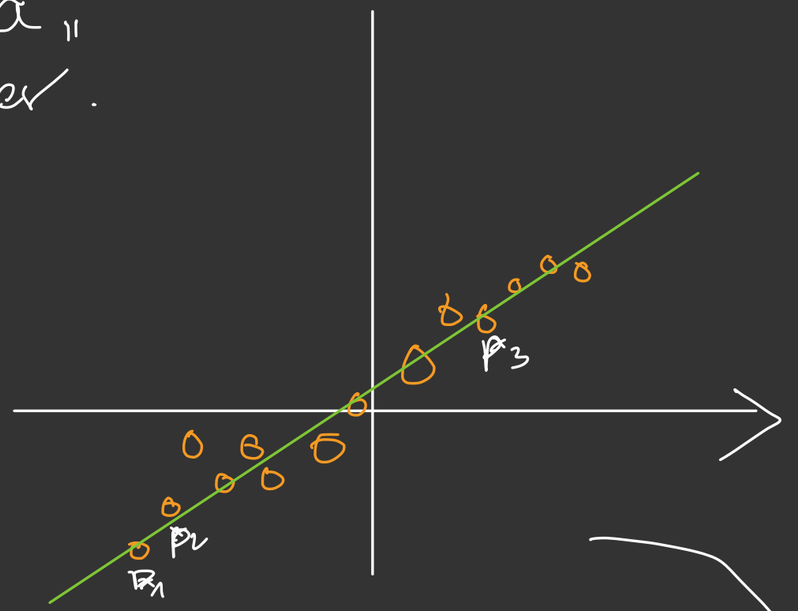
Unsupervised Learning: "Understand data without teacher".

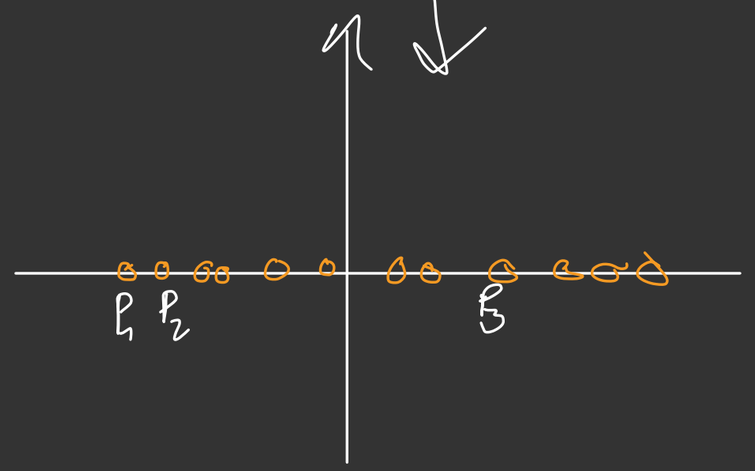
Principal Component Analysis (PCA)

Example: 2D dataset $S = (x_i, y_i)_{i=1}^{n}$

$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \in \mathbb{R}^{n \times 2}$$



PCA



Looks like "almost" a line, can we (as an approximation) represent S in 1D?

Idea: Dimension reduction for
- ▷ interpretability
- ▷ downstream computational savings

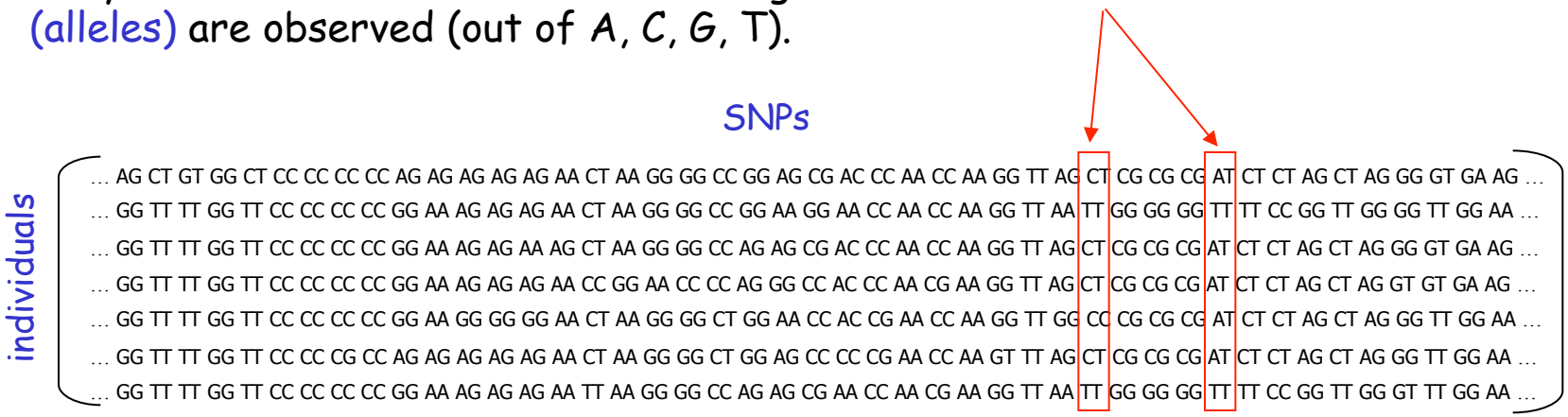
$$\tilde{X} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

Mahoney & Drineas 2015

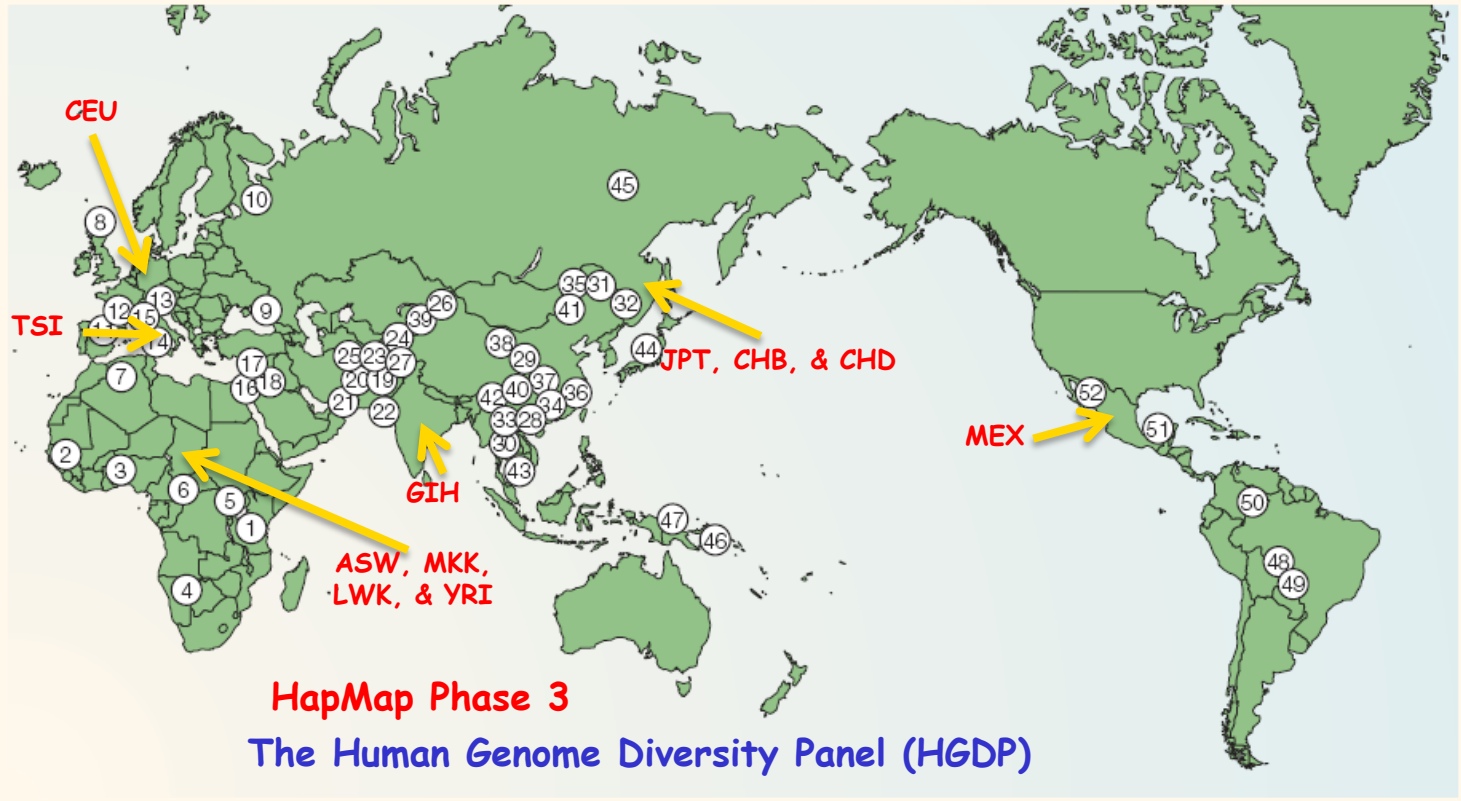
Human genetics

Single Nucleotide Polymorphisms: the most common type of genetic variation in the genome across different individuals.

They are known locations at the human genome where two alternate nucleotide bases (alleles) are observed (out of A, C, G, T).



Matrices including thousands of individuals and hundreds of thousands if SNPs are available.



Africans
1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans
8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians
16 Bedouin
17 Druze
18 Palestinian

Central and Southern Asians
19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur
27 Kalash

Eastern Asians
28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi
43 Cambodian
44 Japanese
45 Yakut

Oceanians
46 Melanesian
47 Papuan

Native Americans
48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

The International HapMap Consortium (2003, 2005, 2007) *Nature*

Mahoney & Drineas 2015

HGDP data

- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data

- 1,207 samples
- 11 populations

We will apply SVD/PCA on the (joint) HGDP and HapMap Phase 3 data.

Matrix dimensions:

2,240 subjects (rows)

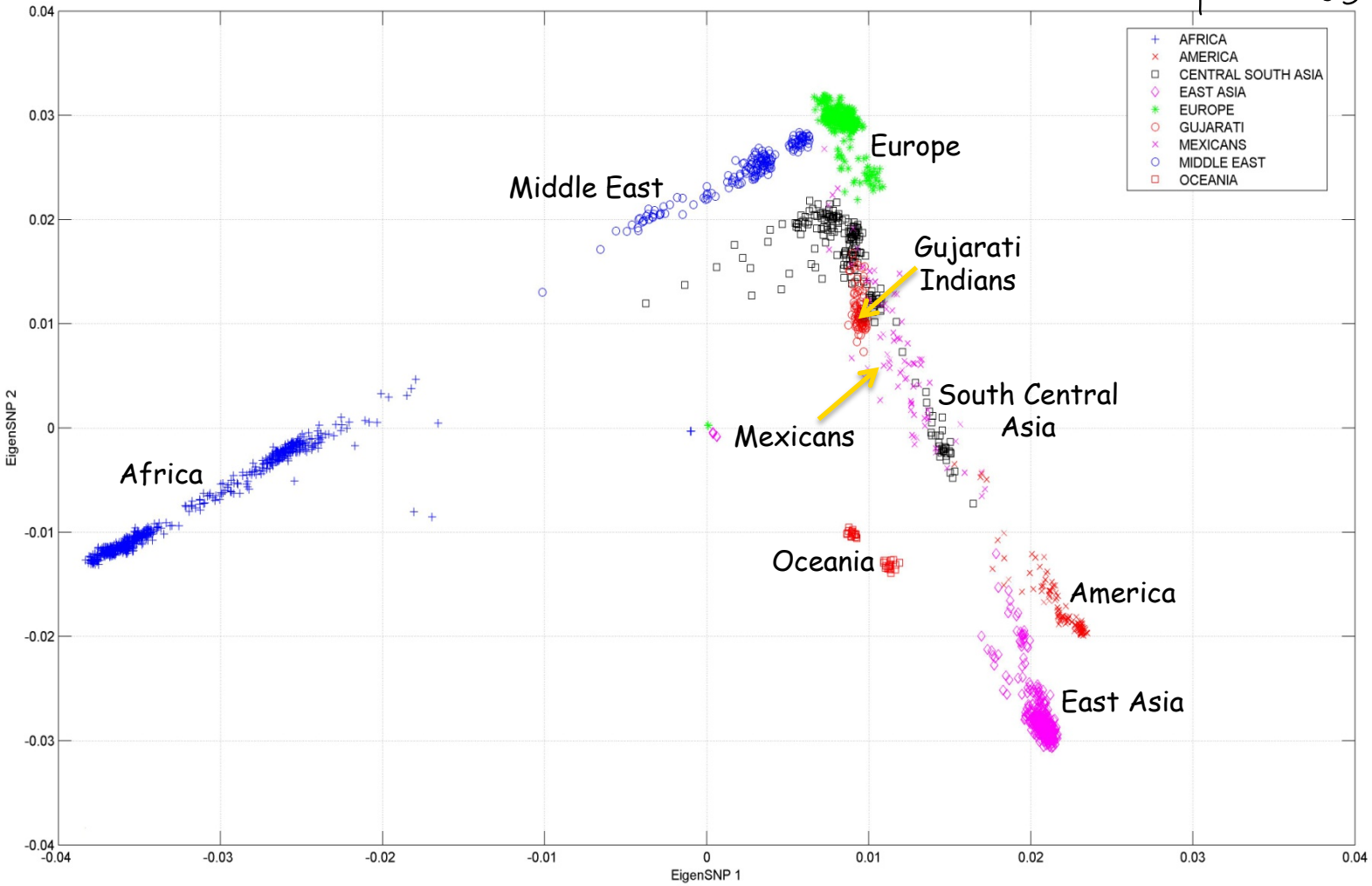
447,143 SNPs (columns)

Dense matrix:

over one billion entries

Paschou, Lewis, Javed, & Drineas (2010) J Med Genet

Mahoney & Drineas 2015



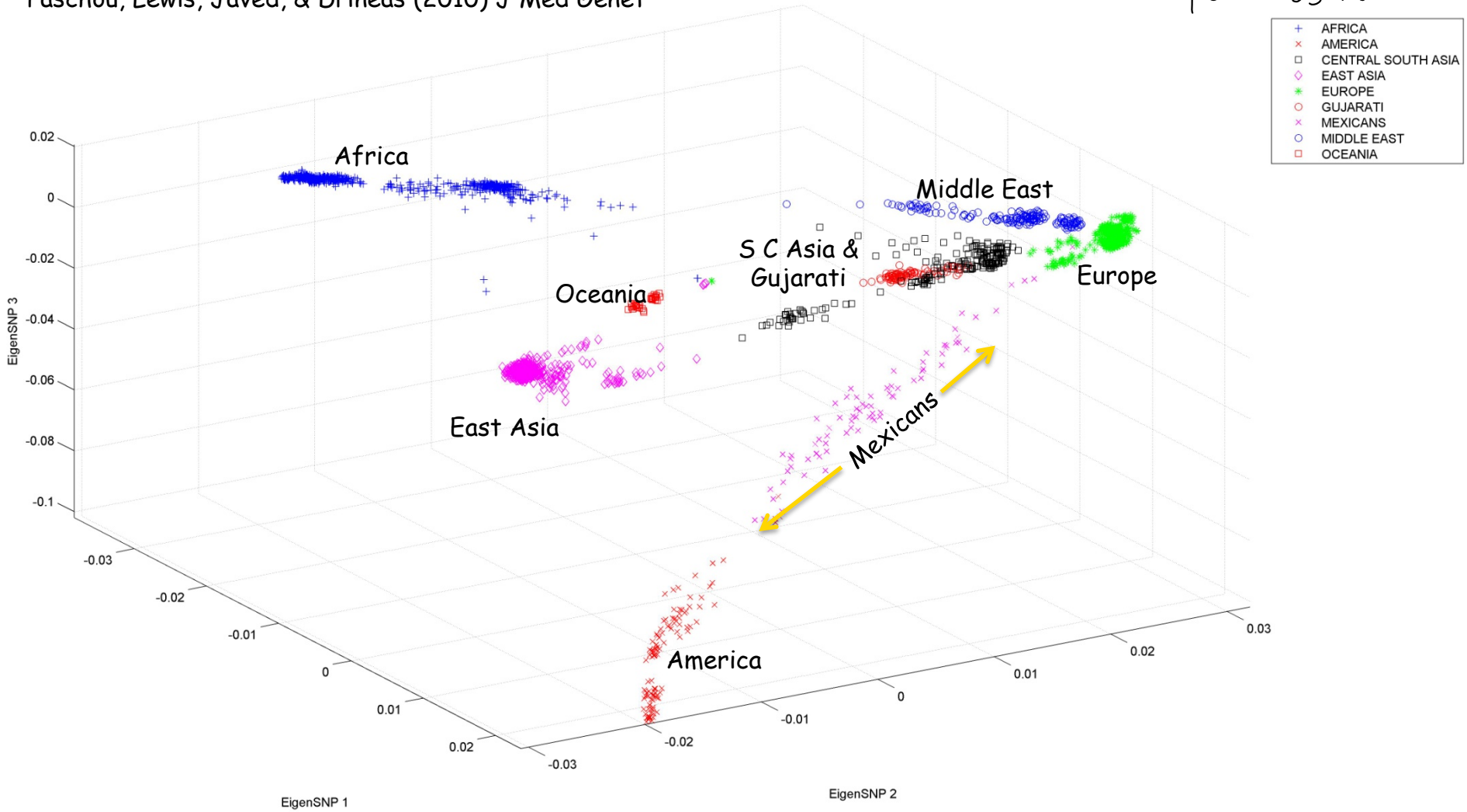
- Top two Principal Components (PCs or eigenSNPs)

(Lin and Altman (2005) *Am J Hum Genet*)

- The figure renders visual support to the "out-of-Africa" hypothesis.
- Mexican population seems out of place: we move to the top three PCs.

Paschou, Lewis, Javed, & Drineas (2010) J Med Genet

Mahoney & Drineas 2015



Not altogether satisfactory: the principal components are linear combinations of all SNPs, and – of course – can not be assayed!

Can we find **actual SNPs** that capture the information in the singular vectors?

Formally: spanning the same subspace.

Setting of PCA: Let $X = \begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_n — \end{bmatrix} \in \mathbb{R}^{n \times k}$

Ex: n: nr. of subjects
k: nr. of SNPs.

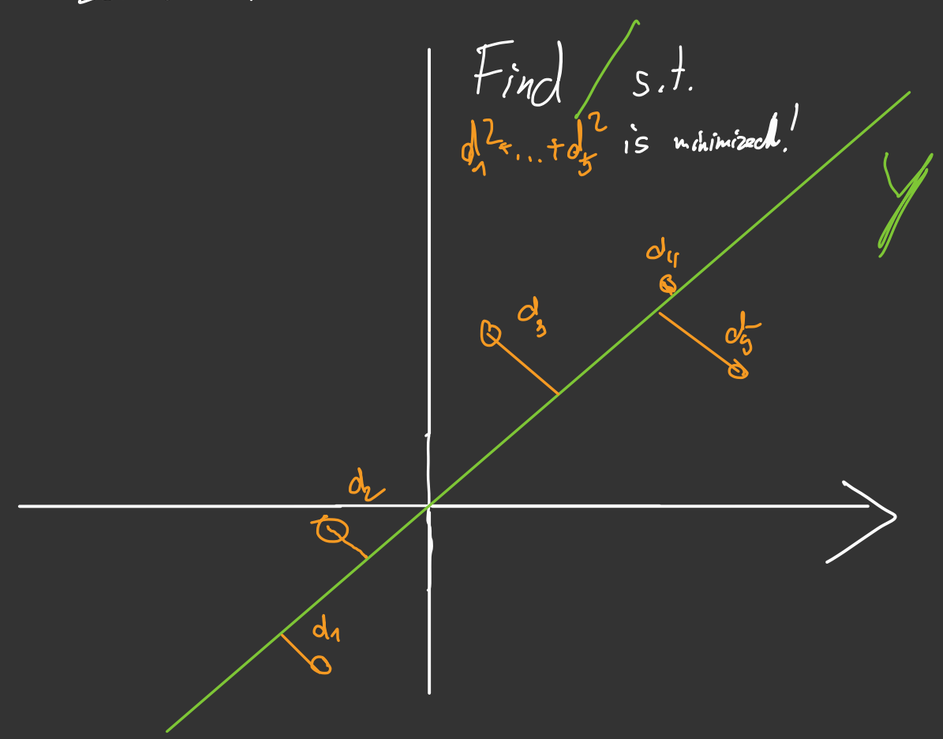
be a matrix of n data points with k features. $\in \mathbb{R}^k$

Assume without loss of generality that X is centered, i.e. $\frac{1}{n}(1,1,\dots,1)^T X = (0,\dots,0)$
(average value of each feature $= 0$).

Goal: For $r << k$, find r-dimensional subspace $Y \subset \mathbb{R}^k$, $\dim(Y) = r$,
with orthonormal basis $V = \begin{bmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{bmatrix} \in \mathbb{R}^{k \times r}$ such that

$$V = \underset{\tilde{V} \in \mathbb{R}^{n \times r}:\ \tilde{V}^T\tilde{V} = I_r}{\operatorname{argmin}} \frac{1}{n} \| X - X\tilde{V}\tilde{V}^T \|_F^2 \qquad (*)$$

"minimizes sum of squares of distances between points and projected points".



Properties of PCA: ▷ Obtain (approximate) low-dimensional representations

$$\tilde{Z} := X \underset{[v_1 \dots v_\tau]}{V} = \begin{bmatrix} - \tilde{z}_1 - \\ \vdots \\ - \tilde{z}_n - \end{bmatrix} \in \mathbb{R}^{n \times \tau}$$

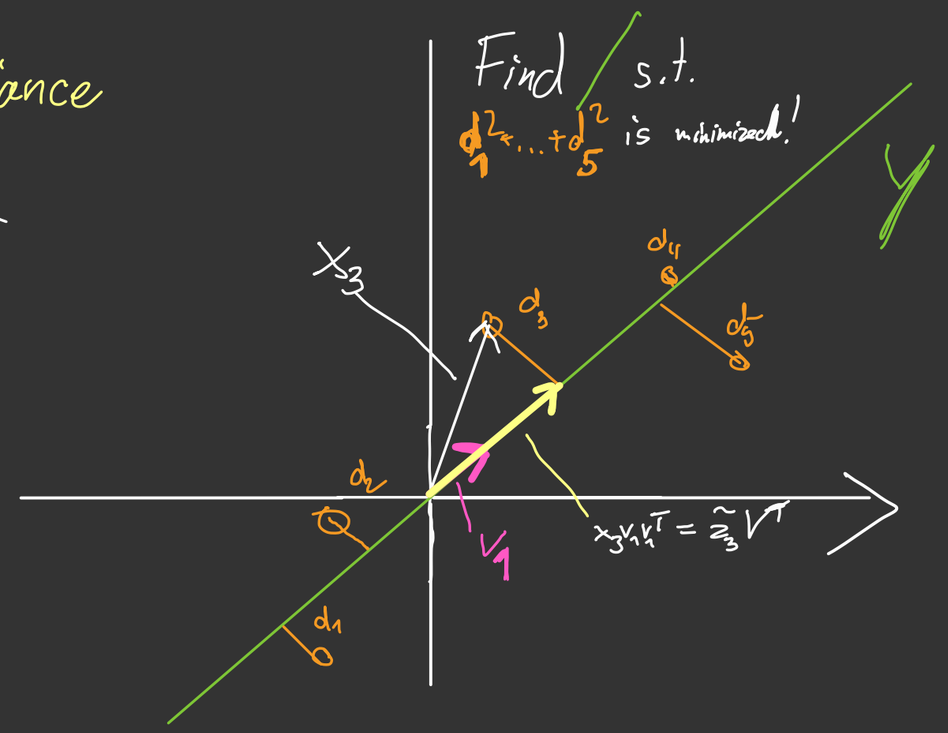
of the data points $\{x_1, \dots, x_n\}$.

Ex: $2240 \times 447143 \;\to\; 2240 \times 3$ (n unchanged, $k \to \tau$)

▷ Memory savings: $O(n\tau)$ instead of $O(nk)$ parameters

▷ $V = \begin{bmatrix} | & & | \\ v_1 & \dots & v_\tau \\ | & & | \end{bmatrix}$: Matrix w/ eigenvectors corresponding to τ largest eigenvalues of empirical covariance matrix

$$M = \frac{1}{n-1} \sum_{i=1}^{n} x_i x_i^T = \frac{1}{n-1} X^T X \qquad \in \mathbb{R}^{k \times k}$$



▷ PCA find r (orthogonal) directions of maximal variance: (Interpretation favored in statistics)

Recall projection perspective of PCA:

$$(*) \quad V = \underset{\tilde{V}\in\mathbb{R}^{n\times r}:\ \tilde{V}^T\tilde{V}=I_r}{\operatorname{argmin}} \frac{1}{n}\|X - X\tilde{V}\tilde{V}^T\|_F^2$$

- $\frac{1}{n}\|X - XVV^T\|_F^2 = \frac{1}{n}\sum_{i=1}^{n} \|x_i^T - x_i^T VV^T\|_2^2$, and

$$\|x_i^T - x_i^T VV^T\|_2^2 = x_i^T x_i - 2x_i^T VV^T x_i + \underbrace{x_i^T VV^T VV^T x_j}_{\overset{=I_r}{=}\ x_i^T VV^T x_j} = x_i^T x_j - x_i^T VV^T x_i$$

- Since $x_i^T x_i$ does not depend on V, rewrite $(*)$ s.t.

$$V = \underset{\tilde{V}:\ \tilde{V}^T\tilde{V}=I_r}{\operatorname{argmax}} \underbrace{\frac{1}{(n-1)}\sum_{i=1}^{n} x_i^T \tilde{V}\tilde{V}^T x_i}_{\frac{1}{(n-1)}\sum_{i=1}^{n} \operatorname{tr}(x_i^T \tilde{V}\tilde{V}^T x_i)}$$

$$\overset{\text{cyclicity of trace}}{=} \frac{1}{(n-1)}\sum_{i=1}^{n} \operatorname{tr}\left(\tilde{V}^T x_i x_i^T \tilde{V}\right) = \operatorname{tr}\left\{\tilde{V}^T\left(\frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T\right)\tilde{V}\right\}$$

$$= \operatorname{tr}\left(\tilde{V}^T M \tilde{V}\right)$$

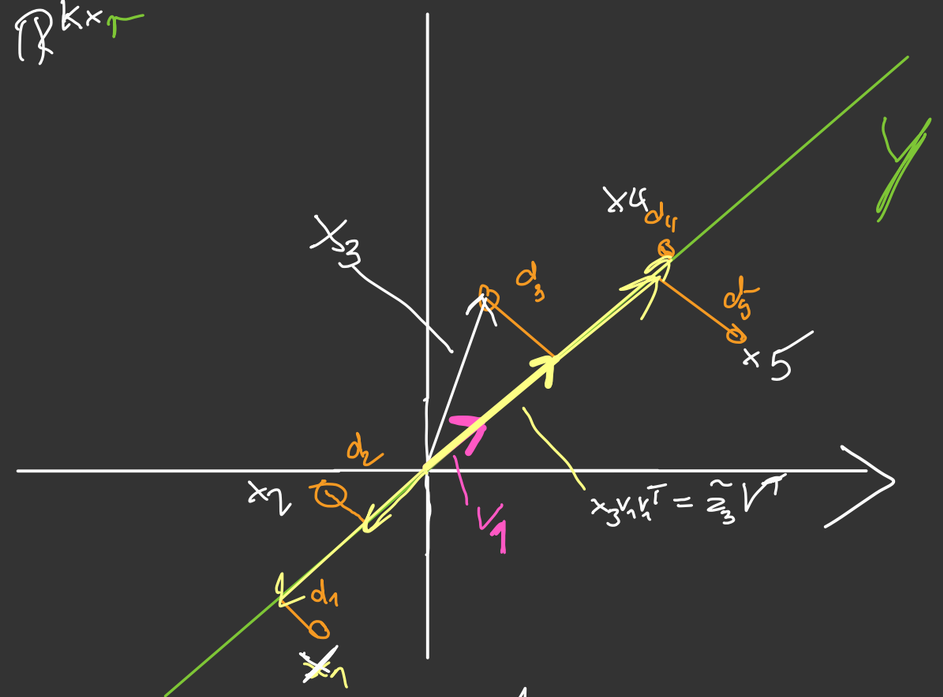
$\Rightarrow$ □ Connection of PCA to first r eigenvectors of M.

- Variance of dataset projected into first PC $\{x_i^T v_1\}_{i=1}^n$: $\Big\}$ if $r=1$

$$\mathrm{Var}_x(x^T v_1) = \mathbb{E}_x[(x^T v_1)^2] = \sum_{i=1}^n \frac{1}{n}\|x_i^T v_1\|^2 = v_1^T \frac{1}{n} \cdots^T v_1 = v_1^T M v_1$$

$\Rightarrow$ PCA: Find Y that
- minimizes sum of squared distances $\sum d_i^2$
- maximizes variance of $\{x_i\}$ in Y.

Terminology:
- Columns v_i of $V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{bmatrix} \in \mathbb{R}^{k \times r}$ "Principal directions / axes"
- Columns Xv_i of $XV \in \mathbb{R}^{n \times r}$: "Principal components" / "Scores"
- λ_i: "Variance explained by i-th PC"
- $\sqrt{\lambda_i} v_i$: "i-th loading"



Recall: $M := \frac{1}{n-1} X^T X$

$= [V \; V_\perp] \begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_r & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} V^T \\ V_\perp^T \end{bmatrix}$

↑ eigendecomposition

Terminology: ▷ Columns v_i of $V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{bmatrix} \in \mathbb{R}^{k \times r}$

"Principal directions / axes"

▷ λ_i : "Variance explained by i-th PC"

▷ Columns Xv_i of $XV \in \mathbb{R}^{n \times r}$: ▷ $\sqrt{\lambda_i} v_i$: "i-th loading"

"Principal components" / "Scores" : $Xv_i = \sqrt{\lambda_i} u_i$

- Instead of eigendecomposition of $M = \frac{1}{n-1} X^T X$, we can compute partial singular value decomposition of X s.t. $X \approx U \Sigma V^T$ ($V \in \mathbb{R}^{r \times k}$, $U \in \mathbb{R}^{n \times r}$)
  - In this case, $\lambda_i = \frac{1}{n-1} \Sigma_{ii}^2$.

⚠ : ▷ All this holds for centered data $X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_n & — \end{bmatrix} \in \mathbb{R}^{n \times k}$.

▷ If $\mu := \frac{1}{n} \mathbb{1}^T X \neq 0$, center data first:

$$X \longmapsto X - \frac{1}{n} \mathbb{1}\mathbb{1}^T X =: \tilde{X},$$

proceed with $\tilde{X}$.

Clustering

Task: Find clusters/subgroups in a dataset $S = \{x_1, \dots, x_n\} \in \mathbb{R}^p$.

- Samples within subgroup similar/homogeneous
- Samples in different subgroups "distant"/heterogeneous from each other.

Ex:
- Customers of a company → grouping for targeted marketing.
- Biology: Find groups of genes.

Q:
- What notion of similarity?
- Precise definition?

Note: Fundamentally different from classification problems!

"**Unsupervised Learning**":
No subset of data with "correct" classification/grouping available.

k-means clustering: [Steinhaus '56, Lloyd '57]

Given n points $x_1, \dots, x_n \in \mathbb{R}^p$, find k centroids $c_1, \dots, c_k \in \mathbb{R}^p$ and a partition $\Gamma_1 \cup \Gamma_2 \cup \dots \cup \Gamma_k = \{1, \dots, n\}$ $(\Gamma_i \cap \Gamma_j = \emptyset \text{ if } i \neq j)$ such that

$$F\left(\{c_i\}_{i=1}^k, \{\Gamma_i\}_{i=1}^k\right) := \sum_{j=1}^{k} \sum_{i \in \Gamma_j} d(x_i, c_j)$$

"k-means objective" (*)

is minimized, where $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a "distance" function

Ex: 1. $d(x,y) = \|x-y\|_2^2$ "squared Euclidean" ← by default in

2. $d(x,y) = \|x-y\|_1$. "k-median".

Observation: K-means with is NP-hard for $k \geq 2$. [Drineas et al. '04].

Lloyd's algorithm (often called "k-means"):

Input: $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^p$, desired nr. of clusters k

1. Initialize $c_1, \dots, c_k \in \mathbb{R}^p$.
   - a) randomly among the $\{x_i\}_{i=1}^n$
   - or
   - b) k-means++ : fancier

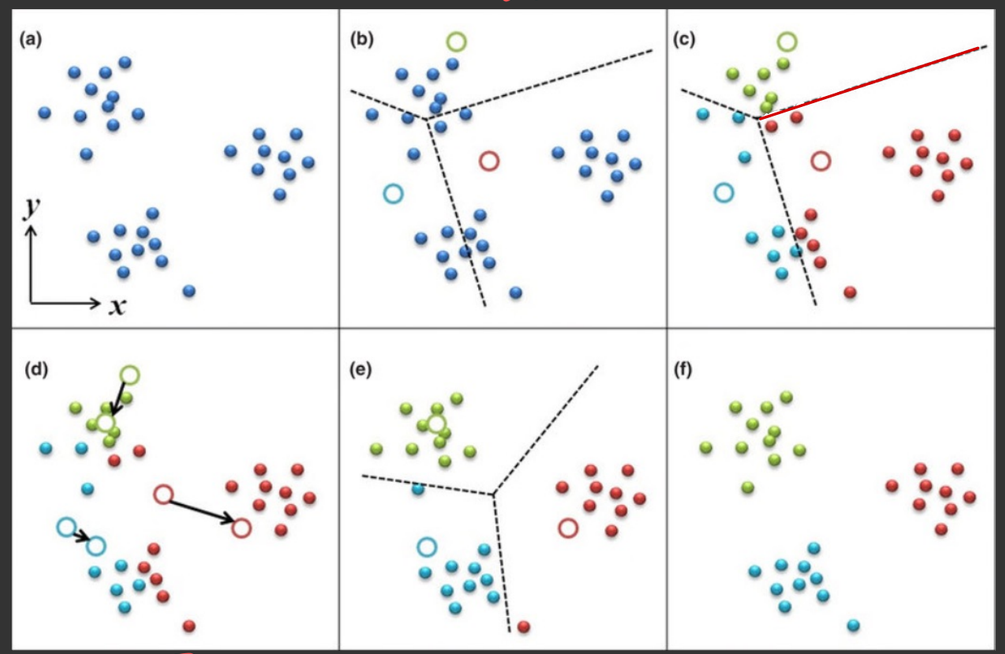
Repeat until convergence:

2. $\forall i = 1, \dots, n$: Assign $x_i \in C_j$ if c_j is closest centroid to x_i among $\{c_\ell\}_{\ell=1}^k$

3. Update $\forall j = 1, \dots, k$:

$$c_j = \underset{c \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i \in C_j} d(x_i, c) = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$$

(the last equality if $d(\cdot,\cdot) = \|\cdot - \cdot\|_2^2$)



- Finds local optimum of (*)
- Works well for "convex" clusters

To consider:

- How to choose nr. of clusters k?
- Which distance to choose (geometry of underlying space...)?
- Initialization: If prior knowledge available, → might be better than random.
  - Needs a lot of pairwise distances. If $n >> 10^5$ or so, slow → "Minibatch KMeans"

Other clustering methods:

- Spectral Clustering: Based on laplacian of similarity graph.
- Hierarchical Clustering: Creates trees