BINF 4211:
Applied Data Mining for Bioinformatics

Mathematical Basis

Xiuxia Du, Ph.D.
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte

Outline

- **Linear algebra**
  - Vector space and subspace
  - Linear transformation
  - Dot product
  - Norm
  - Basis of a vector space
  - Eigenvalue and eigenvector
  - Matrix diagonalization
  - Orthogonal matrix
  - Symmetric matrix

Outline

- **Probability and statistics**
  - Expectation and variance
  - Covariance
  - Correlation
  - Covariance matrix

Linear Algebra

Vector space: definition

A *vector space* is a set V on which two operations $+$ and $\cdot$ are defined, called *vector addition* and *scalar multiplication*.

The operation $+$ must satisfy the following conditions:

Closure: If $\boldsymbol{u}$ and $\boldsymbol{v}$ are any vectors in V, then the sum $\boldsymbol{u} + \boldsymbol{v}$ belongs to V.

(1) *Commutative law*: For all vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in V, $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$

(2) *Associative law*: For all vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ in V, $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$

(3) *Additive identity*: The set V contains an *additive identity* element, denoted by $\mathbf{0}$, such that for any vector $\boldsymbol{v}$ in V, $\mathbf{0} + \boldsymbol{v} = \boldsymbol{v}$ and $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$

(4) *Additive inverses*: For each vector $\boldsymbol{v}$ in V, the equations $\boldsymbol{v} + \boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{x} + \boldsymbol{v} = \mathbf{0}$ have a solution $\boldsymbol{x}$ in V, called an *additive inverse* of $\boldsymbol{v}$, and denoted by $-\boldsymbol{v}$.

Vector space: definition

The operation $\cdot$ (scalar multiplication) is defined between real numbers (or scalars) and vectors, and must satisfy the following conditions:

Closure: If $\boldsymbol{v}$ is any vector in V and c is any real number, then the product $c \cdot \boldsymbol{v}$ belongs to V.

(5) *Distributive law*: For all real numbers c and all vectors $\boldsymbol{u}, \boldsymbol{v}$ in V, $c \cdot (\boldsymbol{u} + \boldsymbol{v}) = c \cdot \boldsymbol{u} + c \cdot \boldsymbol{v}$

(6) *Distributive law*: For all real numbers c, d and all vectors $\boldsymbol{v}$ in V, $(c + d) \cdot \boldsymbol{v} = c \cdot \boldsymbol{v} + d \cdot \boldsymbol{v}$

(7) *Associative law*: For all real numbers c, d and all vectors $\boldsymbol{v}$ in V, $c \cdot (d \cdot \boldsymbol{v}) = (cd) \cdot \boldsymbol{v}$

(8) *Unitary law*: For all vectors $\boldsymbol{v}$ in V, $1 \cdot \boldsymbol{v} = \boldsymbol{v}$

Subspace

- **Definition**

  Let V be a vector space, and let W be a subset of V. If W is a vector space with respect to the operations in V, then W is called a subspace of V.

- **Theorem**

  Let V be a vector space, with operations $+$ and $\cdot$, and let W be a subset of V. Then W is a subspace of V if and only if the following conditions hold.

  (1) W is nonempty: The zero vector belongs to W.

  (2) Closure under $+$: If $\boldsymbol{u}$ and $\boldsymbol{v}$ are any vectors in W, then $\boldsymbol{u} + \boldsymbol{v}$ is in W.

  (3) Closure under $\cdot$: If $\boldsymbol{v}$ is any vector in W, and c is any real number, then $c \cdot \boldsymbol{v}$ is in W.

Linear transformation

- **Definition**

  Let S and T be two vector spaces over the same field. A function $f : S \to T$ is said to be a *linear map* or *linear transformation* if for any two vectors x_1 and x_2 in S and any scalar, the following two conditions are satisfied:

$$f(x_1 + x_2) = f(x_1) + f(x_2)$$

$$f(ax_1) = af(x_1)$$

  This is equivalent to requiring that for any vectors $x_1, x_2, \cdots, x_m$ and scalars $a_1, a_2, \cdots, a_m$:

$$f(a_1x_1 + a_2x_2 + \cdots + a_mx_m) = a_1f(x_1) + a_2f(x_2) + \cdots + a_mf(x_m)$$

Linear transformation

- **Examples of linear transformations**

  - The map $x \longmapsto 2x$ is linear.
  - The map $x \longmapsto x^2$ is nonlinear.
  - The integral yields a linear map from the space of all real-valued integrable functions on some interval to $\mathbb{R}$.

$$\int_a^b (\alpha f(t) + \beta g(t))\, dt = \alpha \int_a^b f(t)dt + \beta \int_a^b g(t)dt$$

  - Differentiation is a linear map from the space of all differentiable functions to the space of all functions.

$$\frac{d}{dt}(\alpha f(t) + \beta g(t)) = \alpha f'(t) + \beta g'(t)$$

Linear transformation

- **Examples of linear transformation**

  - A matrix is special:

    1. An $m \times n$ matrix $\boldsymbol{A}$ defines a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ by sending the column vector $\boldsymbol{X} \in \mathbb{R}^n$ to the column vector $\boldsymbol{AX} \in \mathbb{R}^m$.
    2. In linear algebra, every linear transformation between finite-dimensional spaces can be expressed as a matrix.

Linear transformation

- **Special cases of linear transformations in $\mathbb{R}^2$:**

  1. rotation by θ degrees clockwise

$$\boldsymbol{A} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

  2. scaling by 2 in all directions

$$\boldsymbol{A} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

  3. projection onto the y axis

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Dot product

- **Definition**

  Given two vectors $\boldsymbol{a} = [a_1, a_2, \cdots, a_n]$ and $\boldsymbol{b} = [b_1, b_2, \cdots, b_n]$ in the Euclidean space, their *dot product* is defined as:

$$\boldsymbol{a} \cdot \boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Dot product

- **Geometrical interpretation**

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}|\,|\boldsymbol{b}| \cos(\theta)$$

where θ is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$. Clearly,

$$\boldsymbol{a} \perp \boldsymbol{b} \Leftrightarrow \boldsymbol{a} \cdot \boldsymbol{b} = 0$$

When $\boldsymbol{a} \perp \boldsymbol{b}$, we say that they are *orthogonal* to each other.

Dot product

- **Geometrical interpretation**

  Scalar projection

$$|\boldsymbol{a}| = |\boldsymbol{b}| = 1 \Rightarrow \boldsymbol{a} \cdot \boldsymbol{b} = \cos(\theta)$$
$$|\boldsymbol{b}| = 1 \Rightarrow \boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}| \cos(\theta)$$

  projection of $\boldsymbol{a}$ in the direction of $\boldsymbol{b}$

- The geometric properties rely on the basis (to be described later) being orthonormal, i.e. composed of pairwise perpendicular vectors with unit length.

Dot product

- **Example**

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} [a_{11}, a_{12}, a_{13}] \cdot [x_1, x_2, x_3] \\ [a_{21}, a_{22}, a_{23}] \cdot [x_1, x_2, x_3] \\ [a_{31}, a_{32}, a_{33}] \cdot [x_1, x_2, x_3] \end{bmatrix}$$

Norm

- In linear algebra, functional analysis and related areas of mathematics, a norm is a function that assigns a strictly positive length or size to all vectors in a vector space, other than the zero vector. A seminorm, on the other hand, is allowed to assign zero length to some non-zero vectors.

- A simple example is the 2-dimensional Euclidean space $\mathbb{R}^2$ equipped with the Euclidean norm. Elements in this vector space (e.g., (3, 7)) are usually drawn as arrows in a 2-dimensional cartesian coordinate system starting at the origin (0, 0). The Euclidean norm assigns to each vector the length of its arrow. Because of this, the Euclidean norm is often known as the magnitude.

- A vector space with a norm is called a *normed vector space*. Similarly, a vector space with a seminorm is called a seminormed vector space.

Norm

- **Definition**

  Given a vector space V over a subfield F of the complex numbers, including imaginary numbers or real numbers, a norm on V is a function $p : V \to R$ with the following properties:

  For all $a \in F$ and all $u, v \in V$,

  1. $p(av) = |a|p(v)$, (positive homogeneity or positive scalability)
  2. $p(u + v) \leq p(u) + p(v)$ (triangle inequality or subadditivity).
  3. $p(v) = 0$ if and only if v is the zero vector (positive definiteness).

- The norm of a vector v is usually denoted $\|v\|$.

Different types of norm

- Euclidean

$$\|\mathbf{x}\| := \sqrt{x_1^2 + \cdots + x_n^2}$$

- Taxicab norm or Manhattan norm

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

- Maximum norm

$$\|\mathbf{x}\|_\infty := \max(|x_1|, \cdots, |x_n|)$$

- p-norm

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

Basis of a vector space

A *basis* of a vector space $\boldsymbol{V}$ is a linearly independent subset of $\boldsymbol{V}$ that spans $\boldsymbol{V}$. Specifically, the basis satisfies the following two properties (assume that the vector space $\boldsymbol{V}$ is over the field of $\mathbb{R}^n$):

1. The *linear independence property*: for all $a_1, \cdots, a_n \in \mathbb{R}$, if $a_1\boldsymbol{v_1} + \cdots + a_n\boldsymbol{v}_n = 0$, then $a_1 = \cdots = a_n = 0$.

2. The *spanning* property: for every $\boldsymbol{x}$ in $\boldsymbol{V}$ it is possible to choose $a_1, \cdots, a_n \in \mathbb{R}$ such that $\boldsymbol{x} = a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n$.

The numbers $a_i, i = 1, \cdots, n$ are called the *coordinates* of the vector $\boldsymbol{x}$ with respect to the basis and they are *uniquely* determined. The number of vectors that the basis contains is the *dimension* of the vector space.

Basis of a vector space

- **Example of basis**

1. For $\mathbb{R}^2$, $\boldsymbol{e}_1 = (1, 0)$ and $\boldsymbol{e}_1 = (1, 0)$ form a basis and the dimension of $\mathbb{R}^2$ is 2.
2. For $\mathbb{R}^n$, $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n$ are linearly independent, generate $\mathbb{R}^n$, and thus form a basis. The dimension of $\mathbb{R}^n$ is n.
3. For the vector space of polynomials, a basis is given by $1, x, x^2, \cdots$ and the dimension is countably infinite.

Orthonormal basis

- **Definition**

  For a basis $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$, if:

$$\boldsymbol{v}_i \perp \boldsymbol{v}_j, \quad \text{for } i, j = 1, \cdots, n$$
$$\|\boldsymbol{v}_i\| = 1$$

  then the basis is *orthonormal*.

- Normalization of a vector

$$\boldsymbol{x} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$$

Orthonormal basis

- **Vector transformation between two orthonormal basis**

Let

1. $\boldsymbol{V}_1 = [\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}]$ and $\boldsymbol{V}_2 = [\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t}]$ be two different orthonormal bases of the same space $\mathbb{R}^3$.
2. $\boldsymbol{a}_1 = [a_x, a_y, a_z]$ represent vector $\boldsymbol{a}$ in terms of $\boldsymbol{V}_1$.
3. $\boldsymbol{a}_2 = [a_r, a_s, a_t]$ represent the same vector $\boldsymbol{a}$ in terms of $\boldsymbol{V}_2$.

Then $\boldsymbol{a}_2$ can be obtained from $\boldsymbol{a}_1$ by rotation using a rotation matrix $\boldsymbol{P}$:

$$\boldsymbol{a}_2 = \boldsymbol{P}\boldsymbol{a}_1 = \begin{bmatrix} r_x & r_y & r_z \\ s_x & s_y & s_z \\ t_x & t_y & t_z \end{bmatrix} \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = \begin{bmatrix} \boldsymbol{r} \cdot \boldsymbol{a}_1 \\ \boldsymbol{s} \cdot \boldsymbol{a}_1 \\ \boldsymbol{t} \cdot \boldsymbol{a}_1 \end{bmatrix} = \begin{bmatrix} a_r \\ a_s \\ a_t \end{bmatrix}$$

Eigenvalue and eigenvector

- **Definition**

  Given a linear transformation $\boldsymbol{A}$ of size $n \times n$, a non-zero vector $\boldsymbol{x}$ is defined to be an *eigenvector* of the transformation if it satisfies the equation

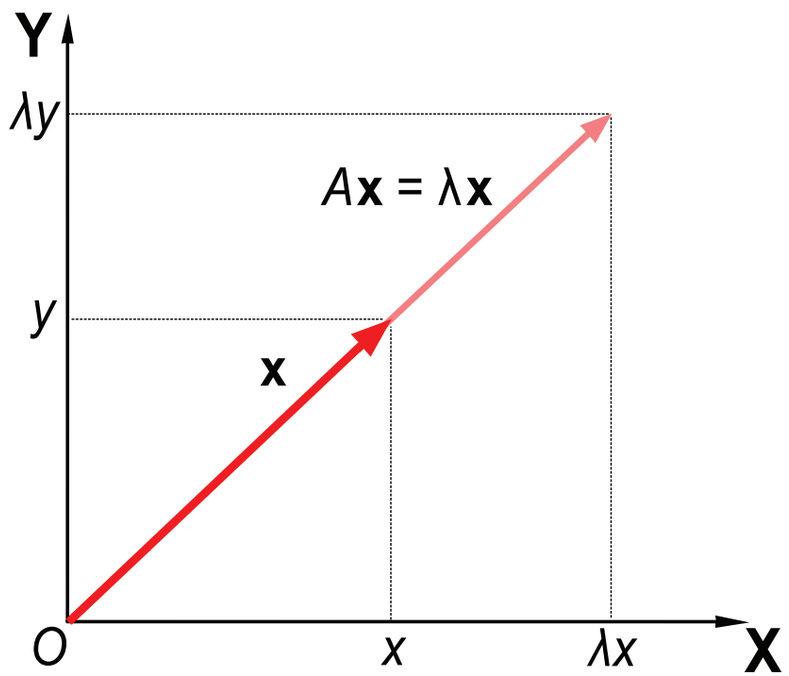
$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

  for some scalar λ. In this situation, the scalar λ is called an *eigenvalue* of $\boldsymbol{A}$ corresponding to the eigenvector $\boldsymbol{x}$.

Eigenvalue and eigenvector

- **Properties**

  1. The eigenvector $\boldsymbol{x}$ has the property that its direction is NOT changed by the transformation $\boldsymbol{A}$. It is scaled by a factor of λ.
  2. Vectors that are not eigenvectors will change direction and magnitude under the transformation. Thus eigenvectors and eigenvalues are special.
  3. For the identity matrix, all non-zero vectors are eigenvectors.



Eigenvalue and eigenvector

- **How to calculate eigenvalues and eigenvectors?**

$$\begin{aligned} \boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} &\Rightarrow \boldsymbol{A} - \lambda\boldsymbol{I}\boldsymbol{v} = \boldsymbol{0} \\ &\Rightarrow (\boldsymbol{A} - \lambda\boldsymbol{I})\,\boldsymbol{v} = \boldsymbol{0} \\ &\Rightarrow \det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0 \end{aligned}$$

The determinant requirement is called the *characteristic equation* and the left-hand-side is called the *characteristic polynomial*.

Let $p(\lambda)$ represent the characteristic polynomial, then $p(\lambda)$ can be factorized as

$$p(\lambda) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2} \cdots (\lambda - \lambda_k)^{n_k} = 0$$

where

$$\sum_{i=1}^{k} n_i = n$$

The set of eigenvalues is called the *spectrum* of $\boldsymbol{A}$.

Eigenvalue and eigenvector

- **Example**

  Compute the eigenvalues and eigenvectors of the following matrix

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

  Solution:

$$\lambda_1 = 1, \boldsymbol{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \lambda_2 = 3, \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Eigenvalue and eigenvector

- **More properties of eigenvalues and eigenvectors**

1. The eigenvectors corresponding to different eigenvalues are *linearly independent*, that is, none of the eigenvectors can be written as a linear combination of other eigenvectors.

2. Each eigenvector is associated with one eigenvalue, but one eigenvalue can be associated with an infinite number of eigenvectors.

3. If $\boldsymbol{x}$ is an eigenvector, then $\alpha\boldsymbol{x}$ is an eigenvector as well corresponding to the same eigenvalue. Together with the zero vector, the eigenvectors of $\boldsymbol{A}$ with the same eigenvalue form a linear subspace of the vector space called an *eigenspace*.

4. An eigenvector cannot be $\mathbf{0}$, but an eigenvalue can be 0.

Eigenvalue and eigenvector

- **Example**

  Compute the eigenvalues and eigenvectors of the following matrix.

$$\begin{bmatrix} 3 & 6 & -8 \\ 0 & 0 & 6 \\ 0 & 0 & 2 \end{bmatrix}$$

Solution:

$$\lambda_1 = 0, \boldsymbol{v}_1 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \quad \lambda_2 = 2, \boldsymbol{v}_2 = \begin{bmatrix} -10 \\ 3 \\ 1 \end{bmatrix}, \quad \lambda_1 = 3, \boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Algebraic and geometric multiplicities

- Given an $n \times n$ matrix A and an eigenvalue λ_i of this matrix, there are two numbers measuring, roughly speaking, the number of eigenvectors belonging to λ_i. They are called *multiplicities*.

    1. The *algebraic multiplicity* of an eigenvalue is defined as the multiplicity of the corresponding root of the characteristic polynomial.

    2. The *geometric multiplicity* of an eigenvalue is defined as the dimension of the associated eigenspace, i.e. number of linearly independent eigenvectors with that eigenvalue.

Algebraic and geometric multiplicities

- Example

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

$\lambda = 1$ is a double root. And there is only one linearly independent eigenvector: $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

- Example

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

$\lambda = 1$ is a double root and the only eigenvalue. But since any non-zero vector can serve as its eigenvector, A has two linearly independent eigenvectors: $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Algebraic and geometric multiplicities

- **Properties**

  1. Both algebraic and geometric multiplicity are integers between (including) 1 and n.
  2. The algebraic multiplicity n_i and geometric multiplicity m_i may or may not be equal, but we always have $m_i \leq n_i$. The simplest case is of course when $m_i = n_i = 1$.
  3. The total number of linearly independent eigenvectors, N_x, is given by summing the geometric multiplicities

$$\sum_{i=1}^{N_\lambda} m_i = N_x.$$

Similar matrices

- **Definition**

  Two $n \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *similar* if there exists matrix $\boldsymbol{P}$ such that

$$\boldsymbol{Y} = \boldsymbol{P}^{-1}\boldsymbol{X}\boldsymbol{P}$$

- Similar matrices have the same:
  1. rank
  2. determinant
  3. trace
  4. eigenvalue
  5. characteristic polynomial

Diagonalizable matrix

- **Diagonalizable matrix**

  In linear algebra, a square matrix A is called *diagonalizable* if it is similar to a diagonal matrix, i.e., if there exists an invertible matrix P such that $P^{-1}AP$ is a diagonal matrix.

- **Under what conditions a matrix will be diagonalizable?**

  An n-by-n matrix A over the field $\mathbb{R}$ is diagonalizable if and only if the sum of the dimensions of its eigenspaces is equal to n, which is the case if and only if there exists a basis of $\mathbb{R}^n$ consisting of eigenvectors of A.

Eigendecomposition of a matrix

- **Definition**

  Let $\boldsymbol{A}$ be a square $n \times n$ matrix with n linearly independent eigenvectors $\boldsymbol{p}_i$, $i = 1, \cdots n$. Then $\boldsymbol{A}$ can be factorized as

$$\boldsymbol{A} = \boldsymbol{P}\Lambda\boldsymbol{P}^{-1}$$

  where

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n]$$

  and $\boldsymbol{\Lambda}$ is the *diagonal matrix* whose diagonal elements are the corresponding eigenvalues.

- From $\boldsymbol{A} = \boldsymbol{P}\Lambda\boldsymbol{P}^{-1}$, we can get $\mathbf{A}\,[\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n] = [\lambda_1\boldsymbol{p}_1, \lambda_2\boldsymbol{p}_2, \cdots, \lambda_n\boldsymbol{p}_n]$.

Eigendecomposition of a matrix

- **Example**

  Diagonalize the following matrix:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$$

Solution: $\lambda_1 = 1$ and $\lambda_2 = 3$

$$\boldsymbol{p}_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \quad \boldsymbol{p}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

So

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} -2 & 0 \\ 1 & 1 \end{bmatrix}^{-1}$$

Eigendecomposition of a matrix

- The diagonal matrix makes the following computations easier:

1. Matrix inverse via eigendecomposition / matrix diagonalization

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^{-1} \quad \text{and} \quad \left[\Lambda^{-1}\right]_{ii} = \frac{1}{\lambda_i}$$

$\boldsymbol{A}$ can be inverted if and only if $\lambda_i \neq 0$ for any i.

2. Determinant

$$\det(\boldsymbol{A}) = \prod_{i=1}^{k} \lambda_i^{n_i}$$

Note that each eigenvalue is raised to the power n_i, the algebraic multiplicity.

Eigendecomposition of a matrix

- The diagonal matrix makes the following computations easier:

3. Power of a matrix

$$\boldsymbol{A}^k = \boldsymbol{P}\boldsymbol{\Lambda}^k\boldsymbol{P}^{-1}, \quad \left[\Lambda^k\right]_{ii} = \lambda_i^k$$

4. Trace

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} n_i\lambda_i$$

Note that each eigenvalue is multiplied by n_i.

Determine diagonalizability

- For an $n \times n$ matrix $\boldsymbol{A}$, if there exists n independent eigenvectors, then:
  1. these eigenvectors can form a basis for the vector space $\mathbb{R}^n$.
  2. matrix $\boldsymbol{A}$ is *diagonalizable*.

- For an $n \times n$ matrix $\boldsymbol{A}$, if there exists $< n$ independent eigenvectors, then:
  1. these eigenvectors can NOT form a basis for the vector space $\mathbb{R}^n$.
  2. matrix $\boldsymbol{A}$ is *NOT diagonalizable*.

Matrix diagonalization

- **Example:** Find a transformation matrix $\boldsymbol{P}$ that can diagonalize

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{bmatrix}$$

Solution:

$$\lambda_1 = 3, \boldsymbol{v}_1 = \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}, \quad \lambda_2 = 2, \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \lambda_3 = 3, \boldsymbol{v}_3 = \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}$$

Let $\boldsymbol{P}$ be the matrix with these eigenvectors as its columns:

$$\boldsymbol{P} = \begin{bmatrix} -1 & 0 & -1 \\ -1 & 0 & 0 \\ 2 & 1 & 2 \end{bmatrix}$$

Then $\boldsymbol{P}$ diagonalizes $\boldsymbol{A}$:

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \begin{bmatrix} 0 & -1 & 0 \\ 2 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -4 & 2 \end{bmatrix} \begin{bmatrix} -1 & 0 & -1 \\ -1 & 0 & 0 \\ 2 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Orthogonal matrix

- **Definition**

  A matrix $\boldsymbol{P}$ is called *orthogonal* if

$$\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$$

  where I is the identity matrix.

- This leads to the equivalent characterization: A matrix $\boldsymbol{P}$ is orthogonal if its transpose is equal to its inverse:

$$\boldsymbol{P}^T = \boldsymbol{P}^{-1}$$

Orthogonal matrix

- If the transformation matrix $\boldsymbol{P}$ is an orthogonal matrix, then the linear transformation $\boldsymbol{Y} = \boldsymbol{P}^T \boldsymbol{X}$ sends every vector into a new vector space spanned by the orthogonal basis (i.e. the columns of $\boldsymbol{P}$).

$$\boldsymbol{Y} = \boldsymbol{P}^T \boldsymbol{X} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n]^T \boldsymbol{X}$$

$$= \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_n \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} \boldsymbol{p}_1 \cdot \boldsymbol{X} \\ \boldsymbol{p}_2 \cdot \boldsymbol{X} \\ \vdots \\ \boldsymbol{p}_n \cdot \boldsymbol{X} \end{bmatrix}$$

- Orthogonal matrices preserves the dot product. So for vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in an n-dimensional real Euclidean space

$$\boldsymbol{u} \cdot \boldsymbol{v} = (\boldsymbol{P}\boldsymbol{u}) \cdot (\boldsymbol{P}\boldsymbol{v})$$

where $\boldsymbol{P}$ is an orthogonal matrix.

Orthogonal matrix

- **Example:** Below are a few examples of small orthogonal matrices and possible interpretations.

  1.

  $$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

  identity transformation

  2.

  $$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

  rotation by θ

  3.

  $$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

  reflection across x-axis

Symmetric matrix

- **Definition**
  A symmetric matrix is a square matrix that is equal to its transpose. Formally, matrix A is symmetric if

$$A = A^T$$

- For any symmetric matrix $\boldsymbol{A}$, there exists an *orthogonal* matrix that can diagonalize $\boldsymbol{A}$. The columns of the orthonormal matrix consists of the eigenvectors of $\boldsymbol{A}$. More precisely:

  A matrix is symmetric if and only if it has an orthonormal basis of eigenvectors.

  The covariance matrix is symmetric!

Probability and Statistics

Expectation

- The expectation of a random variable is intuitively the long-run average value of repetitions of the experiment it represents.

- **Definition for discrete random variables**

  Let X be a discrete random variable with density f. The expectation of X, denoted by $E(X)$, is given by

$$E[X] = \sum_{\text{all } x} x f(x)$$

  provided $\sum_{\text{all } x} |x| f(x)$ is finite. Summation is over all values of X that occur with nonzero probability.

Expectation

- **General definition of expectation for discrete random variables**

  Let X be a discrete random variable with density f. Let $H(X)$ be a random variable. The expectation of $H(X)$, denoted by $E(H(X))$, is given by

$$E[H(X)] = \sum_{\text{all } x} H(x) f(x)$$

  provided $\sum_{\text{all } x} |H(x)| f(x)$ is finite. Summation is over all values of X that occur with nonzero probability.

Expectation

- **Definition for continuous random variables**

  Let X be a continuous random variable with density f. Let $H(X)$ be a random variable. The expected value of $H(X)$, denoted $E[H(X)]$, is given by

$$E[H(X)] = \int_{-\infty}^{\infty} H(x)f(x)dx$$

  provided

$$\int_{-\infty}^{\infty} |H(x)|f(x)dx$$

  is finite.

Expectation

- The expectation is a key aspect of how one characterizes a probability distribution. It is a location parameter.

- **Theorem**

Let X and Y be random variables and let c be a real number. Then

1. $E[c] = c$
2. $E[cX] = cE[X]$
3. $E[X + Y] = E[X] + E[Y]$

Expectation

- Sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance

- Variance measures how far a set of numbers spread out.

- **Definition**

Let X be a random variable with mean μ. The variance of X, denoted by σ^2, is given by

$$\sigma^2 = Var[X] = E[(X - \mu)^2]$$

Variance

- **Properties**
  - Variance is a measure of dispersion of the possible values of the random variable around the expected value.
  - Variance is non-negative.
  - A variance of zero indicates that all the values are identical.
  - A small variance indicates that the data points tend to be very close to the expected value and hence to each other, while a high variance indicates that the data points are very spread out around the mean and from each other.

Variance

- **Computational formula**

$$\sigma^2 = E[X^2] - (E[X])^2$$

- **Theorem**

Let X and Y be random variables and c any real number. Then

  - $Var[c] = 0$
  - $Var[cX] = c^2 Var[X]$
  - If X and Y are independent, then

$$\begin{aligned} Var[X+Y] &= Var[X] + Var[Y] \\ Var[X-Y] &= Var[X] + Var[Y] \end{aligned}$$

Variance

- Sample variance

$$s_x^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2}{n-1}$$

Standard deviation

- Like variance, standard variation is also used to quantify the amount of variation of a set of data values.

- **Definition**

  Let X be a random variable with variance σ^2. The standard variation of X, denoted by σ, is given by

$$\sigma = \sqrt{\sigma^2}$$

- **Properties**

  – Unlike variance, standard deviation is expressed in the same units as the data.

  – A low standard deviation indicates that data points tend to be very close to the mean, while a high standard deviation indicates that the data points are spread out over a wide range of values.

Covariance

- Covariance is a measure of how much two variables change together.

- **Definition**

  The covariance of (X, Y) is defined by

$$cov(X, Y) = \mathbb{E}\left[(X - \mu_x)(Y - \mu_y)\right]$$

- **Computational formula**

$$cov = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

Covariance

- **Properties**
  - The covariance is sometimes called a measure of the *linear dependence* between the two random variables.
  - If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e. the variables tend to show similar behavior, the covariance is a positive number.
  - In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e. the variables tend to show opposite behavior, the covariance is negative.
  - The sign of the covariance therefore shows the tendency in the linear relationship between the variables.
  - The magnitude of the covariance is not that easy to interpret.
  - The normalized version of the covariance, the correlation coefficient, however shows by its magnitude the strength of the linear relation.

Covariance

- **Properties**
  - $cov(X, a) = 0$
  - $cov(X, X) = var(X)$
  - $cov(X, Y) = cov(Y, X)$
  - $cov(aX, bY) = ab \cdot cov(X, Y)$
  - $cov(X + a, Y + b) = cov(X, Y)$
  - $cov(aX + bY, cW + dV) = ac \cdot cov(X, W) + ad \cdot cov(X, V) + bc \cdot cov(Y, W) + db \cdot cov(Y, V)$
  - $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$
  - $\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\ \text{cov}(X, Y)$

Covariance

- **Properties**
  - For the sum of N variables: $Y = \sum_{i=1}^{n} X_i$, we have

$$\mathrm{var}(Y) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2 \sum_{i<j} \mathrm{cov}(X_i, X_j)$$

    or

$$\mathrm{var}(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{cov}(X_i, X_j)$$

  - Therefore, if $\mathrm{cov}(X_i, X_j) = 0, \forall X_i, X_j$ (or we say X_i are uncorrelated), then

$$\mathrm{var}\left(\sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \mathrm{var}(X_i)$$

Correlation

- Two real-valued random variables are said to be *uncorrelated* if their covariance is zero.

- **Definition**

  Pearson's product-moment coefficient

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- The above formula defines the *population correlation coefficient*.

Correlation

- **sample correlation coefficient**

  Substituting estimates of the covariances and variances based on a sample gives the *sample correlation coefficient*, commonly denoted r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlation

- **Alternative formula**

  The Pearson correlation can be expressed in terms of uncentered moments:

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - (E[X])^2}\sqrt{E[Y^2] - (E[Y])^2}}$$

- Alternative formulae for the sample Pearson correlation coefficient:

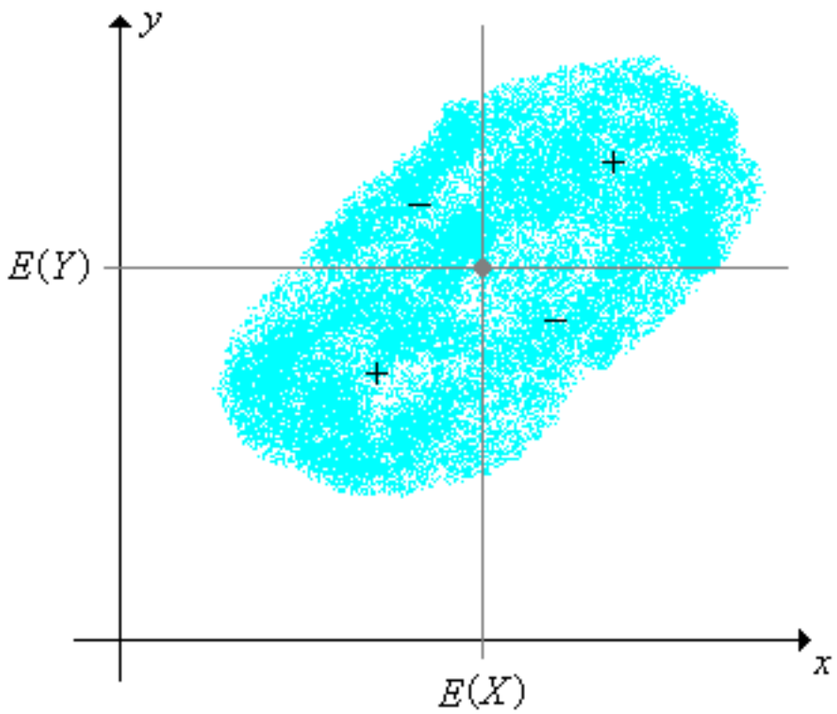
$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Correlation

- **Properties**
  - Correlation is a scaled version of covariance.
  - Correlation and covariance always have the same sign (positive, negative, or 0).
    - When the sign is positive, the variables are said to be *positively correlated*.
    - when the sign is negative, the variables are said to be *negatively correlated*.
    - When the sign is 0, the variables are said to be *uncorrelated*.
  - Correlation is dimensionless.

Correlation

- **Properties**
  - Covariance and correlation measure a certain kind of dependence between the variables.
  - $(E(X), E(Y))$ is the center of the joint distribution of (X, Y) and the vertical and horizontal lines through this point separate $\mathbb{R}$ into four quadrants.
  - $[x - E(X)]\,[y - E(Y)]$ is positive on the first and third quadrants and negative on the second and fourth quadrants.



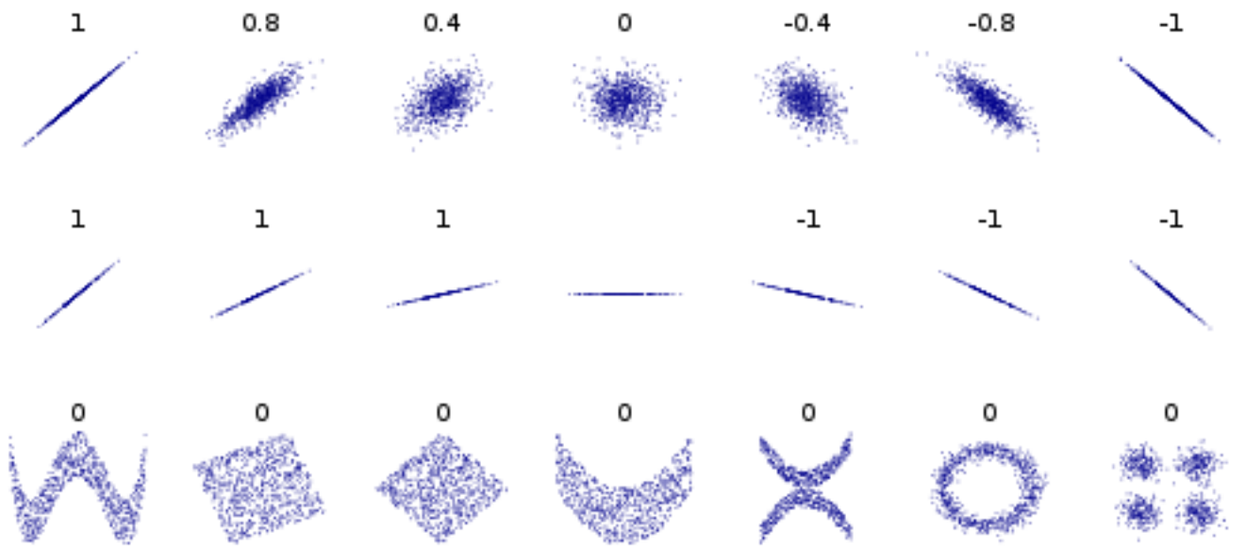
http://www.math.uah.edu/stat/expect/Covariance.html

Correlation

- **Properties**
  - The Pearson correlation is +1 in the case of a perfect positive (increasing) *linear relationship*, -1 in the case of a perfect decreasing (negative) linear relationship, and some value between -1 and 1 in all other cases, indicating the degree of linear dependence between the variables.
  - The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.
  - As it approaches zero there is less of a relationship.

Correlation

- **Properties**
  - The correlation reflects the noisiness and direction of a linear relationship (top).
  - The correlation does NOT reflects the slope of that linear relationship (middle).
  - The correlation does NOT reflect nonlinear relationships (bottom).
  - The correlation for the center figure is undefined because the variance of Y is zero.



https://en.wikipedia.org/wiki/Correlation_and_dependence

Correlation

- **Properties**

  – If two random variables are *independent*, then the Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

    X and Y are independent $\Rightarrow cov(X, Y) = 0 \Rightarrow X$ and Y are uncorrelated

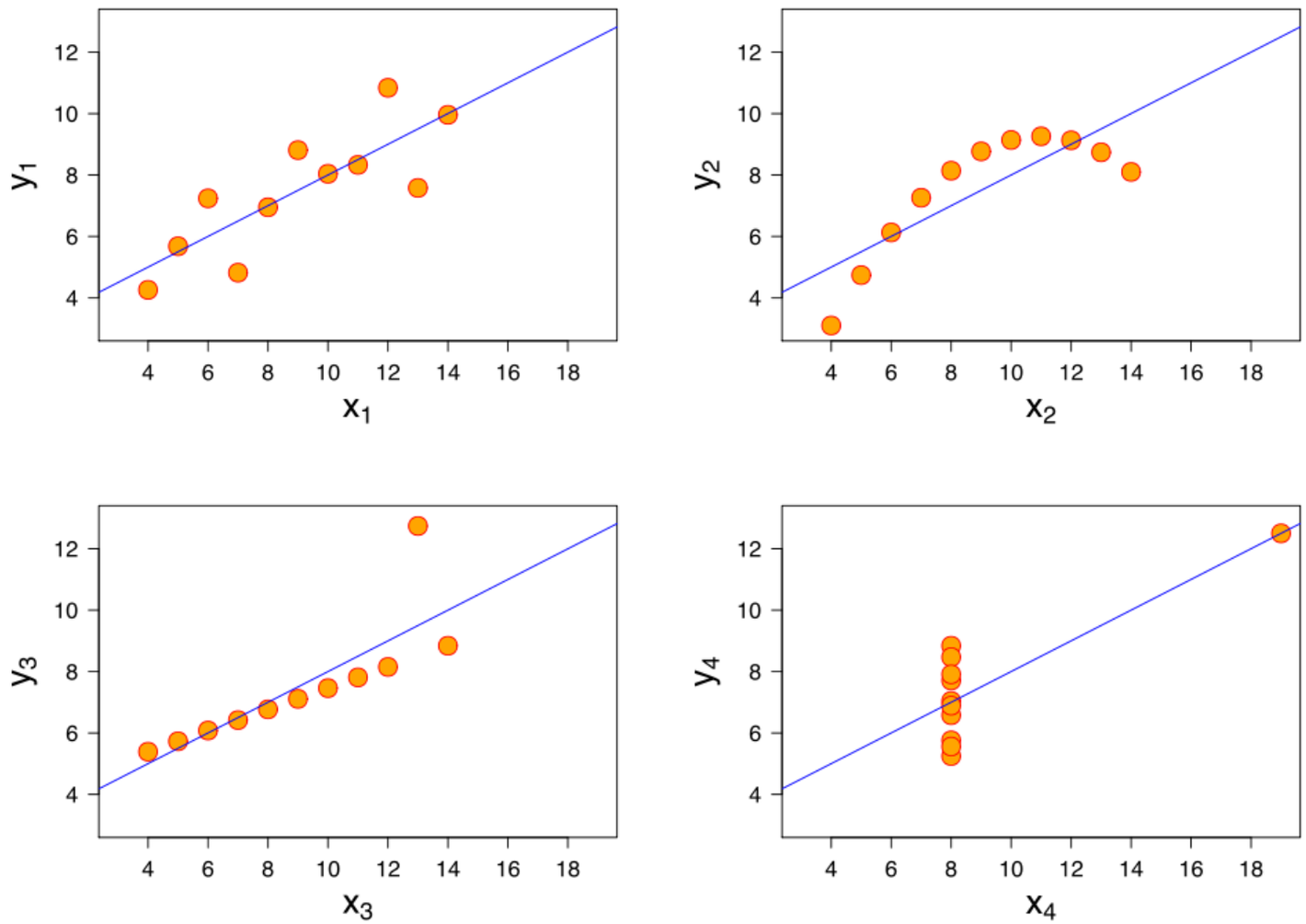
  – Example: Suppose that X is uniformly distributed on the interval $[-1, 1]$ and $Y = X^2$. Then X and Y are uncorrelated even though Y is a function of X (the strongest form of dependence).

Correlation

- **Properties**
  - When X and Y are jointly normal, uncorrelatedness is equivalent to independence.
  - Correlation is invariant to changes in location and scale. That is, we may transform X to $a + bX$ and transform Y to $c + dY$, where a, b, c, and d are constants, without changing the correlation coefficient (this fact holds for both the population and sample Pearson correlation coefficients).
  - Geometric interpretation: For centered data (i.e., data which have been shifted by the sample mean so as to have an average of zero), the correlation coefficient can also be viewed as the cosine of the angle between the two vectors of samples drawn from the two random variables.

Correlation

- **Properties**
  - The Pearson correlation coefficient indicates the strength of a linear relationship between two variables, but its value generally does not completely characterize their relationship.



https://en.wikipedia.org/wiki/Correlation_and_dependence

Correlation

- **Properties**
  - Pearson's correlation and least squares regression analysis

    The square of the sample correlation coefficient, which is also known as the *coefficient of determination*, estimates the fraction of the variance in Y that is explained by X in a *simple linear regression*.

    The total variation in the Y_i around their average value can be decomposed as follows:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2$$

where the $\hat{Y}_i$ are the fitted values from the regression analysis. This can be re-arranged to give

$$1 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y}_i)^2} + \frac{\sum_i (\hat{Y}_i - \bar{Y}_i)^2}{\sum_i (Y_i - \bar{Y}_i)^2}$$

The two summands above are the fraction of variance in Y that is explained by X (right) and that is unexplained by X (left). It can be shown that

$$r(Y, \hat{Y})^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y}_i)^2}{\sum_i (Y_i - \bar{Y}_i)^2}$$

is the proportion of variance in Y explained by a linear function of X.

Covariance matrix

- **Definition**

  For random vectors X and Y (of dimensions $m \times 1$ and $n \times 1$ respectively), the $m \times n$ covariance matrix is

$$\Sigma = cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T] = \mathbb{E}[XY^T] - \mathbb{E}[X][Y]^T$$

$$= \begin{bmatrix} cov\,(X_1, Y_1) & cov\,(X_1, Y_2) & \cdots & cov\,(X_1, Y_n) \\ cov\,(X_2, Y_1) & cov\,(X_2, Y_2) & \cdots & cov\,(X_2, Y_n) \\ \vdots & \vdots & \vdots & \vdots \\ cov\,(X_m, Y_1) & cov\,(X_m, Y_2) & \cdots & cov\,(X_m, Y_n) \end{bmatrix}$$

For a random vector $X = [X_1, X_2, \cdots, X_n]$, the covariance matrix is

$$\Sigma = cov\,(X) = \begin{bmatrix} var\,(X_1) & cov\,(X_1, X_2) & \cdots & cov\,(X_1, X_n) \\ cov\,(X_2, X_1) & var\,(X_2) & \cdots & cov\,(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ cov\,(X_n, X_1) & cov\,(X_n, X_2) & \cdots & var\,(X_n) \end{bmatrix}$$

Covariance matrix

- **Calculating the sample covariance**
  - Let X be a random vector , a row vector whose jth element ($j = 1, 2, \cdots, K$) is one of the random variables.
  - The sample covariance of N observations of K variables is the $K \times K$ matrix $Q = [q_{jk}]$ with the entries given by

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Covariance matrix

- **Calculating the sample covariance**
  - The sample mean and the sample covariance matrix are *unbiased estimates* of the mean and the covariance matrix of the random vector X.
  - The reason the sample covariance matrix has $N-1$ in the denominator rather than N is essentially that the population mean $E(X)$ is not known and is replaced by the sample mean $\bar{X}$.
  - If the population mean $E(X)$ is known, the analogous unbiased estimate is given by

$$q_{jk} = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - E(X_j))(x_{ik} - E(X_k))$$

Covariance matrix

- **Remarks**
  - The linear dependence that the covariance measures between two random variables does not mean the same thing as in the context of linear algebra.
  - A covariance matrix is always symmetric:

$$[cov(X)]^T = cov(X)$$

Correlation

- For a random vector $X = [X_1, X_2, \cdots, X_n]$, the correlation matrix is

$$corr\,(X) = \begin{bmatrix} var\,(X_1) & corr\,(X_1, X_2) & \cdots & corr\,(X_1, X_n) \\ corr\,(X_2, X_1) & var\,(X_2) & \cdots & corr\,(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ corr\,(X_n, X_1) & corr\,(X_n, X_2) & \cdots & var\,(X_n) \end{bmatrix}$$

- Because the correlation matrix does not depend on the scale of the random variables, it is used to express relationships among random variables measured on different scales. If the elements of X are independent, $|corr(X)| = 1$. If the elements of X are dependent, $|corr(X)| = 0$. Thus, the determinant of the correlation matrix may be interpreted as an overall measure of association or non-association.

Thank you!