2. Norm, distance, angle

- norm
- distance
- k-means algorithm
- angle
- hyperplanes
- complex vectors

Euclidean norm

(Euclidean) norm of vector $a \in \mathbf{R}^n$:

$$\begin{aligned} \|a\| &= \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} \\ &= \sqrt{a^T a} \end{aligned}$$

- if $n = 1$, $\|a\|$ reduces to absolute value $|a|$
- measures the magnitude of a
- sometimes written as $\|a\|_2$ to distinguish from other norms, *e.g.*,

$$\|a\|_1 = |a_1| + |a_2| + \cdots + |a_n|$$

Properties

Positive definiteness

$$\|a\| \geq 0 \quad \text{for all } a, \qquad \|a\| = 0 \quad \text{only if } a = 0$$

Homogeneity

$$\|\beta a\| = |\beta| \|a\| \quad \text{for all vectors } a \text{ and scalars } \beta$$

Triangle inequality

$$\|a + b\| \leq \|a\| + \|b\| \quad \text{for all vectors } a \text{ and } b \text{ of equal length}$$

(proof on page 2-7)

Cauchy-Schwarz inequality

$$|a^T b| \leq \|a\| \|b\| \quad \text{for all } a, b \in \mathbf{R}^n$$

moreover, equality $|a^T b| = \|a\| \|b\|$ holds if:

- $a = 0$ or $b = 0$; in this case $a^T b = 0 = \|a\| \|b\|$
- $a \neq 0$ and $b \neq 0$, and $b = \gamma a$ for some $\gamma > 0$; in this case

$$0 < a^T b = \gamma \|a\|^2 = \|a\| \|b\|$$

- $a \neq 0$ and $b \neq 0$, and $b = -\gamma a$ for some $\gamma > 0$; in this case

$$0 > a^T b = -\gamma \|a\|^2 = -\|a\| \|b\|$$

Proof of Cauchy-Schwarz inequality

1. trivial if $a = 0$ or $b = 0$

2. assume $\|a\| = \|b\| = 1$; we show that $-1 \leq a^T b \leq 1$

$$\begin{array}{rcl} 0 & \leq & \|a-b\|^2 \\ & = & (a-b)^T(a-b) \\ & = & \|a\|^2 - 2a^Tb + \|b\|^2 \\ & = & 2(1-a^Tb) \end{array}$$

with equality only if $a = b$

$$\begin{array}{rcl} 0 & \leq & \|a+b\|^2 \\ & = & (a+b)^T(a+b) \\ & = & \|a\|^2 + 2a^Tb + \|b\|^2 \\ & = & 2(1+a^Tb) \end{array}$$

with equality only if $a = -b$

3. for general nonzero a, b, apply case 2 to the unit-norm vectors

$$\frac{1}{\|a\|}a, \quad \frac{1}{\|b\|}b$$

Average and RMS value

let a be a real n-vector

- the *average* of the elements of a is

$$\operatorname{avg}(a) = \frac{a_1 + a_2 + \cdots + a_n}{n} = \frac{\mathbf{1}^T a}{n}$$

- the *root-mean-square* value is the root of the average squared entry

$$\operatorname{rms}(a) = \sqrt{\frac{a_1^2 + a_2^2 + \cdots + a_n^2}{n}} = \frac{\|a\|}{\sqrt{n}}$$

Exercises

- show that $|\operatorname{avg}(a)| \leq \operatorname{rms}(a)$
- show that average of $b = (|a_1|, |a_2|, \ldots, |a_n|)$ satisfies $\operatorname{avg}(b) \leq \operatorname{rms}(a)$

Triangle inequality from Cauchy-Schwarz inequality

for vectors a, b of equal size

$$\begin{aligned}
\|a+b\|^2 &= (a+b)^T(a+b) \\
&= a^Ta + b^Ta + a^Tb + b^Tb \\
&= \|a\|^2 + 2a^Tb + \|b\|^2 \\
&\leq \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \qquad \text{(by Cauchy-Schwarz)} \\
&= (\|a\| + \|b\|)^2
\end{aligned}$$

- taking squareroots gives the triangle inequality
- triangle inequality is an equality if and only if $a^Tb = \|a\|\|b\|$ (see page 2-4)
- also note from line 3 that $\|a+b\|^2 = \|a\|^2 + \|b\|^2$ if $a^Tb = 0$

Outline

- norm
- **distance**
- k-means algorithm
- angle
- hyperplanes
- complex vectors

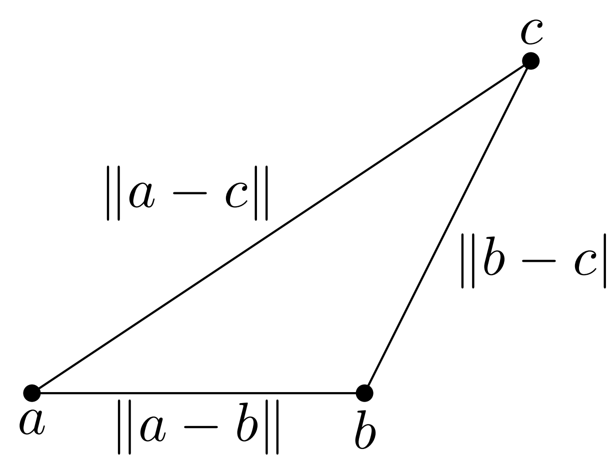
Distance

the (Euclidean) distance between vectors a and b is defined as $\|a - b\|$

- $\|a - b\| \geq 0$ for all a, b and $\|a - b\| = 0$ only if $a = b$

- triangle inequality

$$\|a - c\| \leq \|a - b\| + \|b - c\| \quad \text{for all } a, b, c$$



- RMS deviation between n-vectors a and b is $\mathrm{rms}(a - b) = \dfrac{\|a - b\|}{\sqrt{n}}$

Standard deviation

let a be a real n-vector

- the *de-meaned* vector is the vector of deviations from the average

$$a - \mathrm{avg}(a)\mathbf{1} = \begin{bmatrix} a_1 - \mathrm{avg}(a) \\ a_2 - \mathrm{avg}(a) \\ \vdots \\ a_n - \mathrm{avg}(a) \end{bmatrix} = \begin{bmatrix} a_1 - (\mathbf{1}^T a)/n \\ a_2 - (\mathbf{1}^T a)/n \\ \vdots \\ a_n - (\mathbf{1}^T a)/n \end{bmatrix}$$

- the *standard deviation* is the RMS deviation from the average

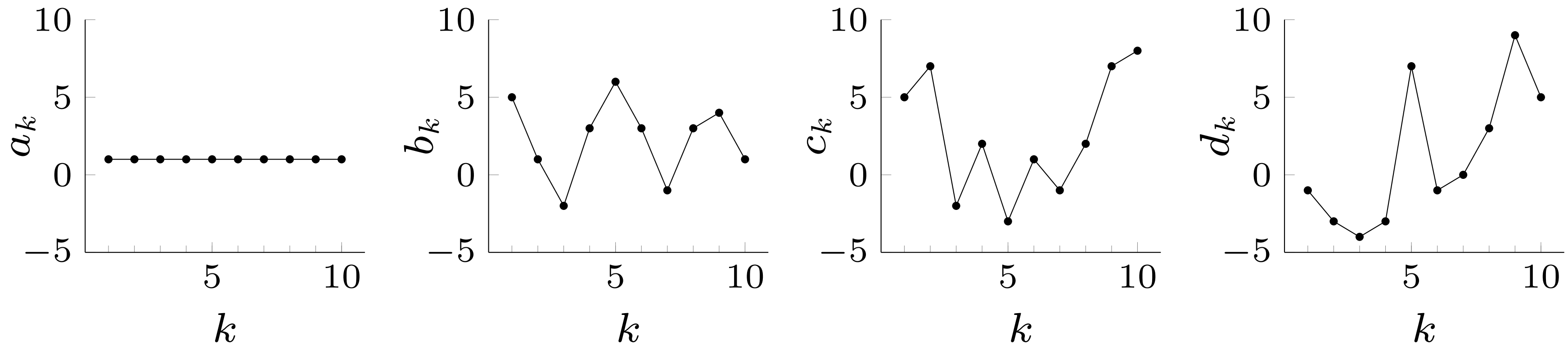
$$\mathrm{std}(a) = \mathrm{rms}(a - \mathrm{avg}(a)\mathbf{1}) = \frac{\left\| a - ((\mathbf{1}^T a)/n)\mathbf{1} \right\|}{\sqrt{n}}$$

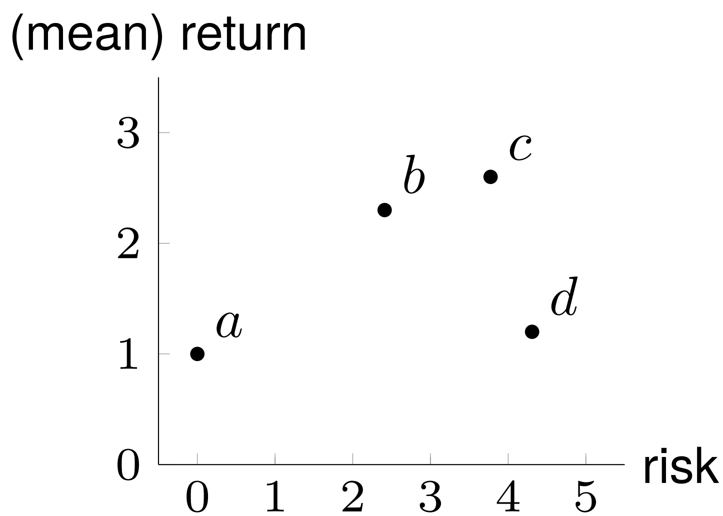
- the de-meaned vector in *standard units* is

$$\frac{1}{\mathrm{std}(a)}(a - \mathrm{avg}(a)\mathbf{1})$$

Mean return and risk of investment

- vectors represent time series of returns on an investment (as a percentage)
- average value is *(mean) return* of the investment
- standard deviation measures variation around the mean, *i.e.*, *risk*





Exercise

show that

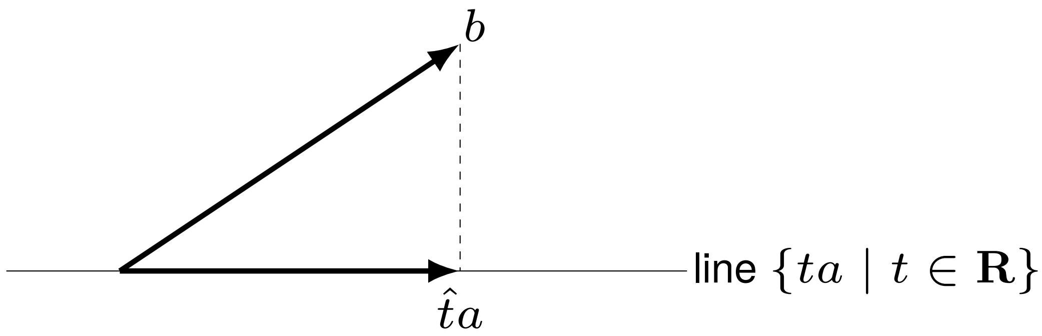
$$\operatorname{avg}(a)^2 + \operatorname{std}(a)^2 = \operatorname{rms}(a)^2$$

Solution

$$\begin{aligned}
\operatorname{std}(a)^2 &= \frac{\|a - \operatorname{avg}(a)\mathbf{1}\|^2}{n} \\
&= \frac{1}{n}\left(a - \frac{\mathbf{1}^T a}{n}\mathbf{1}\right)^T \left(a - \frac{\mathbf{1}^T a}{n}\mathbf{1}\right) \\
&= \frac{1}{n}\left(a^T a - \frac{(\mathbf{1}^T a)^2}{n} - \frac{(\mathbf{1}^T a)^2}{n} + \left(\frac{\mathbf{1}^T a}{n}\right)^2 n\right) \\
&= \frac{1}{n}\left(a^T a - \frac{(\mathbf{1}^T a)^2}{n}\right) \\
&= \operatorname{rms}(a)^2 - \operatorname{avg}(a)^2
\end{aligned}$$

Exercise: nearest scalar multiple

given two vectors $a, b \in \mathbf{R}^n$, with $a \neq 0$, find scalar multiple ta closes to b



Solution

- squared distance between ta and b is

$$\|ta - b\|^2 = (ta - b)^T(ta - b) = t^2 a^T a - 2ta^T b + b^T b$$

  a quadratic function of t with positive leading coefficient $a^T a$

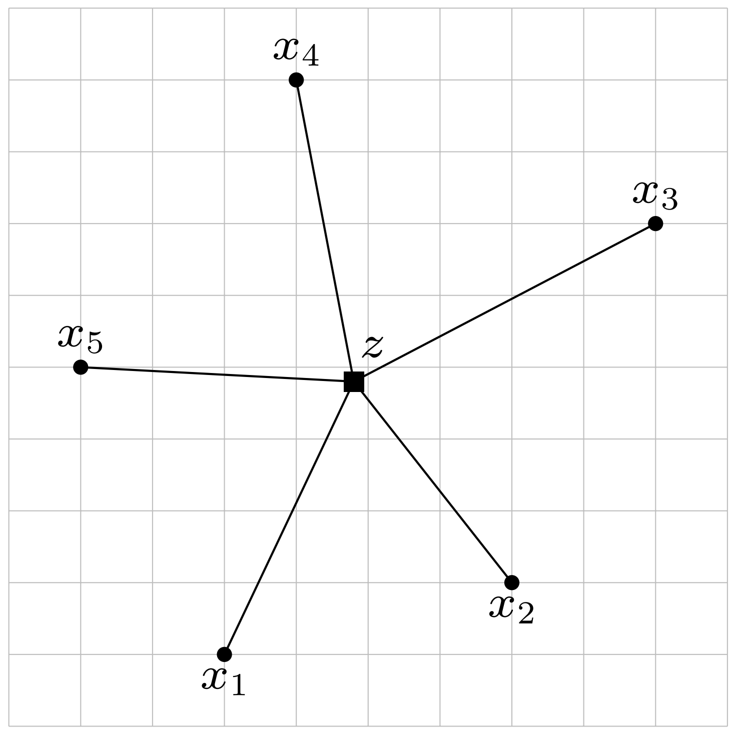
- derivative with respect to t is zero for

$$\hat{t} = \frac{a^T b}{a^T a} = \frac{a^T b}{\|a\|^2}$$

Exercise: average of collection of vectors

given N vectors $x_1, \ldots, x_N \in \mathbf{R}^n$, find the n-vector z that minimizes

$$\|z - x_1\|^2 + \|z - x_2\|^2 + \cdots + \|z - x_N\|^2$$



z is also known as the *centroid* of the points $x_1, \ldots, x_N$

Solution: sum of squared distances is

$$
\begin{aligned}
&\|z - x_1\|^2 + \|z - x_2\|^2 + \cdots + \|z - x_N\|^2 \\
&\quad = \sum_{i=1}^{n} \left((z_i - (x_1)_i)^2 + (z_i - (x_2)_i)^2 + \cdots + (z_i - (x_N)_i)^2 \right) \\
&\quad = \sum_{i=1}^{n} \left(N z_i^2 - 2 z_i \left((x_1)_i + (x_2)_i + \cdots + (x_N)_i \right) + (x_1)_i^2 + \cdots + (x_N)_i^2 \right)
\end{aligned}
$$

here $(x_j)_i$ is ith element of the vector x_j

- term i in the sum is minimized by

$$z_i = \frac{1}{N}((x_1)_i + (x_2)_i + \cdots + (x_N)_i)$$

- solution z is component-wise average of the points $x_1, \ldots, x_N$:

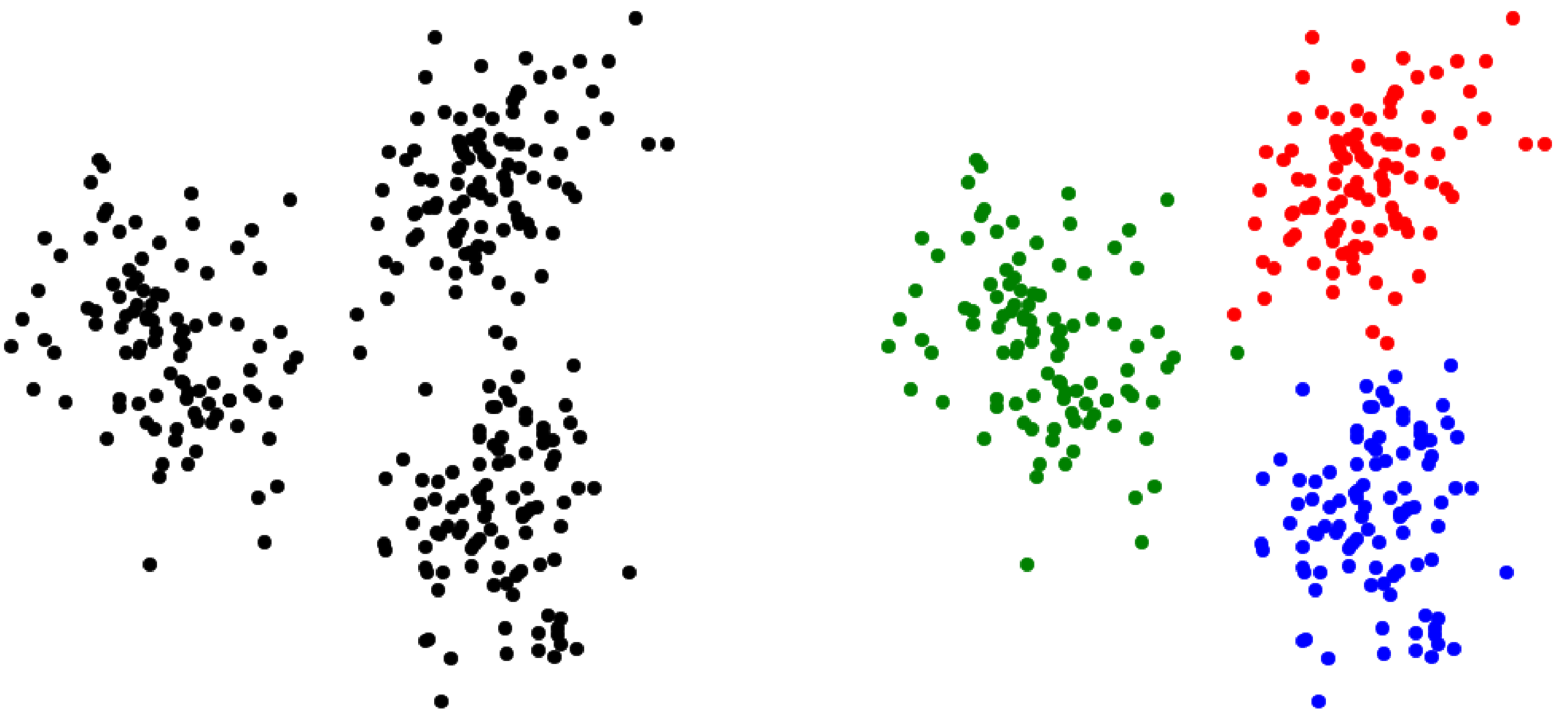
$$z = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)$$

Outline

- norm
- distance
- **k-means algorithm**
- angle
- hyperplanes
- complex vectors

k-means clustering

a popular iterative algorithm for partitioning N vectors $x_1, \ldots, x_N$ in k clusters



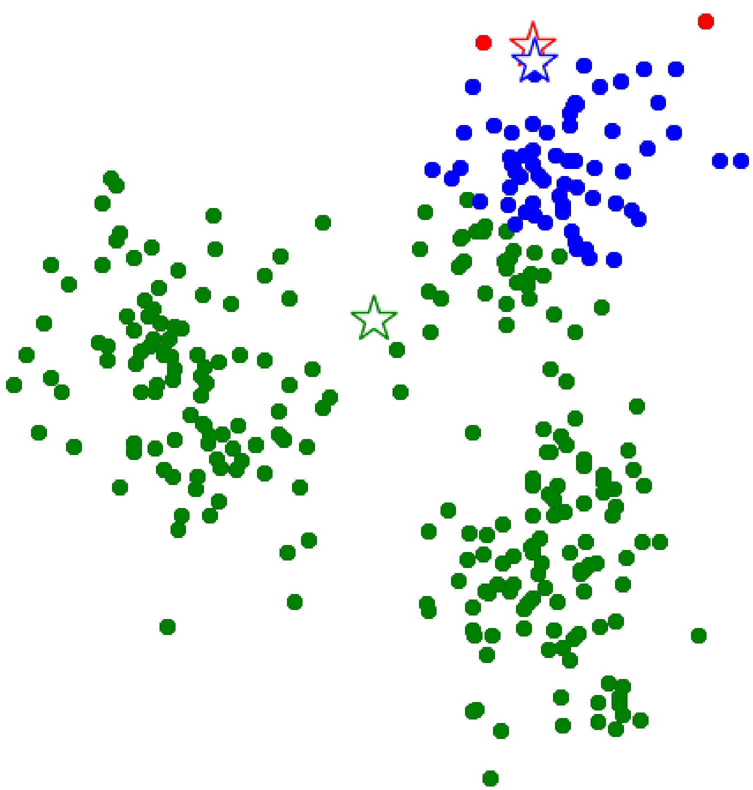
Algorithm

choose initial 'representatives' $z_1, \ldots, z_k$ for the k groups and repeat:

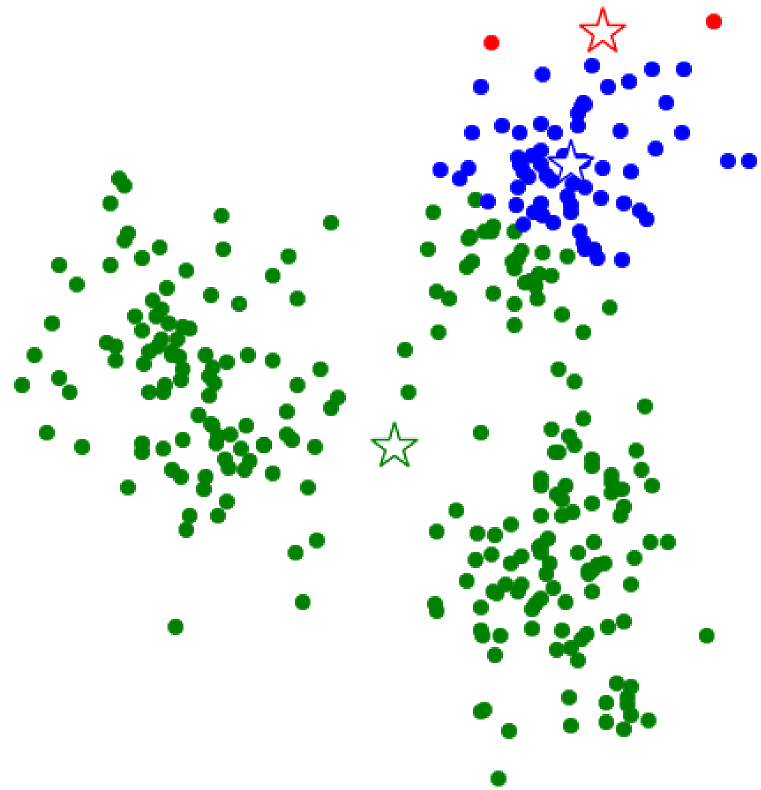
1. assign each vector x_i to the nearest group representative z_j
2. set the representative z_j to the mean of the vectors assigned to it

- as a variation, choose a random initial partition and start with step 2
- initial representatives are often chosen randomly
- solution depends on choice of initial representatives or partition
- can be shown to converge in a finite number of iterations
- in practice, often restarted a few times, with different starting points

Example: first iteration

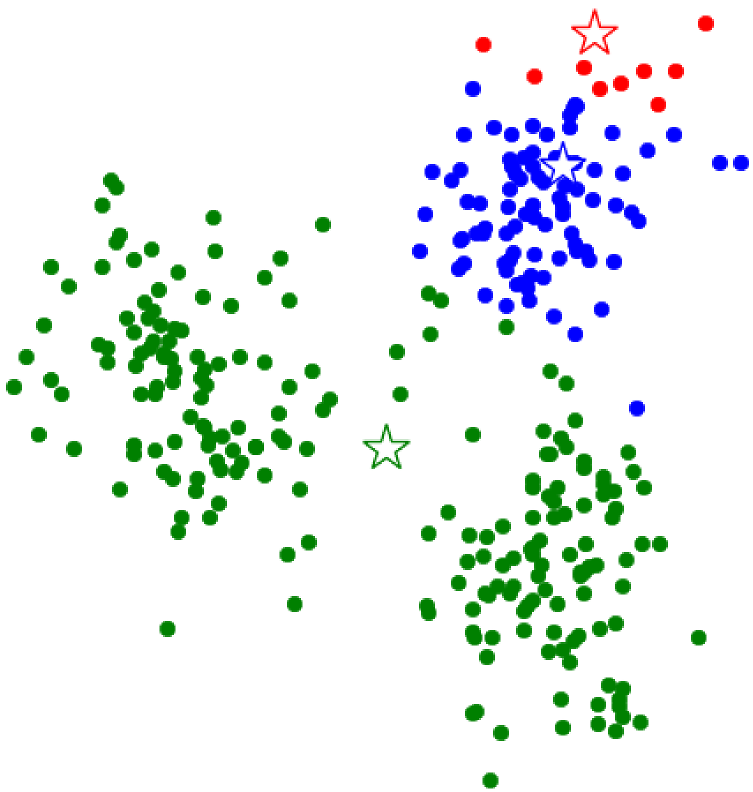


assignment to groups

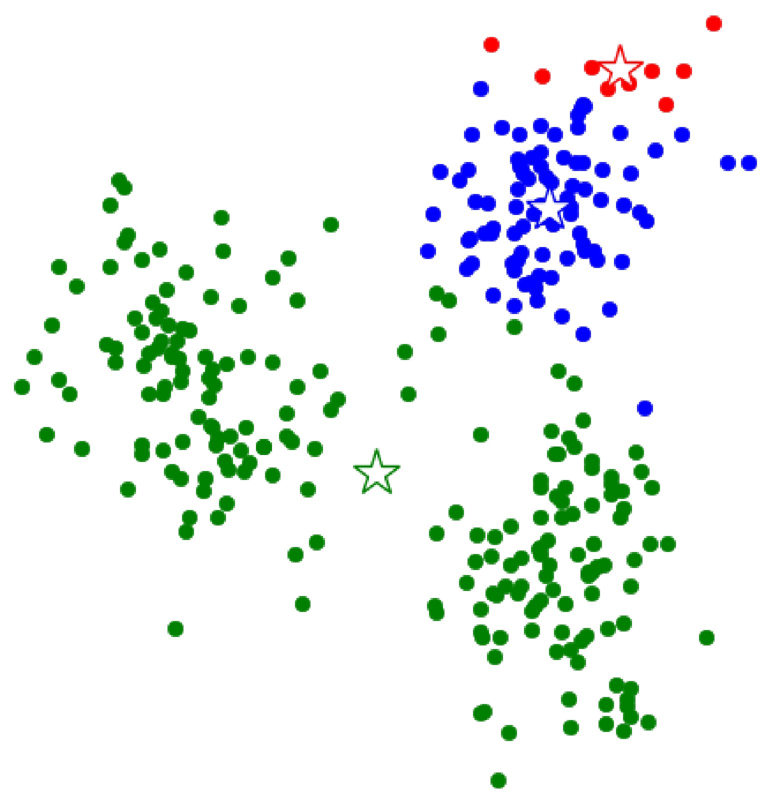


updated representatives

Example: iteration 2

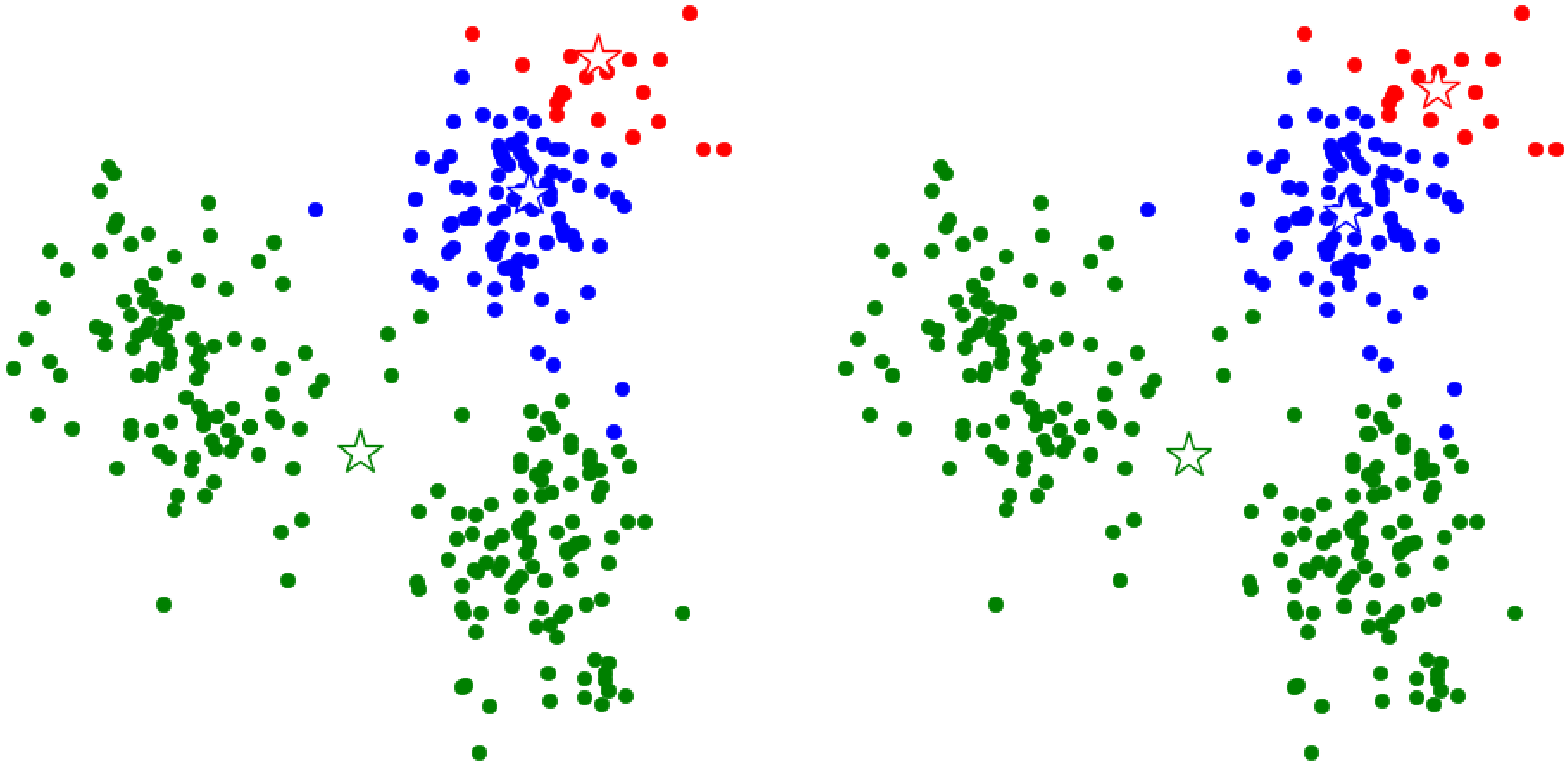


assignment to groups



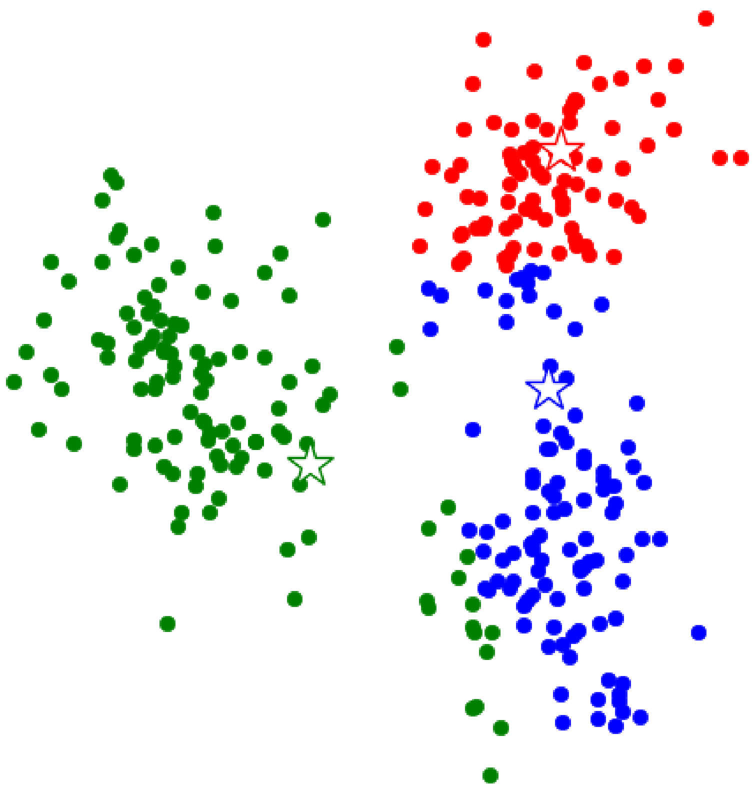
updated representatives

Example: iteration 3

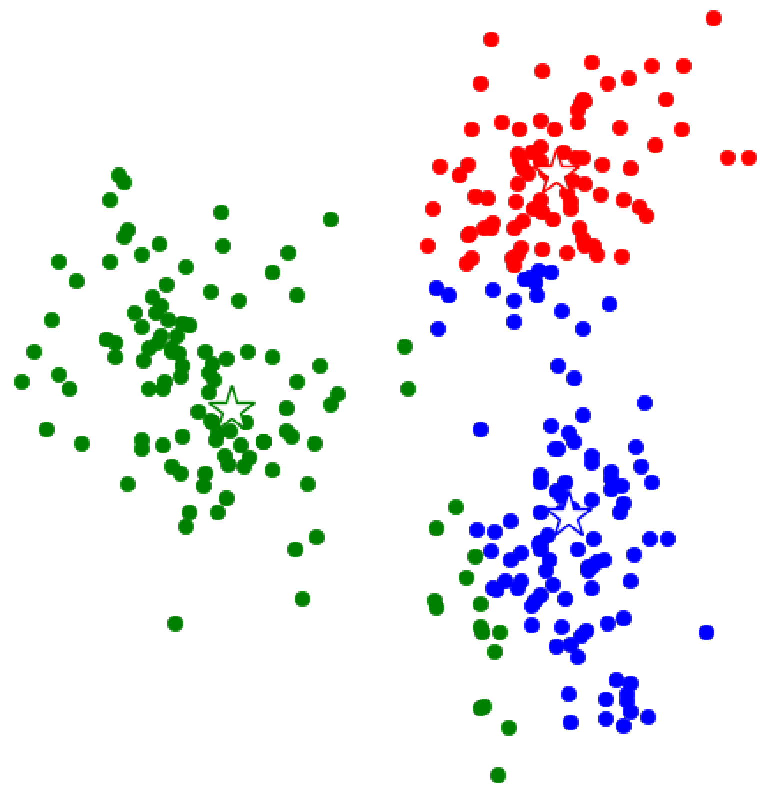


assignment to groups　　　　　updated representatives

Example: iteration 9

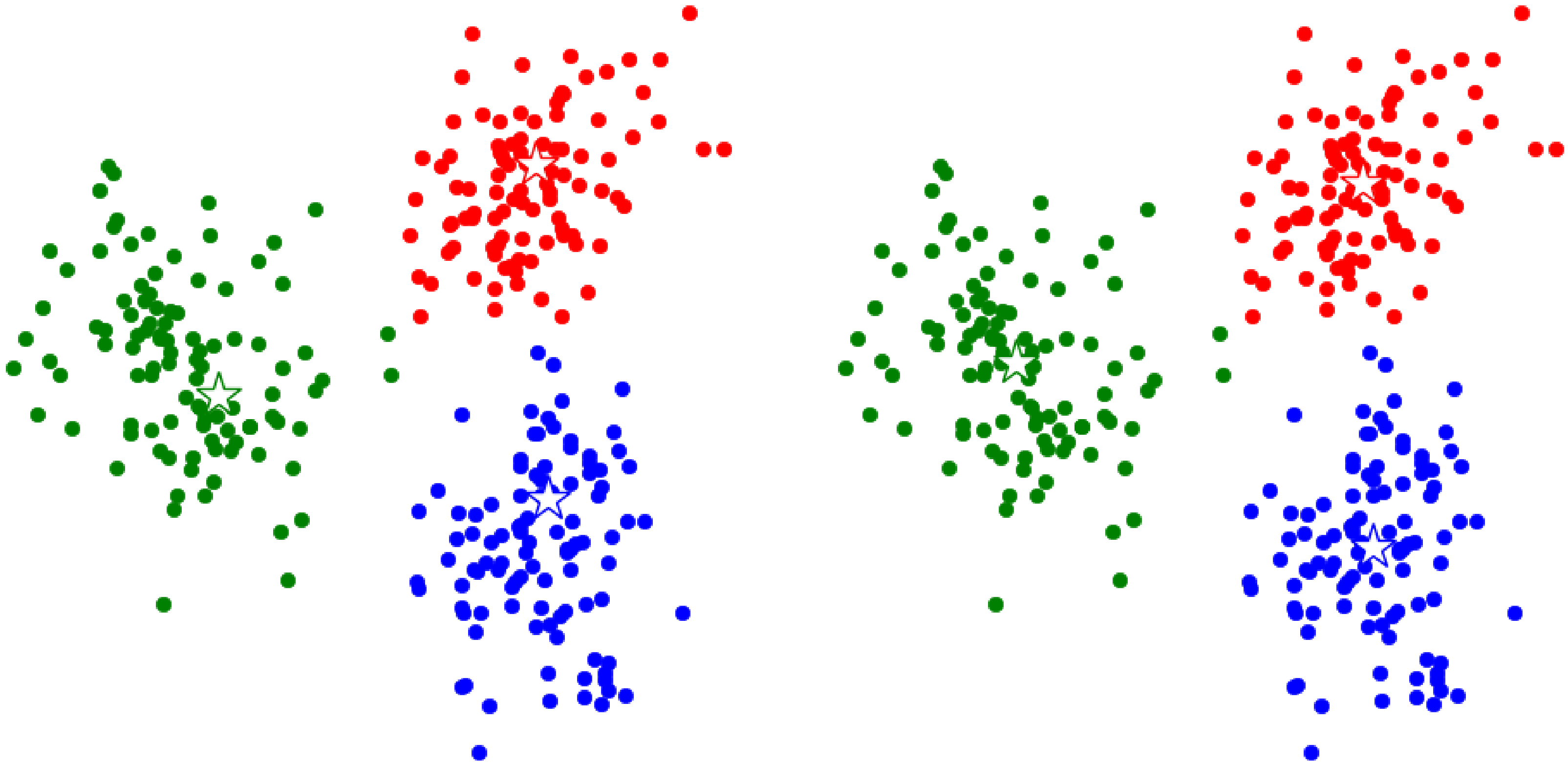


assignment to groups



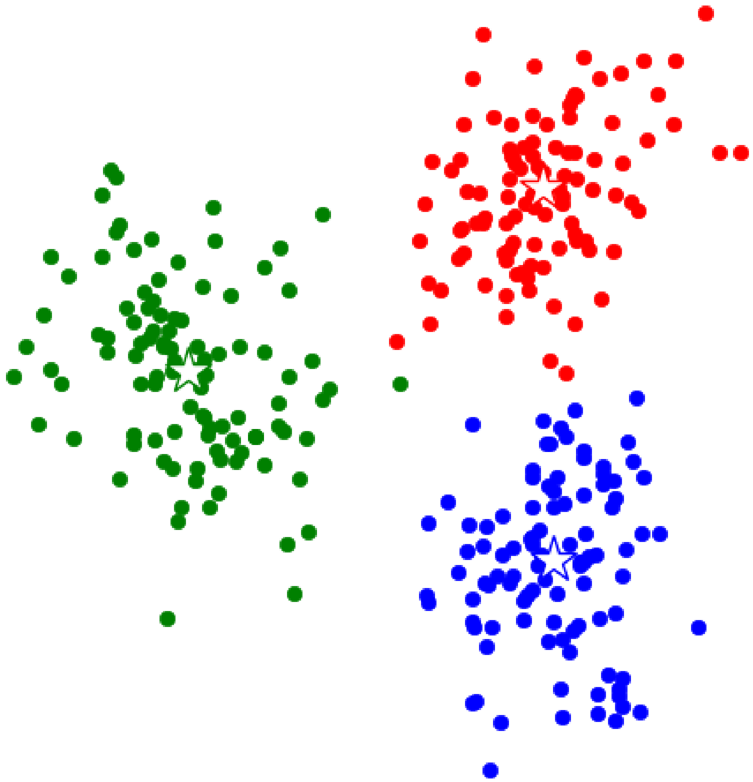
updated representatives

Example: iteration 10

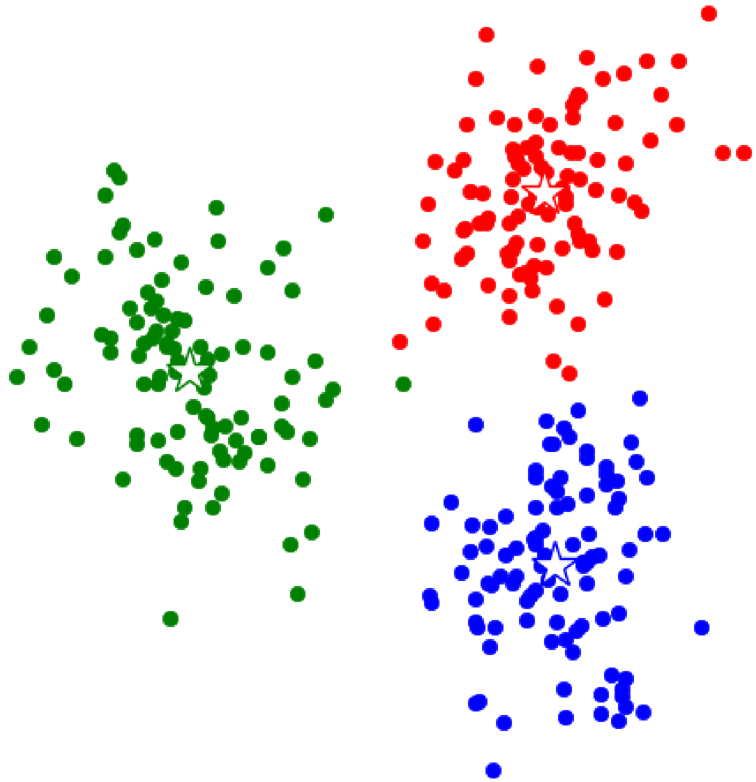


assignment to groups　　　updated representatives

Example: iteration 11

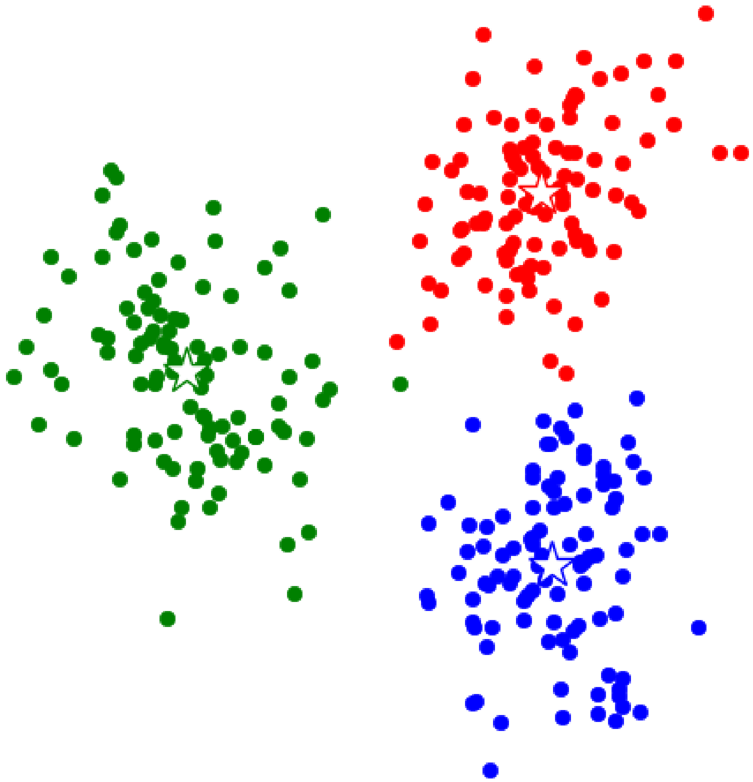


assignment to groups

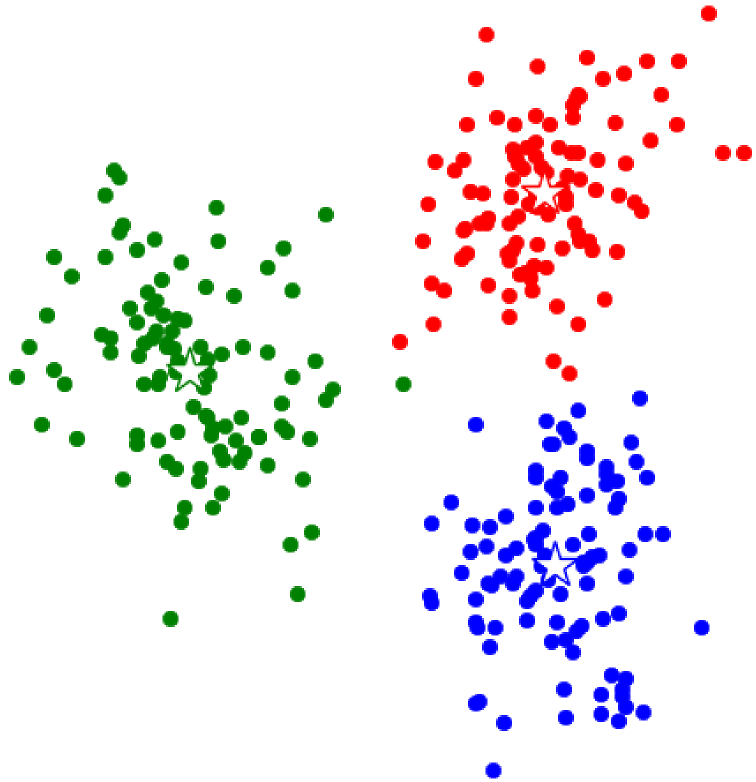


updated representatives

Example: iteration 12



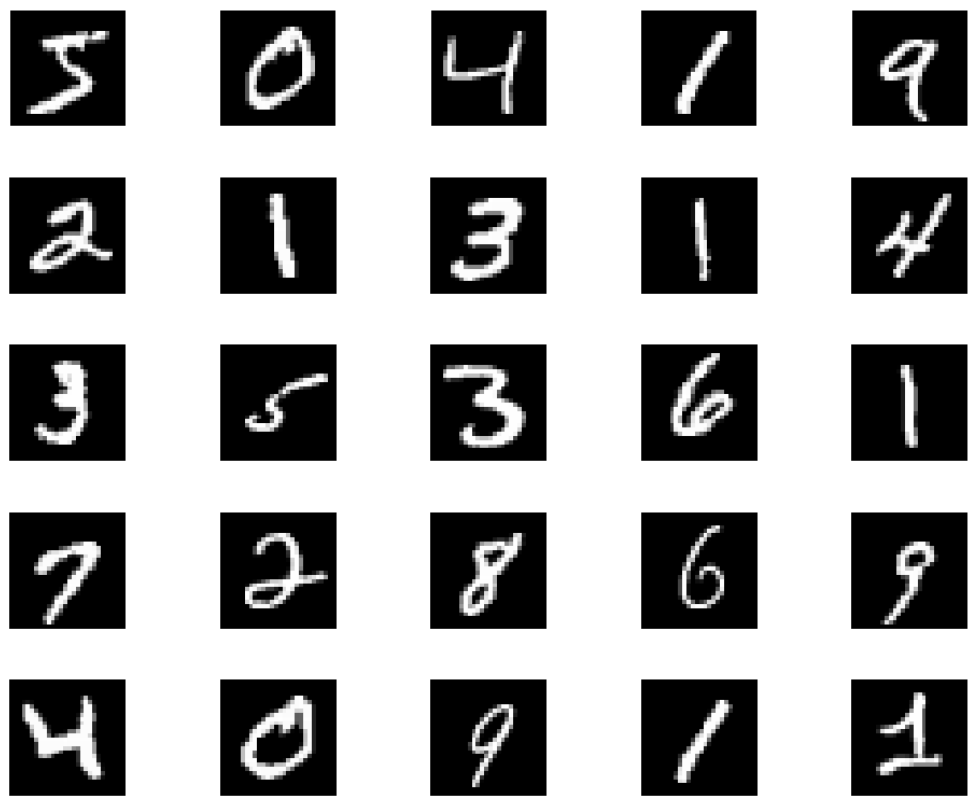
assignment to groups



updated representatives

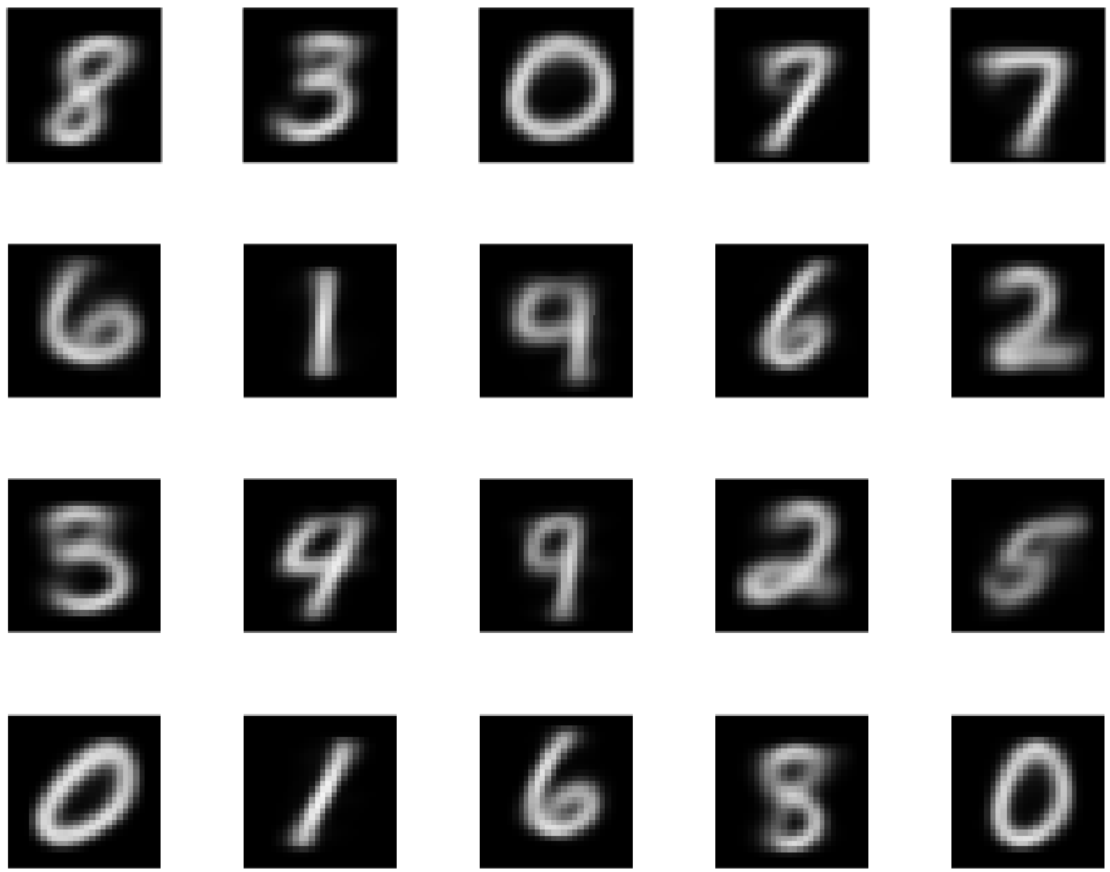
Image clustering

- MNIST dataset of handwritten digits
- $N = 60,000$ grayscale images of size 28×28 (vectors x_i of size $28^2 = 784$)
- 25 examples:



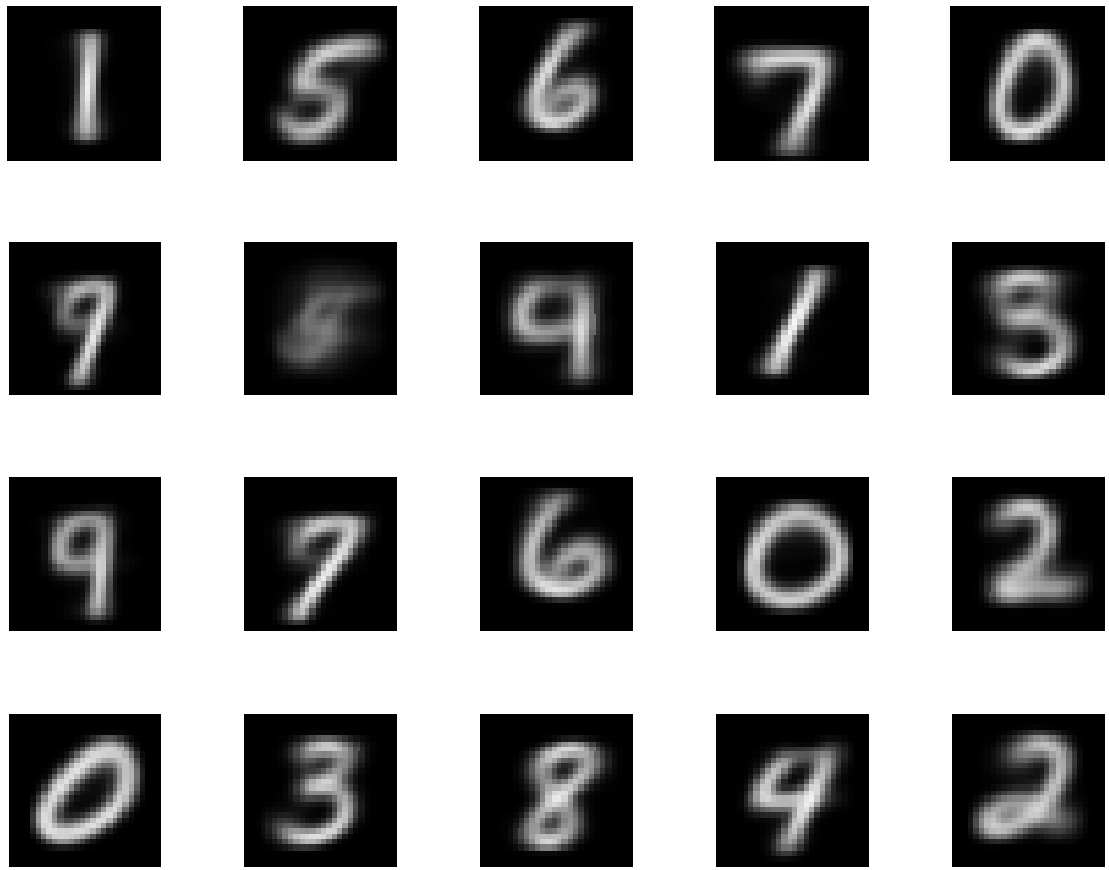
Group representatives ($k = 20$)

- k-means algorithm, with $k = 20$ and randomly chosen initial partition
- 20 group representatives



Group representatives ($k = 20$)

result for another initial partition



Document topic discovery

- $N = 500$ Wikipedia articles, from weekly most popular lists (9/2015–6/2016)
- dictionary of 4423 words
- each article represented by a word histogram vector of size 4423
- result of k-means algorithm with $k = 9$ and randomly chosen initial partition

Cluster 1

- largest coefficients in cluster representative z_1

word	fight	win	event	champion	fighter	...
coefficient	0.038	0.022	0.019	0.015	0.015	...

- documents in cluster 1 closest to representative

  "Floyd Mayweather, Jr", "Kimbo Slice", "Ronda Rousey", "José Aldo", "Joe Frazier", ...

Cluster 2

- largest coefficients in cluster representative z_2

word	holiday	celebrate	festival	celebration	calendar	...
coefficient	0.012	0.009	0.007	0.006	0.006	...

- documents in cluster 2 closest to representative

  "Halloween", "Guy Fawkes Night", "Diwali", "Hannukah", "Groundhog Day", ...

Cluster 3

- largest coefficients in cluster representative z_3

word	united	family	party	president	government	...
coefficient	0.004	0.003	0.003	0.003	0.003	...

- documents in cluster 3 closest to representative

  "Mahatma Gandhi", "Sigmund Freund", "Carly Fiorina", "Frederick Douglass", "Marco Rubio", ...

Cluster 4

- largest coefficients in cluster representative z_4

word	album	release	song	music	single	...
coefficient	0.031	0.016	0.015	0.014	0.011	...

- documents in cluster 4 closest to representative

  "David Bowie", "Kanye West", "Celine Dion", "Kesha", "Ariana Grande", ...

Cluster 5

- largest coefficients in cluster representative z_5

word	game	season	team	win	player	...
coefficient	0.023	0.020	0.018	0.017	0.014	...

- documents in cluster 5 closest to representative

  "Kobe Bryant", "Lamar Odom", "Johan Cruyff", "Yogi Berra", "José Mourinho", ...

Cluster 6

- largest coefficients in representative z_6

word	series	season	episode	character	film	...
coefficient	0.029	0.027	0.013	0.011	0.008	...

- documents in cluster 6 closest to cluster representative

  “The X-Files”, “Game of Thrones”, “House of Cards”, “Daredevil”, “Supergirl”, ...

Cluster 7

- largest coefficients in representative z_7

word	match	win	championship	team	event	...
coefficient	0.065	0.018	0.016	0.015	0.015	...

- documents in cluster 7 closest to cluster representative

  “Wrestlemania 32”, “Payback (2016)”, “Survivor Series (2015)”, “Royal Rumble (2016)”, “Night of Champions (2015)”, ...

Cluster 8

- largest coefficients in representative z_8

word	film	star	role	play	series	...
coefficient	0.036	0.014	0.014	0.010	0.009	...

- documents in cluster 8 closest to cluster representative

  "Ben Affleck", "Johnny Depp", "Maureen O'Hara", "Kate Beckinsale", "Leonardo DiCaprio", ...

Cluster 9

- largest coefficients in representative z_9

word	film	million	release	star	character	...
coefficient	0.061	0.019	0.013	0.010	0.006	...

- documents in cluster 9 closest to cluster representative

  "Star Wars: The Force Awakens", "Star Wars Episode I: The Phantom Menace", "The Martian (film)", "The Revenant (2015 film)", "The Hateful Eight", ...

Outline

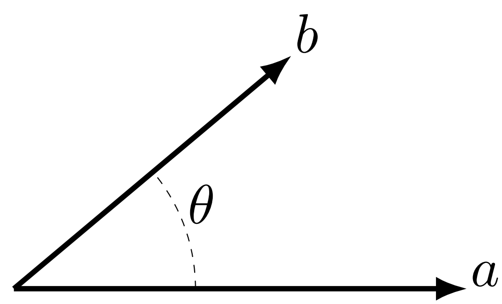
- norm
- distance
- k-means algorithm
- **angle**
- hyperplanes
- complex vectors

Angle between vectors

the angle between nonzero real vectors a, b is defined as

$$\arccos\left(\frac{a^T b}{\|a\|\,\|b\|}\right)$$

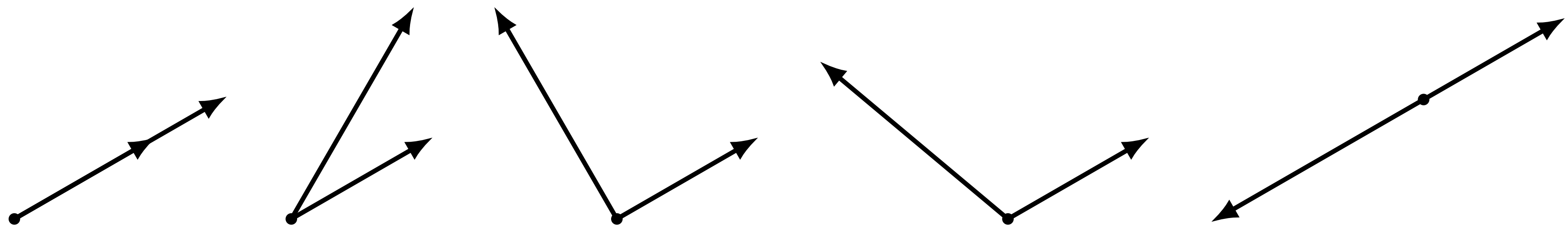
- this is the unique value of $\theta \in [0, \pi]$ that satisfies $a^T b = \|a\|\|b\| \cos\theta$



- Cauchy-Schwarz inequality guarantees that

$$-1 \le \frac{a^T b}{\|a\|\,\|b\|} \le 1$$

Terminology

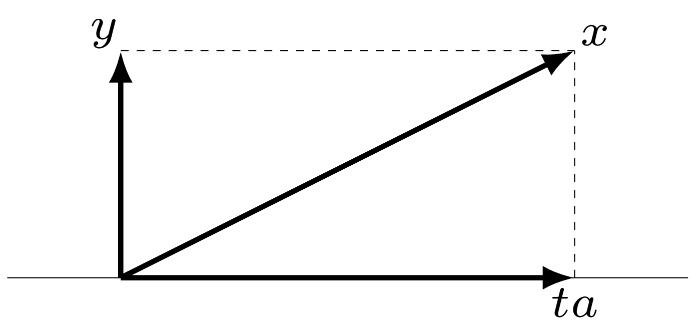


$\theta = 0$	$a^T b = \|a\|\|b\|$	vectors are aligned or parallel
$0 \leq \theta < \pi/2$	$a^T b > 0$	vectors make an acute angle
$\theta = \pi/2$	$a^T b = 0$	vectors are orthogonal ($a \perp b$)
$\pi/2 < \theta \leq \pi$	$a^T b < 0$	vectors make an obtuse angle
$\theta = \pi$	$a^T b = -\|a\|\|b\|$	vectors are anti-aligned or opposed

Orthogonal decomposition

given a nonzero $a \in \mathbf{R}^n$, every n-vector x can be decomposed as

$$x = ta + y \qquad \text{with } y \perp a$$



$$t = \frac{a^T x}{\|a\|^2}, \qquad y = x - \frac{a^T x}{\|a\|^2} a$$

- proof is by inspection
- decomposition (*i.e.*, t and y) exists and is unique for every x
- ta is projection of x on the line through a and the origin (see page 2-12)
- since $y \perp a$, we have $\|x\|^2 = \|ta\|^2 + \|y\|^2$

Correlation coefficient

the *correlation coefficient* between non-constant vectors a, b is

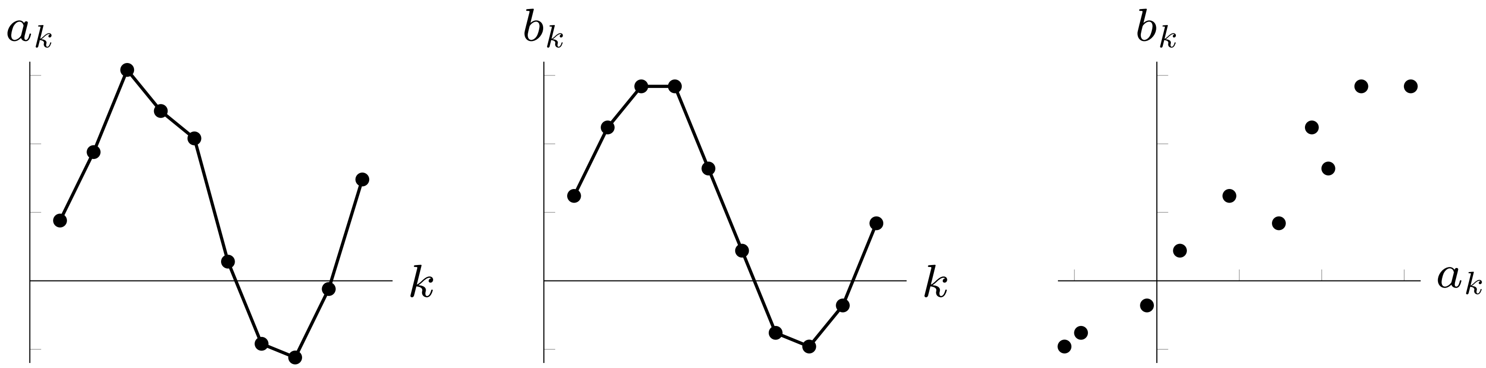
$$\rho_{ab} = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\| \, \|\tilde{b}\|}$$

where $\tilde{a} = a - \text{avg}(a)\mathbf{1}$ and $\tilde{b} = b - \text{avg}(b)\mathbf{1}$ are the de-meaned vectors

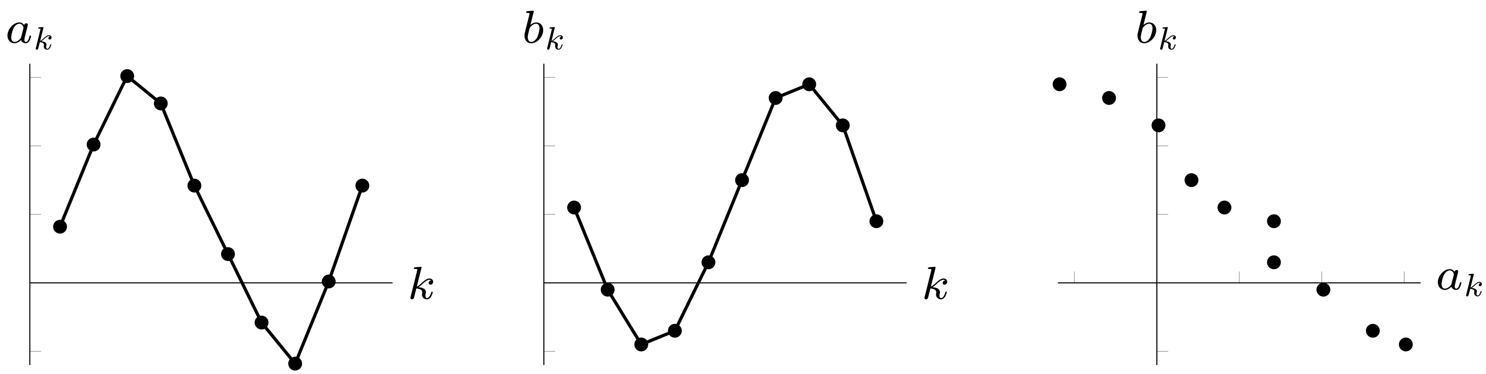
- only defined when a and b are not constant ($\tilde{a} \neq 0$ and $\tilde{b} \neq 0$)
- ρ_{ab} is the cosine of the angle between the de-meaned vectors
- a number between -1 and 1
- ρ_{ab} is the average product of the deviations from the mean in standard units

$$\rho_{ab} = \frac{1}{n} \sum_{i=1}^{n} \frac{(a_i - \text{avg}(a))}{\text{std}(a)} \frac{(b_i - \text{avg}(b))}{\text{std}(b)}$$

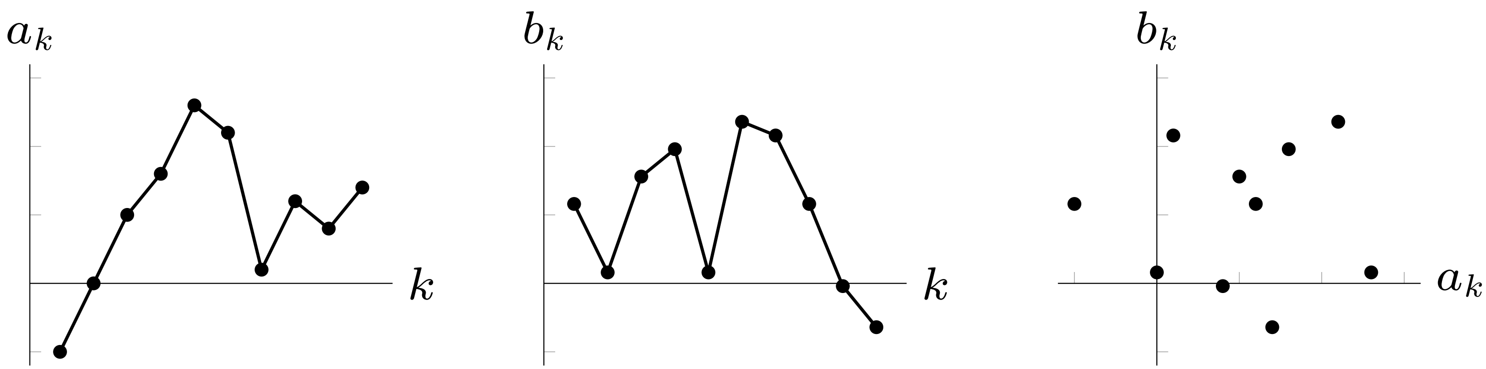
Examples



$$\rho_{ab} = 0.968$$



$$\rho_{ab} = -0.988$$

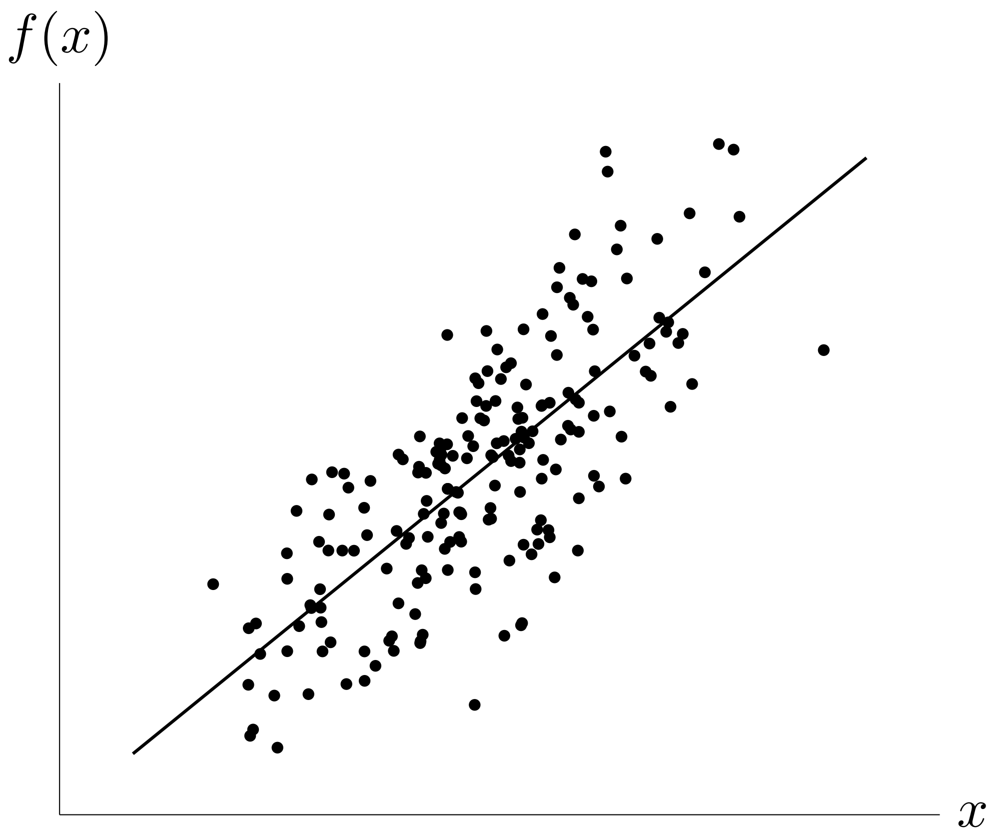


$$\rho_{ab} = 0.004$$

Regression line

- scatter plot shows two n-vectors a, b as n points (a_k, b_k)
- straight line shows affine function $f(x) = c_1 + c_2 x$ with

$$f(a_k) \approx b_k, \quad k = 1, \ldots, n$$



Least squares regression

use coefficients c_1, c_2 that minimize $J = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}\left(f(a_k) - b_k\right)^2$

- J is a quadratic function of c_1 and c_2:

$$\begin{aligned} J &= \frac{1}{n}\sum_{k=1}^{n}\left(c_1 + c_2 a_k - b_k\right)^2 \\ &= \left(nc_1^2 + 2n\,\mathrm{avg}(a)c_1c_2 + \|a\|^2c_2^2 - 2n\,\mathrm{avg}(b)c_1 - 2a^Tbc_2 + \|b\|^2\right)/n \end{aligned}$$

- to minimize J, set derivatives with respect to c_1, c_2 to zero:

$$c_1 + \mathrm{avg}(a)c_2 = \mathrm{avg}(b), \qquad n\,\mathrm{avg}(a)c_1 + \|a\|^2c_2 = a^Tb$$

- solution is

$$c_2 = \frac{a^Tb - n\,\mathrm{avg}(a)\,\mathrm{avg}(b)}{\|a\|^2 - n\,\mathrm{avg}(a)^2}, \qquad c_1 = \mathrm{avg}(b) - \mathrm{avg}(a)c_2$$

Interpretation

slope c_2 can be written in terms of correlation coefficient of a and b:

$$c_2 = \frac{(a - \mathrm{avg}(a)\mathbf{1})^T(b - \mathrm{avg}(b)\mathbf{1})}{\|a - \mathrm{avg}(a)\mathbf{1}\|^2} = \rho_{ab} \, \frac{\mathrm{std}(b)}{\mathrm{std}(a)}$$

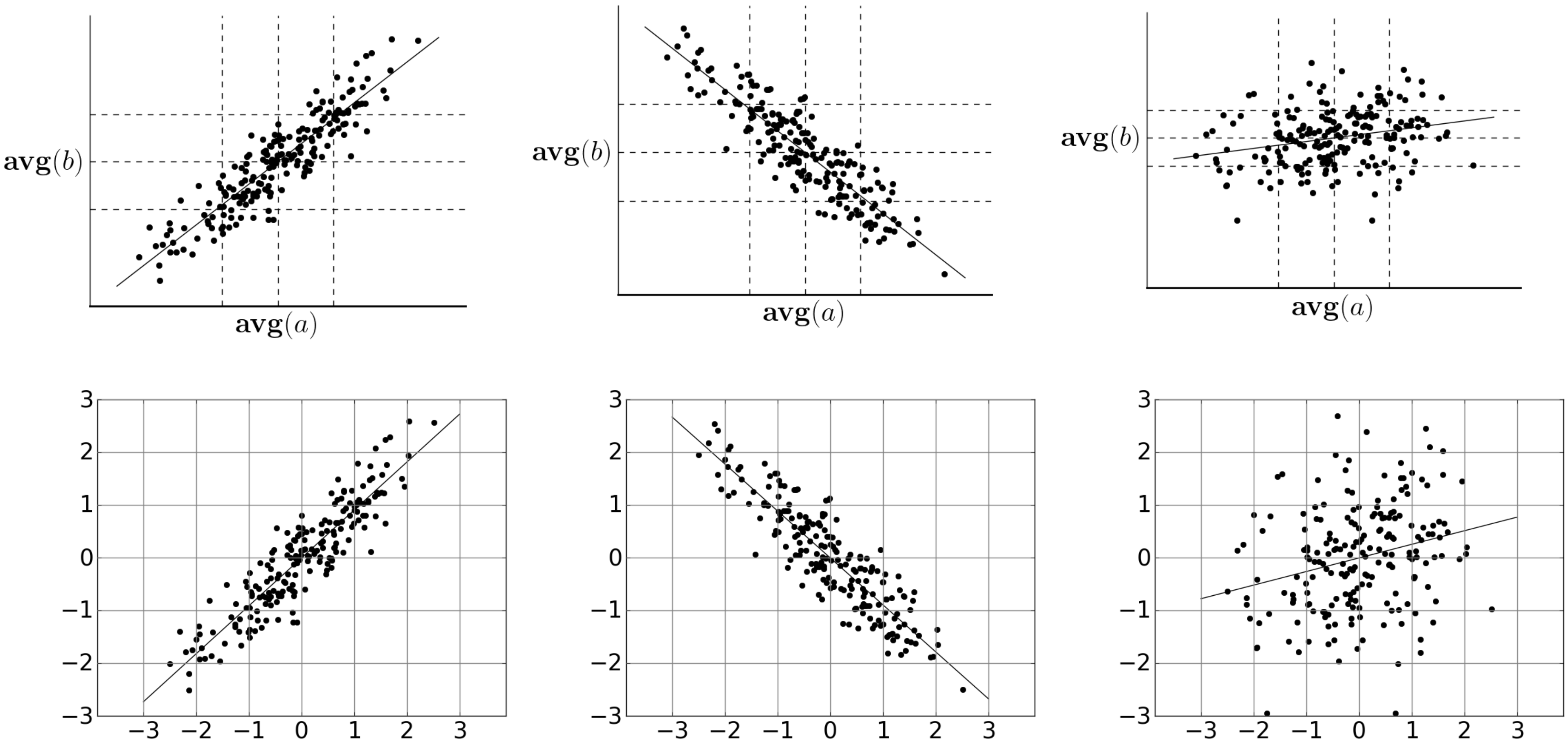
- hence, expression for regression line can be written as

$$f(x) = \mathrm{avg}(b) + \frac{\rho_{ab} \, \mathrm{std}(b)}{\mathrm{std}(a)}(x - \mathrm{avg}(a))$$

- correlation coefficient ρ_{ab} is the slope after converting to standard units:

$$\frac{f(x) - \mathrm{avg}(b)}{\mathrm{std}(b)} = \rho_{ab} \, \frac{x - \mathrm{avg}(a)}{\mathrm{std}(a)}$$

Examples



$\rho_{ab} = 0.91$ $\rho_{ab} = -0.89$ $\rho_{ab} = 0.25$

- dashed lines in top row show average $\pm$ standard deviation
- bottom row shows scatter plots of top row in standard units

Outline

- norm
- distance
- k-means algorithm
- angle
- **hyperplanes**
- complex vectors

Hyperplane

one linear equation in n variables x_1, x_2, ..., x_n:

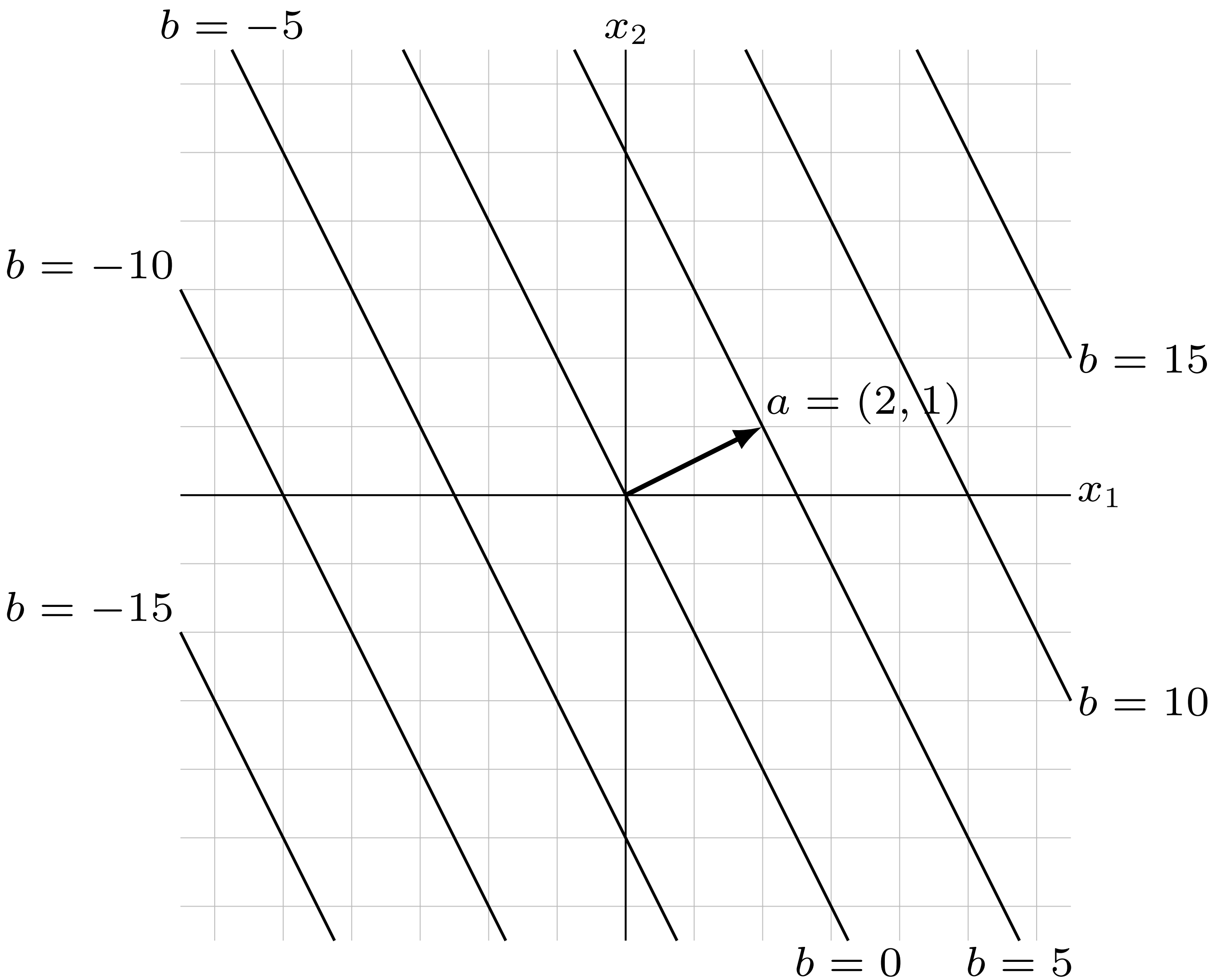
$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

in vector notation: $a^Tx = b$

let H be the set of solutions: $H = \{x \in \mathbf{R}^n \mid a^Tx = b\}$

- H is empty if $a_1 = a_2 = \cdots = a_n = 0$ and $b \neq 0$
- $H = \mathbf{R}^n$ if $a_1 = a_2 = \cdots = a_n = 0$ and $b = 0$
- H is called a *hyperplane* if $a = (a_1, a_2, \ldots, a_n) \neq 0$
- for $n = 2$, a straight line in a plane; for $n = 3$, a plane in 3-D space, ...

Example



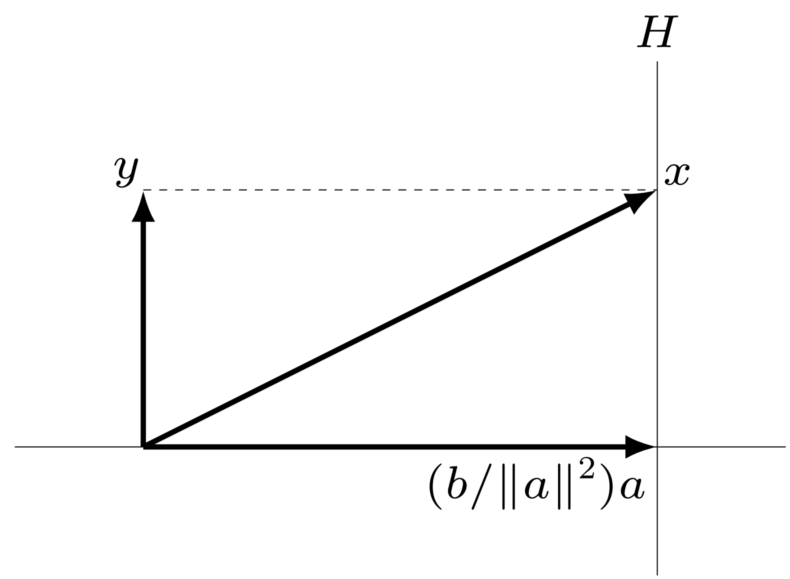
Geometric interpretation of hyperplane

- recall formula for orthogonal decomposition of x with respect to a (page 2-34):

$$x = \frac{a^T x}{\|a\|^2} a + y \quad \text{with } y \perp a$$

- x satisfies $a^T x = b$ if and only if

$$x = \frac{b}{\|a\|^2} a + y \quad \text{with } y \perp a$$



- point $(b/\|a\|^2)a$ is the intersection of hyperplane with line through a

- add arbitrary vectors $y \perp a$ to get all other points in hyperplane

Exercise: projection on hyperplane

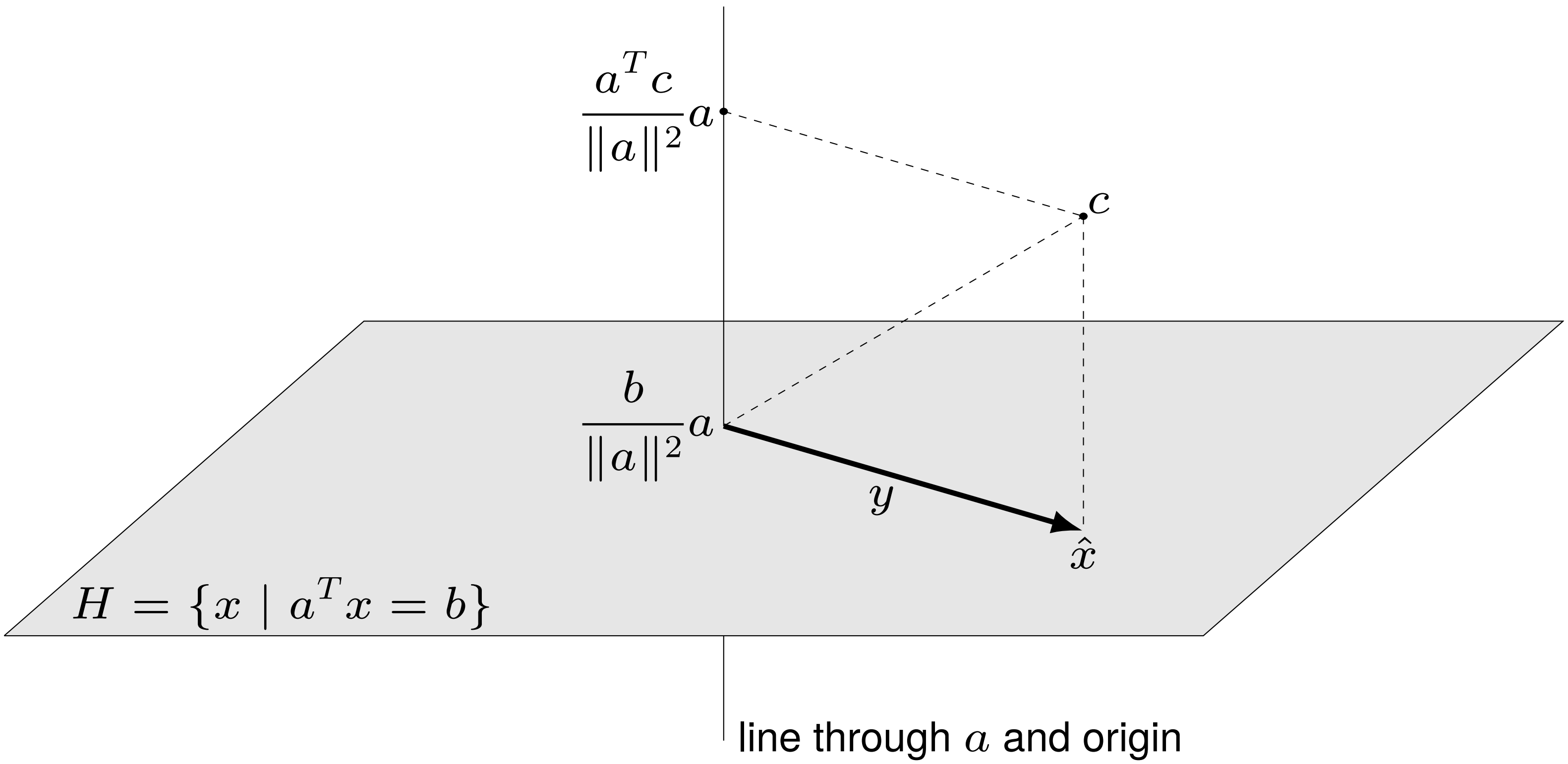
- show that the point in $H = \{x \mid a^T x = b\}$ closest to $c \in \mathbf{R}^n$ is

$$\hat{x} = c + \frac{b - a^T c}{\|a\|^2} a$$

- show that the distance of c to the hyperplane $H = \{x \mid a^T x = b\}$ is

$$\frac{|a^T c - b|}{\|a\|}$$

Solution



$$\hat{x} = c + \frac{b - a^T c}{\|a\|^2} a$$

Solution

- general point x in H is

$$x = \frac{b}{\|a\|^2}a + y, \qquad y \perp a$$

- decomposition of c with respect to a is

$$c = \frac{a^Tc}{\|a\|^2}a + d \quad \text{with } d = c - \frac{a^Tc}{\|a\|^2}a$$

- squared distance between x and c is

$$\|c - x\|^2 = \left\|\frac{a^Tc - b}{a^Ta}a + d - y\right\|^2 = \frac{(a^Tc - b)^2}{\|a\|^2} + \|d - y\|^2$$

(2nd step because $d - y \perp a$); distance is minimized by choosing $y = d$

Kaczmarz algorithm

Problem: find (one) solution of set of linear equations

$$a_1^T x = b_1, \qquad a_2^T x = b_2, \qquad \ldots, \qquad a_m^T x = b_m$$

- here a_1, a_2, …, a_m are nonzero n-vectors
- we assume the equations are solvable (have at least one solution)
- n is huge, so we can only use simple vector operations

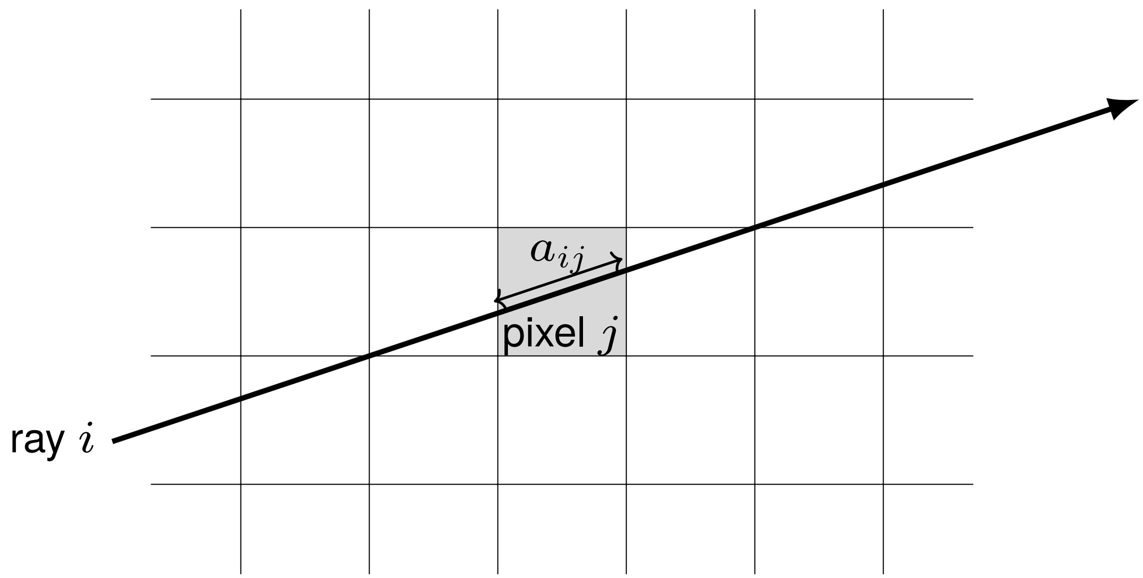
Algorithm: start at some initial x and repeat the following steps

- pick an index $i \in \{1, \ldots, m\}$, for example, cyclically or randomly
- replace x with projection on hyperplane $H_i = \{\tilde{x} \mid a_i^T \tilde{x} = b_i\}$

$$x := x + \frac{b_i - a_i^T x}{\|a_i\|^2} a_i$$

Tomography

reconstruct unknown image from line integrals



- x represents unknown image with n pixels
- a_{ij} is length of intersection of ray i and pixel j
- b_i is a measurement of the line integral $\sum_{j=1}^{n} a_{ij}x_j$ along ray i

Kaczmarz alg. is also known as *Algebraic Reconstruction Technique* (ART)

Outline

- norm
- distance
- k-means algorithm
- angle
- hyperplanes
- **complex vectors**

Norm

norm of vector $a \in \mathbf{C}^n$:

$$\begin{aligned} \|a\| &= \sqrt{|a_1|^2 + |a_2|^2 + \cdots + |a_n|^2} \\ &= \sqrt{a^H a} \end{aligned}$$

- positive definite:

$$\|a\| \geq 0 \quad \text{for all } a, \qquad \|a\| = 0 \quad \text{only if } a = 0$$

- homogeneous:

$$\|\beta a\| = |\beta| \|a\| \quad \text{for all vectors } a\text{, complex scalars } \beta$$

- triangle inequality:

$$\|a + b\| \leq \|a\| + \|b\| \quad \text{for all vectors } a, b \text{ of equal size}$$

Cauchy-Schwarz inequality for complex vectors

$$|a^H b| \leq \|a\|\|b\| \quad \text{for all } a, b \in \mathbf{C}^n$$

moreover, equality $|a^H b| = \|a\|\|b\|$ holds if:

- $a = 0$ or $b = 0$
- $a \neq 0$ and $b \neq 0$, and $b = \gamma a$ for some (complex) scalar γ

- exercise: generalize proof for real vectors on page 2-4
- we say a and b are *orthogonal* if $a^H b = 0$
- we will not need definition of angle, correlation coefficient, ... in $\mathbf{C}^n$