Elements of Computational Biology

Saul Pierotti

October 31, 2019

Introduction and topics

- Linear algebra is one of the main topics for developping algorithms
- Biology ha an high level of noise in its data
- I should know the formula of binomial and normal distribution
- We will study tests like T-Student, ANOVA
- Slides: ww.biocomp.unibo.it/gigi/2019-2020/ECB

Linear algebra

Vectors

- Given a reference system, a vector is represented by its components on the axes
- The span of n vectors is the set of all possible vectors that you can represent by their linear combinations
- If a vector can be expressed as a linear combination of another, it is said to be linear dependent from it
- The basis of a vector space is a set of linearly independent vectors that span the full space
- $\vec{x} \in R^n$ means x is a real vector in an n-dimensional space
- $\vec{x} \in C^n$ means x is a complex vector in an n-dimensional space
- Sum of vectors is done by summing their components or graphically with the parallelogram rule
  - $\vec{c} = \vec{a} + \vec{b} \implies c_i = a_i + b_i \ \forall \ i = 1 \rightarrow n$
  - Difference is the same concept
  - $\forall \ \vec{v} \ \exists \ \vec{0} \ : \ \vec{v} + \vec{0} = \vec{v}$
  - You can only sum vectors in the same vector space
- The norm of a vector $||\vec{v}||$ is its lenght
  - Can be computed with the pitagorean theorem $||\vec{v}|| = \sqrt{\sum_{i=1}^{n} v_i^2}$
  - The norm of the sum is less or equal to the sum of the norm of the components
    - This follows from the geometry of a triangle
  - The scalar product of a norm is the norm of the scalar product
    - $\lambda||\vec{v}|| = ||\lambda\vec{v}||$
- The distance between points in space is the norm of the difference between the vectors defining the points
  - $d(a, b) = ||\vec{a} - \vec{b}||$
- Scalar multiplication
  - $\vec{c} = \lambda\vec{a} \implies c_i = a_i\lambda$
  - A scalar multiplication of a sum is the sum of the scalar multiplications of the components
- Dot product, also called scalar or inner product
  - You can use the notation $< A, B >$
  - It is used in physics to calculate work
  - $\vec{w} = ||\vec{F}|| * ||\vec{s}|| * cos\theta = \sum_{i=1}^{n} F_i * s_i$
  - It is a number, complex or real depending on the vectors (!)
  - It is commutative and distributive
  - $< x, x >= ||\vec{x}||^2$

  - It is positive when the angle is acute
  - No cancelation rule
    - $<A,B>=<A,C> \not\Longrightarrow \vec{B}=\vec{C}$
- Angle between vectors
  - Can be calculated inverting the dot product
- A line passing through the origin can be defined as the set of points orthogonal to a vector $\vec{w}$
  - $w_1x_1 + w_2x_2 = 0$
  - In higher dimensions this describes an hyperplane (an n-1 dimensional object)
- All the point on an hyperplane have the same projection on its defining vector $\vec{w}$
  - The projection p of $\vec{x}$ on $\vec{w}$ is calculated as $\vec{x} * cos\theta$
  - An hyperplane is therefore an object subjected to the constraint $\vec{x} * cos\theta = p$
  - Given that $<\vec{x},\vec{w}>=||\vec{x}|| * ||\vec{w}|| * cos\theta$ we have that $p = \frac{<\vec{x},\vec{w}>}{||\vec{w}||}$
  - If p>0 the hyperplane is in the direction of $\vec{w}$, if it is negative it is in the opposite direction
  - Defining $b = -\frac{p}{||\vec{w}||}$ we have the canonical equation for the hyperplane
    - $w_1x_1 + w_2x_2 + b = 0$ in 2 dimensions
    - $<\vec{w},\vec{x}> + b = W^tX + b = 0$ in n dimensions
  - An hyperplane is useful for subdividing space
- 2 hyperplanes are parallel if their are defined by the same vector $\vec{w}$ allowing for a scaling factor λ
  - $<Y,W>=\lambda <X,W>$
- The distance between parallel hyperplanes is computed as the difference of their projections on $\vec{w}$
  - $d(X,Y) = p_y - p_x = \frac{b_x - b_y}{||W||}$
- The distance of a point A from an hyperplane is the projection of the point on the defining vector $\vec{w}$, minus the projection of the hyperplane on the same vector
  - $D(A,X) = p_a - p_x = \frac{<A,W>+b}{||W||}$
- Hyperplanes are useful for the separation of classes of data
- Every column of a matrix can be thought of as a vector
  - To make the dot product of 2 vectors using matrices you can multiply one vector for the transpose of the second
  - $<\vec{a},\vec{b}>=A * B^t$

Matrices

- A matrix is an array of numbers arranged in a rectangular structure
- The columns of a matrix are the coordinates where the basis vectors land after the transformation
- It has m rows and n columns, it is represented as $A \in R^{m*n}$
  - $A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{,m2} & ... & a_{mn} \end{pmatrix}$
  - The single a_{ij} numbers are called elements
  - The index of an element is always mn, meaning first row and then column
- If n=1, the matrix is called column matrix, which is a vector
- If m=1, the matrix is a row matrix
- A and B are equal if they have the same dimensions and they are equal element by element
  - $A = B \iff a_{ij} = b_{ij}$
- The 0 matrix contains all 0 elements and does not change the matrix it is added to
- The sum is defined as the sum of the respective elements
  - We can sum only matrices of the same dimensions, they are said to be conformable for addition
  - $C = B + A \iff c_{ij} = b_{ij} + a_{ij}$
  - The difference operates in the same way
- Scalar multiplication is performed multiplying all the elements of the matrix for the scalar
  - $C = \lambda A$ implies $c_{ij} = \lambda a_{ij}$
- The negative of A is -A, defined as $-1 * A$
  - $A - A = 0$

- Matrix addition and scalar multiplication are commutative, associative and distributive
- Matrix product is an operation that is defined only if the number of rows of the first matrix is equal to the number of columns of the second (the matrices are conformable for the product)
  - A is of dimensions $m * p$ and B of dimensions $p * n$, if $C = A * B$
  - $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$
  - $C = A * B$ can be computed as row by column product
  - It can be defined only if the number of columns in the first matrix is equal to the number of rows of the second
    - $R^{m*p} * R^{p*n} \implies R^{m*n}$
  - The result is a matrix with the same number of rows as the first, and the same number of columns as the second
  - The product between matrices is NOT commutative (!)
  - $A(B + C) = AB + AC$
  - $(A + B)C = AC + BC$
  - $A(BC) = (AB)C$
  - Be aware!
    - If $AB = 0$ we can NOT conclude that B or C are 0
    - If $AB = AC$ we can NOT conclude that $B = C$
- A square matrix has m=n
- An upper triangular matrix has all the elements below the diagonal equal to 0, and a lower triangular the ones above it
- A diagonal matrix has all the elements outside the diagonal equal to 0
- A diagonal matrix with all 1 elements is the identity matrix I
  - It does not change the square matrix it is multiplied to
  - In this case, $AI = IA = A$
- If AB=BA, A and B are said to commute
  - If A is a square matrix, it commutes with itself and with I
- If AB=-BA, A and B are said to anti-commute
- The transposition of a $n * m$ matrix is a $m * n$ matrix, called A^t, where $[A^t]_{ij} = A_{ji}$
  - A and A^t are always conformable to product, in both directions
  - $(A^t)^t = A$
- A square matrix is symmetric if $A = A^t$, antisymmetric (skew-symmetric) if $A = -A^t$
  - $A + A^t$ is always symmetric
  - $A - A^t$ is always antisymmetric
  - An antisymmetric matrix has a 0 diagonal and antisymmetrical elements otherwise
- The inverse of a matrix A, called A^{-1}, is a matrix such that $A * A^{-1} = A^{-1} * A = I$
  - It is defined only if $det(A) \neq 0$
- An orthogonal matrix has its inverse equal to the transpose, $A^{-1} = A^t$
  - An orthogonal matrix describes a spatial rotation
  - Therefore, $AA^t = A^t A = I$
    - You can check for orthogonality by checking that $A * A^t = I$
- Some properties of transpose and inverse matrices
  - $(AB)^{-1} = B^{-1} * A^{-1}$, but only if $(AB)^{-1}$ exists(!)
  - $(AB)^t = B^t * A^t$
- It is possible to associate a number called determinant to any square matrix
  - $det(A) = |A| \in R$
  - For an order 2 square matrix, that is compute subtracting the product of the second diagonal to that of the first
    - $det(A) = a_{11} * a_{22} - a_{12} * a_{21}$
  - It represents the area of the unit square after the transformation
  - Its sign reflects the orientation of space
    - If it is negative, the transformation flips the axis
- The rank of a transformation is the dimensionality of its output space, called column space
  - The column space of a transformation is the span of the basis vectors defined by its columns

- Some proprieties of determinants
  - If an entire row or column is equal to 0, then the determinant of the matrix is 0
  - $det(A * B) = det(A) * det(B)$
  - The determinant of an orthogonal matrix is either 1 or -1
  - $det(A) = det(A^t)$
- How to compute the inverse of 2*2 matrices
  - Given the definition $A * A^{-1} = I$ if $det(A) \neq 0$
  - It follows $A^{-1} = \frac{1}{det(A)} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$
- A minor M_{ij} of a matrix A is the determinant of any square submatrix of A
- The cofactor of the element a_{ij} is $C_{ij} : C_{ij} = M_{ij} * (-1)^{i+j}$
- To compute the determinant of any matrix you pick any row or column and sum the product of any element in it for its cofactor
  - In a column $det(A) = \sum_{i=1}^{n} a_{ij} * C_{ij}$
  - In a row $det(A) = \sum_{j=1}^{n} a_{ij} * C_{ij}$
  - It is convenient to choose the row or column with most 0 for the computation
  - If 2 rows are identical, det(A)=0
  - If one row is 0, then det(A)=0
  - If you exchange 2 rows, det(A')=-det(A)
  - The determinant of a triangular matrix is the product of the diagonal elements
  - If B is obtained by multiplying every element in a row of A by λ, $det(B) = \lambda det(A)$
  - For any n*n square matrix, $det(\lambda A) = \lambda^n det(A)$
  - If A and B are of the same order, $det(AB) = det(A)det(B)$
- The cofactor matrix of A, called A^c, is a matrix with each element equal to the cofactor of the same element in a
  - $A^c : a^c_{ij} = C_{ij}$
- The adjugate matrix of A is the transpose of its cofactor matrix
  - $A^a = (A^c)^t$
- The inverse matrix can be obtained by dividing the adjugate of a matrix for its determinant
  - $A^{-1} = \frac{1}{|A|} A^a$
- Matrices can represent systems of linear equations
  - The system $\begin{cases} x + y = 7 \\ 3x - y = 5 \end{cases}$ can be represented as $\begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 7 \\ 5 \end{pmatrix}$
  - The system has a solution if the coefficient matrix is invertible

Linear transformations

- A matrix can be thought of as a linear transformation of a vector space
  - $A^{m*n} : R^n \to R^m$
  - A linear transformation is a transformation that preserves linearity and does not move the origin
    - $A(\vec{v} + \vec{u}) = A\vec{v} + A\vec{u}$ and $A(\lambda \vec{v}) = \lambda A \vec{v}$
- A rotation by an angle θ can be describe by the transformation
  - $A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$
- The inverse trasformation takes a transformed vector and restores the original one
  - Sometimes it does not exist (!), when det(A)=0
- If det(A)=0 the transformation squishes space to a lower dimensional vector space
- The composition of the transformations A followed by B is $C = BA \neq AB$
- The scalar product of the transformation of a vector $A\vec{v}$ and the vector $\vec{w}$ is equal to the scalar product of the firts vector with the second vector transformed by the transpose of A
  - $A\vec{v} * \vec{w} = \vec{v} * A^t \vec{w}$
- The null space of a transformation is the set of vectros that get squished to $\vec{0}$ by the trasformation

- $\vec{b} \in Null(A) \iff A\vec{b} = \vec{0}$
- A trivial null space is always $\vec{0}$ itself
- There is a true null space only if $det(A) = 0$
- If $det(A) \neq 0$, the only null space is $\vec{0}$ itself

- The null space of a square matrix can be computed setting up a linear system of equations
  - For a matrix $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$, $det(A) = 0$
  - $A\vec{b} = \vec{0} \implies \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \begin{cases} b_1 + 2b_2 = 0 \\ 2b_1 + 4b_2 = 0 \end{cases} \implies b_1 = -2b_2 \implies \vec{b} = \lambda \begin{pmatrix} -2 \\ 1 \end{pmatrix}$

Eigenstuff

- For the transformation A, if $A\vec{b} = \lambda\vec{b}$, $\vec{b}$ is an eigenvector of A and λ is its eigenvalue
  - An eigenvector of a transformation is a vector that is only rescaled by the transformation
  - An eigenvalue is the scaling factor to which the vector is subjected by the transformation
  - The $\vec{0}$ vector can never be an eigenvector even though $A\vec{0} = \lambda\vec{0}$ is always true
    - On the contrary, it is possible that $\lambda = 0$
- How to find eigenvalues for the matrix A
  - $A\vec{b} = \lambda\vec{b} \implies A\vec{b} - \lambda\vec{b} = 0 \implies (A - \lambda I)\vec{b} = 0$
  - This means that the eigenvectors $\vec{b}$ are the non-trivial null space of the transformation $(A - \lambda I)$
  - It is required that $det(A - \lambda I) = 0$, otherwise there are no eignevectors
  - The equation $det(A - \lambda I) = 0$ is called characteristic equation of A and its root allows to recover the eigenvalues of the matrix
- How to find the eigenvectors for the matrix A
  - Once I have the eigenvalues λ, the eigenvectors can be found by solving $(A - \lambda I)\vec{b} = 0$ for $\vec{b}$, using all the λ
- The number of eigenvalues and eigenvectors (families of linearly dependent eigenvectors) is equal to the dimensions of the vector space
  - I always have n families of eigenvectors in the vector space R^n
  - The families of eigenvectors are orthogonal to each other iff the transformation is symmetric
  - They define a convenient reference frame, even if they are not orthogonal
  - Each eigenvalue scales one of the families of eigenvectors, meaning that it stretches one of the dimensions of the new reference frame
- Note that in a trinagular or diagonal matrix the diagonal elements are its eigenvalues
- The product of the eigenvalues is equal to the determinant of the matrix
  - A non invertible matrix (also called singular matrix) has always at least a 0 eigenvalue
  - The inverse matrix has reciprocal eigenvalues ($\frac{1}{\lambda}$)
  - The eigenvalue of kA is $k\lambda$
  - The eigenvalue of A^n is λ^n
  - Transposition does not change the eigenvalues
- The sum of the eigenvalues is the trace of the matrix
  - The trace of a matrix is the sum of its diagonal elements

Complex field

- Sometimes there can be no real eigenvalues, but there are always solutions in the complex field
  - This happens when the characteristic equation of the matrix has $\Delta < 0$
- A complex number z is written as $z = a + ib$ where $i = \sqrt{-1}$
  - a is called real part
  - b is called imaginary part
- The reference axis of a vector space are NOT necessarily orthogonal to each other (!)
  - But they must be linearly independent

Change of basis and diagonalization

- To go from one system to the other we need the representation of the new basis vectors in term of the previous ones
    - $\hat{i'} = a\hat{i} + b\hat{j}$ and $\hat{j'} = a\hat{i} + b\hat{j}$
    - If $U = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ we have that $\vec{v'} = U\vec{v}$
    - It is possible to go back to the original system of reference using U^{-1}
- If a $n * n$ matrix A has n eigenvectors which are linarly independent, I can write the $n * n$ matrix U containing all the eigenvectors, and use it to convert to a new system of reference
    - The eigenvectors of A will be the new basis vectors
- I can compute A in the new reference frame forming $\Lambda = U^{-1}AU$
    - Given a vector $\vec{v}$, I first convert it to the new reference frame where the eigenvectors are the basis vectors using U, then I apply A and finally I go back to the old system of reference using U^{-1}
    - This new matrix Λ will be diagonal (!)
    - Each column will be made of one eignevector multiplied by its eigenvalue
    - It is good to choose normalized vectors for the change of basis, meaning that their norm should be 1
        - In this case $det(U) = 1$, meaning that areas are preserved by the trasformation
- Why do I want to use eigenvectors as reference frames?
    - Because the components of any vector are only rescaled by the original transformation A in this reference frame
    - This makes much easier to compute transformations
- If a matrix is symmetric ($A = A^t$) its eigenvalues are real and its eigenvectors are orthogonal
    - If the eigenvectors are normalized $U^{-1} = U^t$, therefore $\Lambda = U^tAU$
- In the same way that a linear form can be represented as all the points orthogonal to a vector with a projection p onto it, a matrix can describe a quadratic form

Quadratic forms

- A quadratic form is an equation in more than 1 variable were each term has a variable squared or multiplied to another variable
    - An example is $ax^2 + bxy + cy^2 = 0$
- This is represented as $\vec{x}^tA\vec{x} + b = 0$, where $\vec{x}$ is the vector containing the variables, and A is a matrix of coefficients
    - The vector is multiplied 2 times to reflec the fact that the expression is quadratic
    - The second time the transpose is used in order to allow the product
- By rescaling A, we can obtain the standard form $\vec{x}^tA\vec{x} = 1$
- The matrix that describes a quadratic form is always symmetric
    - If it is not singular (non-invertible), it can be diagonalised as $\Lambda = U^tAU$
    - If $\vec{x'} = U^t\vec{x}$, the quadratic form becomes $\vec{x'}^t\Lambda\vec{x'}$, defined canonical form
- In the canonical form, a $2 * 2$ Λ contains the eigenvalues of the transformation in the diagonal
    - If they are both positive, the quadratic is an ellipse
        - If they are equal, it is a circle
    - If they are of opposite sign, it is an hyperbole
    - If they are both negative, there is no real solution
- In 3d, I can get an ellipsoid, a hyperboloid of 1 sheet or an hyperboloid of 2 sheets

Calculus

Functions

- Calculus is the study of functions
    - Functions are univocal relations between the sets domain and codomain

- The function $f(x) = mx + q$ is a line passing through q at $x = 0$ with slope m
- The inverse of a function correlates $f(x)$ to x
  - It is the reflecion of $f(x)$ on the line $g(x) = x$
- The function $f(x) = a^x$ is an exponential
  - It passes through 1 at $x = 0$
  - $\lim_{x\to-\infty} f(x) = 0$
  - $\lim_{x\to+\infty} f(x) = +\infty$
- The function $f(x) = \log_a(x)$ is a logarthmic function
  - It passes through 1 at $x = 0$
  - Common bases a are 10, 2 and e
  - $\lim_{x\to 0} f(x) = -\infty$
  - $\lim_{x\to+\infty} f(x) = +\infty$
  - Logarithms are useful for performing products
  - $\log_a(xy) = \log_a(x) + \log_a(y)$
  - $\log_a(x^y) = y \log_a(x)$
  - $\log_a(x) = \frac{log_b(x)}{\log_b(a)}$
- Trigonometric functions
  - The cosine is an even function beacuse $\cos(\theta) = \cos(-\theta)$
    - $\cos(\frac{\pi}{2}) = 0$
    - $\cos(0) = 1$
  - The sine is an odd function because $\sin(-\theta) = \sin(-\theta)$
  - Sine and cosine are periodical with a 2π period
  - The secant is the reciprocal of cosine
  - Trigonometric functions can be inverted only in a subdomain
  - They are continous functions
- A function is continuous at a point if the limit at that point is equal to the value of the function at that same point
  - The composition of continuous function is a cpntinuous function
- The intermediate value theorem: a continuous function between two points takes any possible value between them
- Discontinuities can be removed in some cases, but essential discontinuities such as oscillating points, jumps and infinites cannot be removed

Derivatives

- The slope of a line is defined as $\frac{\Delta y}{\Delta x}$
- Therefore, the slope of the secant of a function between two points $f(a)$ and $f(a+h)$ is $\frac{\Delta y}{\Delta x} = \frac{f(a+h)-f(a)}{h}$
- If we try to reduce h as much as possible we obtain the slope of the tangent at point a
  - $m = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$
  - The tangent at a point is an estimation of the rate of change of the function at that point
- The derivative of a function is another function that describes its rate of change, it takes the value of the slope of the tangent of the original function at each point
  - $f'(x)|_a = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$
- A function to be derivable must be continous and must have one-sided derivatives defined at the end-points
  - However, there are functions that are continuous but not derivable
  - Points of non-derivability are cusps, corners, discontinuities and points with vertical tangent
- Some derivatives
  - $\frac{d}{dx}[a] = 0$
  - $\frac{d}{dx}[ax] = a$
  - $\frac{d}{dx}[x^n] = nx^{n-1}$
  - $\frac{d}{dx}[\cos(x)] = -sin(x)$
  - $\frac{d}{dx}[\sin(x)] = cos(x)$

  - $\frac{d}{dx}[e^x] = e^x$
  - $\frac{d}{dx}[a^x] = \frac{d}{dx}[e^{\ln(a)x}] = \ln a * e^{\ln(a)x} = \ln(a) * a^x$
  - $\frac{d}{dx}[\ln(x)] = \frac{1}{x}$
  - $\frac{d}{dx}[\log_a(x)] = \frac{1}{x*\ln(a)}$
- Rules for derivation
  - $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$
  - $\frac{d}{dx}[k * f(x)] = kf'(x)$
  - $\frac{d}{dx}[g(x) * f(x)] = g(x) * f'(x) + g'(x) * f(x)$
  - $\frac{d}{dx}[g(f(x))] = \frac{dg}{df} * \frac{df}{dx}$
  - $\frac{d}{dx}[\frac{g(x)}{f(x)}] = \frac{g(x)*f'(x)+g'(x)*f(x)}{f(x)^2}$
- There are global and local maxima and minima, defined as extremes
  - There are NOT methods to compute global extrema, but only local ones
- A local extreme is referred to an open interval
  - The derivative at that point is 0
    - This is NOT sufficient, it can also be a flexus
  - A minimum has a derivative with positive slope when it intersect the x axis
    - In other words, the second derivative is positive
  - A maximun has a derivative with negative slope when it intersect the x axis
    - In other words, the second derivative is negative
- A critical point of a function is a point where the derivative is 0 or undefined

---

rewied until here

Taylor series

- A line passing in x_a with a slope equal to the derivative at that point approximates the function itself
  - $f(x) - f(x_a) \approx \frac{df}{dx}|_{x_a}(x - x_a)$
- We can approximate a function with a polinomial $P(x)$
- The coefficient a_0 for grade 0 can be found solving
  - $P(x_a) = f(x_a)$
- We can find the other terms by equating the second, third and fourth derivatives
  - $P'(x_a) = f'(x_a)$
  - $P''(x_a) = f''(x_a)$
  - $P'''(x_a) = f'''(x_a)$
  - $P''''(x_a) = f''''(x_a)$
- In general, we can give the Taylor series
  - $P(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3 + ...$

Integrals

- The area under a curve over a partition is called integral
  - This can be computed doing a Riemann sum, meaning that we can sum the area of rectangles under the curve
  - $F(x) = \int_a^b f(x)dx = \lim_{\Delta x \to 0} \sum_{k=1}^n (f(x) * \Delta x)$
- The areas computed by integrals have a sign (!)
- It is possible to solve an integral by substituting a polinomial with the variable u
  - du then becomes $\frac{d}{dx}u\,dx$

Multi-dimensional calculus

- A 2 dimensional function takes 2 inputs x,y and gives the output z

    - $z = f(x, y)$
    - They are usually represented with 3d surfaces or with level curves
- It is not possible to compute single derivatives of the function
- We can fix y and compute the derivative with respect to x in the resulting 2d curve, called partial derivative in x
    - $\frac{\partial f}{\partial x} = f_x(x, y) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x, y) - f(x,y)}{\Delta x}$
- It is possible to take second partial derivatives an mixed partial derivatives
- In mixed partial derivatives the order of derivation is not important (!)
- The set of first partial derivatives is called gradient of the function
    - $\nabla f(x_1, x_2, ...x_n) = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ...\frac{\partial f}{\partial x_n})$
- The square symmetric matrix of all second partial derivatives is called Hessian (H) of $f(x)$
- The minima and maxima of a multivariable function are those points where the gradient of the function is 0
- A minimum has the diagonal elements of the Hessian both positive, a maximum both negative
- If the diagonal of the Hessian has negative and positive elements we have saddle points
- When the Hessian is not diagonal, it can be diagonalized
- The maxima and minima of a multivariable function $f(x, y)$ subjected to the costraint $g(x, y) = 0$ are the points where $g(x, y)$ is tangent to the contour level of $f(x, y)$
    - We can also say that the perpendiculars to both functions are parallel
    - The perpendicular to the contour level is the gradient of the function
    - $\nabla f = \lambda \nabla g$ where λ is the Lagrange multiplier
- The Lagrangian function $L(\vec{(x)}, \lambda)$ is bad
- Information entropy

Probability

- Probability can be interpreted in several ways
    - In a frequency approach, it is the ratio between favourable events and total events
    - It can also be viewed as the confidence in an event happening
- The sample space is the set of all possible outcomes of an experiment
- An event is a subset of the sample space
- A probability is a number between 0 and 1
    - $0 \leq P(E_i) \leq 1$
- The sum of the probabilities of all the outcomes is equal to 1
    - $P(\cup E_i) = 1$
- The probability of on of 2 events happening is equal to the sum of the probabilities minus their intersection
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If two events have NO intersection they are said mutually exclusive
- The intersection of 2 events that are independent and not mutually exclusive is the product of their probabilities
    - $P(A \cap B) = P(A) * P(B)$
- The conditional probability of A given B is represented as $P(A|B)$
    - If the events are independent, $P(A|B) = P(A)$
    - If they are not independent, $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- The intersection of 2 non-independent event is therefore
    - $P(A \cap B) = P(A) * P(B|A) = P(B) * P(A|B)$
- 2 mutually exclusive events cannot be independent, and vice-versa
- The bayes formula: $P(A|B)$ is different from $P(B|A)$
    - $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$
- To count events considering order we use permutations
    - With replacement (same element can be drawn twice)
        - It is the number of events multiplied by itself for a number of times equal to the lenght of the

        sequence
    - k : sequence length, n : possible outcomes $\implies x = n^k$
  - Without replacement (same element cannot be drawn twice)
    - It is the same of the replacement case, but n becomes n-1 at each iteration
    - k : sequence length, n : possible outcomes $\implies x = \frac{n!}{(n-k)!}$
- To count events NOT considering order we use combinations
  - With replacement (same element can be drawn twice)
    - k : sequence length, n : possible outcomes $\implies x = \frac{(n+k-1)!}{n!(k-1)!} = \binom{n+k-1}{n}$
  - Without replacement (same element cannot be drawn twice)
    - It is similar to the permutation without replacement, but with an additional $1/k!$ term, which removes the possible permutations of a sequence of lenght k
    - k : sequence length, n : possible outcomes $\implies x = \frac{n!}{(n-k)!k!} = \binom{n}{k}$

Distributions

- Probabilities can be described with distributions
- The mode of a distribution is the value that occurs with the highest probability
  - It is the peak of the histogram
  - If there are 2 peaks, the distribution is calle bimodal
- The median is the value of the random variable for which $P(x > x_{med}) = P(x < x_{med}) = \frac{1}{2}$
  - The values of x that splits the distribution in quarters is called quartile, in fifths quintile and so on
- The mean or average is the expected value of the random variable x
  - It is calculated as $< x >= E[x] = \mu_x = \frac{1}{n}\sum_{i=1}^{n} f(x_i) * x_i$ where n is the number of elements, x_i is a value of the random variable and $f(x_i)$ is its probability
- The variance is the mean disstance of x from the mean of the distribution
  - $var(x) = \sigma_x^2 = E[(x - \mu_x)^2]$
  - To calculate it easily, we can consider that $var(x) = E[(x - \mu_x)^2] = E[x^2 + 2x\mu_x + \mu_x^2] = E[$
- The standard deviation has the same meaning of the variance, but is more usefull because it has the same dimensionality of the random variable x
  - $\sigma_x = \sqrt{var(x)}$
- The covariance of 2 variables is defined as $cov(x, y) = E[(x - \mu_x)(y - \mu_y)]$
  - It is an extimate of the correlation of the 2 variables
- The binomial distribution
  - The mean is the number of trials times the probability of the favourable event
    - $E[k] = np$
  - The variance is
    - $var(k) = np(1 - p)$
- Maximun likelyhood exstimation
- Poisson distribution
- Probability density
- The normal or gaussian distribution
  - $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- The mean of the normal distribution
  - $E[x] = \int_{+\infty}^{-\infty} x * p(x)dx = \mu$
- The variance of the normal distribution
  - $var(x) = \int_{+\infty}^{-\infty} (x - \mu)^2 * p(x)dx = \sigma^2$
- σ and $-\sigma$ are the flexus points of the curve
- The integral between σ and $-\sigma$ is $\approx 68\%$
- The integral between 2σ and -2σ is $\approx 95\%$
- The integral between 3σ and -3σ is $\approx 99.7\%$
- Any distribution can be normalized by introducing the normalized variable z such that $z = \frac{x-\mu}{\sigma}$

- The normal distribution cannot be integrated analitically, therefore the integral is computed with tables or software
- Proof of why I compute the variance with the degrees of freedom (n-1)
- χ^2 is the distribution of the σs observed from a random sample of a normal distribution
- The Student test compares two means, and is based on a variable k which represents the degrees of freedom
  - It is similar to the normal distribution, but it is more spread
  - It converges to the normal distribution for $k \to +\infty$
- The Gumbel or extreme value distribution is used to model the extreme values of a number of samples of various distributions
  - It was invented to model the distribution of extreme temperatures during the year
  - The pdf is the exponential of an exponential, it decreases very, very fast
    - $pdf = e^{-e^{(x-\mu)/\beta}}$
  - It is used by BLAST for giving the e-value
- The p-value is a number used to disprove an hypothesis
  - We don't have methods to prove an hypothesis
  - It is the probability of obtaining the experimental data if the null hypothesis H_0 holds
    - $p = P(observed|H_0)$
  - If the p-value is sufficiently low, I can reject H_0, if it is not I can NOT reject it, but I can never prove it
    - There are many hypothesis that can be alternative to H_0
  - The threshold used dipends on the significance that we desire
    - In biology it is common to use $p = 0.05$
    - In physics it is used $p = 10^{-10}$ or even lower
  - The critical value is the value for which the area under the curve from the value to the extreme of the distribution is equal to the desired p-value
  - The p-value is currently heavily criticised
- The significance of a test is the probability to reject H_0 if it actually holds
- The power of a test is its ability to reject false H_0
  - It is the rate of false negatives
- The Bayesian approach to data validation is an alternative to the p-value
  - $P(H_0|Observed) = \frac{P(Observed|H_0)*P(H_0)}{P(Observed)}$
  - I need the *a priori* probability of the observation, which usually is not available
- Different tests for different purposes
  - T-Student is used for analyzing difference of means
  - Fisher is used for testing the indipendency of different variables
  - ANOVA is also used for testing the difference of means between sets of experiments
    - In the 2 experiment case, ANOVA converges in the T-Student
  - χ^2 is used for testing variances or to test if 2 distributions are the same or not
- The Fisher test is done by computing a contingecy table
  - From the table I take the various coefficients and compute the probability of obtaining that table under H_0 (no association between the variables)
  - The probability is computed using the binomial
  - The output is a number between 0 and 1 which is the probability of obtaining that table
  - It does NOT give the p-value because it doesn't consider tables that are more extreme of the observed one
- When I do multiple testing on the same data, I can expect to reject $\frac{p}{n}$ null hypothesis wrongly, where n is the number of tests
  - The simplest correction that can be made to preserve the significance is the Bonferroni correction
    - The actual significance to be accepted is the original significance divided by the number of tests
    - Sometimes it is to strict, it fails to reject H_0 that are false
  - Another correction is the Benjamini-Hochberg
- The z test normalizes the observed mean of the sample and cheks the area under the standard normal

distribution outside the computed z
  - The most correct tests check the 2 tails of the distribution, not only one
  - To do this test we need the true σ of the population
- If we don't have the real σ but we use the σ of the sample, the mean follows the t distribution

31/10/19

- One-way ANOVA compares means of different samples (treatments in the ANOVA jargon)
  - The H_0 is that all the means are equal
  - From the different treatments we can compute the means for each treatment and the global mean of all data
  - We can compute the treatment variances and the global variance in the same way
  - The total variance can be partitioned in random variation and in variation between treatments
    - The sum of squares within (SSW) is computed as $\sum_{i,j}(x_{i,j} * \mu_j)^2$
    - The sum of squares between (SSB) is computed as $\sum_j n_j(\mu_j * \mu)^2$
    - It can be proven that the sum of squares total (SST) is equal to the sum of SSW and SSB
      - $SST = \sigma^2 * n_{tot} = \sum_i (x_i - \mu)^2 = SSW + SSB$
  - We can then define the variable f : $f = \frac{MSB}{MSW}$
    - $MSW = SSW/(n_{tot} - n_{treatments})$
    - $MSB = SSB/(n_{treatments} - 1)$
  - The variable f is distributed following the Fisher-Snedecor F distribution
    - There is one distribution for each combination of degrees of freedom within and between
  - The null hypothesis can be rejected when f is high enough
  - If $f \leq 1$ the variability between is equal or lower than the variability within
- Two-way ANOVA can be used to compare the response to different combination of 2 variables
  - It can be used in different treatments defined as different combinations of doses of 2 different drugs
  - It tests for effect of the single variables and for interaction between them
  - If we have the variable A and B, SST is composed of SSW, SSB(A), SSB(B), SSB(A,B)
  - In order to be powerfull, if we increase the number of treatments we need a lot of data
  - It is frequently used for the analisys of expression data
- Correlation can be measured by a coefficient ρ
  - Given the variables x and y, we can compute μ_x and μ_y
  - If we graph a point (μ_x, μ_y) we define the centroid of the data
  - The covariance is a good measure of linear dependency between x and y
    - $cov(x, y) = \frac{1}{n-1}\sum_i (x_i - \mu_x)(y_i - \mu_y)$
  - We can then define the Pearson's correlation coefficient $\rho = \frac{cov(x,y)}{\sigma_x, \sigma_y)}$
    - It is a number between -1 and 1 because it is rescaled by the standard deviations of x and y
    - The Pearson's coefficient tests only linear dependency (!)
      - If $\rho = 0$ we can NOT say that x and y are independent, they can have higher order dependencies
  - The significance of the coefficient is strongly dependent on the number of points
    - The value of the coefficient *per se* does not mean anything, it has always to be tested
    - It can be tested using the Student distribution
  - The Person's coefficient is very sensitive to outlyers
  - Always look at the graph before drawing conclusions, because it can be different fro mwhat you think
- CORRELATION DOES NOT IMPLY CAUSATION
- The Spearman's correlation coefficient measure monotonic dependency
  - It is the Pearson's coefficient of the rank of the variables
  - If we make a ranking of values from the lowes to the highest, the rank of a value is its position in the ranking
  - We have to test the significance also of this coefficient, that depends on the amount of data
- The Matthews correlation coeffcient (MCC) is used for categorical variables

  - We can assign the values 0 and 1 to the categories in both values
  - We can take the Pearson's coefficient of these values, which is the MCC
  - It is used in machine learning in a table real vs predicted
- When the dependency is more complex, we can use the mutual information
  - It will be a topic for next year
- If we have data with a good Pearson correlation, we can define a linear regression
- One technique is to minimize the distance between the points and the line (best fit with least squares)
  - We want to minimize the function $f(a,b) = \sum_i [y_i - (ax_i + b)]^2$
- The same technique can be applied for fitting any polinomial
  - $y = P(x) \sum_{k=0}^{p} a_k * x^k$
- If I use a high degree polinomial I risk to do overfitting
  - I have overfitting when the number of data is of the same order of the number of parameters
  - In overfitting the values of the parameters are often absurd (really high in absolute value), without any physical meaning
- The error can be estimated considering overfitting by giving a penalty for high coefficients
- To test the quality of a model we need to use data not used for building the model itself (!)
- Cross-validation is very dangerous
- Principal component analisys is used with high-dimensional data
  - It reduces the number of dimensions, so to be able to plot the results
  - A good idea to preliminarly decrease the number of variables is to remove the ones with lowest variance
  - If we want to find a better system of reference, we can choose the axes with 0 covariance
- I can build the covariance matrix of the variables
  - The covariance matrix is symmetric and therefore it can be diagonalized easily
  - The change of basis matrix U is orthogonal and therefore represents a rotation
  - $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$
  - If I rank the eigenvalues from highest to lowest, I can choose to use only a subset of dimensions that maximises the variation
- PCA can discriminate only linear dependencies (!) by rotating the frame of reference so to align it with the variation