Inmas Machine Learning Workshop: Session 3

Jan 14 2024

Yesterday: **Supervised Learning:**

- Given labeled examples, find right prediction for unlabeled examples.

This Morning: **Unsupervised Learning:**

- Given data, try to discover similar patterns, structure or sub-spaces within your data
- "Learning without a teacher"

Ex:
- Find categories among pictures on phone app (without supervision)
- Visualize complex data in order to draw further conclusions later (e.g., genetics)

Clustering

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

- Samples within subgroups shall be similar / homogeneous
- Samples in different subgroups shall be "distant" / heterogeneous from each other

Q:
- What notion of similarity makes sense?
- Algorithm / precise definition?

Note: Different from classification since no labels are available.

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

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

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

is minimized.

$d: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a similarity function ← similarity measure

1. $d(x, y) = \|x - y\|_2^2$ ← "squared Euclidean" distance
2. $d(x, y) = \|x - y\|_1$ ← "k-median"

Observation: K-means with 1. is NP-hard for $k \geq 2$ [Drineas et al. '04]

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

1. Initialize $c_1, \dots, c_k \in \mathbb{R}^d$ (for example uniformly at random among $S = \{x_1, \dots, x_n\}$)

Repeat until convergence:

1. $\forall i = 1, \dots, n$: Assign $x_i \in \Gamma_j$ if c_j is closest centroid to x_i among $\{c_j\}_{j=1}^{k}$ (*)
2. Update $\forall j = 1, \dots, k$: $c_j = \underset{c \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{i=\Gamma_j} d(x_i, c) \right\} = \frac{1}{|\Gamma_j|} \sum_{i \in \Gamma_j} x_i$

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

- Finds a local optimum of $F(\cdot)$ from last slide
- Works well for "convex" clusters.

Q:

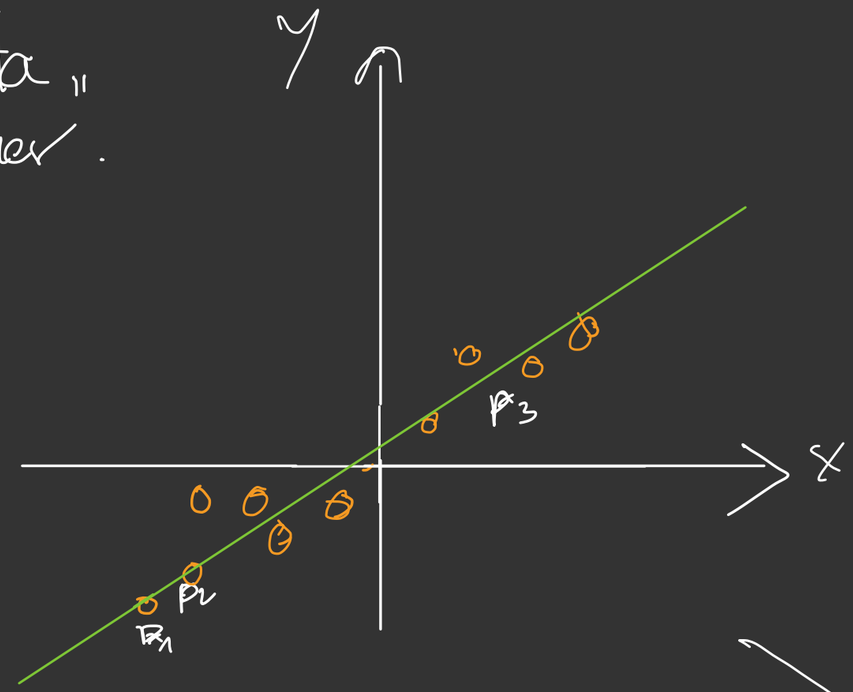
- How to choose nr. of clusters k?
- Choice of $d(\cdot,\cdot)$
- Initialization: If prior knowledge available, use to define initial $\{c_j\}_{j=1}^{k}$.
- ⚠ Needs many evaluations of $d(\cdot,\cdot)$.
  If $n \gg 10^5$, slow → "Minibatch KMeans".

Unsupervised Learning: "Understand data without teacher".

Principal Component Analysis (PCA)

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

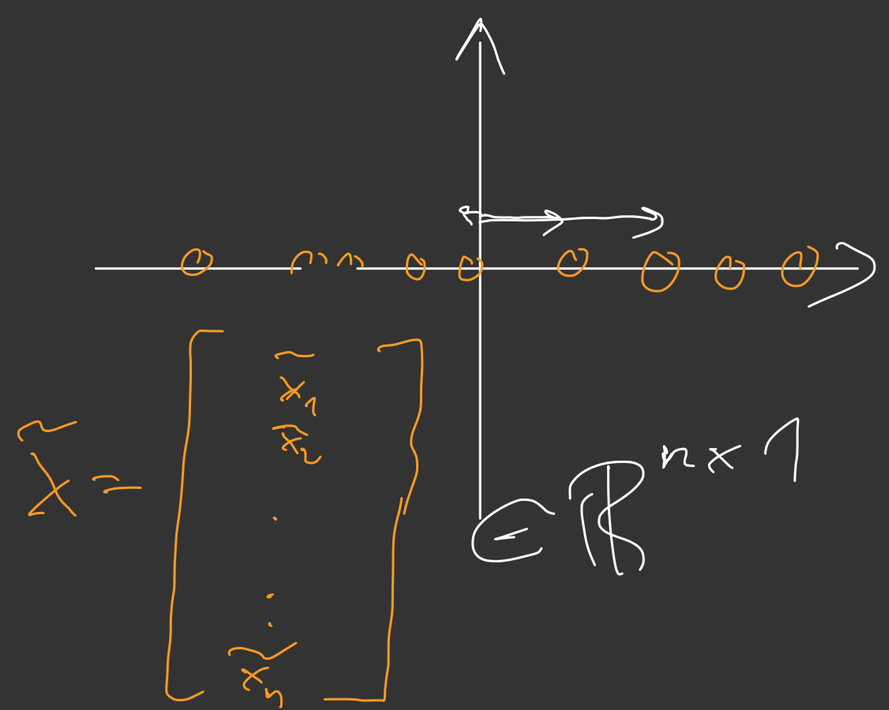
$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \in \mathbb{R}^{n \times 2}$$



PCA

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

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

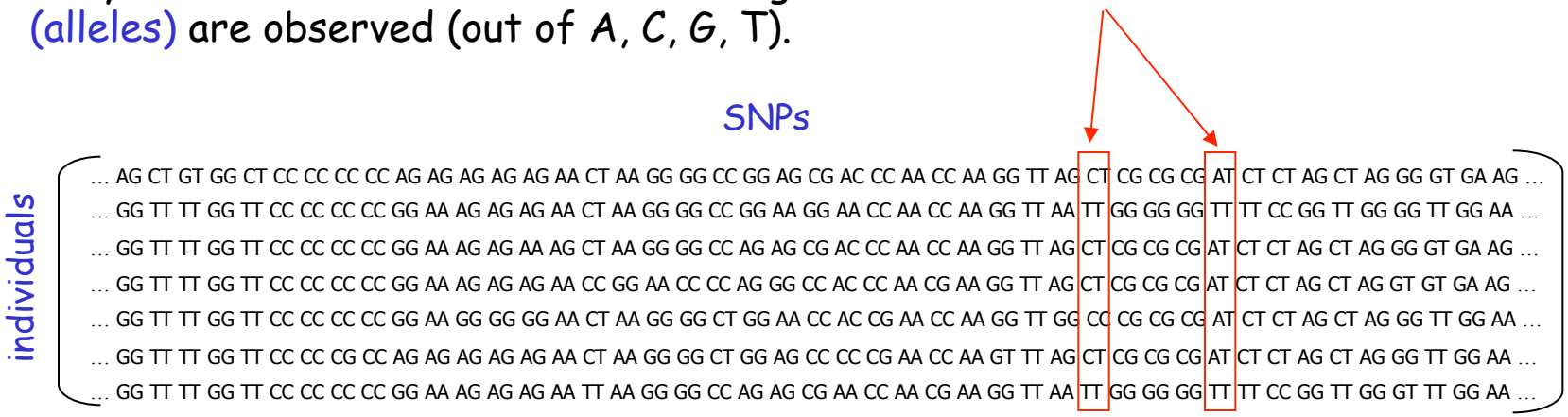


Mahoney & Drineas 2015

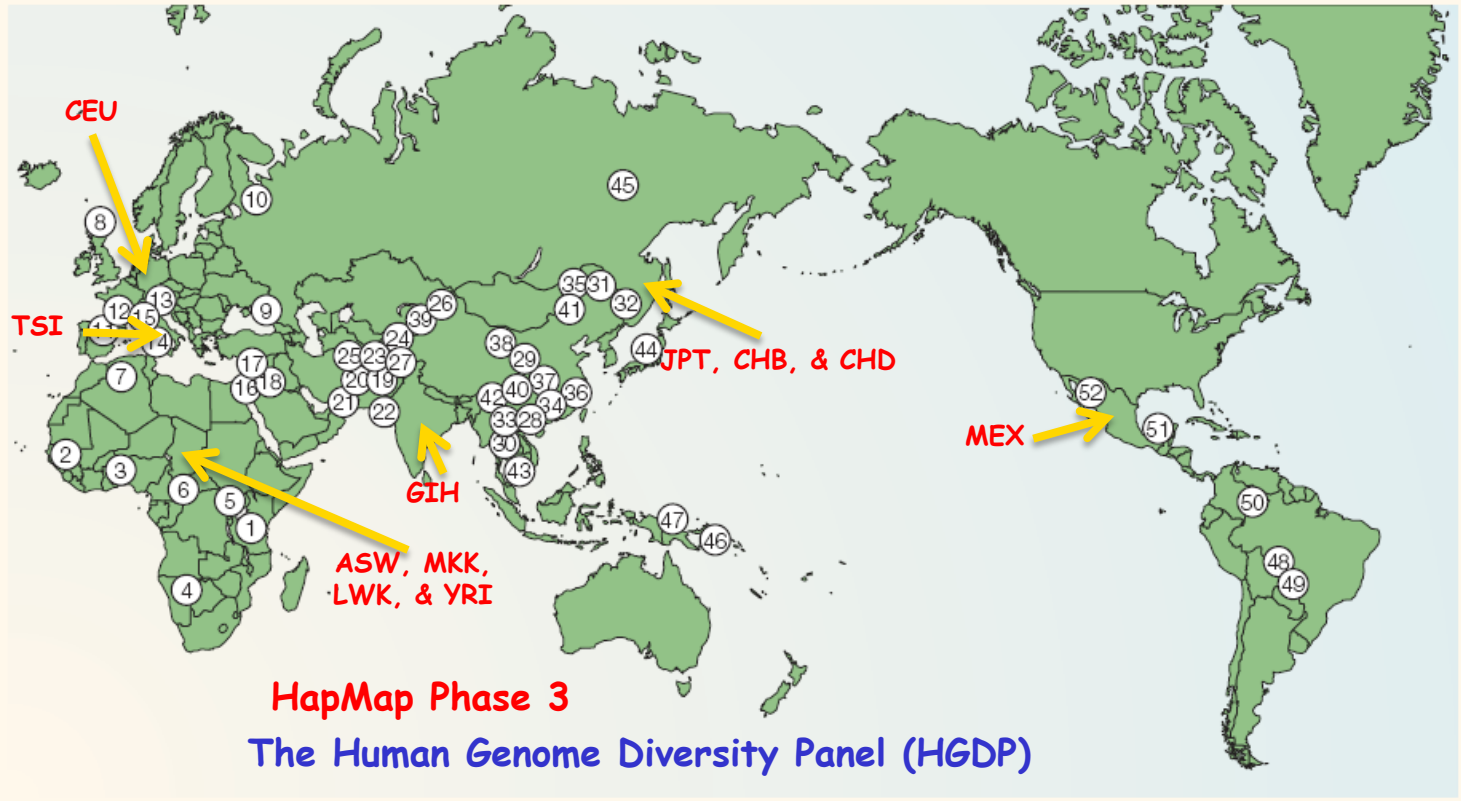
Human genetics

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)

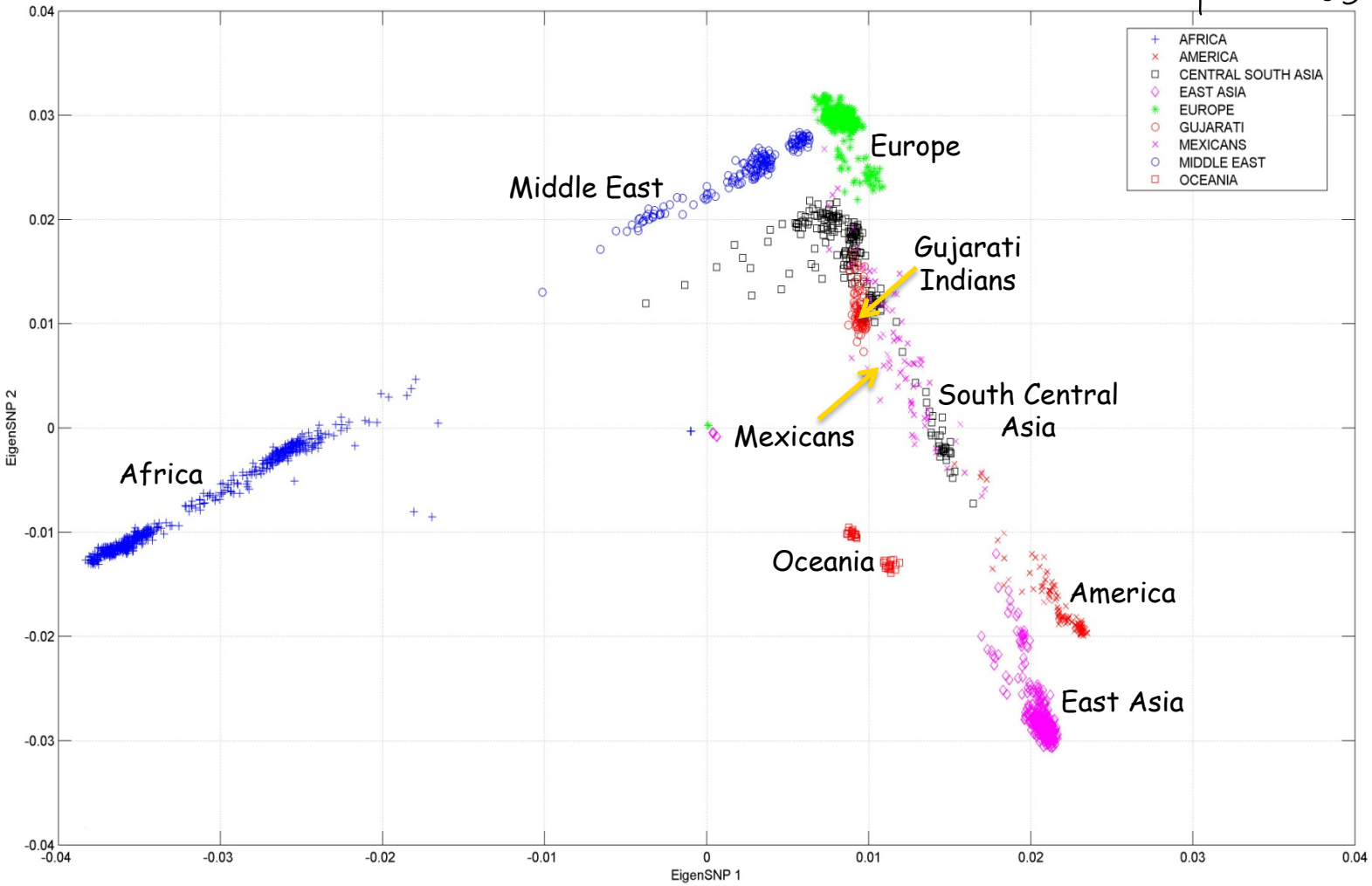
447,143 SNPs (columns)

Dense matrix:

over one billion entries

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

Mahoney & Drineas 2015



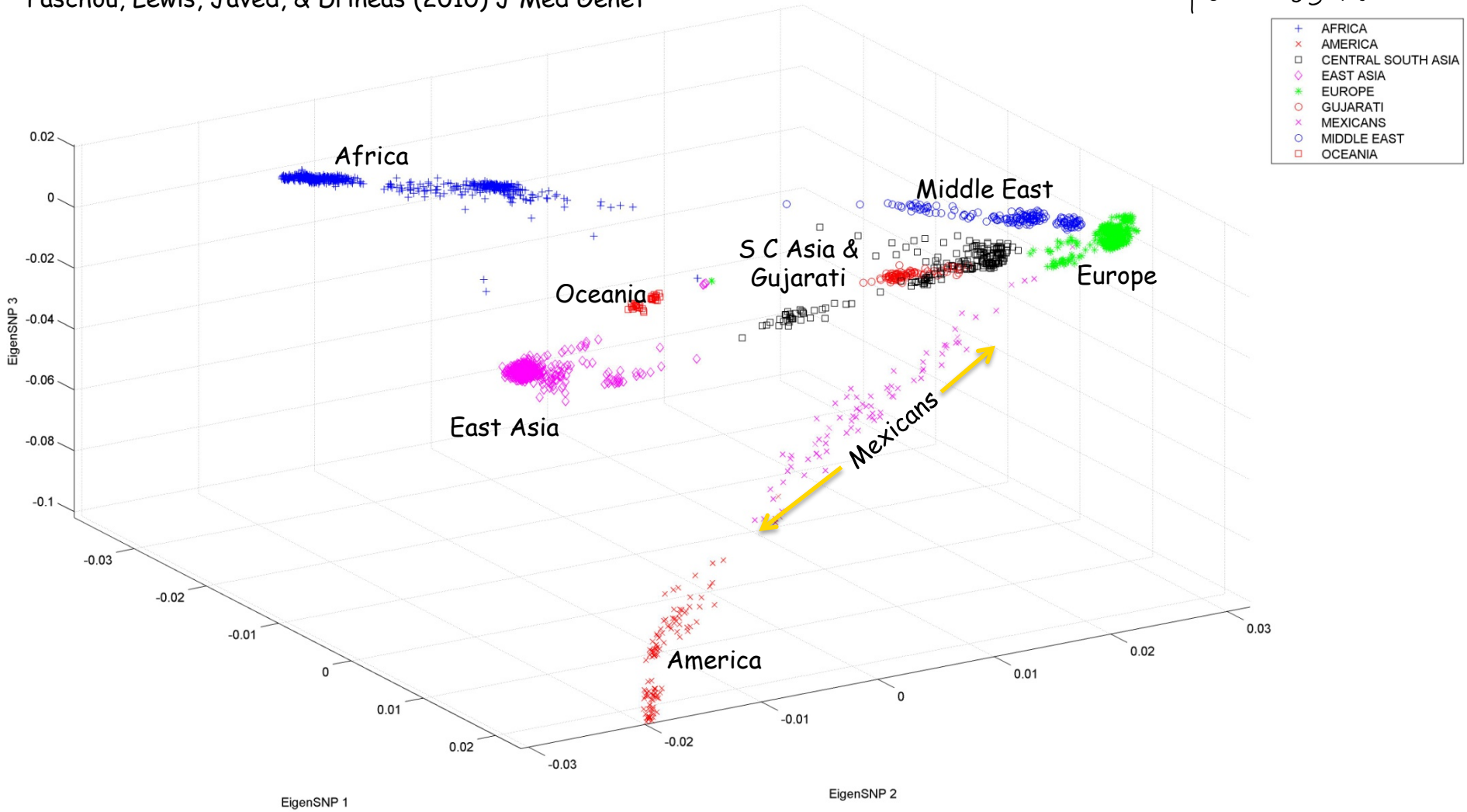
- Top two Principal Components (PCs or eigenSNPs)

(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:

$$X = \begin{bmatrix} - & x_1 & - \\ & \vdots & \\ - & x_n & - \end{bmatrix} \in \mathbb{R}^{n \times d}$$

n: nr of individuals

Ex: d: nr of SNPs.

Assume w.l.o.g. that X is centered, $\frac{1}{n}(1, 1, \dots, 1)^T X = (0, 0, \dots, 0) \in \mathbb{R}^d$

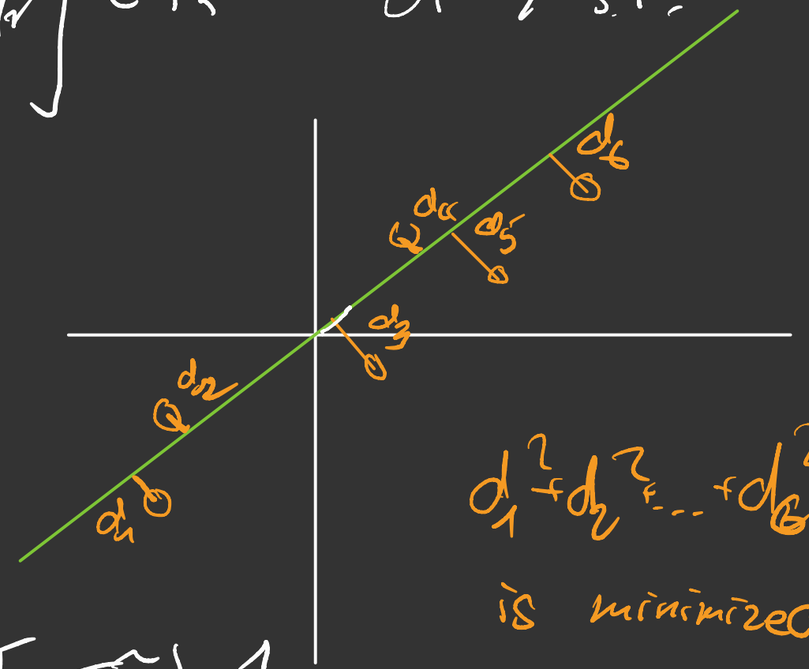
Goal: For some $r \leq d$, find the r-dimensional subspace $Y \subset \mathbb{R}^d$ and an orthonormal basis $V = \begin{bmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times r}$ of Y s.t.

$$V = \tilde{V} \in \operatorname{argmin}_{\tilde{V} \in \mathbb{R}^{d \times r} : \tilde{V}^T \tilde{V} = I_r} \frac{1}{n} \|X - X \tilde{V} \tilde{V}^T\|_F^2$$

"minimize sum of squares of distances between original points and projected points"

$$\frac{1}{n}\|X^T X\|_F^2 - \operatorname{tr}(\tilde{V}^T X X^T \tilde{V}) \cdot \frac{1}{n}$$

$\Rightarrow$ Equivalent to: $V = \operatorname{argmax}_{\tilde{V} : \tilde{V}^T \tilde{V} = I} \operatorname{tr}(\tilde{V}^T \underbrace{\frac{1}{n-1} X X^T}_{: C_X} \tilde{V})$



$d_1^2 + d_2^2 + \dots + d_6^2$ is minimized

Terminology:

- Columns of $V = [v_1, \dots, v_r]$: "Principal components / directions"
- Columns Xv_i of $XV \in \mathbb{R}^{n \times r}$: "Principal components" / "Scores"

sometimes not consistent terminology

- λ_i: i-th eigenvalue of C_x: "Variance explained by i-th PC"
- $\sqrt{\lambda_i}\, v_i$: "i-th loading"

Coming from complementary interpretation of PCA as finding a r-dim. subspace that maximizes the variance of data after projection onto that subspace