

Mathematical foundations for ML - UM6P class

Linear Algebra - Matrix Factorization

05 February 2023

Contents

1 Basic concepts	3
1.1 Matrix Multiplication	3
1.1.1 Vector-Vector Products	3
1.1.2 Matrix-Vector Products	4
1.1.3 Matrix-Matrix Products	5
1.1.4 The Identity Matrix and Diagonal Matrices	8
1.1.5 The Transpose	8
1.1.6 Symmetric Matrices	9
1.1.7 The Trace	9
1.1.8 Norms	10
1.1.9 Linear Independence and Rank	11
1.1.10 The Inverse	12
1.1.11 Orthogonal Matrices	13
1.1.12 Range and Nullspace of a Matrix	14
1.1.13 The Determinant	15

1.1.14	Quadratic Forms and Positive Semidefinite Matrices	19
1.1.15	Eigenvalues and Eigenvectors	20
1.1.16	Eigenvalues and Eigenvectors of Symmetric Matrices	22
2	Singular Value Decomposition	23
2.1	Introduction	23
2.2	Existence of the Singular-Value Decomposition	25
2.2.1	The theorem	25
2.2.2	Existence of SVDs	27
2.2.3	SVDs and Spectral Theory	28
2.3	The Eckhart-Young Theorem	29
2.3.1	The Frobenius Norm	29
2.3.2	Low-Rank Approximation via the SVD	29
2.3.3	Eckhart-Young Theorem	30
2.3.4	Proof of the Eckhart-Young Theorem	30
2.3.5	Choosing k	32
3	Principal Component Analysis	33
3.1	Introduction	33
3.1.1	Goals of PCA	33
3.1.2	Technical Overview	33
3.2	PCA	34

1 Basic concepts

1.1 Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix

$$C = AB \in \mathbb{R}^{m \times p}$$

where

$$C_{ij} = \sum_{k=1}^n A_{ik}B_{kj}$$

Note that in order for the matrix product to exist, the number of columns in A must equal the number of rows in B . There are many ways of looking at matrix multiplication, and we'll start by examining a few special cases.

1.1.1 Vector-Vector Products

Given two vectors $x, y \in \mathbb{R}^n$, the quantity $x^T y$, sometimes called the inner product or **dot** product of the vectors, is a real number given by

$$x^T y \in \mathbb{R} = [\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i$$

Observe that inner products are really just special case of matrix multiplication. Note that it is always the case that $x^T y = y^T x$.

Given vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ (not necessarily of the same size), $xy^T \in \mathbb{R}^{m \times n}$ is called the outer product of the vectors. It is a matrix whose entries are given by $(xy^T)_{ij} = x_i y_j$, i.e.,

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

As an example of how the outer product can be useful, let $\mathbf{1} \in \mathbb{R}^n$ denote an n -dimensional vector whose entries are all equal to 1. Furthermore, consider the matrix $A \in \mathbb{R}^{m \times n}$ whose columns are all equal to some vector $x \in \mathbb{R}^m$. Using outer products, we can represent A compactly as,

$$A = \begin{bmatrix} | & | & & | \\ x & x & \cdots & x \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = x\mathbf{1}^T$$

1.1.2 Matrix-Vector Products

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$, their product is a vector $y = Ax \in \mathbb{R}^m$. There are a couple ways of looking at matrix-vector multiplication, and we will look at each of them in turn.

If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \vdots \\ - & a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}$$

In other words, the i th entry of y is equal to the inner product of the i th row of A and x , $y_i = a_i^T x$.

Alternatively, let's write A in column form. In this case we see that,

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [a_1] x_1 + [a_2] x_2 + \dots + [a_n] x_n$$

In other words, y is a linear combination of the columns of A , where the coefficients of the linear combination are given by the entries of x .

So far we have been multiplying on the right by a column vector, but it is also possible to multiply on the left by a row vector. This is written, $y^T = x^T A$ for $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$. As before, we can express y^T in two obvious ways, depending on whether we express A in terms of its rows or columns. In the first case we express A in terms of its columns, which gives

$$y^T = x^T A = x^T \begin{bmatrix} | & | & \cdots & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & \cdots & | \end{bmatrix} = [x^T a_1 \quad x^T a_2 \quad \cdots \quad x^T a_n]$$

which demonstrates that the i th entry of y^T is equal to the inner product of x and the i th column of A .

Finally, expressing A in terms of rows we get the final representation of the vector-matrix product,

$$\begin{aligned} y^T &= x^T A \\ &= [x_1 \quad x_2 \quad \cdots \quad x_n] \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} \\ &= x_1 [- \quad a_1^T \quad -] + x_2 [- \quad a_2^T \quad -] + \dots + x_n [- \quad a_m^T \quad -] \end{aligned}$$

so we see that y^T is a linear combination of the rows of A , where the coefficients for the linear combination are given by the entries of x .

1.1.3 Matrix-Matrix Products

Armed with this knowledge, we can now look at four different (but, of course, equivalent) ways of viewing the matrix-matrix multiplication $C = AB$ as defined at the beginning of this section.

First, we can view matrix-matrix multiplication as a set of vector-vector products. The most obvious viewpoint, which follows immediately from the definition, is that the (i, j) th entry of C is equal to the inner product of the i th row of A and the j th column of B . Symbolically, this looks like the following,

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | & | & \dots & | \\ b_1 & b_2 & \dots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \dots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \dots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \dots & a_m^T b_p \end{bmatrix}.$$

Remember that since $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, $a_i \in \mathbb{R}^n$ and $b_j \in \mathbb{R}^n$, so these inner products all make sense. This is the most "natural" representation when we represent A by rows and B by columns. Alternatively, we can represent A by columns, and B by rows. This representation leads to a much trickier interpretation of AB as a sum of outer products. Symbolically,

$$C = AB = \begin{bmatrix} | & | & \dots & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ \vdots & & \\ - & b_n^T & - \end{bmatrix} = \sum_{i=1}^n a_i b_i^T.$$

Put another way, AB is equal to the sum, over all i , of the outer product of the i th column of A and the i th row of B . Since, in this case, $a_i \in \mathbb{R}^m$ and $b_i \in \mathbb{R}^p$, the dimension of the outer product $a_i b_i^T$ is $m \times p$, which coincides with the dimension of C . Chances are, the last equality above may appear confusing to you. If so, take the time to check it for yourself!

Second, we can also view matrix-matrix multiplication as a set of matrix-vector products. Specifically, if we represent B by columns, we can view the columns of C as matrix-vector products between A and the columns of B . Symbolically,

$$C = AB = A \begin{bmatrix} | & | & \dots & | \\ b_1 & b_2 & \dots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & \dots & | \\ Ab_1 & Ab_2 & \dots & Ab_p \\ | & | & & | \end{bmatrix}.$$

Here the i th column of C is given by the matrix-vector product with the vector on the right, $c_i = Ab_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection. Finally, we have the analogous viewpoint, where we represent A by rows, and view the rows of C as the matrix-vector product between the rows of A and C . Symbolically,

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} B = \begin{bmatrix} - & a_1^T B & - \\ - & a_2^T B & - \\ \vdots & & \\ - & a_m^T B & - \end{bmatrix}$$

Here the i th row of C is given by the matrix-vector product with the vector on the left, $c_i^T = a_i^T B$.

It may seem like overkill to dissect matrix multiplication to such a large degree, especially when all these viewpoints follow immediately from the initial definition we gave (in about a line of math) at the beginning of this section. However, virtually all of linear algebra deals with matrix multiplications of some kind, and it is worthwhile to spend some time trying to develop an intuitive understanding of the viewpoints presented here.

In addition to this, it is useful to know a few basic properties of matrix multiplication at a higher level:

- Matrix multiplication is associative: $(AB)C = A(BC)$.
- Matrix multiplication is distributive: $A(B + C) = AB + AC$.
- Matrix multiplication is, in general, not commutative; that is, it can be the case that $AB \neq BA$. (For example, if $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times q}$, the matrix product BA does not even exist if m and q are not equal!)

If you are not familiar with these properties, take the time to verify them for yourself. For example, to check the associativity of matrix multiplication, suppose that $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, and $C \in \mathbb{R}^{p \times q}$. Note that $AB \in \mathbb{R}^{m \times p}$, so $(AB)C \in \mathbb{R}^{m \times q}$. Similarly, $BC \in \mathbb{R}^{n \times q}$, so $A(BC) \in \mathbb{R}^{m \times q}$. Thus, the dimensions of the resulting matrices agree. To show that matrix multiplication is associative, it suffices to check that the (i, j) th entry of $(AB)C$ is equal to the (i, j) th entry of $A(BC)$. We can verify this directly using the definition of matrix multiplication:

$$\begin{aligned} ((AB)C)_{ij} &= \sum_{k=1}^p (AB)_{ik} C_{kj} = \sum_{k=1}^p \left(\sum_{l=1}^n A_{il} B_{lk} \right) C_{kj} \\ &= \sum_{k=1}^p \left(\sum_{l=1}^n A_{il} B_{lk} C_{kj} \right) = \sum_{l=1}^n \left(\sum_{k=1}^p A_{il} B_{lk} C_{kj} \right) \\ &= \sum_{l=1}^n A_{il} \left(\sum_{k=1}^p B_{lk} C_{kj} \right) = \sum_{l=1}^n A_{il} (BC)_{lj} = (A(BC))_{ij}. \end{aligned}$$

Here, the first and last two equalities simply use the definition of matrix multiplication, the third and fifth equalities use the distributive property for scalar multiplication over addition, and the fourth equality uses the commutative and associativity of scalar addition. This technique for proving matrix properties by

reduction to simple scalar properties will come up often, so make sure you're familiar with it.

1.1.4 The Identity Matrix and Diagonal Matrices

The identity matrix, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

It has the property that for all $A \in \mathbb{R}^{m \times n}$,

$$AI = A = IA.$$

Note that in some sense, the notation for the identity matrix is ambiguous, since it does not specify the dimension of I . Generally, the dimensions of I are inferred from context so as to make matrix multiplication possible. For example, in the equation above, the I in $AI = A$ is an $n \times n$ matrix, whereas the I in $A = IA$ is an $m \times m$ matrix.

A diagonal matrix is a matrix where all non-diagonal elements are 0 . This is typically denoted $D = \text{diag}(d_1, d_2, \dots, d_n)$, with

$$D_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$$

Clearly, $I = \text{diag}(1, 1, \dots, 1)$.

1.1.5 The Transpose

The transpose of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}$$

We have in fact already been using the transpose when describing row vectors, since the transpose of a column vector is naturally a row vector.

The following properties of transposes are easily verified:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

1.1.6 Symmetric Matrices

A square matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^T$. It is anti-symmetric if $A = -A^T$. It is easy to show that for any matrix $A \in \mathbb{R}^{n \times n}$, the matrix $A + A^T$ is symmetric and the matrix $A - A^T$ is anti-symmetric. From this it follows that any square matrix $A \in \mathbb{R}^{n \times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

$$A = \frac{1}{2} (A + A^T) + \frac{1}{2} (A - A^T)$$

and the first matrix on the right is symmetric, while the second is anti-symmetric. It turns out that symmetric matrices occur a great deal in practice, and they have many nice properties which we will look at shortly. It is common to denote the set of all symmetric matrices of size n as \mathbb{S}^n , so that $A \in \mathbb{S}^n$ means that A is a symmetric $n \times n$ matrix;

1.1.7 The Trace

The trace of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\text{tr}(A)$ (or just $\text{tr } A$ if the parentheses are obviously implied), is the sum of diagonal elements in the matrix:

$$\text{tr } A = \sum_{i=1}^n A_{ii}$$

As described in the CS229 lecture notes, the trace has the following properties (included here for the sake of completeness):

- For $A \in \mathbb{R}^{n \times n}$, $\text{tr } A = \text{tr } A^T$.
- For $A, B \in \mathbb{R}^{n \times n}$, $\text{tr}(A + B) = \text{tr } A + \text{tr } B$.
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$, $\text{tr}(tA) = t \text{tr } A$.
- For A, B such that AB is square, $\text{tr } AB = \text{tr } BA$.
- For A, B, C such that ABC is square, $\text{tr } ABC = \text{tr } BCA = \text{tr } CAB$, and so on for the product of more matrices.

As an example of how these properties can be proven, we'll consider the fourth property given above. Suppose that $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$ (so that $AB \in \mathbb{R}^{m \times m}$ is a square matrix). Observe that $BA \in \mathbb{R}^{n \times n}$ is also a square matrix, so it makes sense to apply the trace operator to it. To verify that $\text{tr } AB = \text{tr } BA$, note that

$$\begin{aligned}\text{tr } AB &= \sum_{i=1}^m (AB)_{ii} = \sum_{i=1}^m \left(\sum_{j=1}^n A_{ij} B_{ji} \right) \\ &= \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ji} = \sum_{j=1}^n \sum_{i=1}^m B_{ji} A_{ij} \\ &= \sum_{j=1}^n \left(\sum_{i=1}^m B_{ji} A_{ij} \right) = \sum_{j=1}^n (BA)_{jj} = \text{tr } BA.\end{aligned}$$

Here, the first and last two equalities use the definition of the trace operator and matrix multiplication. The fourth equality, where the main work occurs, uses the commutativity of scalar multiplication in order to reverse the order of the terms in each product, and the commutativity and associativity of scalar addition in order to rearrange the order of the summation.

1.1.8 Norms

A norm of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

Note that $\|x\|_2^2 = x^T x$.

More formally, a norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity).
2. $f(x) = 0$ if and only if $x = 0$ (definiteness).
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t|f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Other examples of norms are the ℓ_1 norm,

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

and the ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$, and defined as

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$$

Norms can also be defined for matrices, such as the Frobenius norm,

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2} = \sqrt{\text{tr}(A^T A)}$$

Many other norms exist, but they are beyond the scope of this review.

1.1.9 Linear Independence and Rank

A set of vectors $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^m$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors.

Conversely, if one vector belonging to the set can be represented as a linear combination of the remaining vectors, then the vectors are said to be (linearly) dependent. That is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors x_1, \dots, x_n are linearly dependent; otherwise, the vectors are linearly independent. For example, the vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

are linearly dependent because $x_3 = -2x_1 + x_2$.

The column rank of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set. With some abuse of terminology, this is often referred to simply as the number of linearly independent columns of A . In the same way, the row rank is the largest number of rows of A that constitute a linearly independent set.

For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (though we will not prove this), and so both quantities are referred to collectively as the **rank** of A , denoted as $\text{rank}(A)$. The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be full rank.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

1.1.10 The Inverse

The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1} , and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

Note that not all matrices have inverses. Non-square matrices, for example, do not have inverses by definition. However, for some square matrices A , it may still be the case that A^{-1} may not exist. In particular, we say that A is invertible or non-singular if A^{-1} exists and non-invertible or singular otherwise.¹

In order for a square matrix A to have an inverse A^{-1} , then A must be full rank. We will soon see that there are many alternative sufficient and necessary conditions, in addition to full rank, for invertibility.

The following are properties of the inverse; all assume that $A, B \in \mathbb{R}^{n \times n}$ are non-singular:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T} .

As an example of how the inverse is used, consider the linear system of equations, $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, and $x, b \in \mathbb{R}^n$. If A is nonsingular (i.e., invertible), then $x = A^{-1}b$. (What if $A \in \mathbb{R}^{m \times n}$ is not a square matrix? Does this work?)

1.1.11 Orthogonal Matrices

Two vectors $x, y \in \mathbb{R}^n$ are orthogonal if $x^T y = 0$. A vector $x \in \mathbb{R}^n$ is normalized if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being orthonormal).

It follows immediately from the definition of orthogonality and normality that

$$U^T U = I = UU^T.$$

In other words, the inverse of an orthogonal matrix is its transpose. Note that if U is not square - i.e., $U \in \mathbb{R}^{m \times n}$, $n < m$ - but its columns are still orthonormal, then $U^T U = I$, but $UU^T \neq I$. We generally only use the term orthogonal to describe the previous case, where U is square.

Another nice property of orthogonal matrices is that operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$\|Ux\|_2 = \|x\|_2$$

for any $x \in \mathbb{R}^n, U \in \mathbb{R}^{n \times n}$ orthogonal.

1.1.12 Range and Nullspace of a Matrix

The span of a set of vectors $\{x_1, x_2, \dots, x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \dots, x_n\}$. That is,

$$\text{span}(\{x_1, \dots, x_n\}) = \left\{ v : v = \sum_{i=1}^n \alpha_i x_i, \quad \alpha_i \in \mathbb{R} \right\}$$

It can be shown that if $\{x_1, \dots, x_n\}$ is a set of n linearly independent vectors, where each $x_i \in \mathbb{R}^n$, then $\text{span}(\{x_1, \dots, x_n\}) = \mathbb{R}^n$. In other words, any vector $v \in \mathbb{R}^n$ can be written as a linear combination of x_1 through x_n . The projection of a vector $y \in \mathbb{R}^m$ onto the span of $\{x_1, \dots, x_n\}$ (here we assume $x_i \in \mathbb{R}^m$) is the vector $v \in \text{span}(\{x_1, \dots, x_n\})$, such that v is as close as possible to y , as measured by the Euclidean norm $\|v - y\|_2$. We denote the projection as $\text{Proj}(y; \{x_1, \dots, x_n\})$ and can define it formally as,

$$\text{Proj}(y; \{x_1, \dots, x_n\}) = \underset{v \in \text{span}(\{x_1, \dots, x_n\})}{\text{argmin}} \|y - v\|_2$$

The range (sometimes also called the columnspace) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the span of the columns of A . In other words,

$$\mathcal{R}(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\}$$

Making a few technical assumptions (namely that A is full rank and that $n < m$), the projection of a vector $y \in \mathbb{R}^m$ onto the range of A is given by,

$$\text{Proj}(y; A) = \underset{v \in \mathcal{R}(A)}{\text{argmin}} \|v - y\|_2 = A (A^T A)^{-1} A^T y.$$

This last equation should look extremely familiar, since it is almost the same formula we derived in class (and which we will soon derive again) for the least

squares estimation of parameters. Looking at the definition for the projection, it should not be too hard to convince yourself that this is in fact the same objective that we minimized in our least squares problem (except for a squaring of the norm, which doesn't affect the optimal point) and so these problems are naturally very connected. When A contains only a single column, $a \in \mathbb{R}^m$, this gives the special case for a projection of a vector on to a line:

$$\text{Proj}(y; a) = \frac{aa^T}{a^T a} y$$

The nullspace of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$ is the set of all vectors that equal 0 when multiplied by A , i.e.,

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Note that vectors in $\mathcal{R}(A)$ are of size m , while vectors in the $\mathcal{N}(A)$ are of size n , so vectors in $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are both in \mathbb{R}^n . In fact, we can say much more. It turns out that

$$\{w : w = u + v, u \in \mathcal{R}(A^T), v \in \mathcal{N}(A)\} = \mathbb{R}^n \text{ and } \mathcal{R}(A^T) \cap \mathcal{N}(A) = \{\mathbf{0}\}.$$

In other words, $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are disjoint subsets that together span the entire space of \mathbb{R}^n . Sets of this type are called orthogonal complements, and we denote this $\mathcal{R}(A^T) = \mathcal{N}(A)^\perp$.

1.1.13 The Determinant

The determinant of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, and is denoted $|A|$ or $\det A$ (like the trace operator, we usually omit parentheses). Algebraically, one could write down an explicit formula for the determinant of A , but this unfortunately gives little intuition about its meaning. Instead, we'll start out by providing a geometric interpretation of the determinant and then visit some of its specific algebraic properties afterwards.

Given a matrix

$$\begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_n^T & - \end{bmatrix},$$

consider the set of points $S \subset \mathbb{R}^n$ formed by taking all possible linear combinations of the row vectors $a_1, \dots, a_n \in \mathbb{R}^n$ of A , where the coefficients of the linear combination are all between 0 and 1 ; that is, the set S is the restriction of $\text{span}(\{a_1, \dots, a_n\})$ to only those linear combinations whose coefficients $\alpha_1, \dots, \alpha_n$ satisfy $0 \leq \alpha_i \leq 1, i = 1, \dots, n$. Formally,

$$S = \left\{ v \in \mathbb{R}^n : v = \sum_{i=1}^n \alpha_i a_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \dots, n \right\}$$

The absolute value of the determinant of A , it turns out, is a measure of the "volume" of the set S .²

For example, consider the 2×2 matrix,

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$$

Here, the rows of the matrix are

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

The set S corresponding to these rows is shown in Figure 1. For two-dimensional matrices, S generally has the shape of a parallelogram. In our example, the value of the determinant is $|A| = -7$ (as can be computed using the formulas shown later in this section), so the area of the parallelogram is 7 . (Verify this for yourself!)

In three dimensions, the set S corresponds to an object known as a parallelepiped (a threedimensional box with skewed sides, such that every face has the shape of a parallelogram). The absolute value of the determinant of the 3×3 matrix whose rows define S give the three-dimensional volume of the parallelepiped. In even higher dimensions, the set S is an object known as an n -dimensional parallelotope.

Algebraically, the determinant satisfies the following three properties (from which all other properties follow, including the general formula):

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).
2. Given a matrix $A \in \mathbb{R}^{n \times n}$, if we multiply a single row in A by a scalar $t \in \mathbb{R}$, then the determinant of the new matrix is $t|A|$,

$$\left| \begin{bmatrix} - & ta_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} \right| = t|A|.$$

(Geometrically, multiplying one of the sides of the set S by a factor t causes the volume to increase by a factor t .)

3. If we exchange any two rows a_i^T and a_j^T of A , then the determinant of the new matrix is $-|A|$, for example

$$\left| \begin{bmatrix} - & a_2^T & - \\ - & a_1^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} \right| = -|A|.$$

In case you are wondering, it is not immediately obvious that a function satisfying the above three properties exists. In fact, though, such a function does exist, and is unique (which we will not prove here).

Several properties that follow from the three properties above include:

- For $A \in \mathbb{R}^{n \times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n \times n}$, $|AB| = |A||B|$.
- For $A \in \mathbb{R}^{n \times n}$, $|A| = 0$ if and only if A is singular (i.e., non-invertible). (If A is singular then it does not have full rank, and hence its columns are linearly dependent. In this case, the set S corresponds to a "flat sheet" within the n -dimensional space and hence has zero volume.)

- For $A \in \mathbb{R}^{n \times n}$ and A non-singular, $|A^{-1}| = 1/|A|$.

Before giving the general definition for the determinant, we define, for $A \in \mathbb{R}^{n \times n}$, $A_{\setminus i, \setminus j} \in \mathbb{R}^{(n-1) \times (n-1)}$ to be the matrix that results from deleting the i th row and j th column from A . The general (recursive) formula for the determinant is

$$\begin{aligned} |A| &= \sum_{i=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } j \in 1, \dots, n) \\ &= \sum_{j=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } i \in 1, \dots, n) \end{aligned}$$

with the initial case that $|A| = a_{11}$ for $A \in \mathbb{R}^{1 \times 1}$. If we were to expand this formula completely for $A \in \mathbb{R}^{n \times n}$, there would be a total of $n!$ (n factorial) different terms. For this reason, we hardly ever explicitly write the complete equation of the determinant for matrices bigger than 3×3 . However, the equations for determinants of matrices up to size 3×3 are fairly common, and it is good to know them:

$$\begin{aligned} |[a_{11}]| &= a_{11} \\ \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \right| &= a_{11}a_{22} - a_{12}a_{21} \\ \left| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right| &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\quad - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{aligned}$$

The classical adjoint (often just called the adjoint) of a matrix $A \in \mathbb{R}^{n \times n}$, is denoted $\text{adj}(A)$, and defined as

$$\text{adj}(A) \in \mathbb{R}^{n \times n}, \quad (\text{adj}(A))_{ij} = (-1)^{i+j} |A_{\setminus j, \setminus i}|$$

(note the switch in the indices $A_{\setminus j, \setminus i}$). It can be shown that for any nonsingular $A \in \mathbb{R}^{n \times n}$,

$$A^{-1} = \frac{1}{|A|} \text{adj}(A)$$

While this is a nice "explicit" formula for the inverse of matrix, we should note

that, numerically, there are in fact much more efficient ways of computing the inverse.

1.1.14 Quadratic Forms and Positive Semidefinite Matrices

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a quadratic form. Written explicitly, we see that

$$x^T Ax = \sum_{i=1}^n x_i (Ax)_i = \sum_{i=1}^n x_i \left(\sum_{j=1}^n A_{ij} x_j \right) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

Note that,

$$x^T Ax = (x^T Ax)^T = x^T A^T x = x^T \left(\frac{1}{2}A + \frac{1}{2}A^T \right) x$$

where the first equality follows from the fact that the transpose of a scalar is equal to itself, and the second equality follows from the fact that we are averaging two quantities which are themselves equal. From this, we can conclude that only the symmetric part of A contributes to the quadratic form. For this reason, we often implicitly assume that the matrices appearing in a quadratic form are symmetric.

We give the following definitions:

- A symmetric matrix $A \in \mathbb{S}^n$ is positive definite (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$. This is usually denoted $A \succ 0$ (or just $A > 0$), and often times the set of all positive definite matrices is denoted \mathbb{S}_{++}^n .
- A symmetric matrix $A \in \mathbb{S}^n$ is positive semidefinite (PSD) if for all vectors $x^T Ax \geq 0$. This is written $A \succeq 0$ (or just $A \geq 0$), and the set of all positive semidefinite matrices is often denoted \mathbb{S}_+^n .
- Likewise, a symmetric matrix $A \in \mathbb{S}^n$ is negative definite (ND), denoted $A \prec 0$ (or just $A < 0$) if for all non-zero $x \in \mathbb{R}^n$, $x^T Ax < 0$.
- Similarly, a symmetric matrix $A \in \mathbb{S}^n$ is negative semidefinite (NSD), denoted $A \preceq 0$ (or just $A \leq 0$) if for all $x \in \mathbb{R}^n$, $x^T Ax \leq 0$.
- Finally, a symmetric matrix $A \in \mathbb{S}^n$ is indefinite, if it is neither positive semidefinite nor negative semidefinite - i.e., if there exists $x_1, x_2 \in \mathbb{R}^n$ such that $x_1^T Ax_1 > 0$ and $x_2^T Ax_2 < 0$.

It should be obvious that if A is positive definite, then $-A$ is negative definite and vice versa. Likewise, if A is positive semidefinite then $-A$ is negative semidefinite and vice versa. If A is indefinite, then so is $-A$.

One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible. To see why this is the case, suppose that some matrix $A \in \mathbb{R}^{n \times n}$ is not full rank. Then, suppose that the j th column of A is expressible as a linear combination of other $n - 1$ columns:

$$a_j = \sum_{i \neq j} x_i a_i$$

for some $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n \in \mathbb{R}$. Setting $x_j = -1$, we have

$$Ax = \sum_{i=1}^n x_i a_i = 0$$

But this implies $x^T Ax = 0$ for some non-zero vector x , so A must be neither positive definite nor negative definite. Therefore, if A is either positive definite or negative definite, it must be full rank.

Finally, there is one type of positive definite matrix that comes up frequently, and so deserves some special mention. Given any matrix $A \in \mathbb{R}^{m \times n}$ (not necessarily symmetric or even square), the matrix $G = A^T A$ (sometimes called a Gram matrix) is always positive semidefinite. Further, if $m \geq n$ (and we assume for convenience that A is full rank), then $G = A^T A$ is positive definite.

1.1.15 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of A and $x \in \mathbb{C}^n$ is the corresponding eigenvector³ if

$$Ax = \lambda x, \quad x \neq 0.$$

Intuitively, this definition means that multiplying A by the vector x results in a new vector that points in the same direction as x , but scaled by a factor λ . Also note that for any eigenvector $x \in \mathbb{C}^n$, and scalar $t \in \mathbb{C}$, $A(cx) = cAx = c\lambda x = \lambda(cx)$, so cx is also an eigenvector. For this reason when we talk about "the" eigenvector associated with λ , we usually assume that the eigenvector is

normalized to have length 1 (this still creates some ambiguity, since x and $-x$ will both be eigenvectors, but we will have to live with this).

We can rewrite the equation above to state that (λ, x) is an eigenvalue-eigenvector pair of A if,

$$(\lambda I - A)x = 0, \quad x \neq 0$$

But $(\lambda I - A)x = 0$ has a non-zero solution to x if and only if $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e.,

$$|(\lambda I - A)| = 0.$$

We can now use the previous definition of the determinant to expand this expression into a (very large) polynomial in λ , where λ will have maximum degree n . We then find the n (possibly complex) roots of this polynomial to find the n eigenvalues $\lambda_1, \dots, \lambda_n$. To find the eigenvector corresponding to the eigenvalue λ_i , we simply solve the linear equation $(\lambda_i I - A)x = 0$. It should be noted that this is not the method which is actually used

in practice to numerically compute the eigenvalues and eigenvectors (remember that the complete expansion of the determinant has $n !$ terms); it is rather a mathematical argument.

The following are properties of eigenvalues and eigenvectors (in all cases assume $A \in \mathbb{R}^{n \times n}$ has eigenvalues $\lambda_1, \dots, \lambda_n$ and associated eigenvectors x_1, \dots, x_n) :

- The trace of A is equal to the sum of its eigenvalues,

$$\text{tr } A = \sum_{i=1}^n \lambda_i$$

- The determinant of A is equal to the product of its eigenvalues,

$$|A| = \prod_{i=1}^n \lambda_i$$

- The rank of A is equal to the number of non-zero eigenvalues of A .
- If A is non-singular then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector x_i , i.e., $A^{-1}x_i = (1/\lambda_i)x_i$. (To prove this, take the eigenvector equation, $Ax_i = \lambda_i x_i$ and left-multiply each side by A^{-1} .)
- The eigenvalues of a diagonal matrix $D = \text{diag}(d_1, \dots, d_n)$ are just the diagonal entries d_1, \dots, d_n .

We can write all the eigenvector equations simultaneously as

$$AX = X\Lambda$$

where the columns of $X \in \mathbb{R}^{n \times n}$ are the eigenvectors of A and Λ is a diagonal matrix whose entries are the eigenvalues of A , i.e.,

$$X \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}, \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$$

If the eigenvectors of A are linearly independent, then the matrix X will be invertible, so $A = X\Lambda X^{-1}$. A matrix that can be written in this form is called diagonalizable.

1.1.16 Eigenvalues and Eigenvectors of Symmetric Matrices

Two remarkable properties come about when we look at the eigenvalues and eigenvectors of a symmetric matrix $A \in \mathbb{S}^n$. First, it can be shown that all the eigenvalues of A are real. Secondly, the eigenvectors of A are orthonormal, i.e., the matrix X defined above is an orthogonal matrix (for this reason, we denote the matrix of eigenvectors as U in this case).

We can therefore represent A as $A = U\Lambda U^T$, remembering from above that the inverse of an orthogonal matrix is just its transpose.

Using this, we can show that the definiteness of a matrix depends entirely on the sign of its eigenvalues. Suppose $A \in \mathbb{S}^n = U\Lambda U^T$. Then

$$x^T Ax = x^T U\Lambda U^T x = y^T \Lambda y = \sum_{i=1}^n \lambda_i y_i^2$$

where $y = U^T x$ (and since U is full rank, any vector $y \in \mathbb{R}^n$ can be represented in this form). Because y_i^2 is always positive, the sign of this expression depends entirely on the λ_i 's. If all $\lambda_i > 0$, then the matrix is positive definite; if all $\lambda_i \geq 0$, it is positive semidefinite. Likewise, if all $\lambda_i < 0$ or $\lambda_i \leq 0$, then A is negative definite or negative semidefinite respectively. Finally, if A has both positive and negative eigenvalues, it is indefinite.

An application where eigenvalues and eigenvectors come up frequently is in maximizing some function of a matrix. In particular, for a matrix $A \in \mathbb{S}^n$, consider the following maximization problem,

$$\max_{x \in \mathbb{R}^n} x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

i.e., we want to find the vector (of norm 1) which maximizes the quadratic form. Assuming the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, the optimal x for this optimization problem is x_1 , the eigenvector corresponding to λ_1 . In this case the maximal value of the quadratic form is λ_1 . Similarly, the optimal solution to the minimization problem,

$$\min_{x \in \mathbb{R}^n} x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

is x_n , the eigenvector corresponding to λ_n , and the minimal value is λ_n . This can be proved by appealing to the eigenvector-eigenvalue form of A and the properties of orthogonal matrices.

2 Singular Value Decomposition

2.1 Introduction

The best matrices (real symmetric matrices S) have real eigenvalues and orthogonal eigenvectors. But for other matrices, the eigenvalues are complex or the eigenvectors are not orthogonal. If A is not square then $Ax = \lambda x$ is impossible and eigenvectors fail (left side in \mathbf{R}^m , right side in \mathbf{R}^n). We need an idea that succeeds for every matrix.

The **Singular Value Decomposition** fills this gap in a perfect way. In our applications, A is often a matrix of data.

The key point is that we need **two sets of singular vectors**, the u 's and

the v 's. For a real m by n matrix, the n right singular vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ are orthogonal in \mathbf{R}^n and the m left singular vectors $\mathbf{u}_1, \dots, \mathbf{u}_m$ are perpendicular to each other in \mathbf{R}^m .

The connection between n \mathbf{v} 's and m \mathbf{u} 's is not $A\mathbf{x} = \lambda\mathbf{x}$. That is for eigenvectors. For singular vectors, each Av equals σu :

$$\begin{cases} Av_1 = \sigma_1 u_1 \\ \vdots \\ Av_r = \sigma_r u_r \end{cases}$$

$$\begin{cases} Av_{r+1} = 0 \\ \vdots \\ Av_n = 0 \end{cases}$$

I have separated the first r v 's and u 's from the rest. That number r is the rank of A , the number of independent columns (and rows). Then r is the dimension of the column space and the row space. We will have r positive singular values in descending order $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$. The last $n - r$ v 's are in the nullspace of A , and the last $m - r$ u 's are in the nullspace of A^T .

Exercise:

Write the system of equations in matrix form.

Solution:

All of the right singular vectors \mathbf{v}_1 to \mathbf{v}_n go in the columns of V . The left singular vectors \mathbf{u}_1 to \mathbf{u}_m go in the columns of U . Those are square orthogonal matrices ($V^T = V^{-1}$ and $U^T = U^{-1}$) because their columns are orthogonal unit vectors. Then equation (1) becomes the full SVD, with square matrices V and U :

$$AV = U\Sigma \quad (1)$$

i.e,

$$A[\mathbf{v}_1 \dots \mathbf{v}_r \dots \mathbf{v}_n] = [\mathbf{u}_1 \dots \mathbf{u}_r \dots \mathbf{u}_m] \left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & 0 \\ \hline 0 & & & 0 \end{array} \right]$$

Remark

$A\mathbf{v}_k = \sigma_k \mathbf{u}_k$ in the first r columns above. That is the important part of the SVD. It shows the basis of \mathbf{v} 's for the row space of A and then \mathbf{u} 's for the column space. After the positive numbers $\sigma_1, \dots, \sigma_r$ on the main diagonal of Σ , the rest of that matrix is all zero from the nullspaces of A and A^T .

The eigenvectors give $AX = X\Lambda$. But $AV = U\Sigma$ needs two sets of singular vectors.

Moreover, the orthogonality allows us to go from $AV = U\Sigma$ to the usual and famous expression of the SVD:

Exercise:

Show that:

$$A = U\Sigma V^T$$

Solution:

Multiply both sides of $AV = U\Sigma$ by $V^{-1} = V^T$.

The Singular Value Decomposition of A separates A into r pieces of rank 1:

$$A = U\Sigma V^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

2.2 Existence of the Singular-Value Decomposition

2.2.1 The theorem

Theorem 2.2.1. Let \mathbf{A} be a real $m \times n$ matrix of rank r . Write \mathbf{I}_r for the $r \times r$ identity matrix. A singular value decomposition (SVD) of \mathbf{A} is a factorisation $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, where

- \mathbf{U} is an $m \times r$ matrix such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_r$,
- \mathbf{V} is an $n \times r$ matrix such that $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$,
- Σ is an $r \times r$ diagonal matrix $\Sigma = \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix}$, where $\sigma_1 \geq \dots \geq \sigma_r$ are strictly positive.

The columns $\mathbf{u}_1, \dots, \mathbf{u}_r$ of \mathbf{U} , which form an orthonormal set, are called left singular vectors.. The columns $\mathbf{v}_1, \dots, \mathbf{v}_r$ of \mathbf{V} , which also form an orthonormal set, are called right singular vectors. The SVD yields a decomposition of \mathbf{A} as a sum of r rank-one matrices:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \quad (2)$$

The factorisation 2 can equivalently be expressed by the equations:

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i \quad \text{for } i = 1, \dots, r$$

Given an SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, we can expand \mathbf{U} to an $m \times m$ orthogonal matrix $\hat{\mathbf{U}}$ by adding $m - r$ extra columns. We can likewise expand \mathbf{V} to an $n \times n$ orthogonal matrix $\hat{\mathbf{V}}$ by adding $n - r$ extra columns, and furthermore expand Σ to an $m \times n$ matrix by adding extra entries that are all zero. In this case we have $\mathbf{A} = \hat{\mathbf{U}}\hat{\Sigma}\hat{\mathbf{V}}^\top$. We call such a factorisation of \mathbf{A} into a product of an $m \times m$ orthogonal matrix, $m \times n$ nonnegative diagonal matrix, and $n \times n$ orthogonal matrix, a **full SVD** (as opposed to the reduced SVD).

A full SVD expresses the linear transformation represented by \mathbf{A} as a rotation, followed by a scaling, following by another rotation.

2.2.2 Existence of SVDs

Every matrix \mathbf{A} has an SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$. As we shall see, the singular values are uniquely determined by \mathbf{A} .

Exercise:

Every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has an SVD.

Solution: Inductively define an orthonormal sequence of vectors $\mathbf{v}_1, \dots, \mathbf{v}_r \in \mathbb{R}^n$ such that

$$\begin{aligned}\mathbf{v}_1 &:= \arg \max_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\| \\ \mathbf{v}_2 &:= \arg \max_{\mathbf{v} \perp \mathbf{v}_1, \|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\| \\ \mathbf{v}_3 &:= \arg \max_{\mathbf{v} \perp \mathbf{v}_1, \mathbf{v} \perp \mathbf{v}_2, \|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|, \text{ etc.}\end{aligned}$$

where the sequence stops at the first index r such that \mathbf{A} is zero on $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}^\perp$. Furthermore, for $i = 1, \dots, r$, define $\sigma_i := \|\mathbf{A}\mathbf{v}_i\|$ and define the unit vector $\mathbf{u}_i \in \mathbb{R}^m$ by $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$.

The set $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$ is orthonormal by construction. To prove the theorem it suffices to show that $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is an orthonormal set and that $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. For the latter task, write $\mathbf{B} := \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. A direct calculation shows that $\mathbf{A}\mathbf{v}_i = \mathbf{B}\mathbf{v}_i$ for $i = 1, \dots, r$ and that $\mathbf{A}\mathbf{v} = \mathbf{B}\mathbf{v} = 0$ for any vector $\mathbf{v} \in \{\mathbf{v}_1, \dots, \mathbf{v}_r\}^\perp$. It follows that $\mathbf{A} = \mathbf{B}$.

Finally we show that $\mathbf{u}_i^\top \mathbf{u}_j = 0$ for all $1 \leq i < j \leq r$. To this end, define $\mathbf{w} : \mathbb{R} \rightarrow \mathbb{R}^n$ by $\mathbf{w}(t) = \frac{(\mathbf{v}_i + \varepsilon \mathbf{v}_j)}{\sqrt{1+t^2}}$. Since $\mathbf{w}(t)$ is a unit vector orthogonal to $\mathbf{v}_1, \dots, \mathbf{v}_{i-1}$ for all t it must be that $\|\mathbf{A}\mathbf{w}(t)\|^2$ is maximised at $t = 0$. But

$$\|\mathbf{A}\mathbf{w}(t)\|^2 = \frac{\|\sigma_i \mathbf{u}_i + t\sigma_j \mathbf{u}_j\|^2}{1+t^2} = \frac{\sigma_i^2 + 2t\sigma_i\sigma_j (\mathbf{u}_i^\top \mathbf{u}_j) + t^2\sigma_j^2}{1+t^2}$$

and a direct calculation shows that it has zero derivative at $t = 0$ if and only if $\mathbf{u}_i^\top \mathbf{u}_j = 0$.

2.2.3 SVDs and Spectral Theory

In this section we explain the relationship between a singular value decomposition of a matrix \mathbf{A} and the eigenvalues and eigenvectors of the matrices $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$. Note here that $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are square matrices and so it makes sense to talk about their eigenvectors and eigenvalues.

In general a matrix may have many different SVDs. However the following proposition shows that all SVDs involve the same singular values. Thus we may speak of the singular values of a matrix \mathbf{A} .

Exercise: Given any SVD of \mathbf{A} , show that the singular values are the square roots of the nonzero eigenvalues of $\mathbf{A}^\top \mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$ (these matrices have the same eigenvalues).

Solution: We show the result for $\mathbf{A}^\top \mathbf{A}$. Given a full SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, we have

$$\begin{aligned}\mathbf{A}^\top \mathbf{A}\mathbf{V} &= (\mathbf{U}\Sigma\mathbf{V}^\top)^\top (\mathbf{U}\Sigma\mathbf{V}^\top) \mathbf{V} \\ &= \mathbf{V}\Sigma^\top \mathbf{U}^\top \mathbf{U}\Sigma\mathbf{V}^\top \mathbf{V} \\ &= \mathbf{V}\Sigma^\top \Sigma\mathbf{V}^\top \mathbf{V} \\ &= \mathbf{V}\Sigma^2.\end{aligned}$$

It follows that for $i = 1, \dots, r$ each right singular vector \mathbf{v}_i of \mathbf{A} is an eigenvector of $\mathbf{A}^\top \mathbf{A}$ with non-zero eigenvalue σ_i^2 . The remaining columns of \mathbf{V} span the eigenspace of $\mathbf{A}^\top \mathbf{A}$ corresponding to the eigenvalue zero.

One can similarly show that the left singular values of \mathbf{A} are eigenvectors of $\mathbf{A}\mathbf{A}^\top$.

We can see that in any full SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, the columns of \mathbf{U} comprise an orthonormal basis of eigenvectors of $\mathbf{A}\mathbf{A}^\top$ and the columns of \mathbf{V} comprise an orthonormal basis of eigenvectors of $\mathbf{A}^\top \mathbf{A}$. Indeed an alternative way to prove the existence of an SVD of matrix \mathbf{A} is to rely on results about the spectral theory of either of the matrices $\mathbf{A}^\top \mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$.

2.3 The Eckhart-Young Theorem

2.3.1 The Frobenius Norm

The main application of SVD for our purposes is to compute a best low-rank approximation of a given matrix. In order to formalise this notion we introduce the Frobenius matrix norm. The Frobenius norm of an $m \times n$ matrix $\mathbf{A} = (a_{ij})$, denoted $\|\mathbf{A}\|_F$, is defined by

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} = \sqrt{\text{trace}(\mathbf{A}^\top \mathbf{A})} \quad (3)$$

Note that the Frobenius norm of \mathbf{A} is a function of the singular values of \mathbf{A} . Indeed, if $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ then

$$\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{A}^\top \mathbf{A}) = \text{trace}((\mathbf{U}\Sigma\mathbf{V}^\top)^\top (\mathbf{U}\Sigma\mathbf{V}^\top)) = \text{trace}(\Sigma^\top \Sigma) = \sigma_1^2 + \dots + \sigma_r^2$$

2.3.2 Low-Rank Approximation via the SVD

Consider a matrix \mathbf{A} that has an SVD $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Given $k \leq r$ we obtain a rank- k matrix \mathbf{A}_k by "truncating" the SVD after the first k terms:

$$\mathbf{A}_k := \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

The image of \mathbf{A}_k is spanned to the top k left singular vectors. Hence \mathbf{A}_k has rank k . By construction, \mathbf{A}_k has singular values $\sigma_1, \dots, \sigma_k$. Likewise, $\mathbf{A} - \mathbf{A}_k = \sum_{i=k+1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ has singular values $\sigma_{k+1}, \dots, \sigma_r$. Thus

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \sqrt{\sigma_{k+1}^2 + \dots + \sigma_r^2}.$$

Note that σ_{k+1} is the top singular vector of $\mathbf{A} - \mathbf{A}_k$.

2.3.3 Eckhart-Young Theorem

The following result says that \mathbf{A}_k is a best rank- k approximation of \mathbf{A} with respect to the Frobenius norm.

Theorem 2.3.1. *Eckhart-Young Theorem.*

Let \mathbf{A} be a real $m \times n$ matrix.

Then for any $k \in \mathbb{N}$ and any real $m \times n$ matrix \mathbf{B} of rank at most k we have

$$\|\mathbf{A} - \mathbf{A}_k\|_F \leq \|\mathbf{A} - \mathbf{B}\|_F$$

The Eckhart-Young Theorem can also be formulated in terms of orthogonal projections.

Write $\mathbf{P}_k := \sum_{i=1}^k \mathbf{u}_i \mathbf{u}_i^\top$ for the matrix representing the orthogonal projection of \mathbb{R}^m onto the subspace spanned by the top k left singular vectors of \mathbf{A} .

Theorem 2.3.2. *Let \mathbf{A} be a real $m \times n$ matrix. Then for any $k \in \mathbb{N}$ and any $m \times m$ orthogonal projection matrix \mathbf{P} of rank k , we have $\|\mathbf{A} - \mathbf{P}_k \mathbf{A}\|_F \leq \|\mathbf{A} - \mathbf{P} \mathbf{A}\|_F$.*

2.3.4 Proof of the Eckhart-Young Theorem

The key to proving the Eckhart-Young theorem is the following exercise, which gives a lower bound on the singular values of a perturbation of matrix \mathbf{A} by matrix \mathbf{B} of rank at most k . In the exercise, we use the notation $\sigma_i(\mathbf{A})$ to refer to the i -th singular value of a matrix \mathbf{A} . If i greater than the rank of \mathbf{A} then $\sigma_i(\mathbf{A})$ is defined to be 0. The proof of the exercise uses the following "triangle inequality" (whose proof we leave as an exercise): for any two matrices \mathbf{A} and \mathbf{B} of the same dimension, $\sigma_1(\mathbf{A} + \mathbf{B}) \leq \sigma_1(\mathbf{A}) + \sigma_1(\mathbf{B})$.

Exercise:

Show that If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, with \mathbf{B} having rank at most k , then $\sigma_{k+i}(\mathbf{A}) \leq \sigma_i(\mathbf{A} - \mathbf{B})$ for all i .

Solution: We first show the case $i = 1$, i.e., we prove that $\sigma_{k+1}(\mathbf{A}) \leq \sigma_1(\mathbf{A} - \mathbf{B})$.

The kernel of \mathbf{B} has dimension $n - k$ and thus there must exist a unit-length vector \mathbf{w} that lies both in the kernel of \mathbf{B} and in the span of the top $k + 1$ singular vectors $\mathbf{v}_1, \dots, \mathbf{v}_{k+1}$. Then we have

$$\|\mathbf{A}\mathbf{w}\| = \|(\mathbf{A} - \mathbf{B})\mathbf{w}\| \leq \sigma_1(\mathbf{A} - \mathbf{B})\|\mathbf{w}\|$$

On the other hand, from the SVD $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ we have

$$\begin{aligned} \|\mathbf{A}\mathbf{w}\|^2 &= \left\| \sum_{i=1}^{k+1} \sigma_i \mathbf{u}_i (\mathbf{v}_i^\top \mathbf{w}) \right\|^2 \quad \text{since } \mathbf{w} \in \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_{k+1}) \\ &= \sum_{i=1}^{k+1} \sigma_i^2 (\mathbf{v}_i^\top \mathbf{w})^2 \\ &\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\mathbf{v}_i^\top \mathbf{w})^2 \\ &= \sigma_{k+1}^2 \|\mathbf{w}\|^2. \end{aligned}$$

We conclude that

$$\sigma_{k+1}(\mathbf{A})\|\mathbf{w}\| \leq \sigma_1(\mathbf{A} - \mathbf{B})\|\mathbf{w}\|$$

and hence

$$\sigma_{k+1}(\mathbf{A}) \leq \sigma_1(\mathbf{A} - \mathbf{B})$$

Now we do the general case:

$$\begin{aligned}
\sigma_i(\mathbf{A} - \mathbf{B}) &= \sigma_i(\mathbf{A} - \mathbf{B}) + \sigma_1(\mathbf{B} - \mathbf{B}_k) \quad \text{since } \mathbf{B} = \mathbf{B}_k \\
&= \sigma_1(\mathbf{A} - \mathbf{B} - (\mathbf{A} - \mathbf{B})_{i-1}) + \sigma_1(\mathbf{B} - \mathbf{B}_k) \\
&\geq \sigma_1(\mathbf{A} - \mathbf{B} - (\mathbf{A} - \mathbf{B})_{i-1} + \mathbf{B} - \mathbf{B}_k) \\
&= \sigma_1(\mathbf{A} - (\mathbf{A} - \mathbf{B})_{i-1} - \mathbf{B}_k) \\
&\geq \sigma_{i+k}(\mathbf{A})
\end{aligned}$$

Exercise:

Deduce a proof to the Eckhart-Young Theorem

Solution:

We have:

$$\begin{aligned}
\|\mathbf{A} - \mathbf{A}_k\|_F^2 &= \sum_{i=k+1}^r \sigma_i(\mathbf{A})^2 \\
&\leq \sum_{i=1}^{r-k} \sigma_i(\mathbf{A} - \mathbf{B})^2 \\
&\leq \|\mathbf{A} - \mathbf{B}\|_F^2.
\end{aligned}$$

2.3.5 Choosing k

The Eckhart-Young Theorem can help to determine what value of k to take in order to ensure that X_k is a "sufficiently good" approximation of X . In particular it allows to express the relative error a low-rank approximation X_k in terms of the singular values of \mathbf{X} , since

$$\frac{\|X - X_k\|_F^2}{\|X\|_F^2} = \frac{\sigma_{k+1}^2 + \dots + \sigma_r^2}{\sigma_1^2 + \dots + \sigma_r^2}$$

Thus if our goal is to ensure a given bound on the relative error (say at most 0.05), then we can find an appropriate value of k by examining the singular values, instead of proceeding by trial and error and computing X_k for various values of k .

3 Principal Component Analysis

3.1 Introduction

3.1.1 Goals of PCA

Principal components analysis (PCA) is a dimensionality reduction technique that can be used to give a compact representation of data while minimising information loss. Suppose we are given a set of data, represented as vectors in a high-dimensional space. It may be that many of the variables are correlated and that the data closely fits a lower dimensional linear manifold. In this case, PCA finds such a lower dimensional representation in terms of uncorrelated variables called principal components. PCA can also be kernelised, allowing it to be used to fit data to low-dimensional non-linear manifolds. Besides dimensionality reduction, PCA can also uncover patterns in data and lead to a potentially less noisy and more informative representation. Often one applies PCA to prepare data for further analysis, e.g., finding nearest neighbours or clustering.

3.1.2 Technical Overview

In a nutshell, PCA proceeds as follows. We are given a collection of data in the form of n vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^m$. By first translating the data vectors, if necessary, we may assume that the input data are mean centred, that is, $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$. Given a target number of dimensions $k \ll m$, PCA aims to find an orthonormal family of k vectors $\mathbf{u}_1, \dots, \mathbf{u}_k \in \mathbb{R}^m$ that "explain most of the variation in the data". More precisely, for $i = 1, \dots, n$ we approximate each data point \mathbf{x}_i by a linear expression $z_{i1}\mathbf{u}_1 + \dots + z_{ik}\mathbf{u}_k$ for some scalars $z_{i1}, \dots, z_{ik} \in \mathbb{R}$; the goal of PCA is to choose the \mathbf{u}_i so as to optimise the quality of this approximation over all data points. The optimal such vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ are the k principal components: \mathbf{u}_1 is direction of greatest variance in the data, \mathbf{u}_2 is the direction of greatest variance that is orthogonal to \mathbf{u}_1 , etc. To find the principal components we apply a matrix factorisation technique—the singular value decomposition – to the $m \times n$ matrix whose columns are the mean-centred data points \mathbf{x}_i . In the end, representing each data point $\mathbf{x}_i \in \mathbb{R}^m$ by its coordinates $\mathbf{z}_i = (z_{i1}, \dots, z_{ik})$ with respect to the k principal components yields a lower dimensional and (hopefully) more informative representation.

3.2 PCA

In this section we show how the singular value decomposition is used in principal components analysis. In PCA the input is a family $\mathbf{x}_1, \dots, \mathbf{x}_n$ of data points in \mathbb{R}^m . Write $\boldsymbol{\mu} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ for the mean data point. By first replacing \mathbf{x}_i by $\mathbf{x}_i - \boldsymbol{\mu}$ we may assume that the input data are mean centred. Given a target dimension $k \leq m$, our goal is to find points $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n \in \mathbb{R}^m$ such that the reconstruction error $\sum_{i=1}^n \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2$ is minimised subject to the constraint that $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n$ lie in a subspace of \mathbb{R}^m of dimension at most k .

Respectively write the mean-centred data points and their approximants as the columns of two $m \times n$ matrices

$$\mathbf{X} := \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix} \quad \text{and} \quad \tilde{\mathbf{X}} := \begin{pmatrix} | & & | \\ \tilde{\mathbf{x}}_1 & \dots & \tilde{\mathbf{x}}_n \\ | & & | \end{pmatrix}$$

Then the reconstruction error is nothing but $\|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2$. Thus by the Eckhart-Young Theorem an optimal choice of $\tilde{\mathbf{X}}$ is the matrix \mathbf{X}_k , as defined in 3 via a "truncated" SVD. Now recall that if \mathbf{U}_k is the $n \times k$ matrix whose columns are the top k left singular vectors of \mathbf{X} then, writing $\mathbf{Z} := \mathbf{U}_k^\top \mathbf{X}$, we have

$$\mathbf{X}_k = \mathbf{U}_k \mathbf{U}_k^\top \mathbf{X} = \mathbf{U}_k \mathbf{Z}$$

The output of PCA is the pair of matrices \mathbf{U}_k and \mathbf{Z} . The columns of \mathbf{U}_k are the top k left singular vectors, while the columns of \mathbf{Z} give the coefficients that respectively approximate each mean centred data point \mathbf{x}_i as a linear combination of the top k left singular vectors.