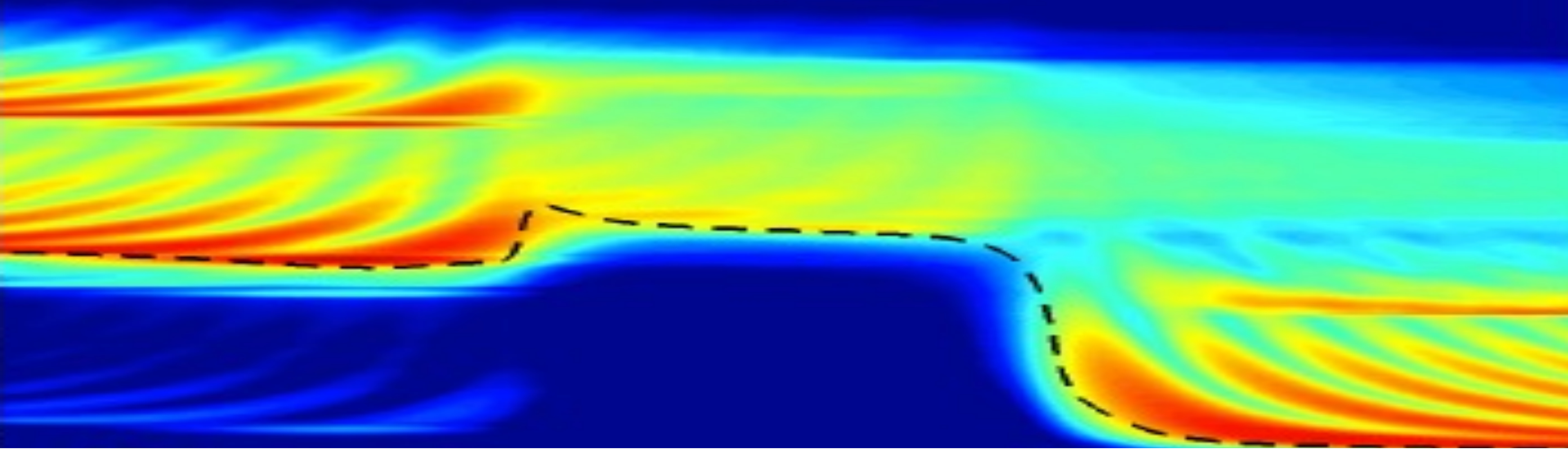


EECS 16B
Designing Information Devices and Systems II
Lecture 24

Prof. Yi Ma

Department of Electrical Engineering and Computer Sciences, UC Berkeley,
yima@eecs.berkeley.edu

Outline

- Singular Value Decomposition (Geometry)
  - Minimum Norm Solution and Optimal Control
- Low-rank Matrix Approximation (Algebra)
- Principal Component Analysis (Statistics)

Computation SVD

$$M : \mathbb{R}^n \to \mathbb{R}^m$$

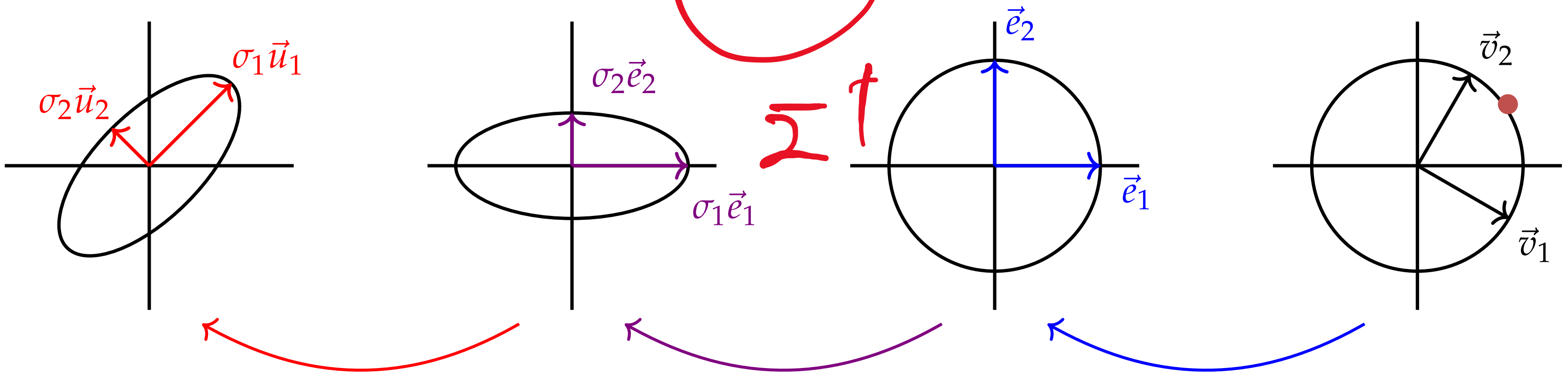
$$[\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n]$$

Interpretation of SVD (Geometry)

$$\boxed{\vec{y} = A\vec{x}:} \quad A = U\Sigma V^\top = [U_r, U_{m-r}] \begin{bmatrix} \Sigma_r & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{bmatrix} \underbrace{\begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix}}_{\mathbb{R}^n}$$

$\vec{x} \in \mathbb{R}^n \longrightarrow \vec{y} \in \mathbb{R}^m$ $\qquad \mathbb{R}^m \qquad \mathbb{R}^n \longrightarrow \mathbb{R}^m$

$$A = \quad U \quad \Sigma \quad V^\top$$



"rotation" scaling "rotation"

Applications of SVD: Minimum Norm Solution

$\min_{\vec{x}} \|\vec{x}\|_2^2$ s.t. $\vec{y} = A\vec{x}$, with $A \in \mathbb{R}^{m\times n}$ and $\mathrm{rank}(A) = m$: $\vec{x}_\star = A^\top(AA^\top)^{-1}\vec{y}$ ← 1

Show: $\vec{x}_\star = A^\dagger \vec{y}\ (= A^\top(AA^\top)^{-1}\vec{y})$.

$\vec{x}_* = A^\dagger \vec{y}$ ← 2

$A^\dagger = A^T(AA^T)^{-1}$

Recall least squares

$\min \|\underbrace{\vec{y} - A\vec{x}}_{\vec{e}\,\perp\, col(A)}\|_2^2$

$$\vec{y} = A\vec{x}_* = \underbrace{AA^T(AA^T)^{-1}}_{I}\vec{y}$$

$$\vec{y} = A(\vec{x}_* + \vec{s})$$

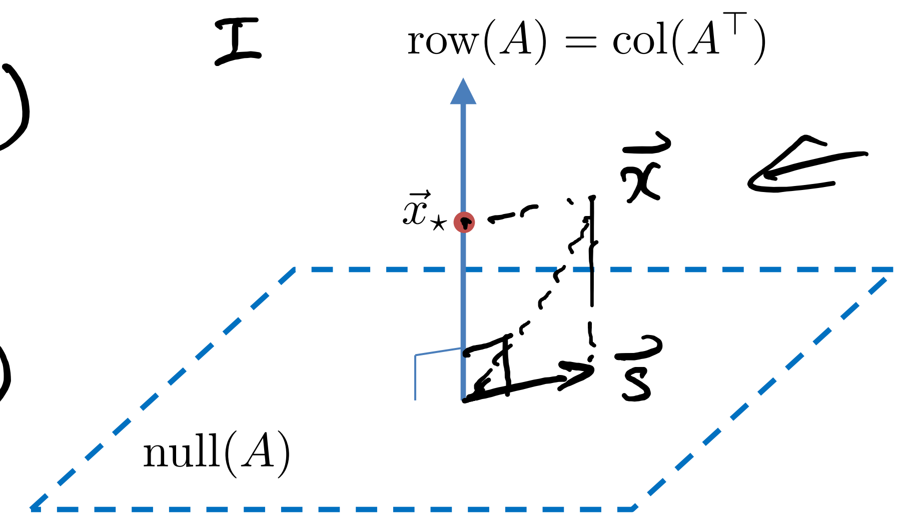
$$\Rightarrow 0 = A\vec{s}$$

$$\vec{s} \in Null(A)$$

if $\vec{x}_* \in col(A^T)$

$$\vec{x}_* = A^T(AA^T)^{-1}\vec{y}$$

↑ HW.



Applications of SVD: Minimum Norm Solution

Optimal Control: $\vec{x}[i+1] = A\vec{x}[i] + Bu[i]$

$\vec{x}[\ell] = A^\ell \vec{x}[0] + \mathcal{C}_\ell \vec{u}[\ell]$ $\quad \mathcal{C}_\ell \doteq \underbrace{[A^{\ell-1}B \mid \cdots \mid AB \mid B]}_{\text{rank}(\mathcal{C}_\ell) = n} \in \mathbb{R}^{n\times\ell}$ $\quad \vec{u}[\ell] = \begin{bmatrix} u[0] \\ \vdots \\ u[\ell-2] \\ u[\ell-1] \end{bmatrix} \in \mathbb{R}^\ell$

$\boxed{}_{n\times\ell}$

objective: $\vec{x}[\ell] = \vec{x}_f \in \mathbb{R}^n$

$\vec{x}_f = A^\ell \vec{x}[0] + \mathcal{C}_\ell \vec{u}[\ell] \Rightarrow \underbrace{\vec{x}_f - A^\ell \vec{x}[0]}_{\vec{y}} = \underbrace{\mathcal{C}_\ell}_{A} \underbrace{\vec{u}[\ell]}_{\vec{x}}$

3

$\min \|\vec{u}[\ell]\|_2^2 = u[0]^2 + \cdots + u[\ell-1]^2$

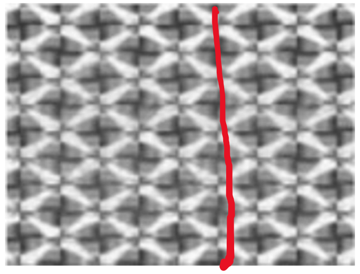
$\vec{u}_*[\ell] \in \text{col}(\mathcal{C}_\ell^T)$ $\quad \vec{u}_*[\ell] = \mathcal{C}_\ell^T \vec{w}$ $\quad \vec{y} = \mathcal{C}_\ell \mathcal{C}_\ell^T \vec{w}$

$\vec{w} = (\mathcal{C}_\ell \mathcal{C}_\ell^T)^{-1} \vec{y} \Rightarrow \vec{u}_*[\ell] = \mathcal{C}_\ell^T (\mathcal{C}_\ell \mathcal{C}_\ell^T)^{-1} (\vec{x}_f - A^\ell \vec{x}[0])$

Low-Rank Approximation (Algebra)

Modeling data as a low-rank matrix:

$$A = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n] \in \mathbb{R}^{m \times n}$$

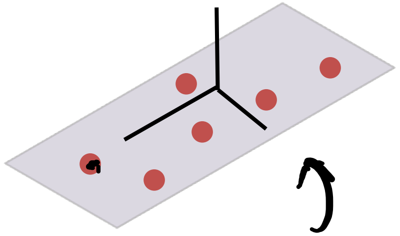


1000×1000

1 mil pixels

entries

$\mathbb{R}^{1000}$



dim = 10.

$$A = \underbrace{U}_{1000 \times 10}\ \underbrace{V^T}_{10 \times 1000}$$

$$= \vec{u}_1 \vec{v}_1^T + \cdots + \vec{u}_{10} \vec{v}_{10}^T$$

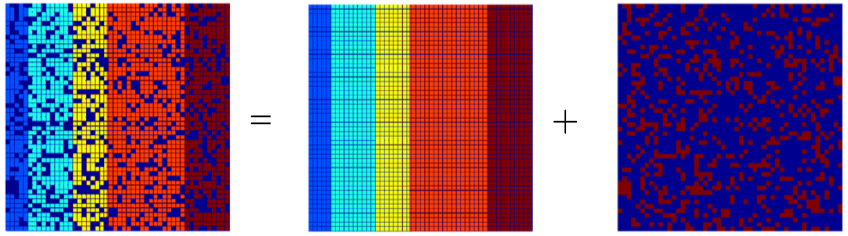
20,000 entries << 1 mil entries

Low-Rank Approximation (Algebra)

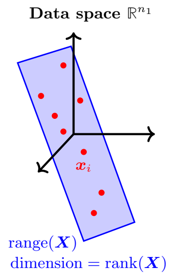
Approximate a matrix by a lower-rank matrix:

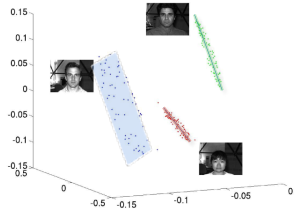
[Beltrami, 1873, Jordan, 1874]

$$A = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n] \in \underline{\mathbb{R}^{m \times n}}$$



A B N





$\mathrm{rank}(A) = r = \min(m, n)$

$$A = \underset{\text{low-rank}}{\underline{B}} + \underset{\text{small noise}}{\underline{N}}$$

$\mathrm{rank}(B) = l << \mathrm{rank}(A) = r$

$$\|M\|_F^2 = \sum_{i,j} |m_{ij}|^2$$

problem: Find a low-rank B s.t.

$$\min \|A - B\|_F^2 \quad (\text{small})$$

Low-Rank Approximation: Eckart-Young Theorem

Approximate a matrix $A \in \mathbb{R}^{m\times n}$ with rank $r \leq \min\{m, n\}$ by a lower-rank matrix.

$$A = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n] = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top = \underbrace{\sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top}_{B_\star} + \underbrace{\sum_{i=\ell+1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top} \quad \text{with } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$$

Theorem [Eckart-Young 1936]: The optimal solution to the low-rank approximation problem:

$$\min_{B \in \mathbb{R}^{m\times n}} \|A - B\|_F^2 \quad \text{subject to} \quad \text{rank}(B) = \ell$$

$$B = \boxed{}\;\boxed{\ell}$$

is given by: $B_\star = A_\ell = \sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top$.

$$\|M\|_F^2 = \text{tr}(M M^\top) = \text{tr}(M^\top M)$$

$$\|A - B_\star\|_F^2 = \Big\| \sum_{i=\ell+1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top \Big\|_F^2$$

$$= \text{tr}\Big(\sum_{i=\ell+1}^{r} \sigma_i \vec{u}_i \underbrace{\vec{v}_i^\top \vec{v}_i}_{1} \sigma_i \vec{u}_i^\top \Big) = \text{tr}\Big(\sum_{i=\ell+1}^{r} \sigma_i^2 \vec{u}_i \vec{u}_i^\top \Big)$$

$$= \sum_{i=\ell+1}^{r} \sigma_i^2$$

$$\| B_* \|_F^2 = \sum_{i=1}^{l} \sigma_i^2$$

$$\| A \|_F^2 = \sum_{i=1}^{r} \sigma_i^2$$

Low-Rank Approximation: Rank Minimization

Approximate a matrix $A \in \mathbb{R}^{m\times n}$ with rank $r \leq \min\{m,n\}$ by a lower-rank matrix.

$$A = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n] = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top = \underbrace{\sum_{i=1}^{\ell^*} \sigma_i \vec{u}_i \vec{v}_i^\top}_{} + \sum_{i=\ell+1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$$

Rank minimization problem: $\min_{B\in\mathbb{R}^{m\times n}} \underline{\operatorname{rank}(B)}$ subject to $\underline{\|A-B\|_F^2 \leq \epsilon^2}$?

$$\|A-B\|_F^2 = \sum_{i=\ell+1}^{r} \sigma_i^2 \leq \epsilon^2$$

$\min \ \ell = \operatorname{rank}(B)$?

$$\sigma_1^2 \geq \sigma_2^2 \cdots \geq \sigma_{\ell^*}^2 \geq \underbrace{\sigma_{\ell+1}^2 \cdots \sigma_{r-1}^2 \geq \sigma_r^2}_{\leq \epsilon^2}$$

$$\sum_{i=\ell^*}^{r} \sigma_i^2 > \epsilon^2$$

$$\sum_{i=\ell^*+1}^{r} \sigma_i^2 \leq \epsilon^2$$

Low-Rank Approximation: Model Selection

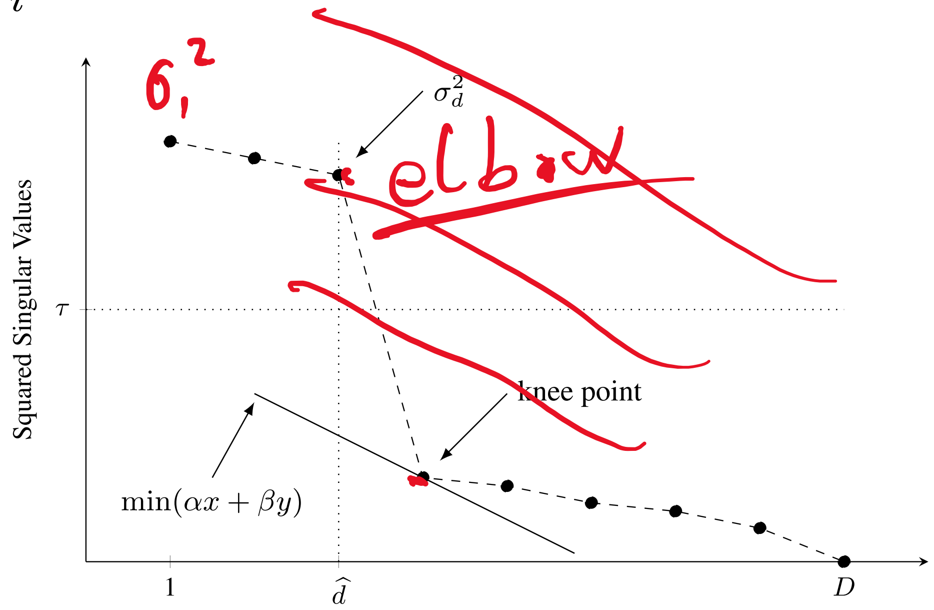
Approximate a matrix $A \in \mathbb{R}^{m\times n}$ with rank $r \le \min\{m,n\}$ by a lower-rank matrix.

$$A = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n] = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top = \sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top + \sum_{i=\ell+1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$$

Selecting a good tradeoff between rank and residual:

1. $\min_{B\in\mathbb{R}^{m\times n}} \ \mathrm{rank}(B) = d \quad \text{subject to} \quad \sigma_{d+1}^2 \le \tau?$

2. $\min_{B\in\mathbb{R}^{m\times n}} \ \alpha \cdot \mathrm{rank}(B) + \beta \cdot \sigma_{d+1}^2?$



Principal Component Analysis (Statistics)

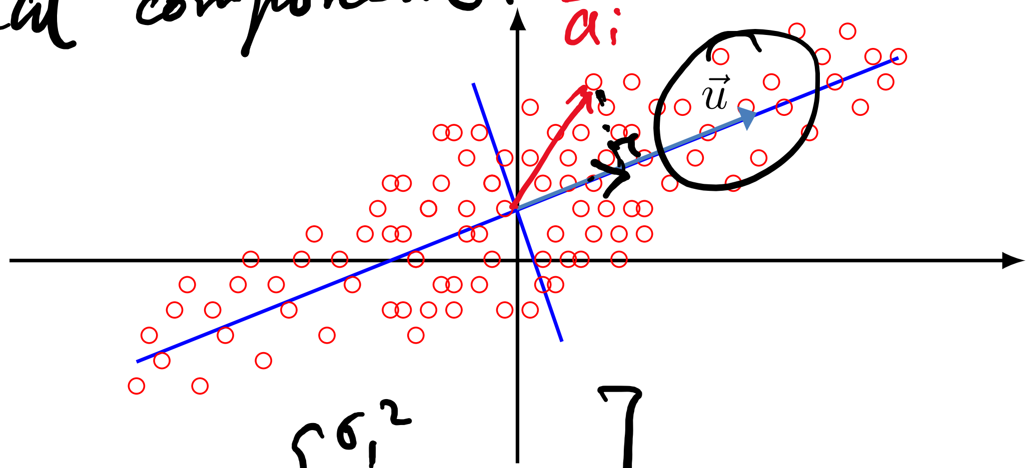
Problem [Pearson, 1901, Hotelling, 1933]: given $A = [\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n] \in \mathbb{R}^{m\times n}$ $\vec{\mu} = \frac{1}{n}(\vec{a}_1 + \vec{a}_2 + \cdots + \vec{a}_n) = \mathbf{0}$

find a normal vector $\|\vec{u}\|_2 = 1$ such that $\max\limits_{\vec{u}} \|\vec{u}^\top A\|_2^2 = \|\vec{u}\vec{u}^\top A\|_2^2$.

$\mathbb{R}^2$

$\vec{a}_i$ n samples $\in \mathbb{R}^m$ $\vec{u}$ — principal component.

$\vec{a}_i$



$$\|\vec{u}^T A\|_2^2 = \sum_{i=1}^{n} (\vec{u}^T \vec{a}_i)^2$$

$$= \|\vec{u}\,\underline{\vec{u}^T A}\|_2^2 \qquad \underline{\max} \|\vec{u}^T A\|_2^2 = \vec{u}^T A A^T \vec{u}$$

$$= \underline{\vec{u}^T U}\, \Lambda\, \underline{U^T \vec{u}} \qquad \Lambda = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix}$$

$$U^T \vec{u} = \vec{u}_*' = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\vec{u}_* = \vec{u}_1$$

$$= \vec{u}'^T \Lambda \vec{u}' = \lambda_1 \qquad \|\vec{u}'\|_2^2 = 1 \qquad \vec{u}_*' = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \qquad \lambda_1 \geq \lambda_2 \cdots \geq \lambda_r$$

Principal Component Analysis (Statistics)

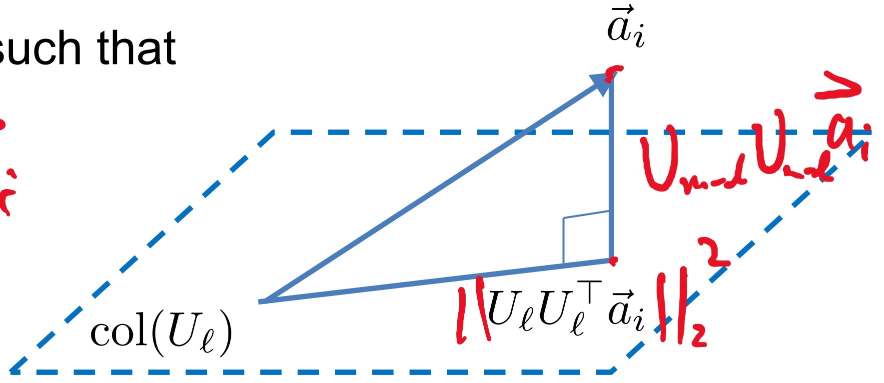
Problem [Pearson, 1901, Hotelling, 1933]**:** $A = [\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n] \in \mathbb{R}^{m \times n}$ find $\vec{u}$ such that $\max_{\vec{u}} \|\vec{u}^\top A\|_2^2$.

Multiple principal components: $U_\ell = [\underline{\vec{u}_1}, \ldots, \underline{\vec{u}_\ell}] \in \mathbb{R}^{m \times \ell}$ orthogonal such that

$$\underline{\max_{U_\ell} \|U_\ell U_\ell^\top A\|_F^2}$$

$$= \sum_{i=1}^{n} \|U_\ell U_\ell^\top \vec{a}_i\|_2^2$$

$$\underline{U_\ell U_\ell^\top \vec{a}_i}$$



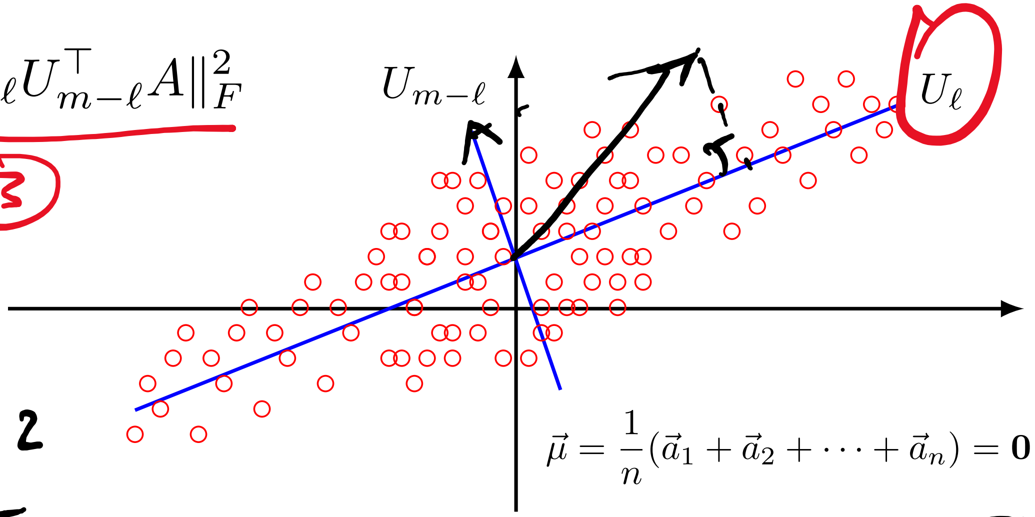
$$A = U \Sigma V^\top \qquad U = [\underline{U_\ell}, U_{m-\ell}] \in \mathbb{R}^{m \times m}$$

Principal Component Analysis (Statistics)

$a^2+b^2=c^2$

$U=[U_\ell, U_{m-\ell}]\in\mathbb{R}^{m\times m}$ orthogonal $\quad \|A\|_F^2 = \|UU^\top A\|_F^2 = \|U_\ell U_\ell^\top A\|_F^2 + \|U_{m-\ell}U_{m-\ell}^\top A\|_F^2$

$$\underbrace{\max_{U_\ell} \|U_\ell U_\ell^\top A\|_F^2}_{1} \Leftrightarrow \underbrace{\min_{U_\ell} \|A - U_\ell U_\ell^\top A\|_F^2}_{2} \Leftrightarrow \underbrace{\min_{U_{m-\ell}} \|U_{m-\ell}U_{m-\ell}^\top A\|_F^2}_{3}$$



$$\|A\|_F^2 = \|\underbrace{UU^T}_{I}A\|_F^2$$

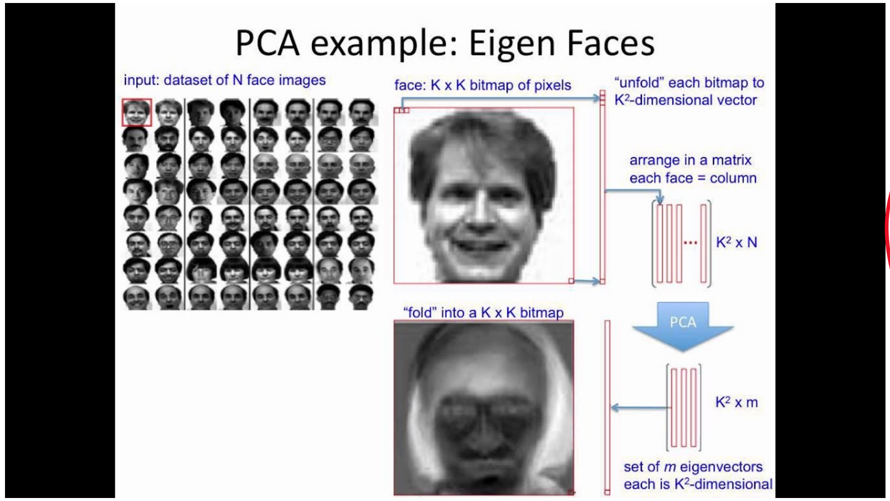
$$= \underline{\|U_\ell U_\ell^T A\|_F^2} + \underbrace{\|U_{m-\ell}U_{m-\ell}^T A\|_F^2}_{(I-U_\ell U_\ell^T)}$$

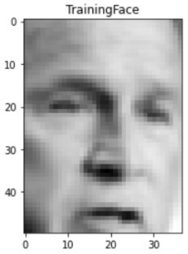
$$\underbrace{UU^T}_{I} = \underline{U_\ell U_\ell^T + U_{m-\ell}U_{m-\ell}^T}$$

$\mathrm{col}(U_\ell)$ — principal subspace

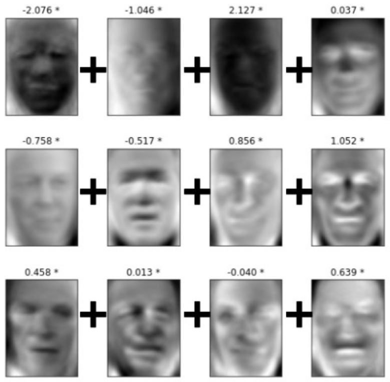
Applications of PCA

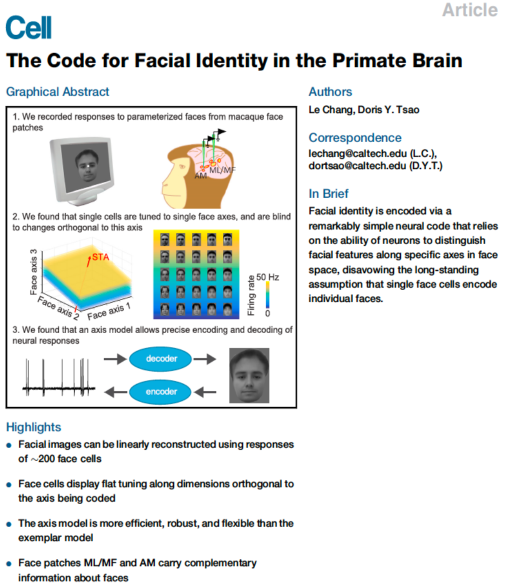
- Eigenfaces [Turk & Pentland 1991]:





=




Article

Cell

The Code for Facial Identity in the Primate Brain

Graphical Abstract

1. We recorded responses to parameterized faces from macaque face patches

2. We found that single cells are tuned to single face axes, and are blind to changes orthogonal to this axis

3. We found that an axis model allows precise encoding and decoding of neural responses

Authors

Le Chang, Doris Y. Tsao

Correspondence

lechang@caltech.edu (L.C.), dortsao@caltech.edu (D.Y.T.)

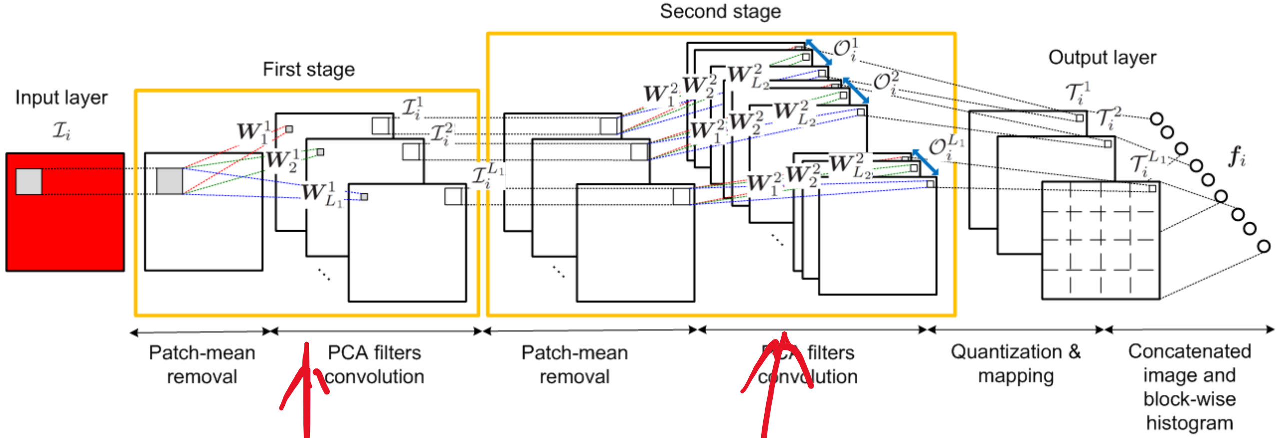
In Brief

Facial identity is encoded via a remarkably simple neural code that relies on the ability of neurons to distinguish facial features along specific axes in face space, disavowing the long-standing assumption that single face cells encode individual faces.

Highlights

- Facial images can be linearly reconstructed using responses of ~200 face cells
- Face cells display flat tuning along dimensions orthogonal to the axis being coded
- The axis model is more efficient, robust, and flexible than the exemplar model
- Face patches ML/MF and AM carry complementary information about faces

- PCANet [Chan & Ma et. al. 2015]:

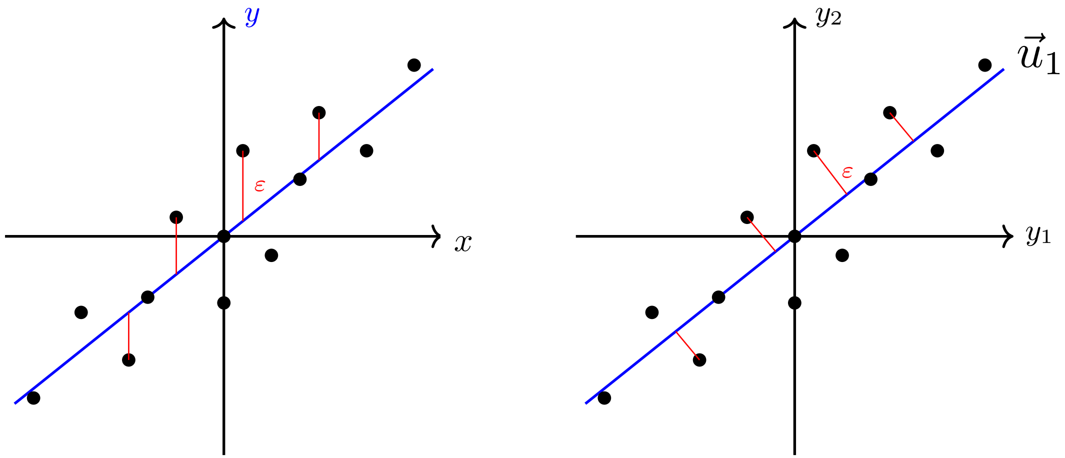


Recognition rates (%) on FERET dataset.

Probe sets	Fb	Fc	Dup-I	Dup-II	Avg.
LBP [18]	93.00	51.00	61.00	50.00	63.75
DMMA [25]	98.10	98.50	81.60	83.20	89.60
P-LBP [21]	98.00	98.00	90.00	85.00	92.75
POEM [26]	99.60	99.50	88.80	85.00	93.20
G-LQP [27]	99.90	100	93.20	91.00	96.03
LGBP-LGXP [28]	99.00	99.00	94.00	93.00	96.25
sPOEM+POD [29]	99.70	100	94.90	94.00	97.15
GOM [30]	99.90	100	95.70	93.10	97.18
PCANet-1 (Trn. CD)	99.33	99.48	88.92	84.19	92.98
PCANet-2 (Trn. CD)	99.67	99.48	95.84	94.02	97.25
PCANet-1	99.50	98.97	89.89	86.75	93.78
PCANet-2	99.58	100	95.43	94.02	97.26

Least Squares (Regression) versus PCA

Prediction versus Correlation



Supervised versus Unsupervised

Least Squares (Regression) versus PCA

Example: $A = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 1 & 1 \end{bmatrix}$

Prediction versus Correlation