# SANS2020-21 Matrix Factorizations

Jorge Garcia Vidal

May 2021

## 1 Some basic facts that you probably know about vectors and linear maps.

### 1.1 Vector spaces and sub-spaces

A vector space $V$ is a set of objects called vectors, which can be added and multiplied by numbers (scalars). In our course the scalars will be real or complex numbers. Vector addition and scalar multiplication operations must satisfy certain requirements:

- $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$
- $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$
- $\exists \mathbf{0} \in V$ such that $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$, $\forall \boldsymbol{v} \in V$.
- $\forall \boldsymbol{v} \in V, \exists {-\boldsymbol{v}} \in V$ such that $\boldsymbol{v} + (-\boldsymbol{v}) = \mathbf{0}$.

Moreover, for $a, b \in \mathbb{R}$:

- $a(b\boldsymbol{v}) = (ab)\boldsymbol{v}$
- $1\boldsymbol{v} = \boldsymbol{v}$.
- $a(\boldsymbol{u} + \boldsymbol{v}) = a\boldsymbol{u} + a\boldsymbol{v}$, and $(a + b)\boldsymbol{v} = a\boldsymbol{v} + b\boldsymbol{v}$

If $S$ is a subset of $V$ which is closed respect the operations of sum of vectors and multiplication by an scalar, and that itself is a vector space, we say that $S$ is a *vector subspace* of $V$.

### 1.2 Linear combinations and independence

If we have a set of vectors $\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$ a *linear combination* of these vectors is an expression of the form $\boldsymbol{v} = \sum_i a_i \boldsymbol{v_i}$ for some scalars $a_i$.

If we have a set of vectors belonging to a vector space, $\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$, the set of all linear combinations of these vectors is a vector subspace. This subspace is called span$\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$.

One vector $\boldsymbol{w}$ is *linearly independent* of a set of vectors $\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$ when $\boldsymbol{w}$ cannot be expressed as linear combination of the vectors $\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$, or in other words, when $\boldsymbol{w} \notin \text{span}\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$. A set of vectors $\{\boldsymbol{v_1}, ..., \boldsymbol{v_k}\}$ are linearly independent when the only linear combination that produces the vector $\mathbf{0}$ is the one with all coefficients equal to zero.

## 1.3 Bases and dimension

There are many examples of vector spaces: $\mathbb{R}^n$, $\mathbb{C}^n$, or $P_n(\mathbb{R})$, the set of polynomials of $n$ degree with real coefficients. The vector sum and scalar multiplication in these spaces are the ones you have studied during your high-school.

These spaces are examples of *finite dimensional* vector spaces. This means that there is a set of linearly independent vectors of $V$, $\{\boldsymbol{u_i}\}_{i=1,...,n}$ (a *base* of $V$) such that all other vectors of $V$ can be expressed as $\boldsymbol{v} = \sum_{i=1}^{n} v_i \boldsymbol{u_i}$. A vector space has in general an infinite number of possible bases, but the number of elements in each of those basis is always the same. We say that this number $n$ the *dimension* of $V$, $\dim(V) = n$. Moreover, we can use these coefficients to represent $\boldsymbol{v}$ as a column vector:

$$\boldsymbol{v} = [v_1, v_2, ..., v_n]^t = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}. \tag{1}$$

Same definitions apply to vector subspaces.

Some vector spaces have *infinite* dimensions, for instance, $P_\infty(\mathbb{R})$, the set of polynomials of an arbitrary order, $C[0, 1]$, the set of continuous functions defined on the interval $[0, 1]$, or $L^2[0, 1]$, the set of functions $f$ for which $\int_0^1 |f|^2 d\mu < \infty$. In this course we will deal with the finite dimensional case only. The infinite dimensional case is studied in *functional analysis*, and is important, for instance, when dealing with stochastic processes.

## 1.4 Scalar product, orthogonality and norm

The *scalar product* of two column vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is the scalar:

$$< \boldsymbol{u}, \boldsymbol{v} >= \boldsymbol{u}^t \boldsymbol{v}.$$

Two vectors are *orthogonal* when its scalar product is zero.

The *norm* (length) of a vector $\boldsymbol{v}$ is defined as the non-negative number:

$$||\boldsymbol{v}|| = +\sqrt{< \boldsymbol{v}, \boldsymbol{v} >}.$$

## 1.5 Linear maps and matrices

If $V$ and $W$ are vector spaces, we define a *linear map* $L$ as an application $L : V \to W$ for which $\forall \boldsymbol{u}, \boldsymbol{v} \in V$, and $\forall a, b \in \mathbb{R}$,

$$L(a\boldsymbol{u} + b\boldsymbol{v}) = aL(\boldsymbol{u}) + bL(\boldsymbol{v}).$$

Assume that $\{\boldsymbol{v_i}\}_{i=1,...,n}$ is a base of a $n$ dimensional space $V$, while $\{\boldsymbol{w_i}\}_{i=1,...,m}$ is a base of a $m$ dimensional space $W$. The image of the $\boldsymbol{u_i}$ base vector by the linear application $L$, i.e. $L(\boldsymbol{u_i})$, is a vector of $W$, that we can express in the base $\{\boldsymbol{w_i}\}_{i=1,...,m}$:

$$L(\boldsymbol{u_i}) = a_{1,i}\boldsymbol{w_1} + a_{2,i}\boldsymbol{w_2} + ... + a_{m,i}\boldsymbol{w_m}.$$

An $n \times m$ matrix $A$ is an arrangement of these numbers $a_{i,j}$ into an $n \times m$ array $A = [a_{i,j}]$. For an arbitrary vector $\boldsymbol{v}$, expressed as a column vector in the base $\{\boldsymbol{v_i}\}_{i=1,...,n}$, the product $A\boldsymbol{v}$ gives a result the vector $\boldsymbol{w}$, which is the image of the vector $\boldsymbol{v}$ by the linear map $L$ expressed in the base $\{\boldsymbol{w_i}\}_{i=1,...,m}$.

This is a bit confusing at first. Let's think on the following example: $\{1, x, x^2\}$ is a possible base of $P_2(\mathbb{R})$. Using this base, we can represent the polynomial $p(x) = a + bx + cx^2$ by the $\mathbb{R}^3$ vector $p = [a, b, c]^t$. Let us define the linear map $\frac{d}{dx} : P_2(\mathbb{R}) \to P_2(\mathbb{R})$ that assigns to a polynomial $p(x)$ its derivative (another polynomial). If we represent the polynomials in the image and domain sets by the same base $\{1, x, x^2\}$, we can thus represent the linear map by the matrix :

$$D = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \tag{2}$$

We will always use matrices to represent linear maps.

Very often, we express a matrix $A$ as an arrangement of its columns (or its rows) considered as vectors. For instance if $A = [a_{i,j}]$ and let us define the $n$ column vectors $\boldsymbol{a_i} = [a_{i,1}, ..., a_{i,m}]^t$ for $i = 1, ..., n$. We can write the matrix as:

$$A = [\boldsymbol{a_i}, ..., \boldsymbol{a_n}]$$

Similarly, we can write $A$ using its row vectors $\boldsymbol{r_i}^t = [a_{1,i}, ..., a_{n,i}]$:

$$A = \begin{bmatrix} \boldsymbol{r_1}^t \\ \vdots \\ \boldsymbol{r_m}^t \end{bmatrix} \tag{3}$$

If we have a matrix $A$ and a column vector $\boldsymbol{v} = [v_1, ..., v_n]^t$, we can express its matrix times vector product as

$$A\boldsymbol{v} = [\boldsymbol{a_1}, ..., \boldsymbol{a_n}] \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i=1,..,n} v_i \boldsymbol{a_i}.$$

## 1.6 Matrix rank

The image of all vectors of a subspace $S$ of $V$ by means of a matrix $A$ is also a subspace of $W$.

In the special case when the subspace $S$ is $V$ itself, the generated subspace is $\text{span}\{\boldsymbol{a_i}, ..., \boldsymbol{a_n}\}$. The dimension of this subspace is *the rank* of the matrix $A$, $\text{rank}(A)$, and it is the number of linearly independent columns.

Very surprisingly, this number is also the number of linearly independent *rows* of the matrix $A$, meaning that $A$ *and* $A^t$ *have the same rank.*

## 1.7 Eigenvectors and eigenvalues

Let $A$ be a square real matrix (although the matrix $A$ has real components, in this section is better to think that vectors can have complex coefficients and that scalars are also in general complex numbers).

A non zero vector $\boldsymbol{v}$ is an *eigenvector*, and the scalar $\lambda$ is an *eigenvalue* of the matrix $A$ when:

$$A\boldsymbol{v} = \lambda\boldsymbol{v}.$$

Eigenvalues must fulfill the condition: $A\boldsymbol{v} - \lambda\boldsymbol{v} = (A - \lambda I)\boldsymbol{v} = \boldsymbol{0}$. As $\boldsymbol{v}$ is non-zero, this is only possible if $\text{rank}(A) < n$, which is equivalent to the condition $\det(A - \lambda I) = 0$. This determinant is general a polynomial on $\lambda$ of degree $n$, meaning that has $n$ complex roots (if we count multiplicity of roots). For very small matrices we can find the roots of this polynomial to compute eigenvalues. For larger matrices there are more computationally efficient methods.

Once we know the eigenvalues, we can find the associated eigenvectors by solving the undetermined system of linear equations $(A - \lambda I)\boldsymbol{v} = \boldsymbol{0}$. We can set, for instance, the condition $||\boldsymbol{v}|| = 1$ to find unique solutions (up to the sign).

In the case of eigenvalues of multiplicity larger than 1, we can have several linearly independent associated eigenvectors. The dimension of the generated subspace must be less of equal the multiplicity of the root. When this dimension is strictly lower that the multiplicity of $\lambda$, we say that the matrix is *defective*. You can find detailed discussions on this in any text of linear algebra. We will be interested mainly in symmetric matrices, which are never defective.

## 1.8 Some important types of matrices

- An square $n \times n$ matrix $\Lambda$ with 0 off-diagonal elements is a *diagonal matrix*. The product $\Lambda \boldsymbol{v}$ produces an stretching of the different components of the vector $\boldsymbol{v}$ according with the corresponding values of the diagonal. If some components of the diagonal of $\Lambda$ are zero, the product collapses to zero the corresponding components of $\boldsymbol{v}$. The inverse of $\Lambda$ is obtained by simply inverting the diagonal elements (if any of those elements is zero, then the matrix is not invertible).

- A square matrix $Q$ with columns that are orthonormal vectors (i.e. orthogonal and with norm 1) is called an *orthonormal matrix* (or very often simply *orthogonal* matrix, as we will assume the norm 1 condition). The product $Q\boldsymbol{v}$ produces a rotation, a reflection, or a combination of both operations, on the vector $\boldsymbol{v}$. Orthonormal matrices are always invertible and $Q^{-1} = Q^t$.

- An square matrix $A$ that fulfils $A = A^t$ is a *symmetric matrix.*

- If $\boldsymbol{v}$ is a column vector of $V$, the square $n \times n$ matrix $\boldsymbol{v}\boldsymbol{v}^t$ is a *projection matrix* (do not confuse with the scalar product $\boldsymbol{v}^t\boldsymbol{v}$ which is a number). Its rank is 1. If we have an arbitrary vector $\boldsymbol{u}$ in $V$, the product of the projection matrix and the vector $\boldsymbol{u}$ is a vector which is the orthogonal projection of the vector $\boldsymbol{u}$ on the line that follows the direction of $\boldsymbol{v}$ (i.e. in the 1-dimensional subspace generated by the vector $\boldsymbol{v}$): $\boldsymbol{v}\boldsymbol{v}^t u =< \boldsymbol{v}, \boldsymbol{u} > \boldsymbol{v}$.

## 1.9 An $n \times m$ matrix of rank $r$ maps an sphere of dimension $n$ into an ellipsoid of dimension $r$

A basic fact of linear algebra is the following: Assume that we have an $n \times m$ matrix $A$ of with rank($A$)=$r$. Assume that we compute the products $y = Ax$, where $x$ is an $n$-dimensional vector that lies in a sphere of radios 1 in the space $R^n$. Then the locus of all the generated vector $y$ lies in an ellipsoid of dimension $r$ embedded in the space $R^m$.

As an example, assume a $2 \times 2$ matrix $A$ of rank 2. If we compute $y = Ax$ for $x = (cos(\theta), sin(\theta))^t$ with $\theta \in [0, 2\pi)$, the vector $y$ will lie in an ellipsoid centered in the origin in $R^2$. If rank($A$)=1, the ellipsoid will collapse one of its dimensions, resulting in a segment that crosses the origin.

## 1.10 Matrix factorization

In this course we deal with two important matrix factorizations:

- $S = Q\Lambda Q^t$, for symmetric matrices.

- $A = U\Sigma V^t$, Singular Value Decomposition (SVD) for general matrices.

We are interested in the first factorization as covariance matrices are symmetric, and we are interested in the second factorization as it allow to approximate clouds of points in high-dimensional spaces by clouds of points in lower dimensional spaces.

## 1.11 Diagonalization of symmetric matrices

Assume that $S$ is an $n \times n$ *symmetric matrix*, i.e. $S = S^T$.

The *spectral theorem* tell us that:

- $S$ has $n$ real eigenvalues, $\lambda_i$, (counting possible multiciplities)
- The $n$ associated eigenvectors, $\boldsymbol{q_i}$ are orthonormal.
- These matrices are not defective.

Let us prove it for the case in which the matrix has non-repeated eigenvalues. This result can be extended for the repeated eigenvalues case by using continuity arguments.

### 1.11.1 Eigenvectors are orthogonal

Assume that $S\boldsymbol{v} = \lambda\boldsymbol{v}$ and $S\boldsymbol{u} = \mu\boldsymbol{u}$ different eigenvalues $\lambda$ and $\mu$. We have:

$$\boldsymbol{v}^t S\boldsymbol{u} = \mu\boldsymbol{v}^t\boldsymbol{u},$$

and

$$\boldsymbol{u}^t S\boldsymbol{v} = \lambda\boldsymbol{u}^t\boldsymbol{v}.$$

but $\boldsymbol{u}^t S\boldsymbol{v} = \boldsymbol{v}^t S^t\boldsymbol{u} = \boldsymbol{v}^t S\boldsymbol{u}$ as $S$ is symmetric, and $\boldsymbol{u}^t\boldsymbol{v} = \lambda\boldsymbol{v}^t\boldsymbol{u}$, which implies $\lambda\boldsymbol{u}^t\boldsymbol{v} = \mu\boldsymbol{u}^t\boldsymbol{v}$, and from this we get $\boldsymbol{u}^t\boldsymbol{v} = 0$.

### 1.11.2 Eigenvalues are real

Assume that $\lambda$ is a complex eigenvalue. As $S$ is real, $\lambda^*$ (i.e. its complex conjugate) must also be an eigenvalue:

$$S\boldsymbol{v} = \lambda\boldsymbol{v},$$

and:

$$S\boldsymbol{v}^* = \lambda^*\boldsymbol{v}^*.$$

Then we have:

$$\boldsymbol{v}^+ S\boldsymbol{v} = \lambda\boldsymbol{v}^+\boldsymbol{v} = \lambda$$

and:

$$\boldsymbol{v}^t S \boldsymbol{v}^* = \lambda^* \boldsymbol{v}^t \boldsymbol{v}^* = \lambda^*.$$

But $\boldsymbol{v}^t S \boldsymbol{v}^* = (\boldsymbol{v}^t S \boldsymbol{v}^*)^t = \boldsymbol{v}^+ S \boldsymbol{v}^t$, meaning $\lambda = \lambda^*$.

As a consequence, we can write $S$ as: $S = Q \Lambda Q^t$, where $Q$ is an $n \times n$ matrix with columns the eigenvectors of $S$, $Q = [\boldsymbol{q_1}, ..., \boldsymbol{q_n}]$, and $\Lambda$ is a diagonal matrix with diagonal elements $\Lambda_{i,i} = \lambda_i$.

An alternative way of expressing this is by the formula: $S = \sum_i \lambda_i \boldsymbol{q_i} \boldsymbol{q_i}^t$. Recall that the terms $\boldsymbol{q_i} \boldsymbol{q_i}^t$ are *rank* 1 matrices.

If $rank(S) = n$, the eigenvalues $\lambda_i$ must be different from 0. In this case $S$ has an inverse, which is also a symmetric matrix, and: $S^{-1} = Q \Lambda^{-1} Q^t = \sum_i \frac{1}{\lambda_i} \boldsymbol{q_i} \boldsymbol{q_i}^t$.

When all eigenvalues $S$ are positive, we say that the symmetric matrix is *definite positive*. In this case, $S^{-1}$ is also a symmetric definite positive matrix. These matrices somehow play the role of positive numbers in the matrix world.

The variance-covariance matrix $\Sigma = \mathbb{E}[(X_1, .., X_n)(X_1, ..., X_n)^t]$ of multivariate gaussian distributions is a definite positive matrix. The inverse of the covariance matrix, called *precision matrix* is also definite positive.

## 1.12 Geometric interpretation for symmetric matrices: Rotation/Reflection, Stretching, Rotation/Reflection$^{-1}$

We know that matrix multiplication can be interpreted as linear map composition, meaning that the geometric interpretation of $S\boldsymbol{v} = Q \Lambda Q^t \boldsymbol{v}$ for an arbitrary vector $\boldsymbol{v}$ is:

- Rotate the coordinate system to align it with the set of vectors $\boldsymbol{q_i}$ which form the columns of the matrix $Q$. The vector $\boldsymbol{v}$ in this new coordinate system has the expression $Q^t \boldsymbol{v}$ (note that the vectors $\boldsymbol{q_i}$ expressed in the new coordinate system have the expression $Q^t \boldsymbol{q_i} = [0, ..., 1, ...0]$ as we expect).
- Stretch each component of the resulting vector $Q^t \boldsymbol{v}$ according with the diagonal elements of the matrix $\Lambda$, obtaining the vector $\Lambda Q^t \boldsymbol{v}$. This stretching causes in general a change of direction of the vector, but if $\boldsymbol{v}$ is aligned with a vector $\boldsymbol{q_i}$, it does not change its direction.
- Apply the inverse rotation to the coordinate system. If we express the resulting $\Lambda Q^t \boldsymbol{v}$ in this new coordinate system we obtain $Q \Lambda Q^t \boldsymbol{v}$.

For instance a positive definite matrix $S$ will map an $n$-dimensional sphere of radius 1 to an $n$-dimensional ellipsoid with axis given by its eigenvectors $\boldsymbol{q_i}$, and axis lengths given by its eigenvalues $\boldsymbol{\lambda_i}$, (think on this).

# 2 Multivariate Gaussian distribution

Recall that we say that $\boldsymbol{X}$ follows a multivariate gaussian distribution of parameters $\boldsymbol{\mu}$ and $\Sigma$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\Sigma$ is a *positive definite matrix*, if the joint probability density function of $\boldsymbol{X}$ is of the form:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

## 2.1 The quadratic form $\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$

If $S$ is a definite positive matrix, then the graph of the quadratic form

$$z = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

is a $n+1$ dimensional paraboloid that takes always values of $z$ which are non-negative and the only point at which $z = 0$ is $\boldsymbol{\mu}$.

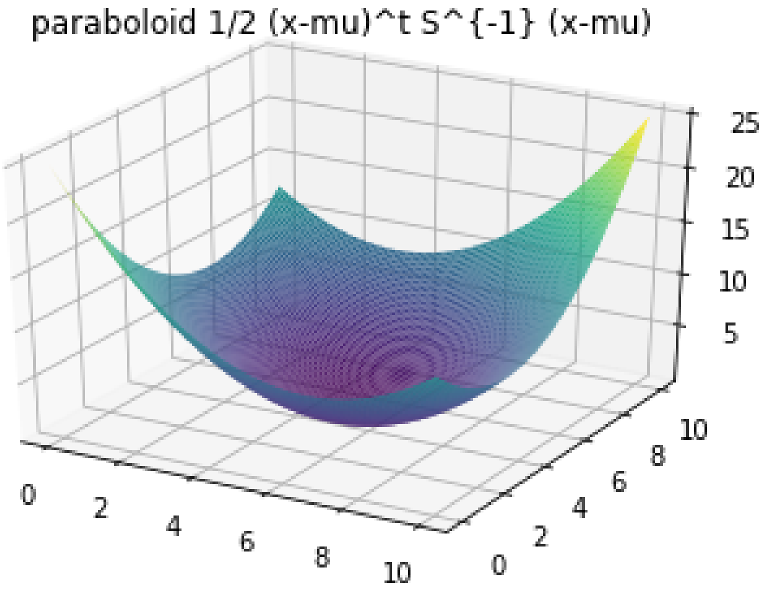
For instance, let $\Sigma = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$. To diagonalize $\Sigma$, we find the eigenvalues and orthonormal eigenvectors:

$$\Sigma = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \tag{4}$$

Meaning that:

$$\Sigma^{-1} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}. \tag{5}$$

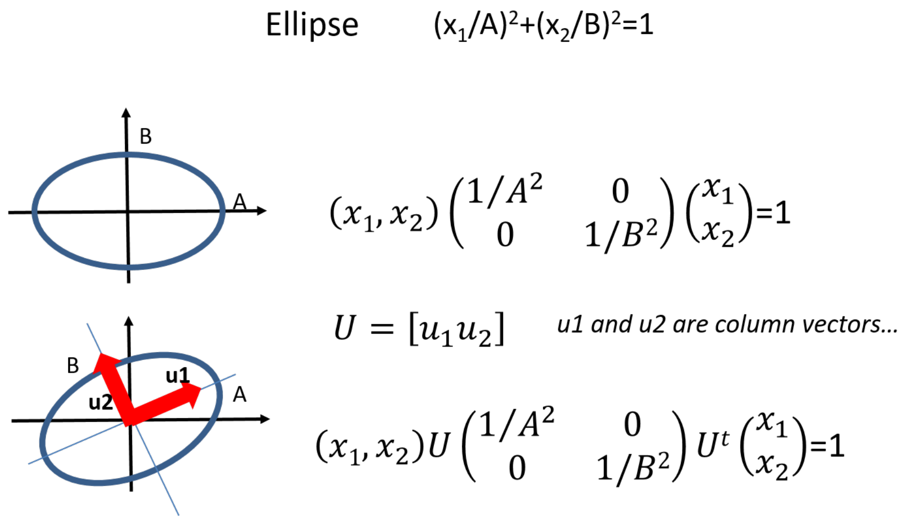
Here we plot the 3-d parabole $\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ for $\boldsymbol{\mu} = [5,5]^t$:



## 2.2 Isocontour lines $(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2$

The equation of an ellipsoid centered at the origin and of semi-axis given by $A_i$ oriented according with the orthonormal vectors $\boldsymbol{q_i}$ is

$$\boldsymbol{x}^t Q \begin{bmatrix} \frac{1}{A_1^2} & 0 & 0 & ... \\ 0 & \frac{1}{A_2^2} & 0 & ... \\ 0 & 0 & \frac{1}{A_3^2} & ... \\ ... & ... & ... & ... \end{bmatrix} Q^t \boldsymbol{x} = 1$$



Ellipse $(x_1/A)^2+(x_2/B)^2=1$

$$(x_1,x_2)\begin{pmatrix}1/A^2 & 0\\ 0 & 1/B^2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=1$$

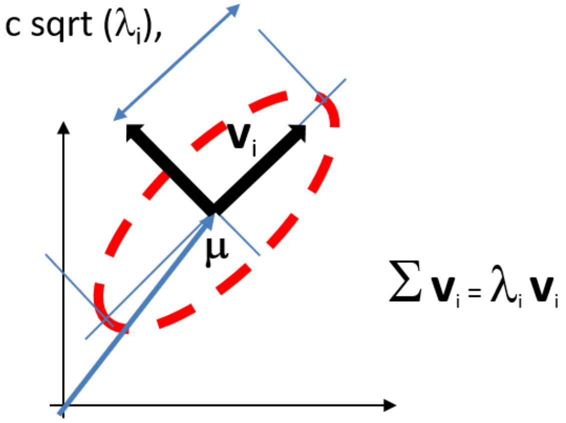
$U=[u_1 u_2]$ *u1 and u2 are column vectors…*

$$(x_1,x_2)U\begin{pmatrix}1/A^2 & 0\\ 0 & 1/B^2\end{pmatrix}U^t\begin{pmatrix}x_1\\ x_2\end{pmatrix}=1$$

If we plot the geometrical locus of the points that fulfill the equation:

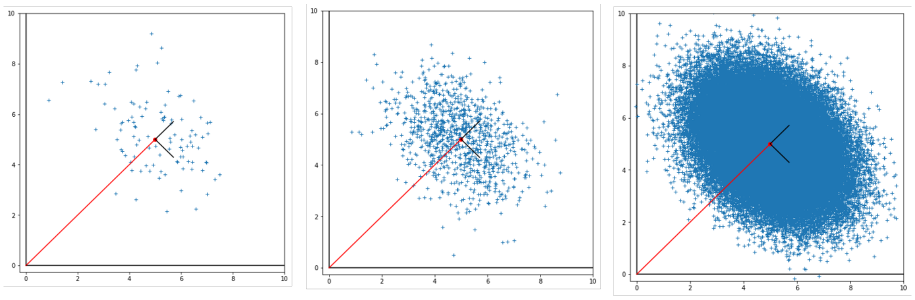
$$(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=c^2$$

we would obtain an $n$ dimensional ellipsoid, centered at the point $\boldsymbol{\mu}$, with axis aligned with the (orthonormal) eigenvectors $\boldsymbol{q_i}$ of the matrix $\Sigma$, and with semi-axis length in the axis pointed by $\boldsymbol{q_i}$ equal to $c\sqrt{\lambda_i}$, where $\lambda_i$ is the eigenvalue associated with $\boldsymbol{q_i}$:



## 2.3 Datasets generated from independent sampling of a multivariate gaussian distribution

If perform a number of independent sampling of a multivariate gaussian distribution, we will obtain clouds of points following the previously described ellipsoids:



N=100 points N=1000 points N=100000 points

## 2.4 The term $|\Sigma|^{1/2}$

The determinant of a matrix is the product of its eigenvalues, meaning that $|\Sigma|^{1/2} = (\prod_i \lambda_i)^{1/2}$.

# 3 The trace operator

For a square matrix $A$ we define the trace as:

$$trace(A) = \sum_k a_{k,k}$$

## 3.1 The trace of a matrix is the sum of its eigenvalues counting multiplicities

As we know, the eigenvalues $\lambda_k$ are the solutions of the equation:

$$det(A - \lambda I) = (-1)^n \lambda^n + (-1)^{n-1} trace(A) \lambda^{n-1} + ... = 0.$$

The eigenvalues $\lambda_k$ are the roots of the polynomial in $\lambda$, meaning that we have:

$$(-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2)... = 0.$$

and from this we obtain:

$$\sum_k \lambda_k = trace(A)$$

## 3.2 The trace operator is cyclic

Let $A$ and $B$ are in general non square matrix of sizes $n \times m$ and $m \times n$. The diagonal elements of $P = AB$ can be found as:

$$p_{k,k} = \sum_{i=1,\ldots,m} a_{k,i} b_{i,k}.$$

while that for $Q = BA$ we have:

$$q_{i,i} = \sum_{k=1,\ldots,n} a_{i,k} b_{k,i}.$$

We have then

$$trace(AB) = \sum_{k=1,..,n} \sum_{i=1,\ldots,m} a_{k,i} b_{i,k} = \sum_{i=1,..,m} \sum_{k=1,\ldots,n} a_{k,i} b_{i,k} = trace(BA).$$

If we have now three arbitrary matrices of the rigth sizes to produce an square matrix in its product $ABC$ we have:

$$trace(ABC) = trace((AB)C) = trace(C(AB)) = trace(CAB).$$

### 3.3 Expressing $\frac{1}{2n}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x}$ as the trace of the product of two matrices

Assume that $\boldsymbol{x}$ is a vector with 0 mean (otherwise, we would use $\boldsymbol{x}-\boldsymbol{m}$ instead). Scalars are special cases of square matrices, meaning that

$$\frac{1}{2n}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x} = trace(\frac{1}{2n}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x}) = trace(\frac{1}{2n}\boldsymbol{x}\boldsymbol{x}^t\Sigma^{-1}) = \frac{1}{2}trace(S_n\Sigma^{-1}),$$

where $S_n = \frac{1}{n}\boldsymbol{x}\boldsymbol{x}^t$ is the sample covariance-variance matrix.

### 3.4 Derivative of $log(|\Sigma|)$ and $tr(S_n\Sigma^{-1})$

## 4 The Singular Value Decomposition (SVD)

The diagonalization of symmetric matrix is an extremely important result, that tells us that a symmetric matrix has as a ”core” a diagonal matrix with real diagonal elements. A similar result can be generalized for some other square matrices, but it cannot be applied to general (possible non-square) matrices.

There is however another factorization that can be applied to *general* matrices, even for non-square matrices which is known as Singular Value Decomposition (SVD).

As we have seen if $S$ is a symmetric $n \times n$ matrix, we can find a set of orthonormal vectors $\boldsymbol{u}_k$, which are left and rigth eigenvectors of $S$ associated with the real eigenvalues $\lambda_k$, meaning that they fulfill the equations:

$$S\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k, \quad k \in \{1, .., n\}.$$

The SVD generalizes these equalities for a general $n \times m$ (i.e. $n$ rows and $m$ columns) matrix $A$ of rank $r \leq min(n, m)$. The idea is to find a set of $r$ *orthonormal* vectors $\boldsymbol{v}_k$ of $R^m$, $r$ orthonormal vectors $\boldsymbol{u}_k^t$ of $R^m$, and *positive* real numbers $\sigma_k$ that fulfill the equations:

$$A\boldsymbol{v}_k = \sigma_k \boldsymbol{u}_k, \quad k \in \{1, .., n\}$$

Surprisingly, *we will see this is always possible.* The values $\sigma_k$ are called *singular values* of $A$, while $\boldsymbol{v}_k$ and $\boldsymbol{u}_k^t$ are the right and left *singular vectors* of $A$. We will always assume $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0$

The previous expression can be written using matrix notation as:

$$A\ [\boldsymbol{v}_1, ..., \boldsymbol{v}_r] = [\boldsymbol{u}_1, ..., \boldsymbol{u}_r]\ diag(\sigma_1, ..., \sigma_r).$$

As $\sigma_k \geq 0$, we see that if such two collections of vectors exists, we must have that vectors $\boldsymbol{u}_k$ are an orthonormal basis of $Col(A)$, while the vectors $\boldsymbol{v}_k$ must be an orthonoomal basis of $Col(A^t)$.

Assume now that we complete the previous collections of orthogonal vectors, with an orthonormal basis of $Ker(A)$, $\boldsymbol{v}_{r+1}, ..., \boldsymbol{v}_n$, and an orthonormal basis of $Ker(A^t)$, $\boldsymbol{u}_{r+1}, ..., \boldsymbol{u}_m$. We also define the matrix $\Sigma$ (do not confuse with the variance-covariance matrix of multidimensional gaussian distributions) as:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & ... & 0 & ... \\ 0 & \sigma_2 & 0 & ... & 0 & ... \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_r & 0 & ... \\ 0 & 0 & ... & ... & 0 & ... \\ ... & ... & ... & ... & & \end{bmatrix}$$

while $V = [\boldsymbol{v}_1, ..., \boldsymbol{v}_n]$ and $U = [\boldsymbol{u}_1, ..., \boldsymbol{u}_m]$.

With these definitions, we obtain:

$$AV = U\Sigma.$$

Multiplying on the right by $V^t$, we obtain the final expression for the SVD:

$$A = U\Sigma V^t.$$

At this point, however, we do not know yet whether we can find suitable vectors $\boldsymbol{v}_k$, $\boldsymbol{u}_k$ and positive real values $\sigma_k$ that fulfill the initial equations. Now we can prove that these values exists.

$A^tA$ is a symmetric rank $r$ semi-definite positive matrix (prove it). Using the previous expression for the SVD we obtain

$$A^tA = V\Sigma U^t U \Sigma V^t = V\Sigma^2 V^t.$$

This corresponds to diagonalization expression of a symmetric matrix, meaning that the vectors $\boldsymbol{v}_k$ are the right eigenvectors of $A^tA$ associated with the eigenvalues $\lambda_k$. $r$ of these eigenvalues must be positive (as the rank of this matrix is $r$ and the matrix is semi-definite positive), which implies that the singular values are $\sigma_k^2 = \lambda_k$, $k \in \{1...r\}$.

If the eigenvalues $\lambda_k$ are non repeated, the set of orthonormal vectors $\boldsymbol{v}_k$ is unique (up to the sign). For repeated eigenvalues, we must find and orthonormal basis for the corresponding eigenspace.

We must check now that the vectors $\boldsymbol{u}_k$ defined as $A\boldsymbol{v}_k = \sigma_k \boldsymbol{u}_k$ are also orthonormal:

$$\boldsymbol{u}_i^t \boldsymbol{u}_j = \frac{1}{\sigma_i \sigma_j} \boldsymbol{v}_i^t A^t A \boldsymbol{v}_j = \frac{\sigma_j^2}{\sigma_i \sigma_j} \boldsymbol{v}_i^t \boldsymbol{v}_j = 0, \quad i \neq j.$$

Analogously, if we multiply $AA^t$ we obtain:

$$AA^t = U\Sigma\Sigma U^t = U\Sigma^2 U^t,$$

meaning that the vectors $\boldsymbol{u}_k$ are eigenvectors associated to the positive real eigenvalues $\lambda_k$ of the semi definite positive matrix $AA^t$. In case of repeated eigenvalues, we must pick the vectors $\boldsymbol{u}_k = \frac{1}{\sigma_k} A\boldsymbol{v}_k$ as the orthornomal basis of the eigenspace.

I other words, *we can always find the suitable singular vectors and singular values for any matrix $A$.*

## 4.1 Economy SVD

The expression $A = U\Sigma V^t$ can be written as:

$$A = U\Sigma V^t = [\sigma_1 \boldsymbol{u}_1, ..., \sigma_r \boldsymbol{u}_r, \boldsymbol{0}, ..., \boldsymbol{0}][\boldsymbol{v}_1, ..., \boldsymbol{v}_n]^t =$$

$$\sum_{k=1,..r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t + \sum_{k=r+1,..n} \boldsymbol{0}\ \boldsymbol{v}_k^t = \sum_{k=1,..r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t.$$

Defining $\Sigma_r = diag(\sigma_1, ..., \sigma_r)$, $U_r = [\boldsymbol{u}_1, ..., \boldsymbol{u}_r]$, and $V_r = [\boldsymbol{v}_1, ..., \boldsymbol{v}_r]$, we obtain:

$$A = U_r \Sigma_r V_r^t$$

which is known as the *economy SVD.*

### 4.2 The SVD as the sum of $r$ matrices of rank $1$

Note that the previous expression tells us that the matrix $A$ can be written as the sum of $r$ matrices of rank 1:

$$A = \sum_{k=1,..r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t.$$

This will be a key fact to find the best low rank approximation of a matrix by using the SVD.

## 5 Geometric interpretation: Rotation/Reflection in $R^n$, Stretching, Rotation/Reflection in $R^m$

### 5.1 General case

### 5.2 Visualizing the SVD when $n = 2, m = 1$ or $n = 1, m = 2$