we are using the $a^T_{i,j}$ to denote the $(i,j)^{\text{th}}$ element of $\mathbf{A}^T$. If we want to explicitly look at the elements of $\mathbf{A}^T$, then the transpose of (4) can be written as

$$\mathbf{A}^T = \begin{pmatrix} a_{1,1} & \cdots & a_{p,1} \\ \vdots & \ddots & \vdots \\ a_{1,n} & \cdots & a_{n,p} \end{pmatrix}. \tag{14}$$

We should note the relationship between the transpose and vectors. Namely, the transpose of a column vector (4) is a row vector (6), and similarly, the transpose of a row vector is a column vector.

Similar to (9), we can represent $\mathbf{A}$ as a vector. Only this time, we will represent it as a column vector, where the i^{th} entry is the transpose of the i^{th} column vector of $\mathbf{A}$, which is a row vector. This notation is given by

$$\mathbf{A}^T = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{pmatrix}, \tag{15}$$

where $\mathbf{a}_i$ is the i^{th} column vector of $\mathbf{A}$.

The transpose has several useful properties that we will take advantage of when working with the SVD. To start, we will show that for any matrix $\mathbf{A}_{n\times p}$, we have $(\mathbf{A}^T)^T = \mathbf{A}$. That is, taking the transpose of the transpose returns the original matrix. This is easy to see; recall from (13) that $a^T_{i,j} = a_{j,i}$. Then applying the transpose again gives us $(a^T)^T_{i,j} = a^T_{j,i} = a_{i,j}$, which shows that $(\mathbf{A}^T)^T = \mathbf{A}$. □

Next we will show that for two matrices $\mathbf{A}_{n\times p}$ and $\mathbf{B}_{p\times q}$, we have $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. We will prove this by showing that the $(i,j)^{\text{th}}$ element of $(\mathbf{AB})^T$ and $\mathbf{B}^T\mathbf{A}^T$ are equal; that is, $(ab)^T_{i,j} = (b^T a^T)_{i,j}$ for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,q\}$. Recall from (12) that the elements of the product of two matrices are given by

$$(ab)_{i,j} = \sum_{k=1}^{p} a_{i,k} b_{k,j}.$$

Then applying the transpose to this gives us

$$(ab)^T_{i,j} = (ab)_{j,i} = \sum_{k=1}^{p} a_{j,k} b_{k,i} = \sum_{k=1}^{p} b^T_{i,k} a^T_{k,j} = (b^T a^T)_{i,j},$$

which shows that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. □

2.4.3 Symmetric Matrices

A matrix $\mathbf{A}$ is *symmetric* if it is equal to its transpose, i.e. $\mathbf{A} = \mathbf{A}^T$. That is, $a_{i,j} = a_{j,i}$ for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$. Note that in order for a matrix to be symmetric, it must be square. There are some important properties that symmetric matrices have, but we will show them later on when orthogonal matrices and eigenvectors are introduced.

2.4.4 The Main Diagonal and Diagonal Matrices

The *main diagonal* of a matrix $\mathbf{A}$, denoted as $\text{diag}(\mathbf{A})$, is the collection of elements $a_{i,j}$ such that $i = j$, that is, all elements of the form $a_{i,i}$. For a matrix $\mathbf{A}_{n\times p}$, we have

$$\text{diag}(\mathbf{A}) = \{a_{1,1}, \dots, a_{\min\{n,p\},\min\{n,p\}}\}. \tag{16}$$

The notation of $\min\{n,p\}$ may be slighlty confusing, but all it means is that the maximum amount of elements that can be on the main diagonal of a matrix is the smaller amount between n and p. To illustrate this, the main diagonal of matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are highlighted below in orange. We have three separate cases:

1. $\mathbf{A}_{n\times n}$ is a square matrix.
2. $\mathbf{B}_{n\times p}$ is a matrix such that $n > p$. That is, the matrix has more rows than columns.
3. $\mathbf{C}_{n\times p}$ is a matrix such that $n < p$. That is, the matrix has more columns than rows.

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \cdots & a_{p,p} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,p} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} & \cdots & a_{n,p} \end{pmatrix}$$

We can see that if $n > p$ (more rows than columns), then the maximum number of elements in $\text{diag}(\mathbf{A})$ is p, and if $n < p$ (more columns than rows), then the maximum number of elements in $\text{diag}(\mathbf{A})$ is n.

A *diagonal matrix*, generally denoted as $\mathbf{D}$, is a matrix whose only non-zero entries are elements along its main diagonal. That is, $d_{i,j} = 0$ if $i \neq j$.

When you multiply a diagonal matrix by itself, $\mathbf{D}^2$, the diagonal of the resulting matrix has the squared values of the original matrix, and zeros in every other entry. That is, $\mathbf{D}^2$ is also diagonal, and

$$(d^2)_{i,j} = \begin{cases} (d_{i,i})^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}. \tag{17}$$

This is easy to see using (12), as the elements of $\mathbf{D}^2$ are given by

$$(d^2)_{i,j} \sum_{k=1}^{n} d_{i,k} d_{k,j} = d_{i,i} d_{i,j} + \sum_{\substack{k=1 \\ k \neq i}}^{n} d_{i,k} d_{k,j} = d_{i,i} d_{i,j}.$$

If $i = j$, then $(d^2)_{i,j} = (d_{i,i})^2$, but if $i \neq j$, then $(d^2)_{i,j} = d_{i,i} \cdot 0 = 0$, which shows (17).

One other thing to note is that all diagonal matrices are symmetric. That is, for a diagonal matrix $\mathbf{D}$, we have $\mathbf{D} = \mathbf{D}^T$.

2.4.5 The Identity Matrix

A special case of the diagonal matrix is the *identity matrix*, denoted as $\mathbf{I}$, is a square diagonal matrix whose elements along the diagonal are all equal to 1. That is,

$$\iota_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}, \tag{18}$$

and $\text{diag}(\mathbf{I}) = \{1, \ldots, 1\}$. It is important to stress that, unlike other diagonal matrices, the identity matrix must be square.

The identity matrix is named the way it is because of the fact that multiplying any matrix $\mathbf{A}_{n \times p}$ by it returns $\mathbf{A}$. In other words,

$$\mathbf{I}_{n \times n} \mathbf{A}_{n \times p} = \mathbf{A}_{n \times p} \mathbf{I}_{p \times p} = \mathbf{A}_{n \times p}. \tag{19}$$

This is easy to see using (12) and (18), since

$$(\iota a)_{i,j} = \sum_{k=1}^{n} \iota_{i,k} a_{k,j} = \iota_{i,i} a_{i,j} + \sum_{\substack{k=1 \\ k \neq i}}^{n} \iota_{i,k} a_{k,j} = a_{i,j}$$

and

$$(a\iota)_{i,j} = \sum_{k=1}^{p} a_{i,p} \iota_{p,j} = a_{i,j} \iota_{j,j} + \sum_{\substack{k=1 \\ k \neq j}}^{p} \iota_{i,k} a_{k,j} = a_{i,j}.$$

2.4.6 Matrix Inversion

Given a square matrix $\mathbf{A}_{n \times n}$, its *inverse*, denoted as $\mathbf{A}^{-1}$, is defined as the matrix such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{20}$$

In other words, if you multiply a matrix by its inverse, you get the identity matrix.

Note that we mentioned the matrix $\mathbf{A}$ must be square in order to be inverted. This technically is not entirely true, since a non-square matrix can have a *left inverse* and a *right inverse*. However, for the purposes of SVD, the only matrices we will be inverting will be square, so we will not worry about this.

In general, computing the inverse of a matrix by hand is difficult, and no formulas to do so will be presented here; any software program such as `R` or `MATLAB` can compute the inverse. For the purposes of SVD, some inverse matrices will be of interest to us. However, those will be orthogonal matrices (to be covered soon), which are easy to take the inverse of.

2.4.7 Inner Products, Norms, Orthogonal Vectors, and Unit Vectors

For two vectors $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_i, \ldots, w_n$, both of order n, the *inner product* between them, denoted by $\langle \mathbf{v}, \mathbf{w} \rangle$, is defined as

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{w} = \sum_{i=1}^{n} v_i w_i \tag{21}$$

The inner product is also commonly referred to as the *dot product* and denoted $\mathbf{v} \cdot \mathbf{w}$, but we avoid that notation here. One thing to note that inner products are commutative, since

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{w} = \sum_{i=1}^{n} v_i w_i = \sum_{i=1}^{n} w_i v_i = \mathbf{w}^T \mathbf{v} = \langle \mathbf{w}, \mathbf{v} \rangle.$$

We are so interested in the inner product of two vectors since it gives us a method to determine whether or not the vectors are *orthogonal*. Two vectors are orthogonal if they are *perpendicular*, which is true if

$$\langle \mathbf{v}, \mathbf{w} \rangle = 0. \tag{22}$$

For a vector $\mathbf{v} = (v_1, \ldots, v_n)$, the *Euclidean norm*, or *ℓ_2-norm*, notated as $\|\mathbf{v}\|$, is defined as

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\sum_{i=1}^{n} v_i^2}. \tag{23}$$

That is, to calculate the ℓ_2-norm of a vector $\mathbf{v}$, we simply take the square root of the inner product of the vector with itself. The reader should note, then, that $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v} = \langle \mathbf{v}, \mathbf{v} \rangle$.

The ℓ_2-norm is used to measure the "length" of a vector. The reason we are so interested in the ℓ_2-norm is because it provides us a method to generate a *unit vector*, which is a vector with length 1 (i.e. an ℓ_2-norm equal to 1). Given a vector $\mathbf{v} = (v_1, \ldots, v_n)$, it is easy to construct a corresponding unit vector, denoted as $\mathbf{u} = (u_1, \ldots, u_n)$, by dividing $\mathbf{v}$ by $\|\mathbf{v}\|$; that is,

$$\mathbf{u} = \frac{\mathbf{v}}{\|\mathbf{v}\|}. \tag{24}$$

In other words, each element of $\mathbf{u}$ is just each element of $\mathbf{v}$ divided by $\|\mathbf{v}\|$, i.e. $u_i = v_i / \|\mathbf{v}\|$ for $i \in \{1, \ldots, n\}$.

Finally, two vectors $\mathbf{v}$ and $\mathbf{w}$ are *orthonormal* if they are orthogonal and both unit vectors. That is, if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ and $\|\mathbf{v}\| = \|\mathbf{w}\| = 1$, then $\mathbf{v}$ and $\mathbf{w}$ are orthonormal.

2.4.8 Outer Products

Given two n-vectors $\mathbf{v}$ and $\mathbf{w}$, their outer product, denoted as $\mathbf{v} \otimes \mathbf{w}$, is defined as

$$\mathbf{v} \otimes \mathbf{w} = \mathbf{v}\mathbf{w}^T. \tag{25}$$

Note the similarity to the inner product (21). Given the dimensions of $\mathbf{v}$ $(n \times 1)$ and $\mathbf{w}^T$ $(1 \times n)$, the outer product will result in a square matrix, whose elements are given by

$$(v \otimes w)_{i,j} = v_i w_j. \tag{26}$$

2.4.9 Orthogonal Matrices

An *orthogonal matrix*, generally denoted as $\mathbf{U}$, is a matrix whose row vectors and column vectors make up a set of mutually orthonormal (*not* orthogonal) vectors. That is, for $\mathbf{U}$ to be orthogonal, all of its row vectors must be orthonormal to each other and all of its column vectors must be orthonormal to each other. While a non-square matrix may possibly be orthogonal, there are some subtle technicalities that we will not pursue further, and from now on we will assume that orthogonal matrices are square.

One useful property about orthogonal matrices is that their inverse is given by their transpose. That is, for an orthogonal matrix $\mathbf{U}_{n \times n}$, we have $\mathbf{U}^{-1} = \mathbf{U}^T$. This property is extremely helpful because it makes determining the inverse very easy.

To do this, we will show that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. We will first use (9) and (15) two write this product as a product of vectors, so

$$\mathbf{U}^T\mathbf{U} = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix},$$

where $\mathbf{u}_i$ is the i^{th} column vector of $\mathbf{U}$. The elements of this product, which is an outer product, are given by

$$(u^T u)_{i,j} = \mathbf{u}_i^T \mathbf{u}_j.$$

We will now take advantage of the fact that the vectors are orthonormal. For any two of the vectors $\mathbf{u}_i$ and $\mathbf{u}_j$, we have $\mathbf{u}_i^T\mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$, since they are orthogonal. We also have $\mathbf{u}_i^T\mathbf{u}_i = \|\mathbf{u}_i\|^2 = 1$, since they are unit vectors. Then from this, we have

$$(u^T u)_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}.$$

These are the same elements as the identity matrix (18), which shows that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Finally, multiply both sides of this equation by $\mathbf{U}^{-1}$, and we have $\mathbf{U}^T = \mathbf{U}^{-1}$. That is, for an orthogonal matrix, its inverse is its transpose. □

2.4.10 Gramian Matrices

Given a matrix $\mathbf{A}_{n\times p}$, its two corresponding *Gramian matrices* are given by $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. From now on we will refer to the matrix $\mathbf{A}\mathbf{A}^T$ as the *left Gramian matrix* and $\mathbf{A}^T\mathbf{A}$ as the *right Gramian matrix*.

One useful property of the Gramian matrices are that they are always symmetric. This is easy to see, since

$$(\mathbf{A}\mathbf{A}^T)^T = (\mathbf{A}^T)^T\mathbf{A}^T = \mathbf{A}\mathbf{A}^T$$

and

$$(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T(\mathbf{A}^T)^T = \mathbf{A}^T\mathbf{A}.$$

For a right Gramian matrix, we can look at the $(i,j)^{\text{th}}$ element as an inner product as the i^{th} and j^{th} column vectors of $\mathbf{A}$. We have

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix} \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix},$$

where $\mathbf{a}_i$ is the i^{th} column vector of $\mathbf{A}$. The elements of this outer product are then given by

$$(a^T a)_{i,j} = \mathbf{a}_i^T \mathbf{a}_j. \tag{27}$$

Since the inner product is commutative, we have $(a^T a)_{i,j} = (a^T a)_{j,i} = (a^T a)_{i,j}^T$, which is another indication that the right Gramian matrix is symmetric.

2.5 Eigenvalues and Eigenvectors

Given a *square* matrix $\mathbf{A}_{n\times n}$, the non-zero vector $\mathbf{v}$ is an *eigenvector* of $\mathbf{A}$ if it satisfies the equation

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \tag{28}$$

where we say the scalar λ is the corresponding *eigenvalue*. What this means is that if we multiply a matrix $\mathbf{A}$ by an eigenvector $\mathbf{v}$, the resulting vector is the same as just taking the original eigenvector and scaling it up by a factor of λ.

To solve for the eigenvectors and their corresponding eigenvalues, we will first determine the eigenvalues. To start, we can see that $\lambda\mathbf{v} = \lambda\mathbf{I}\mathbf{v}$, since multiplying any vector or matrix by the identity matrix does not change its value. Next, subtract this from both sides of the equation, which gives us $\mathbf{A}\mathbf{v} - \lambda\mathbf{I}\mathbf{v} = \mathbf{0}$, where $\mathbf{0}$ is the null vector. Finally factor out $\mathbf{v}$ to get

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}. \tag{29}$$

Since we required $\mathbf{v} \neq \mathbf{0}$, the only way that (29) holds is for $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, where $\det(\cdot)$ is the *determinant*. In this review we did not discuss matrix inversion, so the reasoning as to why the determinant must equal 0 will be unclear to some.

The importance of (28) for understanding SVD cannot be understated, as there is a strong connection between eigenvalues/eigenvectors and singular values/singular vectors.

Some of the special matrices mentioned in Section 2.4 have very useful properties. The first we would like to examine is that for a symmetric matrix $\mathbf{A}$, the eigenvectors corresponding to *different* eigenvalues *will always be orthogonal*. To see this, let $\mathbf{v}_i$ and $\mathbf{v}_j$ be the i^{th} and j^{th} eigenvectors of a symmetric matrix $\mathbf{A}_{n\times p}$ with corresponding eigenvalues λ_i and λ_j, and assume that $\lambda_i \neq \lambda_j$. We can see that

$$\lambda_i\langle\mathbf{v}_i, \mathbf{v}_j\rangle = \lambda_i\mathbf{v}_i^T\mathbf{v}_j = (\lambda_i\mathbf{v}_i)^T\mathbf{v}_j = (\mathbf{A}\mathbf{v}_i)^T\mathbf{v}_j = \mathbf{v}_i^T\mathbf{A}^T\mathbf{v}_j = \mathbf{v}_i^T(\mathbf{A}^T\mathbf{v}_j)$$

where we used (28). Now, since $\mathbf{A}$ is symmetric, we have $\mathbf{A} = \mathbf{A}^T$, and so our equation becomes

$$\mathbf{v}_i^T(\mathbf{A}^T\mathbf{v}_j) = \mathbf{v}_i^T(\mathbf{A}\mathbf{v}_j) = \mathbf{v}_i^T(\lambda_j\mathbf{v}_j) = \lambda_j\mathbf{v}_i^T\mathbf{v}_j = \lambda_j\langle\mathbf{v}_i, \mathbf{v}_j\rangle,$$

and this shows that $\lambda_j\langle\mathbf{v}_i, \mathbf{v}_j\rangle = \lambda_j\langle\mathbf{v}_i, \mathbf{v}_j\rangle$. We can re-arrange this to get $(\lambda_i - \lambda_j)\langle\mathbf{v}_i, \mathbf{v}_j\rangle = 0$, and since $\lambda_i \neq \lambda_j$, we have $\lambda_i - \lambda_j \neq 0$, and so $\langle\mathbf{v}_i, \mathbf{v}_j\rangle = 0$, making the two eigenvectors orthogonal. □

3 The Singular Value Decomposition

3.1 Singular Values and Singular Vectors

For a given matrix $\mathbf{A}_{n\times p}$, a non-negative scalar σ is a *singular value*, with a corresponding *left singular vector* $\mathbf{u}$ and *right singular vector* $\mathbf{v}$ if they satisfy the two equations

$$\begin{aligned}\mathbf{A}\mathbf{v} &= \sigma\mathbf{u}\\ \mathbf{A}^T\mathbf{u} &= \sigma\mathbf{v}.\end{aligned} \tag{30}$$

4 Application to Statistics and Data Science

Now that the theory of SVD has been introduced, we will now examine *how* it is used for applications in statistics and data analysis.

4.1 Preliminary Information

4.1.1 The Data Matrix

For a sample of collected data, several *observations* (also called *data points*) are made, where data is collected for each one. For each observation, there are several *predictors* (also known as *features* or *variables*) that are recorded for each one. In general, there will be n observations and p predictors, and we will let $x_{i,j}$ denote the value of the j^{th} predictor for the i^{th} observation. We will be storing out data in a *data matrix*, denoted as $\mathbf{X}_{n\times p}$, where $x_{i,j}$ is the $(i,j)^{\text{th}}$ entry in the matrix. In other words,

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}. \tag{31}$$

The i^{th} row of $\mathbf{X}$, which we will denote as $\mathbf{n}_i$, can be viewed as the information for the i^{th} observation, while the j^{th} column, denoted as $\mathbf{p}_j$, can be viewed as the information for the j^{th} predictor. As a result, we can also write $\mathbf{X}$ in two different ways:

1. As a *row vector* (with n elements), where the i^{th} element is $\mathbf{p}_i$.
2. As a *column vector* (with p elements), where the i^{th} element is $\mathbf{n}_i$.

In other words,

$$\mathbf{X} = \begin{pmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_p \end{pmatrix} = \begin{pmatrix} \mathbf{n}_1 \\ \vdots \\ \mathbf{n}_n \end{pmatrix}. \tag{32}$$

We will be more interested in the first alternative expression. For notation, if we need to look at the individual elements of $\mathbf{p}_j$, we will write it as $\mathbf{p}_j = (x_{1,j}, \ldots, x_{n,j})$. Similarly for $\mathbf{n}_i$, we have $\mathbf{n}_i = (x_{i,1}, \ldots, x_{i,p})^T$.

4.1.2 Covariance and the Covariance Matrix

Given two predictors $\mathbf{p}_i$ and $\mathbf{p}_j$, it is often of interest to figure out how related they are to each other. One way to do this is to measure how much each predictor varies with each other. In other words, we would like to know how changes in the first variable are related to changes in the second.

To do this, we will use the *sample covariance* between the two predictors. Denoted as $\text{Cov}(\mathbf{p}_i, \mathbf{p}_j)$, it is defined as

$$\text{Cov}(\mathbf{p}_i, \mathbf{p}_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{k,i} - \mu_i)(x_{k,j} - \mu_j). \tag{33}$$

The use of $n-1$ in the denominator as opposed to n is known as *Bessel's correction*. The terms μ_i and μ_j are the *sample means* for the two predictors $\mathbf{p}_i$ and $\mathbf{p}_j$, respectively. To calculate the sample mean for a predictor, simply take each element of the predictor and divide by n. That is,

$$\mu_i = \frac{1}{n} \sum_{k=1}^{n} x_{k,i}. \tag{34}$$

We can also express (33) as an inner product between the two predictors. This is given by

$$\text{Cov}(\mathbf{p}_i, \mathbf{p}_j) = \frac{(\mathbf{p}_i - \boldsymbol{\mu}_i)^T (\mathbf{p}_j - \boldsymbol{\mu}_j)}{n-1} = \frac{\langle \mathbf{p}_i - \boldsymbol{\mu}_i, \mathbf{p}_j - \boldsymbol{\mu}_j \rangle}{n-1}. \tag{35}$$

Here, the notation $\boldsymbol{\mu}_i$ represents the *sample mean vector* for $\mathbf{p}_i$, and is given by $\boldsymbol{\mu}_i = \mu_i \mathbf{1}$. In other words, $\boldsymbol{\mu}_i$ is a vector where every element is μ_i. Note that because inner products are commutative, we have $\text{Cov}(\mathbf{p}_i, \mathbf{p}_j) = \text{Cov}(\mathbf{p}_j, \mathbf{p}_i)$.

A special case of the sample covariance is the *sample variance* of a predictor, denoted as $\text{Var}(\mathbf{p}_i)$, which is just the covariance between the predictor and *itself*. That is,

$$\text{Var}(\mathbf{p}_i) = \text{Cov}(\mathbf{p}_i, \mathbf{p}_i). \tag{36}$$

It is often of interest for us to *center* the data, the process of setting the sample average to zero for each predictor (i.e. $\mu_i = 0$ for all $i \in \{1, \ldots, n\}$). To do this, we simply have to take each predictor $\mathbf{p}_i$ and subtract $\boldsymbol{\mu}_i$. In other words, the *centered predictors* are given by

$$\mathbf{p}_i' = \mathbf{p} - \boldsymbol{\mu}_i. \tag{37}$$

In order to avoid unnecessary notation, and the fact that it is never a computational issue to center the data, *from now on we are going to assume that every data matrix* $\mathbf{X}$ *is centered, such that* $\mu_i = 0$ *for all* $i \in \{1, \ldots, n\}$. Note that using a centered data matrix as opposed to the original *does not* actually change the covariance.

Given this, we can re-express (33) and (35) as

$$\text{Cov}(\mathbf{p}_i, \mathbf{p}_j) = \frac{1}{n-1} \sum_{k=1}^{n} x_{k,i} x_{k,j} = \frac{\mathbf{p}_i^T \mathbf{p}_j}{n-1} = \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{n-1}. \tag{38}$$

It will be useful for us to store the covariances between several predictors in a *covariance matrix*, denoted as $\mathbf{\Sigma}$. The $(i,j)^{\text{th}}$ element of $\mathbf{\Sigma}$, denoted as $\Sigma_{i,j}$, is the covariance between the two predictors $\mathbf{p}_i$ and $\mathbf{p}_j$ (i.e. $\Sigma_{i,j} = \text{Cov}(\mathbf{p}_i, \mathbf{p}_j)$), so

$$\mathbf{\Sigma} = \begin{pmatrix} \text{Cov}(\mathbf{p}_1, \mathbf{p}_1) & \cdots & \text{Cov}(\mathbf{p}_1, \mathbf{p}_p) \\ \vdots & \ddots & \vdots \\ \text{Cov}(\mathbf{p}_p, \mathbf{p}_1) & \cdots & \text{Cov}(\mathbf{p}_p, \mathbf{p}_p) \end{pmatrix}. \tag{39}$$

We can see that the elements along the main diagonal of $\mathbf{\Sigma}$ are the variances of each of the predictors. Also, since covariance is commutative, we have

$$\Sigma_{i,j} = \text{Cov}(\mathbf{p}_i, \mathbf{p}_j) = \text{Cov}(\mathbf{p}_j, \mathbf{p}_i) = \Sigma_{j,i} = \Sigma_{i,j}^T,$$

so the covariance matrix is always symmetric. Using (38), we can re-write the elements of $\mathbf{\Sigma}$ as

$$\Sigma_{i,j} = \text{Cov}(\mathbf{p}_i, \mathbf{p}_j) = \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{n-1} = \frac{\mathbf{p}_i^T \mathbf{p}_j}{n-1}.$$

We know that $\mathbf{p}_i^T \mathbf{p}_j$ is the $(i,j)^{\text{th}}$ element of the right Gramian matrix of the data matrix, $\mathbf{X}^T\mathbf{X}$. Therefore, by accounting for the scaling by $1/(n-1)$, we have

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}. \tag{40}$$

This gives us a very convenient way to calculate the covariance between all p predictors in a data set $\mathbf{X}$ with n observations.

Eigenvectors $\mathbf{V}$ of the covariance matrix are principal directions. Projections of the data on these eigenvectors are principal components; these projections are given by $\mathbf{US}$.