Computing for Data Sciences
Lecture 4

Introduction:

The Singular value decomposition (SVD) is a factorization of a real or complex matrix. It has many useful applications in signal processing and statistics. Formally, the singular value decomposition of an $m \times n$ real or complex matrix M is a factorization of the form $M = U\Sigma V^T$, where U is an $m \times m$ real or complex unitary matrix, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and V^T (the conjugate transpose of V, or simply the transpose of V if V is real) is an $n \times n$ real or complex unitary matrix. The diagonal entries $\Sigma_{i,j}$ of Σ are known as the singular values of M. The m columns of U and the n columns of V are called the left-singular vectors and right-singular vectors of M, respectively.

SVD is primarily used for dimension reduction which is converting data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions convey much more information.

Consider an example:

Suppose you have a list of 100 movies and 1000 people and for each person, you know whether they like or dislike each of the 100 movies. So for each instance (which in this case means each person) you have a binary vector of length 100 (position i is 0 if that person dislikes the i'th movie, otherwise 1.

You can perform your machine learning task on these vectors directly.. but instead you could decide upon 5 genres of movies and using the data you already have, figure out whether the person likes or dislikes the entire genre and, in this way reduce your data from a vector of size 100 into a vector of size 5 [position i is 1 if the person likes genre i]

The vector of length 5 can be thought of as a good representative of the vector of length 100 because most people might be liking movies only in their preferred genres. Though this may not be an exact representation since there will be people who might hate all the 5 genres but the rest of them more or less can be confined using this method.

This is mainly done because the reduced vector conveys most of the information in the larger one while consuming a lot less space and being faster to compute with.

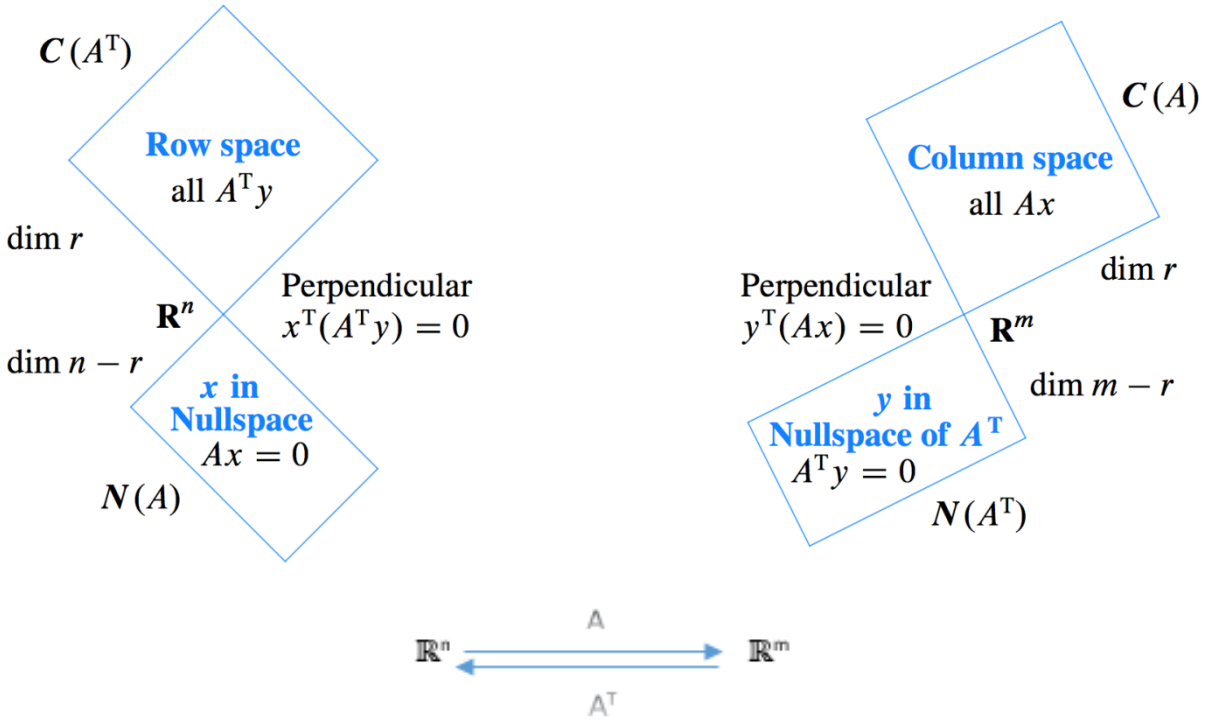


Figure 1. Fundamental picture of Linear Algebra

Fig.1 is used to explain the equation Ax = b. The first step sees Ax (matrix times vector) as a combination of the columns of A. Those vectors Ax fill the column space C (A). When we move from one combination to all combinations (by allowing every x), a subspace appears. Ax = b has a solution exactly when b is in the column space of A

NOTE: We choose a matrix of rank one, $= xy^T$. When $m = n = 2$, all four fundamental subspaces are lines in $\mathbb{R}^2$

The dimensions obey the most important laws of linear algebra: $dim\, R(A) = dim\, R(A^T)$ and $dim\, R(A) + dim\, N(A) = n$.

The row space has dimension r, the nullspace has dimension $n - r$. Elimination identifies r pivot variables and $n - r$ free variables. Those variables correspond, in the echelon form, to columns with pivots and columns without pivots. They give the dimension count r and $n - r$.

Every x in the nullspace is perpendicular to every row of the matrix, exactly because $Ax = 0$

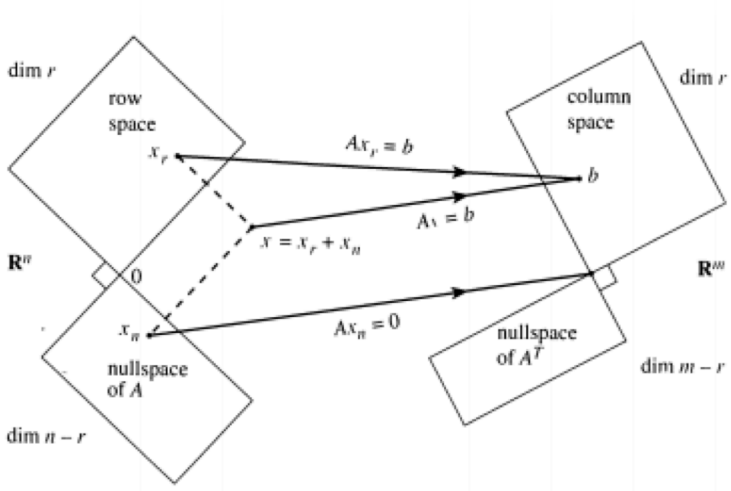


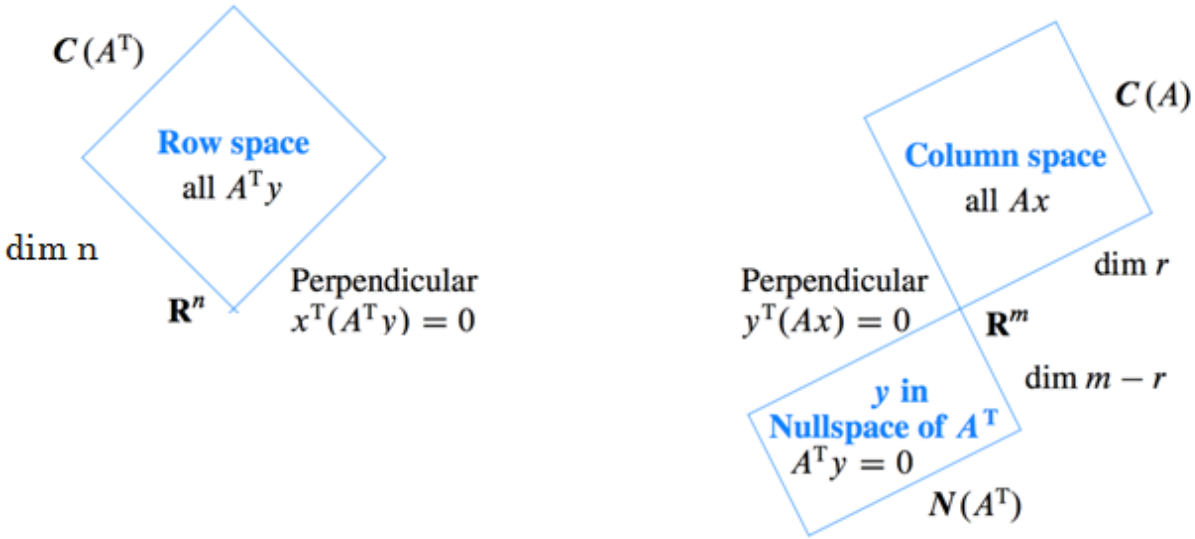
Figure 1. The action of A: Row space to column space, nullspace to zero.

Fig.2

From Fig.2 we can see that A takes x into the column space, nullspace goes to the zero vector. Nothing goes elsewhere in the left nullspace. With b in the column space, $Ax = b$ can be solved. There is a particular solution x, in the row space. The homogeneous solutions x, forms the nullspace. The general solution is $x_r + x_n$. The particularity of x_r is that it is orthogonal to every x_r.

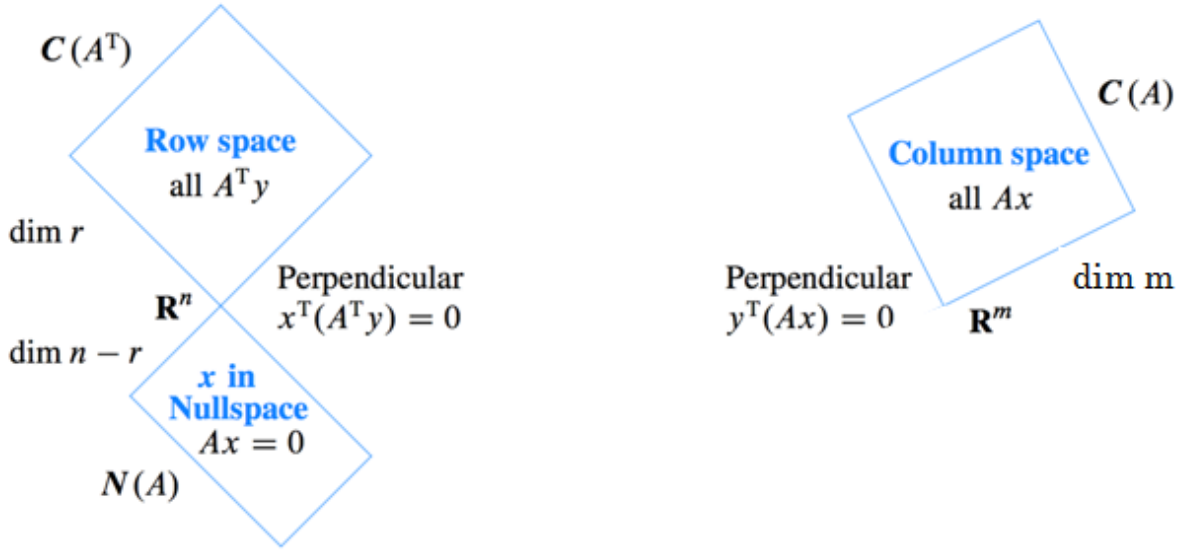
Visualizing the Fundamental Picture of Linear Algebra in terms of different types of Matrices.

- Matrices which behave as Injective function (1 to 1).



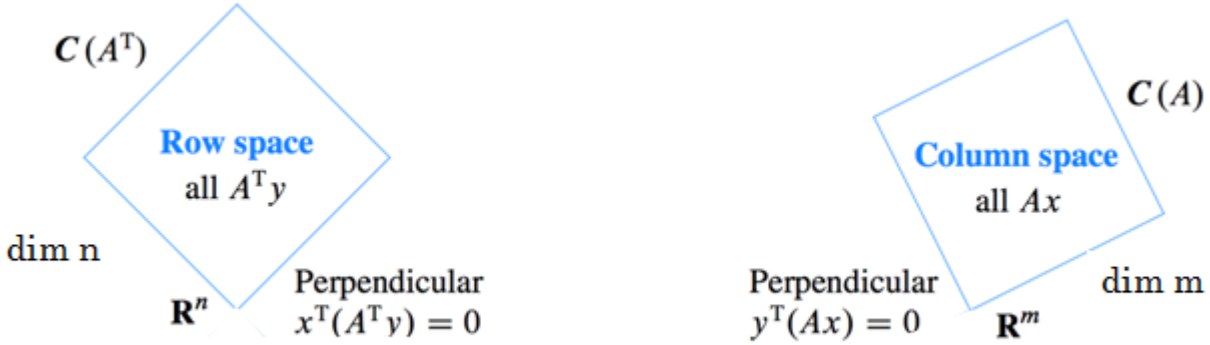
In the above figure it can be seen that all the elements of Row space of A map to Column space of A. There is only one element in $C(A^T)$ which maps to zero. There is no null space of A, only there is zero in Row space of A that maps to zero.

- Matrices which behave as Surjective function (Onto).

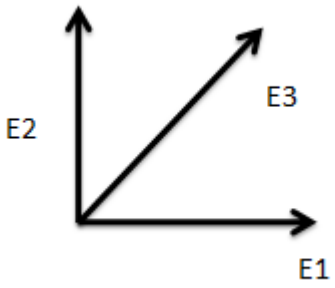


In the above picture it can be seen that the null space of A^T does not exists. This ensures that the range of matrix we are considering does not have any element that can be mapped from

- Matrices which behave as Bijective function (Onto).



The span of a finite, non-empty set of vectors $\{u_1, u_2, \ldots, u_n\}$ is the set of all possible linear combinations of those vectors, i.e. the set of all vectors of the form $c_1u_1 + c_2u_2 + \ldots + c_nu_n$ for some scalars $c_1, c_2, \ldots, c_n$.



The three vectors E1, E2, E3 present in R^2 span the whole R^2 . This can be represented as Span{E1,E2,E3}= R^2

The vectors E1,E2,E3 can be the representation of the space R^2. We don't need all the three vectors to represent our space. The space R^2 can be spanned by 2 vectors either E1, E2 or E2, E3 or E1, E3.

Span{E1,E2}= R² *Span{E2,E3}= R²* *Span{E1,E3}= R²*

This type of representation where only n-linearly independent vectors map the n-dimensional vector space is know is Minimal Representation. The minimum set of vectors that span the space are called Basis. ex *{E1,E2} ,{E2,E3}.*

A basis of a vector space V is defined as a subset $v_1, \ldots, v_n$ of vectors in V that are linearly independent and vector space span V. Consequently, if $(v_1, v_2, \ldots, v_n)$ is a list of vectors in V, then these vectors form a basis if and only if every $v \in V$ can be uniquely written as

$$v = a_1 v_1 + a_2 v_2 + \ldots + a_n v_n,$$

where a_1, ..., a_n are scalars. A vector space V will have many different bases, but there is always the same number of basis vectors in each of them. The number of basis vectors in V is called the dimension of V. Every spanning list in a vector space can be reduced to a basis of the vector space. The bases need not be orthogonal to each other they must be linearly independent.

Greedy Approach of finding bases of Vector

The greedy approach involves picking n linearly independent vectors such that they are linearly independent. These vectors can be the basis of n dimensional space. There are two types of greedy Algorithm for finding bases of vectors.

Grow Algorithm

```
def GROW(V)
  S = ∅
  repeat while possible:
      find a vector v in V that is not in Span S, and put it in S.
```

The algorithm stops when there is no vector to add, at which time S spans all of V. Thus, if the algorithm stops, it will have found a generating set.

Shrink Algorithm

```
def SHRINK(V)
  S = some finite set of vectors that spans V
  repeat while possible:
      find a vector v in S such that Span (S − {v}) = V, and remove v from S.
```

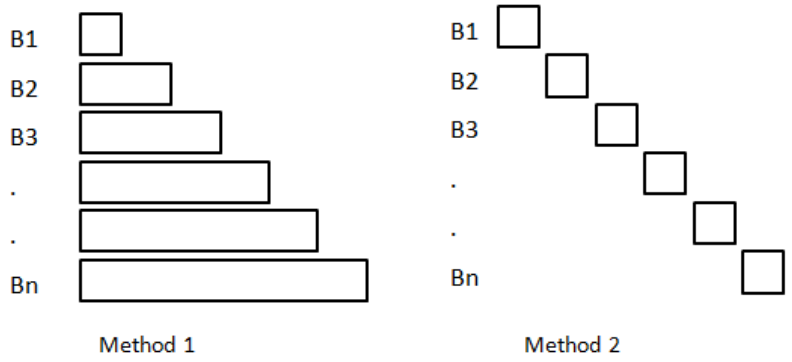
The algorithm stops when there is no vector whose removal would leave a spanning set. At every point during the algorithm, S spans V, so it spans V at the end. Thus, if the algorithm stops, the algorithm will have found a generating set.

Choice by Induction Approach of finding bases of Vector

Method 1: Pick a basis $B1 \in \Re^1$ and then $B2 \in \Re^2$ Such that $B2 \notin Span(B1)$. Then pick a basis $B3 \in \Re^3$ such that $B3 \notin Span(B1, B2)$. Then proceed in similar manner to find the other basis. Finally find the $Bn \in \Re^n$ such that $Bn \notin Span(B1, B2, ...Bn-1)$.

Method 2: Pick a vector *Bi* such that *Bi* is orthogonal to all the bases vector selected before it (*B1, B2... Bi-1*). This will ensure that all the vectors are linearly independent. Being orthogonal means the inner product of two vectors is zero. $\langle Bi, B1 \rangle = 0$ $\langle Bi, B2 \rangle = 0$ ..... $\langle Bi, Bi-1 \rangle = 0$

The orthonormal basis vectors means, the basis vectors are orthogonal and their size=1, i.e their norm=1. $\|B1\| = 1$, $\|B2\| = 1$, $\|B3\| = 1$..... $\|Bn\| = 1$



Representation of basis vectors produced by the above two methods.

A standard basis is a orthonormal basis in which each vector in the basis has only one nonzero entry, and that entry is equal to 1. The vectors are usually denoted $\mathbf{e}_i$ with $i=1$, ..., n, with n representing the dimension of the vector space that is spanned by this basis. For example, in the real vector space $\mathbb{R}^3$, the standard basis is the set of vectors $\{\{1,0,0\},\{0,1,0\},\{0,0,1\}\}$.

Up to this point, we have talked about basis generation but nothing was said about bases for the four subspaces. Those bases can be constructed from an echelon form-the output from elimination. This construction is simple, but the bases are not perfect. A really good choice, in fact a "canonical choice" that is close to unique, would achieve much more. On top of it, we make two requirements:

1. The basis vectors are orthonormal.
2. The matrix with respect to these bases is diagonal.

If $v_1, v_2, \ldots, v_r$ is the basis for the row space and $u_1, u_2, \ldots, u_r$ is the basis for the column space, then $Av_i = \sigma_i u_i$. That gives a diagonal matrix Σ. We can further ensure that $u_i > 0$. Orthonormal bases are no problem-the Gram-Schmidt process is available. But a diagonal form involves eigenvalues. In this case they are the eigenvalues of A^TA and AA^T. Those matrices are symmetric and positive semidefinite, so they have nonnegative eigenvalues and orthonormal eigenvectors (which are the bases). Starting from $A^TAv_i = \sigma_i v_i$, and here are the key steps:

$v_i{}^T A^T A v_i = \sigma_i v_i{}^T v_i$, so that $\|Av_i\| = u_i$

$AA^TAv_i = \sigma^2 Av_i$ so that $u_i = Av_i/\sigma_i$ is a unit eigenvector of AA^T.

All these matrices have rank r. The r positive eigenvalues σ_i^2 give the diagonal entries σ_i of Σ. The whole construction is called the singular value decomposition (SVD). It amounts to a factorization of the original matrix A into $U\Sigma V^T$, where

1. U is an m by m orthogonal matrix. Its columns $u_1, \ldots, u_r, , \ldots, u_m$ are basis vectors for the column space and left nullspace.

2. C is an m by n diagonal matrix. Its nonzero entries are $\sigma_1 > 0, \ldots, \sigma_r > 0$.

3. V is an n by n orthogonal matrix. Its columns $v_1, v_2, \ldots, v_r$ are basis vectors for the row space and nullspace.

The equations $Av_i = \sigma_i u_i$ mean that $AV = U\Sigma$. Then multiplication by V^T gives $A = U\Sigma V^T$. When A itself is symmetric, its eigenvectors ui make it diagonal: $A = U\Lambda U^T$. The singular value decomposition extends this spectral theorem to matrices that are not symmetric and not square. The eigenvalues are in Λ, the singular values are in Σ. The factorization $A = U\Sigma V^T$ joins $A = LU$ (elimination) and $A = QR$ (orthogonalization) as a beautifully direct statement of a central theorem in linear algebra.

Think of the rows of an $n \times d$ matrix A as n data points in a d-dimensional space and consider the problem of finding the best k-dimensional subspace with respect to the set of points. Here best means minimize the sum of the squares of the perpendicular distances of the points to the subspace. We begin with a special case where the subspace is 1-dimensional, namely a line through the origin. The best fitting k-dimensional subspace is found by repeated applications of the best fitting line algorithm, each time finding the best fitting line perpendicular to the subspace found so far. When k reaches the rank of the matrix, a decomposition of the matrix, called the Singular Value Decomposition (SVD), is obtained from the best fitting lines.

The singular value decomposition of a matrix A is the factorization of A into the product of three matrices, $A = UDV^T$; where the columns of U and V are orthonormal and the matrix D is diagonal with positive real entries. In many applications, a data matrix A is close to a low rank matrix and a low rank approximation to A is desired. The singular value decomposition of A gives the best rank k approximation to A, for any k .

The singular value decomposition is defined for all matrices, whereas the more commonly used eigenvector decomposition requires the matrix A be square and certain other conditions on the matrix to ensure orthogonality of the eigenvectors. In contrast, the columns of V in the singular value decomposition, called the right-singular vectors of A, always form an orthonormal set with no assumptions on A. The columns of U are called the left-singular vectors and they also form an orthonormal set. A simple consequence of the orthonormality is that for a square and invertible matrix A, the inverse of A is VDU^T.

Example

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} = \frac{\begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix}}{\sqrt{10}} \begin{bmatrix} \sqrt{50} & 0 \\ 0 & 0 \end{bmatrix} \frac{\begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}}{\sqrt{5}} = U\Sigma V^T.$$

All four subspaces are 1-dimensional. The columns of A are multiples of $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ in U. The rows are multiples of [1 2] in V^T. Both A^TA and AA^T have eigenvalues 50 and 0. So the only singular value is $\sigma_1 = \sqrt{50}$.

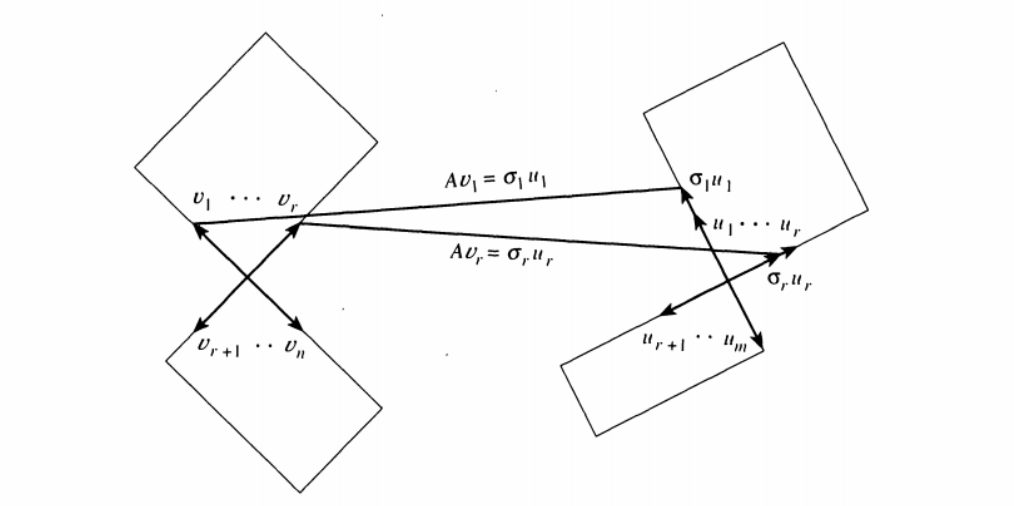


Figure: Orthonormal bases that diagonalize A.

Factorizations of Matrices:

By bringing together three important ways to factor the matrix A, we can find solutions to various problems.

1. PLU Factorization

Here,

P is a so-called permutation matrix, L is lower triangular, U is upper triangular.

We will start with LU Factorisation:

LU-Factorization:- The nonsingular matrix A has an LU-factorization if it can be expressed as the product of a lower-triangular matrix L and an upper triangular matrix U:

$$A = LU$$

Condition for factorization: square matrix A

Permutation Matrix:

A permutation σ on $\{1, \ldots, n\}$ has an associated permutation matrix

$$\mathbf{M}_\sigma = [m_{ij}], \text{ where } m_{ij} = \begin{cases} 1, & j = \sigma(i) \\ 0 & \text{otherwise} \end{cases}$$

So if we look at our σ = {2, 4, 1, 3}, we get matrix

$$\mathbf{M}_\sigma = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

PLU Matrix:

Decomposition of Matix $A = PLU$

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -8 & 8 & 1 \\ 2 & -2 & 0 \end{bmatrix}$$

PLU decomposition of A is

$$\begin{bmatrix} 0 & 1 & 0 \\ -8 & 8 & 1 \\ 2 & -2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

LU and PLU substitutions are mainly used in solving linear equations of the type: $Ax = b$

Method for solving;

- Find the PLU decomposition of A i.e. $A = PLU$
- The equation to be solved becomes;

$$PLUx = b;$$

$$LUx = P^T b$$

- Calculate $P^T b$ and use forward substitution to solve for y in $Ly = P^T b$. Then, having found y, use backward substitution to solve for x in $Ux = y$.
-

2. Decompositions based on eigenvalues

This methodis also called *spectral decomposition.*

- Condition for decomposition: square matrix A with distinct eigenvectors (not necessarily distinct eigenvalues).
- Decomposition:

  $A = VDV^{-1}$,

  where D is a diagonal matrix formed from the eigenvalues of A, and the columns of V are the corresponding eigenvectors of A.

  An n-by-n matrix A always has n eigenvalues, which can be ordered (in more than one way) to form an n-by-n diagonal matrix D and a corresponding matrix of nonzero columns V that satisfies the eigenvalue equation $AV = VD$. If the n eigenvectors (not necessarily

eigenvalues) are distinct (that is, none is equal to any of the others), then V is invertible, implying the decomposition $A = VDV^{-1}$.

- One can always normalize the eigenvectors to have length one (see definition of the eigenvalue equation). If A is real-symmetric, V is always invertible and can be made to have normalized columns. Then the equation $VV^T = I$ holds, because each eigenvector is orthogonal to the other. Therefore the decomposition (which always exists if A is real-symmetric) reads as: $A = VDV^T$
- The condition of having n distinct eigenvalues is sufficient but not necessary. The necessary and sufficient condition is for each eigenvalue to have geometric multiplicity equal to its algebraic multiplicity.
- The eigen decomposition is useful for understanding the solution of a system of linear ordinary differential equations or linear difference equations. For example, the difference equation $x_{t+1} = Ax_t$ starting from the initial condition $x_0 = c$ is solved by $x_t = A^t c$, which is equivalent to $x_t = VD^tV^{-1}c$, where V and D are the matrices formed from the eigenvectors and eigen values of A. Since D is diagonal, raising it to power D^t, just involves raising each element on the diagonal to the power t. This is much easier to do and to understand than raising A to power t, since A is usually not diagonal.

3. Factorization of the original matrix A into $U\sum V^T$

Where,

i. U is an m by m orthogonal matrix. Its columns $u_1, \ldots, u_r, , \ldots, u_m$ are basis vectors for the column space and left null space.

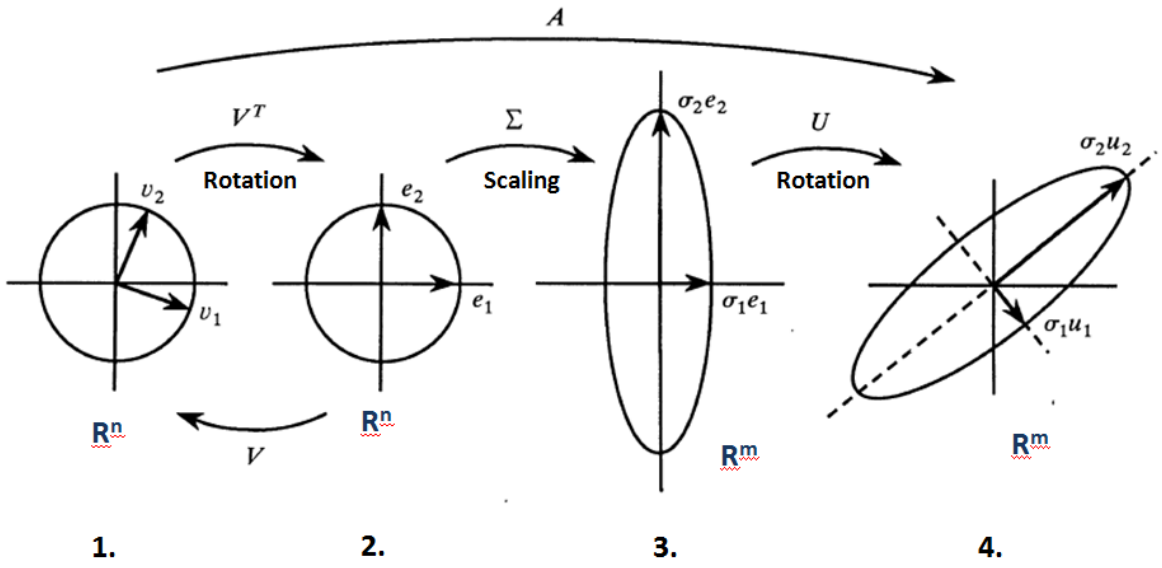
ii. $\sum$ is an m by n diagonal matrix. Its nonzero entries are $\sigma_1 > 0, \ldots, \sigma_r > 0$.

iii. V is an n by n orthogonal matrix. Its columns $v_1, .., v_r, \ldots, v_m$ are basis vectors for the row space and null space.

This factorization gives us good bases for the subspaces and a $\sum$ matrix that provides information about the most significance action of A.

Diagonalisation effect of A

This figure shows the diagonalization of A. Basis vectors go to .basis vectors (principal axes). A circle goes to an ellipse. The matrix is factored into $U\sum V^T$



1. V^T rotates the basis vectors and brings them to the axes.

2. The $\sum$- singular value matrix scales the matrix primarily in the direction of principle component, making it look like an ellipsoid. The stretching along each axes is independent of the other.

3. U matrix is an orthonormal basis matrix. It provides the effect of rotation in its own way. This operation does not provide any scaling effect.