Introduction to Applied Linear Algebra
Stephen Boyd & Lieven Vandenberghe
(Summarize version)

Quang Huy Nguyen
Department of Pharmacology
Dainam University

I. VECTORS

1 Vectors

1.1. Vectors

Vectors are usually written as vertical array (the first two notations) or numbers separated by column and surrounded by parentheses (the last one).

$$\begin{pmatrix}0\\1\end{pmatrix} or \begin{bmatrix}0\\1\end{bmatrix} or(0,\ 1)$$

The elements (or entries, coefficients, components) of a vector are the values in the array. The size (or dimension, length) of the vector is the number of elements it contains. A vector of size n is called an n-vector
The ith element of the vector a is denoted a_i, where the subscript i is an integer index
The numbers or values of the elements in a vector are called *scalars.*

Block or stacked vectors
vectors by concatenating or stacking two or more vectors, as in

$$\begin{bmatrix}b\\c\\d\end{bmatrix}$$

where b is an m-vector, c is an n-vector and d is a p-vector and a is a (m+n+p)-vector

Subvectors
b,c and d are subvector or slices of a. Colon notations is used to denote subvectors. $a_{r:s}$ is the vector, then $a_{r:s}$ is a vector of size s-r+1, with entries $a_r...a_s$.

The subcript r:s is called the *index range.*

Indexing
$(a_i)_j$ to refer to jth entry of a_i, the ith vector in our list

Zero vectors - Unit vectors - Ones vector
a, Zero vectors is an n-vector with all elements equal to 0, written as 0_n
b, Unit vectors is an n-vector with all elements equal to 0, except one element which is equal to one, written as e_j, where jth is the position of element which is one
c, Ones Vector is an n-vector with all elements equal to 1, written as 1_n

Sparsity
A vector is said to be *sparse* if many of its entries are zero, its *sparsity pattern* is the set of indices of nonzero entries. The number of the nonzero entries of an n-vector x is denoted **nnz**(x)

EXAMPLE
•*Location and displacement.* A 2-vector can represent a position or location in a 2-dimensional space i.e, a plane (Oxy), as shown in figure 1.1, and the same for 3-vector in a 3-D space, Oxyz. The entries of the vector give the coordinates of the position/ location.
A vector can also represent the velocity or acceleration, at a given time, of a point that moves in a plane or 3-D space

•*Time series.* An n-vector can represent a *time series* or *signal*, that is the value of some quantity at different times (The entries in a vector that represents a time series called *samples*, when the quantity is sth measured.
When a vector represents a time series, plot x_i versus i with lines connecting consecutive time series values

•*Occurence or subsets.* Occurence = 1, No Occurence = 0 => n-vector with its entries equal to either 0 or 1 called Boolean

•*Features or attributes.* a vector collects together n different quantities that pertain to a single thing or object. A quantities can be measured or derived from the object => such a vector are called a *feature vector*, and its entries are called the *features* or *attributes.* e.g, 6-vector f could give the age, height, weight, blood pressure, and gender of a patient admitted to a hospital.

**Other Examples can be encountered in page 6-11.*

1.2. Vector addition

$$\begin{pmatrix}0\\7\end{pmatrix} + \begin{pmatrix}1\\2\end{pmatrix} = \begin{pmatrix}1\\9\end{pmatrix}$$

Note: vectors must have the same size. Vector Substraction is similar. The result of vector substraction is called the diffirentce of the two vector
Properties: For any vector a,b and c of the same size:
•Vector addition is *commutative.* a + b = b + a
•Vector addition is *associative.* (a + b) + c = a + (b + c)

EXAMPLE
•*Displacements.* vectors a and b represent displacement, the sum a + b is the net displacement found by first displacing by a, then displacing by b (figure 1.3). If the vector p represent a position and the vector a represent a displacement, then p+a is the position of the point p, displaced by a (figure 1.4).

•*Displacements between two points.* vector p and q represent the positions of two points in 2-D or 3-D space, p-q is the displacement vector from p to q (figure 1.5)

•*Feature differences.* f and g are n-vectors that give n feature values for two items, the different vector d = f - g give the difference in feature values for two objects. e.g, $d_7 = 0$ means the two objects have the same value for feature 7.

•*Time series.* a and b represent time series of the same quantity, such as daily profit at two different stores, a + b represents a time series which is the total daily profit at the two store.

**Other Examples can be encountered in page 14*

1.3. Scalar-vector multiplication

a vector is multiplied by a scalar(i.e., number)

$$(-2)\begin{bmatrix}1\\9\end{bmatrix} = \begin{bmatrix}1\\9\end{bmatrix}(-2) = \begin{bmatrix}-2\\-18\end{bmatrix}$$

Properties For any scalar α, β, γ and any vectors a and b
•*commutative.* αa = aα
•*associative.* $(\beta\gamma)$a = a$(\beta\gamma)$
•*left distributive.* $(\beta + \gamma)$a = βa + γa. (The same for right-distribuitive property)

•*left distributive in another version.* (β(a+b) = βa + βb). (The same for right-distribuitive property)

EXAMPLE

•*Displacements.* vector a represents displacement, and $\beta > 0$, βa is a displacement in the same direction of a, with its magnitude scaled by β. $\beta < 0$, βa represents a displacement in the opposite direction of a, with its magnitude scaled by β (figure 1.6).

*Other Examples can be encountered in page 16

Linear combinations $a_1 ... a_m$ are n-vectors, and $\beta_1 ... \beta_m$ are the scalars

$$\beta_1 a_1 + ... + \beta_m a_m$$

is called a linear combination of the vectors $a_1 ... a_m$ and the scalars $\beta_1 ... \beta_m$ are called the coefficients of the linear combination.

Linear combination of unit vectors any n-vector b is linear combination of the standard unit vectors e_i

$$b = b_1 e_1 + ... + b_m e_m$$

b_i is the ith entry of b. In other words, the coefficients are the entries of the vector b

Special Linear combinations the linear combination with $\beta_1 = ... = \beta_m = 1$, given by $a_1 + ... + a_m$, is the *sum* of vectors, and the linear combination with $\beta_1 = ... = \beta_m = 1/\text{m}$, given by $(a_1 + ... + a_m)/\text{m}$, is the *average* of vectors, the linear combination with $\beta_1 + ... + \beta_m = 1$, the combination is called *affine* combination, if β_i is non-negative, it is called a *convex* combination, a *mixture*, or a *weighted average.*

EXAMPLE

•*Displacement.* When vectors represent displacements, a linear combination is the sum of the scaled displacements. (figure 1.7)

•*Line and segment.* When a and b are different n-vectors, the affine combination c =$(1 - \theta)a + \theta b$, where θ is the scalar, describes a point on the *line* passing through a and b. When $0 \leq \theta \leq 1$, c is the convex combination of a and b, and is said to lie on the *segment* between a and b (figure 1.8)

1.4. Inner Product

The i*nner product* (also called *dot product*) of two n-vectors is defined as the scalar

$$a^T b = a_1 b_1 + ... + a_n b_n.$$

Some other notations of Inner product: <a, b>, <a | b>, (a, b) and a·b

Properties For any a,b, and c are n-vectors, and γ is a scalar
•*Commutativity.* a^Tb=b^Ta
•*Associativity with scalar multiplication.* $(\gamma a)^T$b=$\gamma(a^T$b) (1)
•*Distributivity with vector addition.* $(a+b)^T$c = a^Tc + b^Tc (2)
(1) + (2) => $(a+b)^T(c+d) = a^Tc + a^Td + b^Tc + b^Td$)

GENERAL EXAMPLES textit

•Unit vector. e_i^Ta = a_i textit

•Sum. $1^T a = \sum_{i=1}^n a_i$
•*Average.* **avg**(x) = $\mu = (\frac{1}{n})^T a = \frac{\sum_{i=1}^n a_i}{n}$

•*Sum of squares.* $a^T a = \sum_{i=1}^n a_i^2$

•*Selective Sum.* Let b be a vector all of whose entries are either 0 or 1.
$b^T a = k$, k is the sum of the elements in a for which $b_i = 1$

Block vectors. If a and b are two *k-blocks vectors* (called them *conform*)

$$a^T b = [a_1^T b_1 + ... + a_k^T b_k] \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_k \end{bmatrix} = a_1^T b_1 + ... + a_k^T b_k$$

Applications
•***Co-occurrence.*** If a and b are n-vectors, with each of their elements is either 0 or 1, then $a^T b$ gives the total number of indices for which a and b are both 1, that is, the total number of co-occurrences.

a=(0,1,1,1,1,1,0) and b =(0,0,1,0,1,0,0) => $a^T b = 2$ (figure 1.9)

•***Probability and expected values.*** let n-vector p has non-negative entries

that sum to one, so it describes a set of proportions among n items, or a set of probabilities of n outcomes, one of which must occur. Let f be another n-vector, f_i as the value of some quantity if outcome i occurs. Then $f^T p$ gives the expected value/mean of the quantity, under probabilities(or fractions) given by p

• ***Polymonial evaluation.*** $p(x) = c_1 + c_2 x + ... + c_{n-1} x_{n-1} + c_n x_n$
Assumption:
$c = (c_1, ..., c_n)$ is the n-vector, which represents the coefficient of polymonial p of degree n-1 or less.
$z = (1, t, t^2, ..., t^{n-1})$ is the n-vector of powers of t, let t be a number
$=> p(t) = c^T z$, the value of the polymonial p at the point t.

Other applications can be encountered in page 20 - 22

2 Linear functions

2.1. Linear functions

Function notation. f: $R^n \to R$ means f is a function maps the n-vector x to $f(x)$, which is scalar, denotes the value of the function f at x (In the notation $f(x)$, x is referred to as the argument of the function). We can also interpret f as a function of n scalar arguments, the entries of the vector argument, in which case we write $f(x)$ as

$$f(x) = f(x_1, ..., x_n)$$

we refer to $x_1, ..., x_n$ as the arguments of f.
For example: : $R^4 \to R$ is a function f that maps $x = (x_1, x_2, x_3, x_4)$ to $f(x) = x_1 + x_2 + x_4^2$

The inner product function. let a be an n-vector. A scalar-valued function f of n-vectors

$$f(x) = a^T x = a_1 x_1 + ... + a_n x_n \quad (2.1)$$

for any n-vector x => this function gives the inner product of its n-vector argument x with some (fixed) n-vector a.

Superposition and Linearity. (2.1) satifies the property

$$f(\alpha x + \beta y) = a^T(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad (2.2)$$

for all n-vector x, y, and all scalars $\alpha, \beta \Rightarrow$ the property is called *superposition* and a function f is called *linear*. If a function f is linear, superposition extends to linear combinations of anu number of vectors, and not just linear combinations of two vectors:

$$f(\alpha_1 x_1 + ... + \alpha_k x_k) = \alpha_1 f(x_1) + ... + \alpha_k f(x_k)$$

for any n vector $x_1, ..., x_k$ and any scalar $\alpha_1, ..., \alpha_k$. A function f: $R^n \to R$ is linear if it satisfies the following properties:

•*Homogeneity.* $f(\alpha x) = \alpha f(x)$

•*Additivity.* $f(x + y) = f(x) + f(y)$

Inner product representation of a linear function. inner product of its argument with some fixed vector is linear $\Leftrightarrow$ a function is linear, then function can be expressed as the inner product of its argument with some fixed vector. (2.2) holds for all n-vectors x, y and all scalars α, β => $\exists a$, which is an n-vector, such that $f(x) = a^T x$ for all x => a^T is the inner product representation of f. To see this, using linear combination of unit vector: $x = e_1 x_1 + ... + e_n x_n$. If f is linear, then by multi-term superposition:

$$f(x) = f(x_1 e_1 + ... + x_n e_n) = x_1 f(e_1) + ... + x_n f(e_n) = a^T x \text{ (2.3)},$$

with $a = (f(e_1), ..., f(e_n))$ (UNIQUE $\Leftrightarrow$ only one vector a for all x)

EXAMPLE

•*Average.* **avg**(x) $= \overline{x} = f(x) = (x_1 + ... + x_n)/n$

•*Maximum.* $f(x) = max(x_1 + ... + x_n)$ is not a linear function (except when n=1)

Affine functions. f: $R^n \to R$ is *affine* function if and only if can be expressed as $f(x) = a^T x + b$ for some n-vector a and scalar b (constant, called *offset*).
=>Any affine scalar-valued function $f(x)$ will satisfies "restricted superposition" property for all n-vector x,y and all scalars α, β that satisfy $\alpha + \beta = 1$ $\Leftrightarrow$ any scalar-valued function that satisfies the restricted superposition will be affine,

$$f(\alpha x + \beta y) = a^T(\alpha x + \beta y) + b = \alpha a^T x + \beta a^T y + (\alpha + \beta)b =$$
$$\alpha(a^T x + b) + \beta(a^T x + b) = \alpha f(x) + \beta f(x)$$

(In the second "=" we use $\alpha + \beta = 1$).
An analog of the formular (2.3) is

$$f(x) = f(0) + x_1(f(e_1) - f(0)) + ... + x_n(f(e_n) - f_(0)),$$

which holds when f is affine, and x is any n-vector.
For an affine function, we know $n + 1$ numbers $f(0), f(e_1), ..., f(e_n)$, we can predict (or reconstruct or evaluate) $f(x)$ for any n-vector x and show how the n-vector a and constant b in the representation $f(x) = a^T x + b$ can be found

from the function $f : a_i = f(e_i) - f(0)$, and $b = f(0)$.
In some contexts, affine functions are called linear when x is a scalar, perhaps because its graph is line. When $\beta \neq 0$, f is not a linear function of x (figure 2.1)

a civil engineering example (page 33). an interesting example which is easy to understand how valued-scalar function $f(x) = a^T x$ works in reality :)

2.2. Taylor approximation

f: $R^n \to R$ is differentiable, which means that its partial derivatives exist. Let z be an n-vector. The (first-order) *Taylor approximation* of f near (or at) the point z is the function $\widetilde{f}(x)$ of x defined as

$$\widehat{f}(x) = f(z) + \frac{\partial f}{\partial x_1}(z)(x_1 - z_1) + ... + \frac{\partial f}{\partial x_1}(z)(x_n - z_n),$$

where $\frac{\partial f}{\partial x_i}(z)$ denotes the partial derivetive of f with respect to its ith argument, evaluated at the n-vector z. Sometimes $\widehat{f}$ is written with a second vector argument, as $\widehat{f}(x; z)$, to show the point z at which the approximation is developed. The first term in the Taylor approximation is a constant the other terms can be interpreted as the contributions to the change in the function value (from $f(z)$) due to changes in the components of x (from z).
Evidently, $\widehat{f}$ is an affine function of x (called the linear approxiation of f near z. Written using inner product notation as

$$\widehat{f}(x) = f(z) + \nabla f(z)^T (x - z) \text{ (2.4)},$$

where $\nabla f(z)$ is an n-vector, the *gradient* of f (at the point z),

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(z) \\ . \\ . \\ . \\ \frac{\partial f}{\partial x_n}(z) \end{bmatrix}$$

The first term in the Taylor approximation (2.4) is the constant $f(z)$, which is the value of the function when $x = z$. The second term is the inner product of the gradient of f at z and the *deviation* or *perturbation* of x from z, i.e., $x - z$

2.3. Regression model

The affine function of x given by

$$\widehat{y} = x^T \beta + v,$$

where β is an n-vector and v is a scalar, is called a *regression model*. The entries of x are called *regressors*, and $\widehat{y}$ is called the *prediction*, since the regression model is typically an approximation or prediction of some true value y, which is called the *dependent variable, outcome, or label*.
The vector β is called *weight vector* or *coefficient vector*, the scalar v is called *offset* or *intercept* in the regression model. Together, β and v are called the *parameters* in the regression model. $\widehat{y}$ is to emphasize that it's an *estimate* or *prediction* of some outcome y
β_i is the amount by which $\widehat{y}$ increase (if $\beta_i > 0$) when feature i increase by one (with all other features the same). If β_i is small, the prediction $\widehat{y}$ doesn't depend too strongly on feature i. The offset v is the value of $\widehat{y}$ when all features have the value 0
An example for the regression model with its features are Boolean, for example, the lifespan of a person in some group:

$$\widehat{y} = (x_1, x_2, x_3)^T(\beta_1, \beta_2, \beta_3) + v$$

where:
Features/ attributes x of a person in the model:
$x_1 = 1$ or 0 is a male or female, respectively
$x_2 = 1$ or 0 is the person has type II diabetes or not, respectively
$x_3 = 1$ or 0 is the person smokes cigarettes or not, respectively
Entries of the coefficient β in the model:
β_1 is the increase in estimated lifespan if the person is male
β_2 is the increase in estimated lifespan if the person is diabetic
β_3 is the increase in estimated lifespan if the person smokes cigarettes.
Offset:
v is the regression model estimate for the lifespan of a female nondiabetic non-smoker
=> If the model fits real data, β_i would be negative, meaning that they decrease the regression model estimate of lifespan.

House price regression model.(Pages 39-41)

III. Norm and distance

3.1. Norm

The *Euclinde norm* of an n-vector x denoted $\|x\|$:

$$\|x\| = \sqrt{x_1^2 + ... + x_n^2}$$

can also be expressed as $\|x\| = \sqrt{x^T x}$. Sometimes written with a subscript 2 as $||x||_2$, the subscript 2 means entries x are raised to the second power.
A vector is *small* if its norm is a small number (numerical values of the norm that qualify for small or large depend on the particular application and context.)

Properties of norm. x and y are vectors of same size and β is a scalar.

•*Nonnegative homogeneity.* $||\beta x|| = |\beta|||x||$

•*Triangle inequality.* $||x + y|| \leq ||x| + ||y||$. *Another name is subadditivity.*

•*Nonnegative.* $||x|| \geq 0$

•*Definiteness.* $||x|| = 0$ *if* $x = 0$
The last two properties together are called *positive definiteness*

General norm. Any real-valued function of an n-vector that satisfies the four properties listed above is called a (general) norm.

Root-mean-square value (RMS).

$$\textbf{rms}(\text{x}) = \sqrt{\frac{x_1^2 + ... + x_n^2}{n}} = \frac{||x||}{\sqrt{n}}$$

The RMS value of a vector x is useful when comparing norms of vectors with different dimensions; tell us what a 'typical' value of entry $|x_i|$ is. For example, the norm of 1_n, $||x|| = \sqrt{n}$, but its RMS value is 1, independent of n. If all the entries of a vector are the same, k, then the RMS value of the vector is k.

Norm of a sum.

$$||x + y|| = \sqrt{||x||^2 + 2x^T y + ||y||^2}$$

Norm of block vectors. d = [a,b,c] (where a, b, c are vectors), we have:

$$||[a, b, c]|| = \sqrt{||a||^2 + ||b||^2 + ||c||^2}$$
$$<=> ||d||^2 = d^T d = a^T a + b^T b + c^T c = ||a||^2 + ||b||^2 + ||c||^2$$

Chebyshev inequality. x is an n-vector and k of its entries satisfy $|x| \geq a$, where $a \geq 0$. Then k of its entries satisfy $x^2 i \geq a^2$

$$||x||^2 = x_1^2 + ... + x_n^2 \geq ka^2,$$

since k of the numbers in the sum are at least a^2, and the other $n - k$ numbers are nonnegative => $k \leq \frac{||x||^2}{a^2}$, which called the *Chebyshev inequality.* No entry of a vector can be larger in magnitude than the norm of the vector.
The inequality is easier for us to interpret in term of the RMS value of a vector:

$$\frac{k}{n} = (\frac{\mathbf{rms}(x)}{a})^2,$$

where k is the number of entries of x with absolute value at least a. Left-hand side is the fraction of entries of the vector are at least a in absolute value. The right-hand side is the inverse square of the ratio of a to **rms**(x) =¿ the inequality states that not too many of the entries of a vector can be much bigger (in absolute value) than its RMS value. (converse statement: At least one entry of a vector has absolute value as large as the RMS value of the vector).

3.2. Distance

Euclidean distance. between two vectors a and b as the norm of their difference:

$$\mathbf{dist}(\text{a,b}) = ||a - b||$$

holds for vectors of any dimension. If a and b are n-vectors, RMS value of difference, $\frac{||a-b||}{\sqrt{n}}$, as the *RMS deviation* between the two vectors.
The distance between x and y is small <=> 'close' or 'nearby' and large <=> 'far'.

Triangle inequality. a triangle in two or three dimensions, whose vertices have coordinates a, b and c. The lengths of the sides are the distances between the vertices,

$$\mathbf{dist}(\text{a,b}) = ||a - b||, \mathbf{dist}(\text{b,c}) = ||b - c||, \mathbf{dist}(\text{a,c}) = ||a - c||$$

figure 3.1 tell us the length of any side of a triangle cannot exceed the sum of the lengths of the other two sides

$$||a - c|| \leqslant ||a - b|| + ||b - c||$$

EXAMPLE.

•*Feature distance.* If x and y represent vectors of n features of two objects, $||x - y||$ is called the *feature distance*, and gives a measure of how different the objects are (in term of their feature values)

•*RMS prediction error.* Let y be n-vector, then the difference $y - \widehat{y}$ is called the *prediction error* and its RMS value $\mathbf{rms}(y - \widehat{y})$ is called the *RMS prediction error*. If its value is small (compared to **rms**(y)) the prediction is good

•*Nearest neighbor.* Let $z_1, ..., z_m$ be m vectors, and that x is another n-vector. We say z_j is the *nearest neighbor* of x (among $z_1, ..., z_m$) if

$$||x - z_j|| \leq ||x - z_i||, i = 1, ..., m. \text{ (Figure 3.2)}$$

Units for heterogeneous vector entries. If we want the different entries to have approximately equal status in determining distance, their numerical values should be approximately of the same magnitude. For this reason units for different entries in vectors are often chosen in such a way that their typical numerical value are similar in magnitude, so that the different entries play similar roles in determining distance. (an example is in pages 51-52)

3.3. Standard deviation

For any n-vector x, the vector $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ is called the associated de-meaned vector. The de-meaned vector is useful for understanding how the entries of a vector deviate from their mean value. It is zero if all the entries in the original vector x are the same.
The standard deviation of an n-vector x is defined as the RMS value of the de-meaned vector $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$

$$\mathbf{std}(\text{x}) = \sigma = \sqrt{\frac{(x_1 - \mathbf{avg}(x))^2 + ... + (x_n - \mathbf{avg}(x))^2}{n}} = \frac{||x - (\frac{1^T x}{n})\mathbf{1}||}{\sqrt{n}}$$

The standard deviation of a vector x tells us the typical amount by which its entries deviate from their average value. The standard value is small when the entries of the vector are nearly the same.

Average, RMS value, and standard deviation. the relationship of them is

$$\mathbf{rms}(x)^2 = \sigma^2 + \mu^2$$

EXAMPLES Shown in page 54.

Chebyshed inequality for standard deviation. The Chebyshed inequality can be expressed in term of the mean and standared deviation: If k is the number of entries of x that satisfy $|x_i - \mathbf{avg}(x)| \geq a$, then $\frac{k}{n} \leq (\frac{\mathbf{std}(x)}{a})^2$

Properties of standard deviaiton.
•*Adding a constant.* For any vector x and any number a, $\mathbf{std}(x + a) = \mathbf{std}(x)$

•*Multiplying by a scalar.* For any vector x and any number a, $\mathbf{std}(ax) = |a|\mathbf{std}(x)$, $|a|$ is the absolute value of a

Standardization. For any vector x, $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ as the de-meaned version of x, since it has mean value zero. If devide by the RMS value of $\tilde{x}$:

$$\text{z} = \frac{1}{\mathbf{std}(x)}(x - \mathbf{avg}(x)\mathbf{1}) = \frac{\tilde{x}}{\mathbf{std}(x)}$$

This vector is called the *standardized* version of x. It has mean zero, and standard deviation one. Its entries are called the *z-scores* associated with the original emtries of x
For example, If an n-vector x gives the values of some medical test of n patients admitted to a hospital, the standardized values of z-scores tell us how high or low, compared to the population, that patient's value is. A value $z_6 = -3.2$ means that patient 6 has a very low value of the measurement; whereas $z_{22} = 0.3$ says that patients 22's value is quite close to the average value.

3.4. Angle

Cauchy-Schwarz inequality. For any n-vectors a and b

$$|a^T b| < ||a||\ ||b||$$

Verification of triangle inequality. Let a and b be any vectors

$$||a + b|| \leq (||a|| + ||b||)^2$$

Angle between vectors. the *angle* between two non-zero vectoes a, b is defined as

$$\theta = \angle(a, b) = \arccos(\frac{a^T b}{||a||.||b||})$$

where arccos denotes the inverse cosine, normalized to lie in the interval $[0, \pi]$ => θ as the unique number between 0 and *pi*

Properties.
•*symmetric function* $\angle(a, b) = \angle(b, a)$

•$\angle(\alpha a, \beta b) = \angle(a, b)$

Acute and obtuse angles.
•$\angle(a, b) = \frac{\pi}{2} = 90^o$, i.e $a^T b = 0$ or $a \perp b$ => a and b are *orthogonal* (a zero vector is orthogonal to any vector)

•$\angle(a, b) = 0^o$, i.e $a^T b = ||a||.||b||$ => a and b are *aligned*

•$\angle(a, b) = 180^o$, i.e $a^T b = -||a||.||b||$ => a and b are *anti-aligned*

•$\angle(a,b) < \frac{\pi}{2} = 90^o$, i.e $a^Tb > 0$ => a and b make *acute angle*

•$\angle(a,b) > \frac{\pi}{2} = 90^o$, i.e $a^Tb < 0$ => a and b make *obtuse angle*

EXAMPLES Shown in page 58.

Norm of sum via angles. For vectors x and y

$$||x+y||^2 = ||x||^2 + 2x^Ty + ||y||^2 = ||x||^2 + 2||x||.||y||cos\theta + ||y||^2$$

Where $\theta = \angle(x,y)$. Making several observations.
•x and y are aligned => $||x+y|| = ||x|| + ||y||$

•x and y are orthogonal => $||x+y||^2 = ||x||^2 + ||y||^2$ => *Pythagorean theorem*

Correlation coefficient. a and b are n-vectors, with associated de-meaned vectors

$$\tilde{a} = a - \mathbf{avg}(a)\mathbf{1} \ ; \ \tilde{b} = b - \mathbf{avg}(b)\mathbf{1}$$

These de-meaned vectors are not zero, we define their *correlation coefficient* as

$$\rho = \frac{\tilde{a}^T\tilde{b}}{||\tilde{a}||.||\tilde{b}||}$$

=> $\rho = cos\theta$, where $\theta = \angle(\tilde{a},\tilde{b})$. Also, with $u = \frac{\tilde{a}}{\mathbf{std}(a)}$ and $v = \frac{\tilde{b}}{\mathbf{std}(b)}$. We have

$$\rho = \frac{u^Tv}{n}$$

(We use $||u|| = ||v|| = n$).
•The Cauchy-Schwarz inequality tells us the correlation coefficient ranges between [-1, 1] (or expressed as %).
•The correlation coefficient tells us how the entries in the two vectors vary together.
+ High correlation (say, $\rho = 80\%$) => typically $a_i > \frac{a_i}{n}$ and $b_i > \frac{b_i}{n}$, $i = (1,...,n)$
+ $\rho = 1$ => $\tilde{a}$ and $\tilde{b}$ are aligned. <=> each is a positive multiple of the other
+ $\rho = 0$ => the vectors are *uncorrelated* <=> a vector with all entries equal is uncorrelated with any vector.
+ $\rho = -1$ => $\tilde{a}$ and $\tilde{b}$ are negative multiples of each other
•The correlation coefficient is often used when the vectors represent time series. For example, let a and b be the two time series of rainfall in two nearby locations over some time interval. We expect they are highly correlated (say, $\rho > 0.8$) at

the same time.

Standard deviation of sum.

$$\textbf{std}(\text{a+b}) = \sqrt{\textbf{std}(a)^2 + 2\rho\textbf{std}(a)\textbf{std}(b) + \textbf{std}(b)^2}$$

+ $\rho = 1 \Rightarrow \textbf{std}(a+b) = \textbf{std}(a) + \textbf{std}(b)$
+ $\rho = 0 \Rightarrow a$ and b are uncorrelated $\Rightarrow \textbf{std}(a+b) = \sqrt{\textbf{std}(a)^2 + \textbf{std}(b)^2}$
+ $\rho = -1 \Rightarrow \textbf{std}(a+b) = |\textbf{std}(a) - \textbf{std}(b)|$

Units for heterogeneous vector entries. The general rule of thumb is to choose units for different entries so the typical vector entries have similar sizes or ranges of values.

IV. Clustering

4.1. Clustering

We have p n-vectors, $x_1, .., x_p$. The goal of *clustering* is to group or partition the vectors into k groups or clusters, with the vectors in each group close to each other. Normally, $k < p$

EXAMPLES
•*Patient clustering.* Let x_i be feature vectors associated with p patients admitted to a hospital, a clustering algorithm partitions the p patients into k groups of similar patients

**Other Examples can be encountered in pages 70,71*

4.2. A clustering objective

Specifying the cluster assignments. by saying which cluster each vector belongs to. 2 ways to express:
+ We label the groups $1, ..., k$, and specify a clustering or assignment of the p given vectors to groups using an p-vector c_i where c_i is the group (number) that the vector x_i is assigned to. Example: $p = 5$ vectors, $k = 3$ groups, c=(3,1,1,1,2) $\Rightarrow$ x_1 is assigned to group 3, x_2, x_3, x_4 is assigned to group 1 and x_5 is assigned to group 2.
+ Let G_j be the set of indices corresponding to group j

$$G_j = \{i | c_i = j\} \Rightarrow \text{all indices } i \in \text{Group } j$$

Above example: $G_1 = \{2, 3, 4\}$, $G_2 = \{5\}$, $G_3 = \{1\}$

Group representatives.each of groups we associate a *group representatives*

n-vectors, denote $z_1, ..., z_k$. These representatives can be any n-vetors: not need to be one of the given vectors.