Dimensionality Reduction

Internet Analytics (COM-308)

Prof. Matthias Grossglauser
School of Computer and Communication Sciences

EPFL

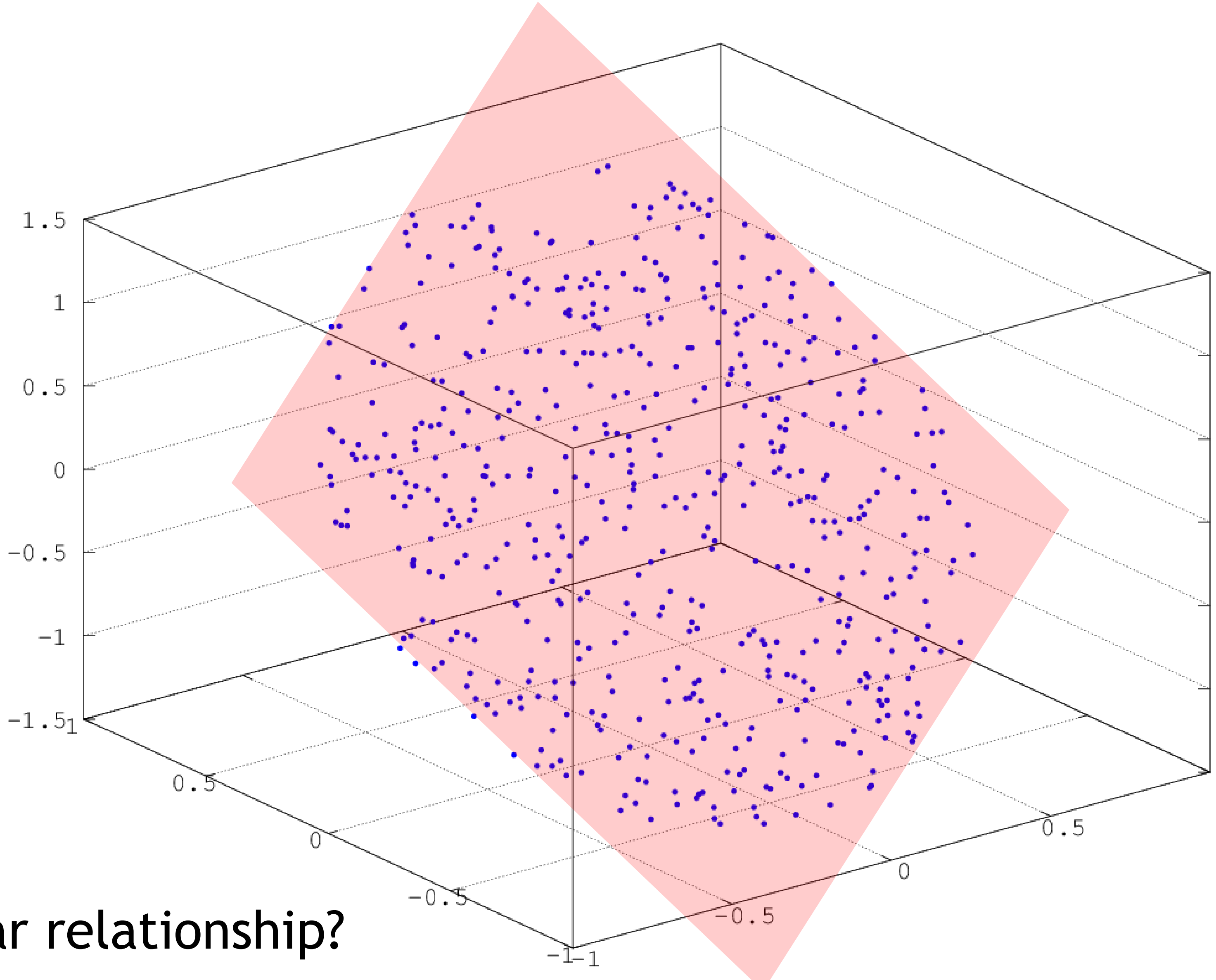
Overview

- Introduction and motivation
- Singular Value Decomposition (SVD)
  - Every matrix has a SVD
  - Intuition
  - Applications in dimensionality reduction
- Principal Component Analysis (PCA)
  - Visualization and exploration
  - Goal: find low-dimensional projection that represents data well
- Comments on Multi-Dimensonal Scaling (MDS) and non-linear embedding

What is dimensionality reduction?

- Goal: find "structure" in high-dimensional data
  - Structure means: patterns, dependencies, clusters,…
- Motivating example:
  - Stock price analysis: we want to understand the structure of the stock market
  - One data point X_i: stock quotes for one day
  - 1000 stocks: dimension of full space ($m = 1000$)
  - n data points
  - Is there structure, i.e., exact or approximate relationships?
  - In other words: does data "concentrate in" a subspace of $\mathbb{R}^m$?

Example: 3d data with 2d structure



Linear relationship?

Case study: Smartvote dataset

- smartvote pre-electoral opinions of the 2011 parliamentary elections
  - 2,985 candidates (82.4% of all candidates)
  - 229,133 citizens (~9% of total turnout)
- Examples of questions:
  - “Should Switzerland embark on negotiations in the next four years to join the EU?”
  - “How much should the public transport budget be?”
- Possible answers
  - strongly disagree - disagree - agree - strongly agree
  - less - no change - more

Case study: Smartvote dataset

Applications of dim reduction

- Visualization & interpretation
  - Useful first step in data analysis
- Discover hidden correlations, laws, mechanisms
- Noise reduction
  - For example, data could be truly low-dimensional, but noise is high-dimensional
- Efficiency: compression & processing
  - Many algorithms are hard in high dimensions (“the curse of dimensionality”)
  - E.g., nearest neighbor

Spectral theorem

- Theorem:
  - A real symmetric matrix X can be factored as
    $$X = QDQ^T,$$
    where Q is orthogonal ($Q^{-1} = Q^T$) and D is diagonal.
- Convention:
  - Write diagonal values in decreasing order
  - $D = diag(\lambda_1, \lambda_2, \dots \lambda_n)$
- Def: positive definite:
  - All $\lambda_i > 0$
  - $x^T X x > 0$ for all nonzero vectors x
- Def: positive semidefinite (PSD):
  - All $\lambda_i \geq 0$
  - $x^T X x \geq 0$ for all vectors x

Singular Value Decomposition (SVD)

- Theorem:
  - Any real $n \times m$ matrix X can be factored as
    $$X = U\Sigma V^T,$$
    where
    U is $n \times n$ and orthogonal,
    V is $m \times m$ and orthogonal, and
    Σ is $n \times m$ diagonal
- Proof:
  - $X^T X$ is symmetric and positive semidefinite
  - Apply spectral theorem to $X^T X$
    - There exists orthogonal V such that $V^T X^T X V = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$
    - D is diagonal and positive

SVD: existence (cont.)

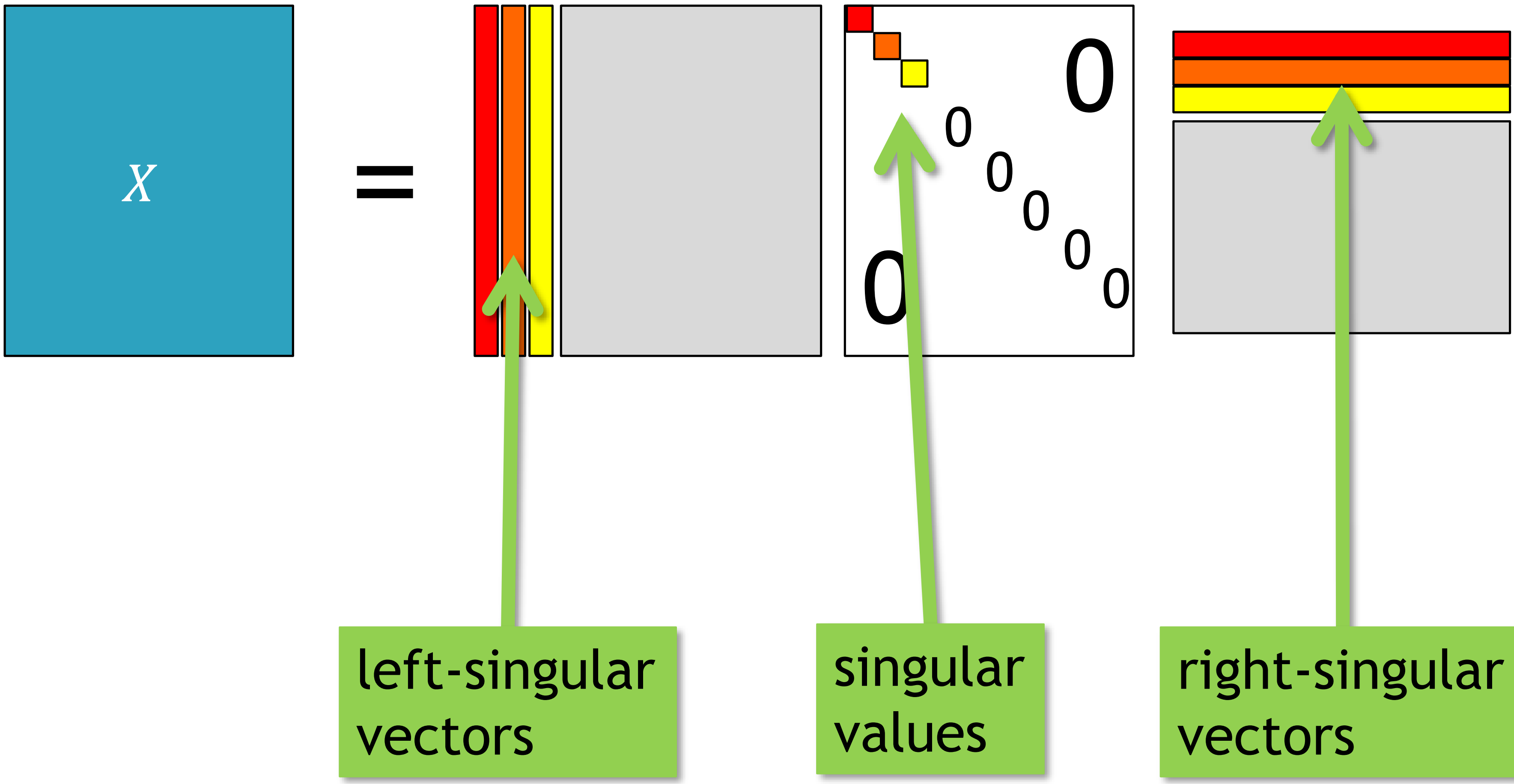
- Proof (cont):
  - $D = diag(\lambda_1, \lambda_2, \ldots \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$
  - $r = rank(X)$
  - $\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} X^T X \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$
  - This shows that
    $V_1^T X^T X V_1 = D$,
    and that
    $V_2^T X^T X V_2 = 0$; this implies $XV_2 = 0$ (null space of X)
  - Also: V orthogonal → $VV^T = I = V_1 V_1^T + V_2 V_2^T$
  - $v_i^T X^T X v_j = X v_i \circ X v_j = \begin{cases} \lambda_j & i = j \ (i, j \leq r) \\ 0 & \text{otherwise} \end{cases}$

SVD: existence (cont.)

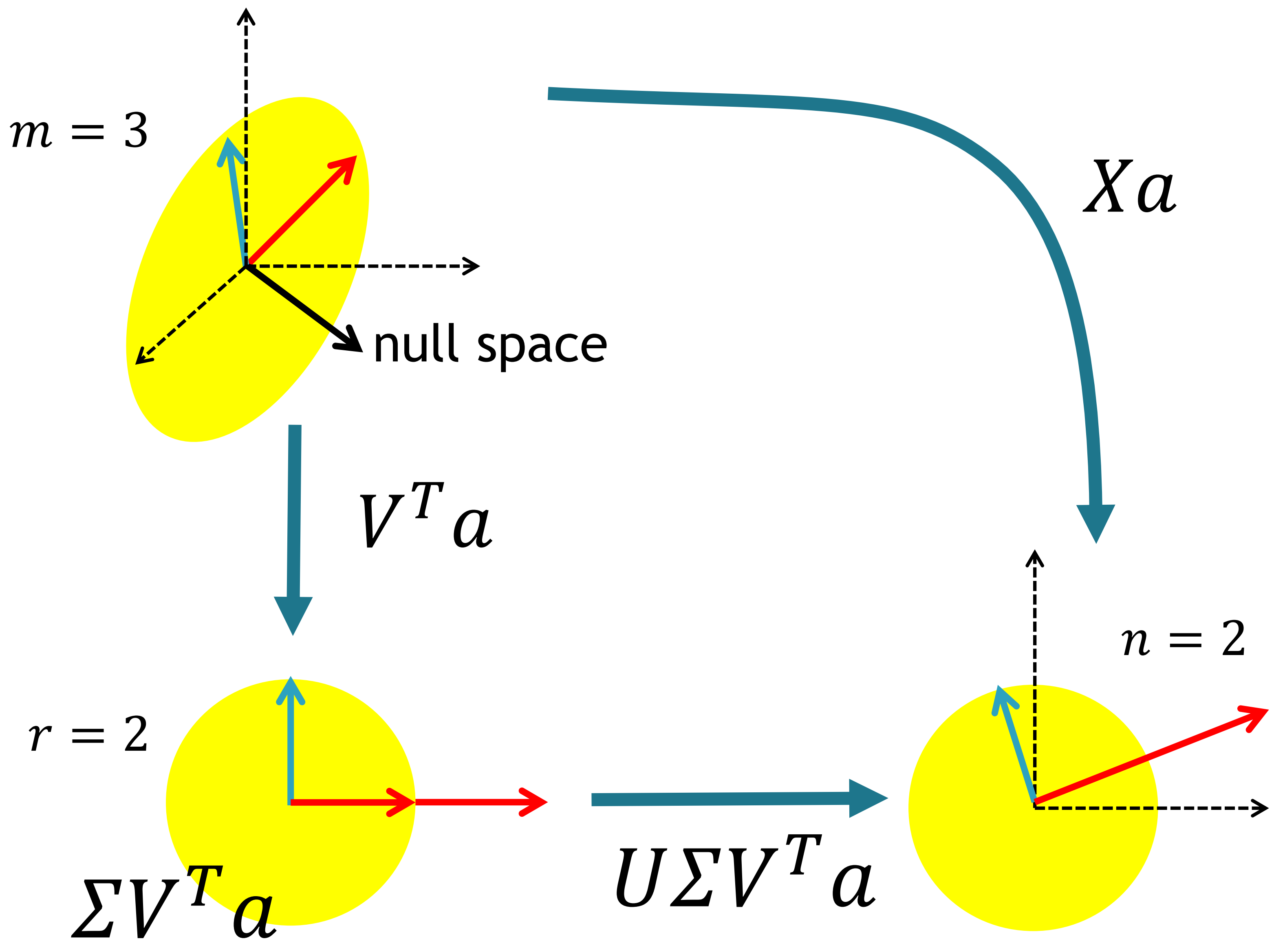
- Proof (cont.):
  - Let $\sigma_j = \sqrt{\lambda_j}$
  - Let $\Sigma = \begin{bmatrix} \sqrt{D} & 0 \\ 0 & 0 \end{bmatrix}$ the $n \times m$ matrix with σ_j on the diagonal (otherwise 0)
  - Set $U_1 = XV_1D^{-\frac{1}{2}}$
    - Note $u_j = \dfrac{1}{\sigma_j} X v_j$ are orthonormal
  - Complete remaining vectors $U_2 = [u_{r+1}, \dots, u_n]$ to have orthonormal basis of $\mathbb{R}^n$
  - $U\Sigma V^T = \begin{bmatrix} XV_1D^{-\frac{1}{2}} & U_2 \end{bmatrix} \begin{bmatrix} \sqrt{D} & 0 \\ 0 & 0 \end{bmatrix} [V_1 \;\; V_2]^T = XV_1V_1^T =$

    $= X\left(I - V_2V_2^T\right) = X$ (because $XV_2 = 0$)

SVD $U\Sigma V^T$



SVD: geometric interpretation of $U\Sigma V^T a$



Singular Value Decomposition (SVD)

- Alternative definition:

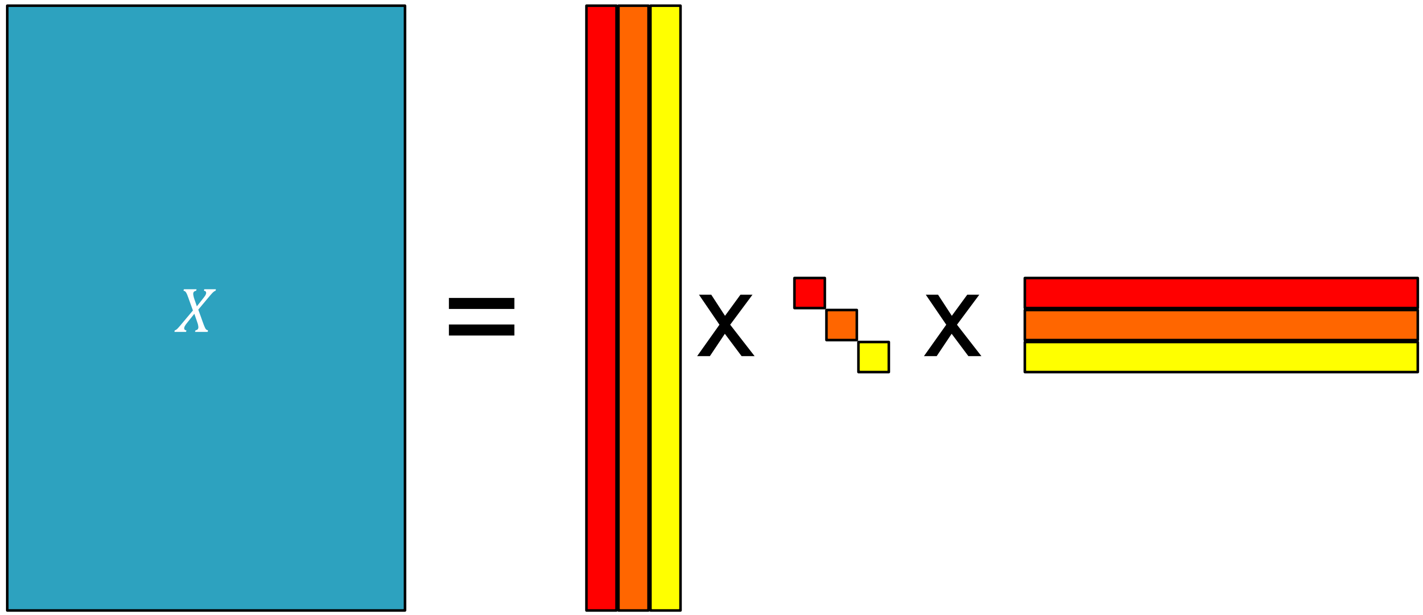
$$X = U\Sigma V^T$$

where:

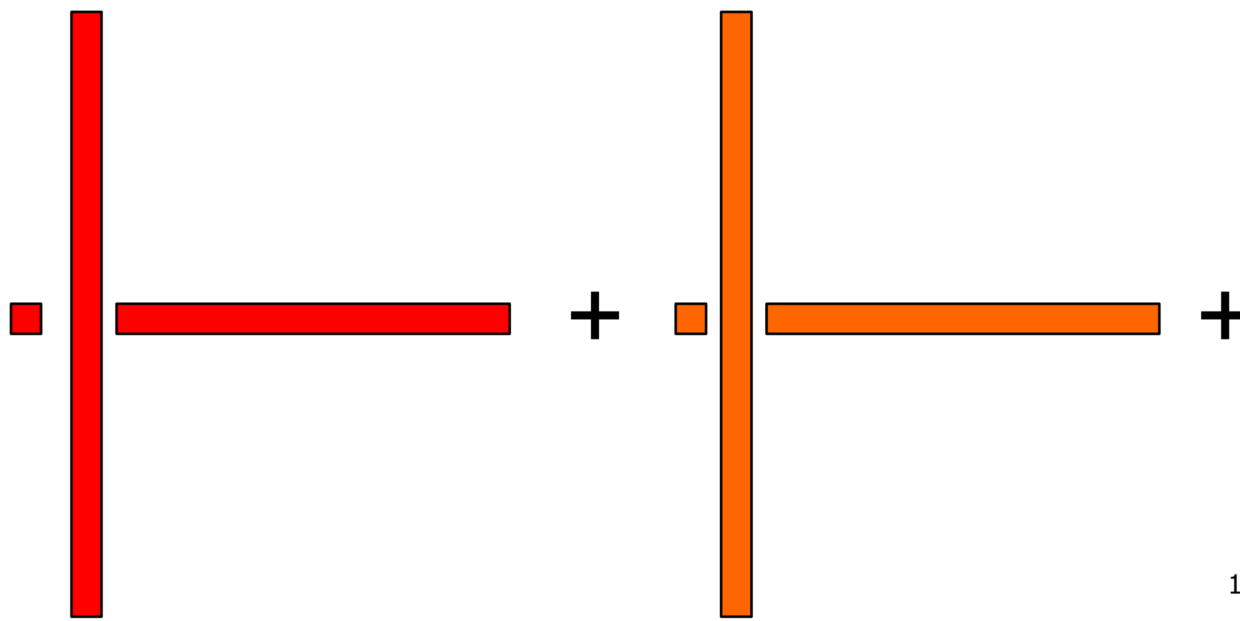
- $r = rank(X)$
- U is column-orthonormal ($n \times r$) (“tall”)
  - $U^T U = I$
- V^T is row-orthonormal ($r \times m$) (“fat”)
  - $V^T V = I$
- Σ is diagonal ($r \times r$)
  - Singular values of X

SVD: low-rank approximation

$$X = U\Sigma V^T$$



$$X = \sum_{1}^{r} \sigma_i U_i V_i^T =$$



Singular Value Decomposition (SVD)

- Goal:
  - Find low-dimensional latent space that "explains" data
- Motivating example: survey
  - We have $n = 5$ individuals and $m = 4$ questions
  - Each person answers questions in a range (e.g., -5 to 5)
  - Represent as a matrix: $X = \begin{bmatrix} 5 & 0 & 0 & -4 \\ -4 & -1 & 0 & 4 \\ -5 & 5 & 5 & 5 \\ 0 & 4 & 5 & 0 \\ 5 & -5 & -5 & -5 \end{bmatrix}$
- Latent space/concepts/hidden variables:
  - Some people are similar, and some questions are similar
  - Question: how many "degrees of freedom" or "dimensions" does the system have?

Singular Value Decomposition (SVD)

- $$U = \begin{bmatrix} -0.30 & 0.54 & -0.12 & 0.78 & 0 \\ 0.24 & -0.54 & -0.72 & 0.35 & 0 \\ 0.62 & 0.11 & 0.23 & 0.21 & 0.71 \\ 0.26 & 0.63 & -0.60 & -0.43 & 0 \\ -0.62 & -0.11 & -0.23 & -0.21 & 0.71 \end{bmatrix}$$

- $$V = \begin{bmatrix} -0.55 & 0.49 & -0.07 & 0.67 \\ 0.44 & 0.53 & 0.72 & 0.05 \\ 0.47 & 0.54 & -0.69 & -0.09 \\ 0.53 & -0.42 & -0.06 & 0.73 \end{bmatrix}$$

- $\Sigma = \mathrm{diag}(\mathbf{16}, \mathbf{7}.\mathbf{7}, 0.9, 0.5)$

SVD: Interpretation

- Reformulation as sum of outer products:

$$X = \sum_{i=1}^{r} \sigma_i U_i V_i^T$$

- σ_i: strength of concept i
- U_i: influence of concept i on “people”
- V_i : influence of concept i on “questions”

SVD: Best rank(r)-approximation

- Frobenius norm:
  - $||X||_F^2 = \sum_{i,j} X_{i,j}^2$
- Theorem:
  - Let X be any matrix, and $X = U\Sigma V^T$ its SVD
  - Let $X' = \sum_{i=1}^{r} \sigma_i U_i V_i^T$ a rank(r)-approximation of X
  - Then $||X - X'||_F^2$ is smallest possible for rank=r
- Intuition:
  - X' captures the most important dimensions of the linear map
- Criterion for r:
  - Often, try to capture ~ 80-90% of "energy" in X, i.e., of $||X||_F^2$

Best rank(r)-approx: example

- $X = \begin{bmatrix} 5 & 0 & 0 & -4 \\ -4 & -1 & 0 & 4 \\ -5 & 5 & 5 & 5 \\ 0 & 4 & 5 & 0 \\ 5 & -5 & -5 & -5 \end{bmatrix}$
- $X'_1 = \sigma_1 U_1 V_1^T = \begin{bmatrix} 2.7 & -2.1 & -2.3 & -2.6 \\ -2.1 & 1.7 & 1.8 & 2.0 \\ -5.5 & 4.4 & 4.7 & 5.3 \\ -2.3 & 1.8 & 2.0 & 2.2 \\ 5.5 & -4.4 & -4.7 & -5.3 \end{bmatrix}$
- $X'_2 = \sum_{i=1}^{2} \sigma_i \, U_i V_i^T = \begin{bmatrix} 4.7 & 0.06 & -0.04 & -4.3 \\ -4.2 & -0.5 & -0.4 & 3.8 \\ -5.1 & 4.8 & 5.1 & 4.9 \\ 0.1 & 4.4 & 4.6 & 0.1 \\ 5.1 & -4.8 & -5.1 & -4.9 \end{bmatrix}$

Principal Component Analysis (PCA)

- Data matrix X:
  - Row: data point (n)
  - Columns: dimensions (m)
- Goal:
  - Explain relationships between variables
- Approach:
  - Low-dimensional representation conserving “variability”

Stock A	Stock B	Stock C	Stock D

n

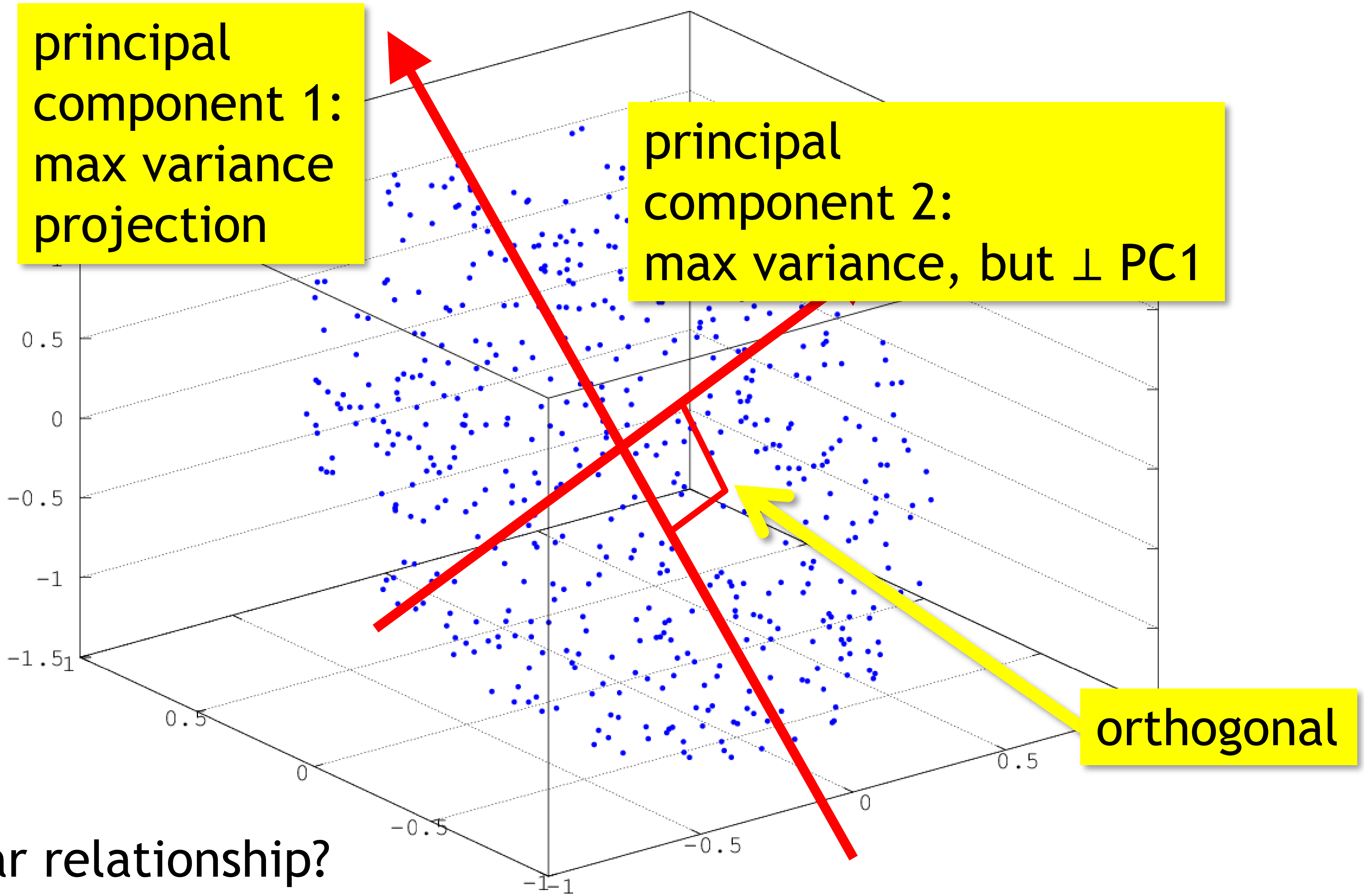
m

PCA

hemos restado la media

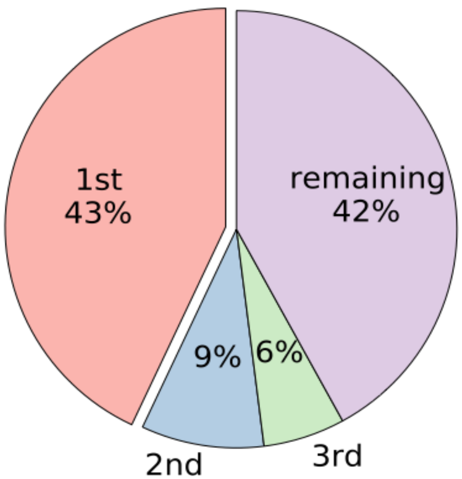
- $\frac{1}{n}X^TX$: covariance matrix (X centered)
  - $\left(X^TX\right)_{ij}$: inner (scalar) product of variables i and j
  - Large value = strongly correlated dimensions
- Eigenpairs: (v_i, λ_i) of $X^TX = V\Lambda \mathrm{V}^{\mathrm{T}}$
  - v_i: ith eigenvector (unit)
  - λ_i: ith-largest eigenvalue
  - Choose a dimension $d \ll m$
  - Define $V = [v_1, v_2, \ldots, v_d]$
  - Define $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d]$
- $Y = XV$: points of X projected on new space
  - Note: $Y^TY = V^TX^TXV = V^TV\Lambda V^TV = \Lambda$ → principal components are decorrelated

Example: 3d data with 2d structure

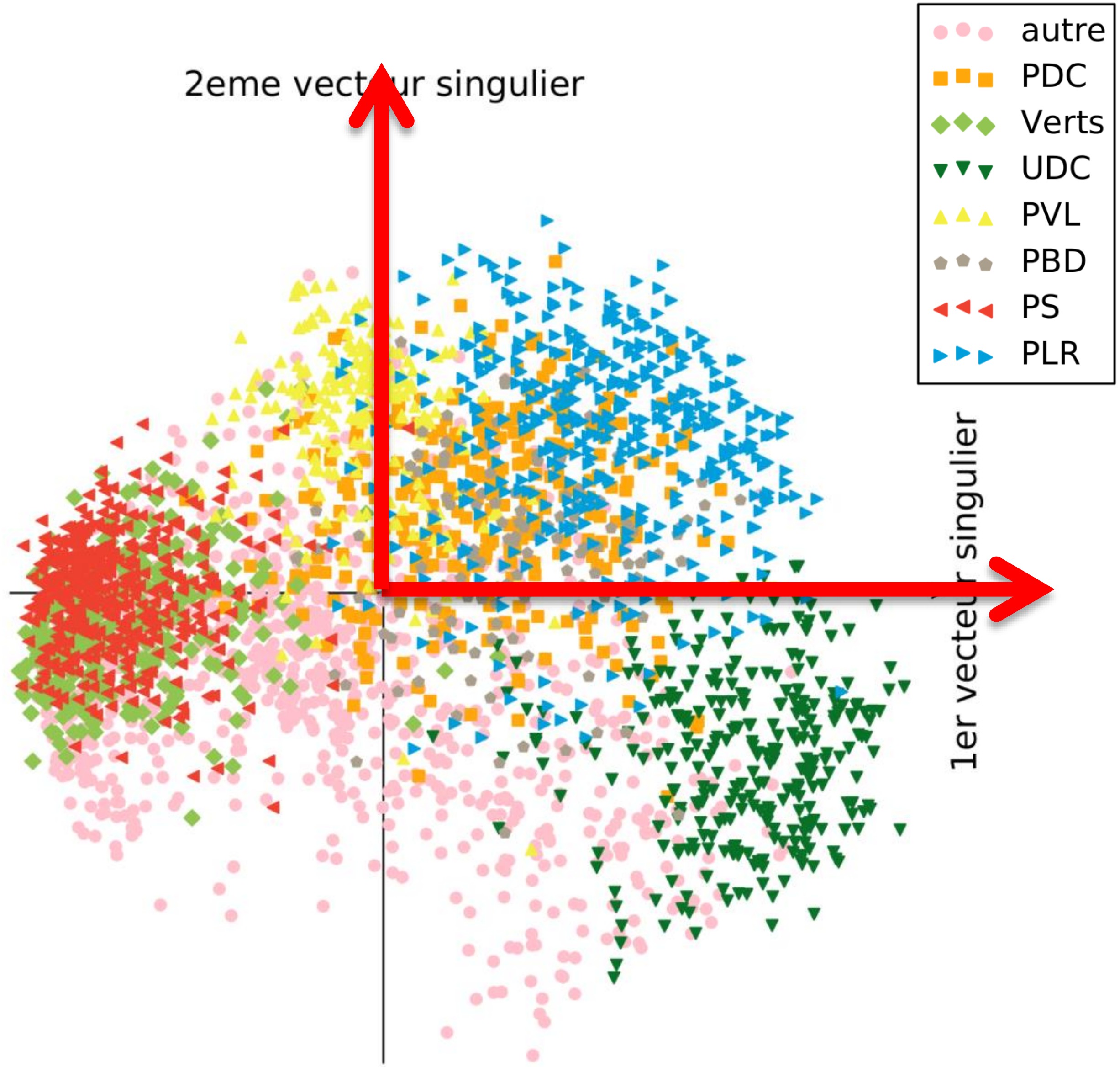


Linear relationship?

Case study: PCA on smartvote data



3 PCs capture
~ 60% of variance



Principal component v_1

1st axis
- Seriez-vous favorable à ce que le **droit de vote** au niveau communal soit instauré pour les **étrangers** qui vivent en Suisse depuis au moins dix ans et ce, dans toute la Suisse?
- Approuveriez-vous que la **concurrence fiscale** entre les **cantons** soit plus limitée?
- Soutenez-vous l'initiative populaire qui souhaite que le **salaire** le plus élevé au sein d'une **entreprise** ne puisse pas être plus de douze fois supérieur au salaire le plus bas versé par la même entreprise. (initiative 1:12)?
- Une initiative populaire souhaite instaurer une **caisse maladie** unique et publique pour l'assurance de base. Êtes-vous favorable à ce projet?

Social questions («égalité»)

Principal component v_2

2[nd] axis
- Approuvez-vous des engagements de soldats armés (pour l'autoprotection) de l'**armée** suisse à l'**étranger** dans le cadre de missions de maintien de la paix de l'ONU ou de l'OSCE?
- Êtes-vous en faveur d'un accord de **libre-échange** agricole avec l'**UE** ?
- Êtes-vous favorable à l'accord sur la **libre circulation** des personnes existant avec l'UE?
- Une imposition centrale sur les quantités dans la production laitière doit-elle être réinstaurée en Suisse à la place du **libre marché** laitier?

Economics, globalisation («liberté»)

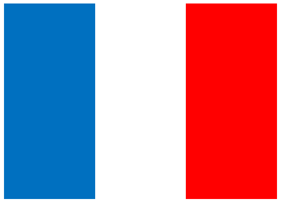
Principal component v_3

3rd axis
- Seriez-vous favorables à ce que l'**euthanasie** active directe soit légalement possible par le biais d'un médecin en Suisse?
- Les couples **homosexuels** sous le régime du partenariat enregistrés devraient-ils pouvoir adopter des enfants?
- La Suisse possède des règles relativement strictes concernant la **procréation** médicalement assistée. Celles-ci devrait-elles être assouplies?
- La consommation ainsi que la possession pour la consommation personnelle de **drogues** dures et douces doivent-elles être légalisées?

Society, ethics («fraternité»)

In other words: PCA produces the French flag ;)

Observation:
- Principal components correspond to clearly interpretable political and ideological dimensions



PCA: Covariance vs correlation matrix

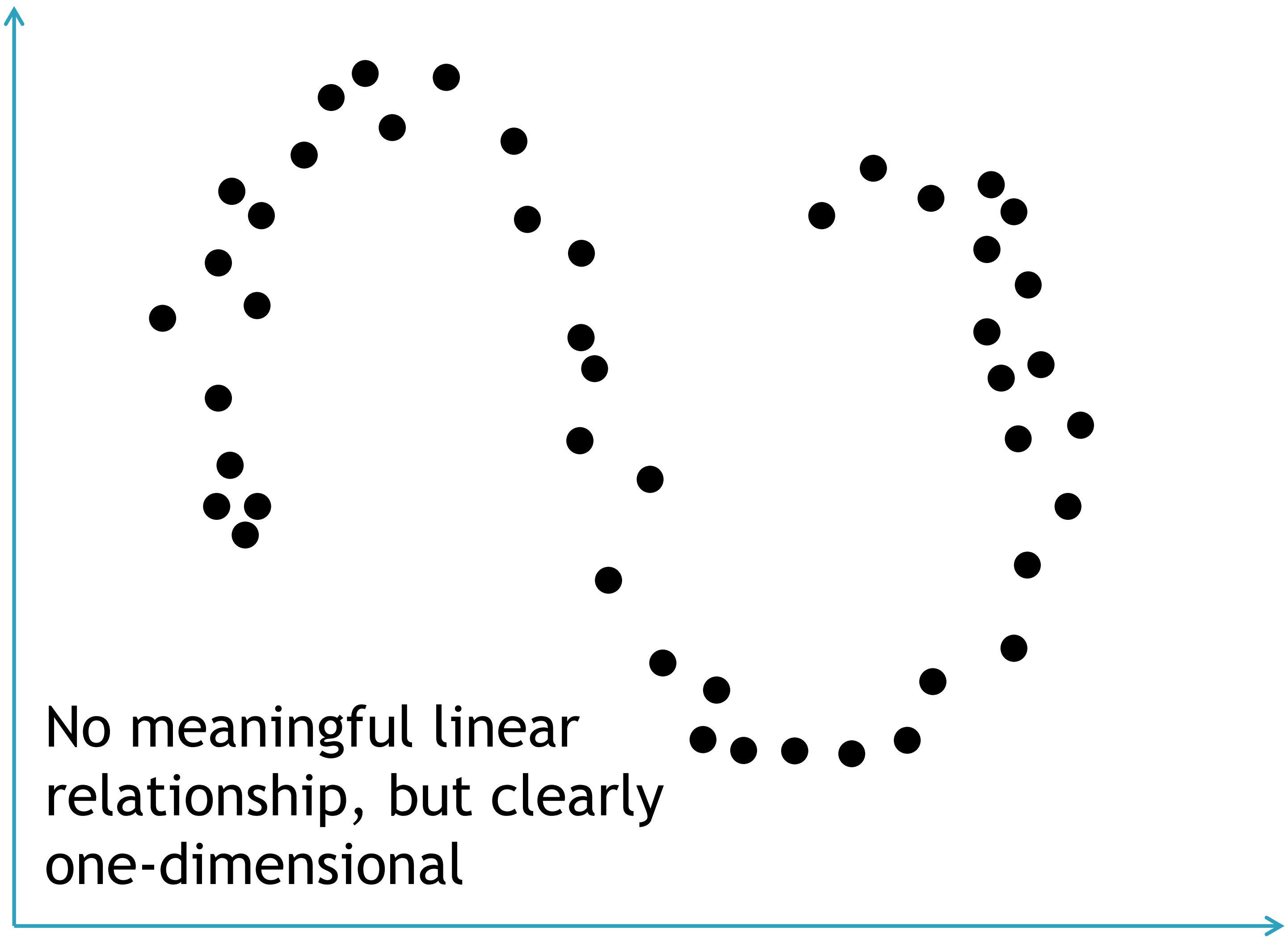
- Assume X centered, i.e., $1_n X = 0_m$
  - If not, do not forget to center it first!
- Covariance matrix: $\frac{1}{n} X^T X$
- Correlation matrix R:
  - $R_{ij} = \frac{X_i^T X_j}{\sqrt{(X_i^T X_i)(X_j^T X_j)}}$
  - Normalized, $-1 \leq R_{ij} \leq 1$
  - Advantage: unit/range independent
  - Good when different dimensions are numerically very different, or even in different units
- Ultimately scenario-dependent
  - Considered a drawback of PCA

Multidimensional Scaling (MDS)

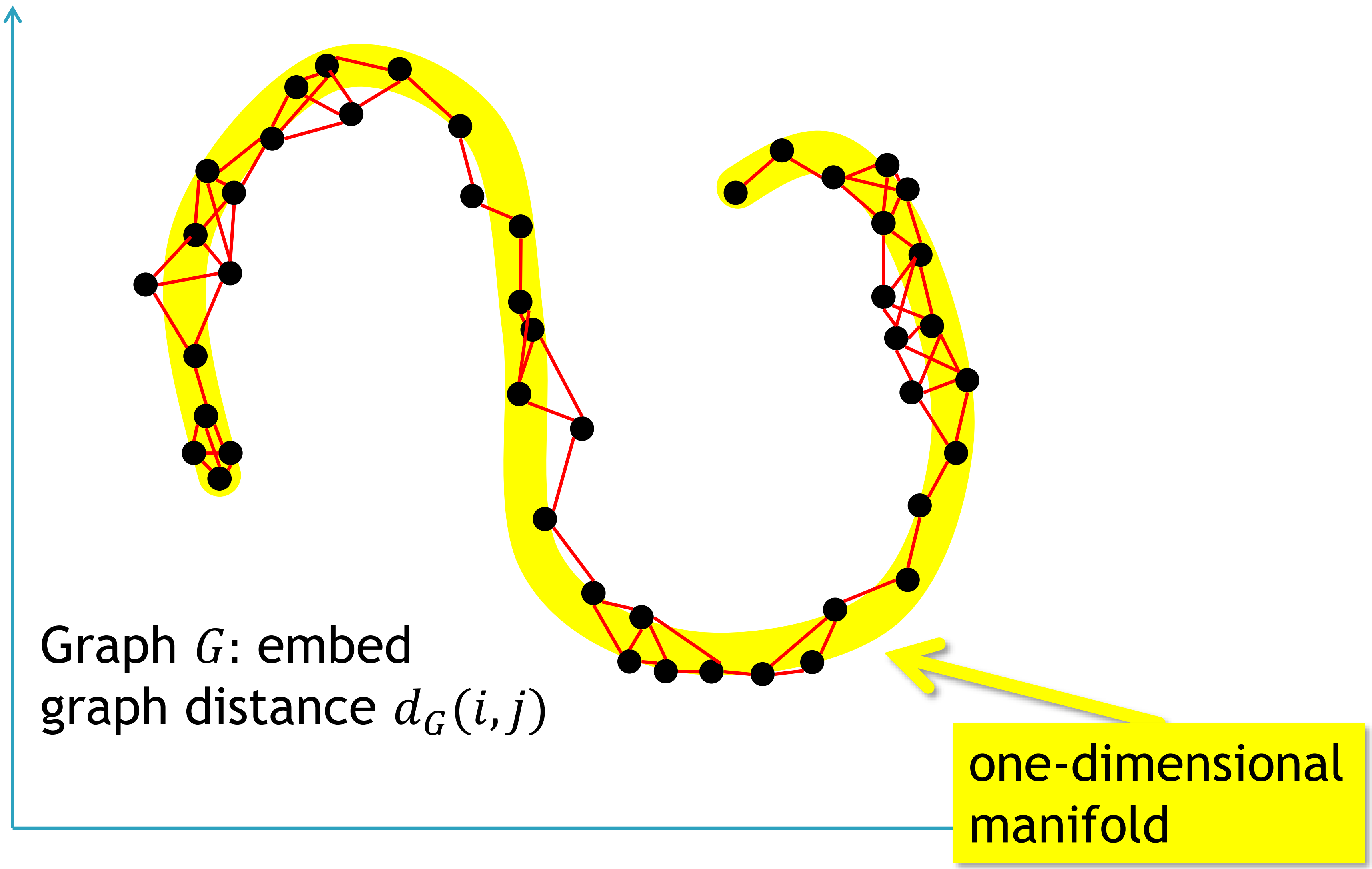
- PCA: two strong assumptions
  - Linear relationships among dimensions
  - Orthogonal principal components
- Often low-dimensional structure exists, but above assumptions are too strong
- Generalization: MDS
  - PCA: find structure in data $\{X_i\}$
  - MDS: Find structure in metric space (distance function): $d(X_i, X_j)$
  - Choice of distance function allows to generalize (Euclidean → PCA)

In MDS you can choose the distance function, and allows MDS to makes sense of data when PCA fails

Non-linear embedding: motivation

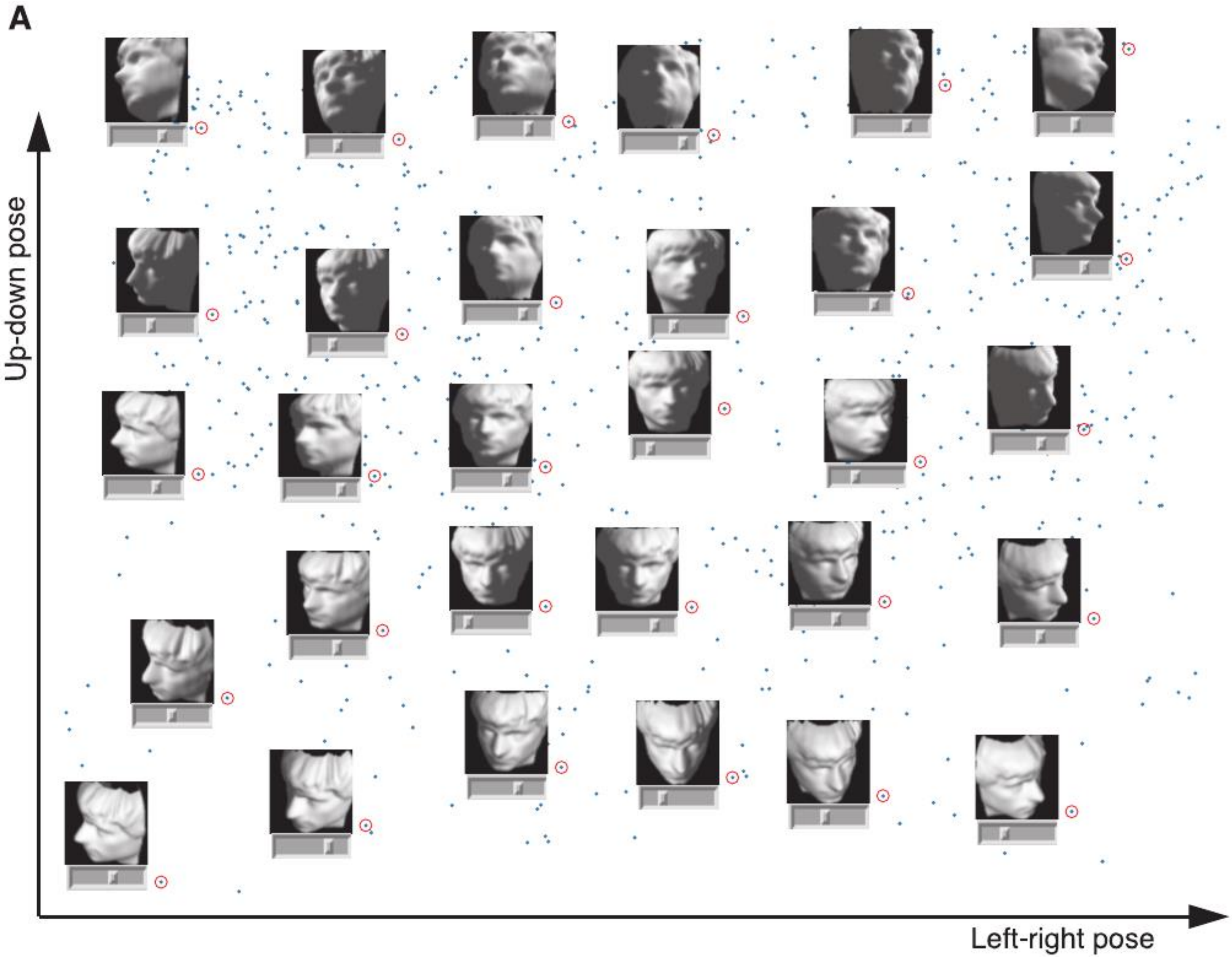


Isomap: approximate geodesic distance



Isomap: example

Source: [Tenenbaum et al.]



Stochastic Neighborhood Embedding

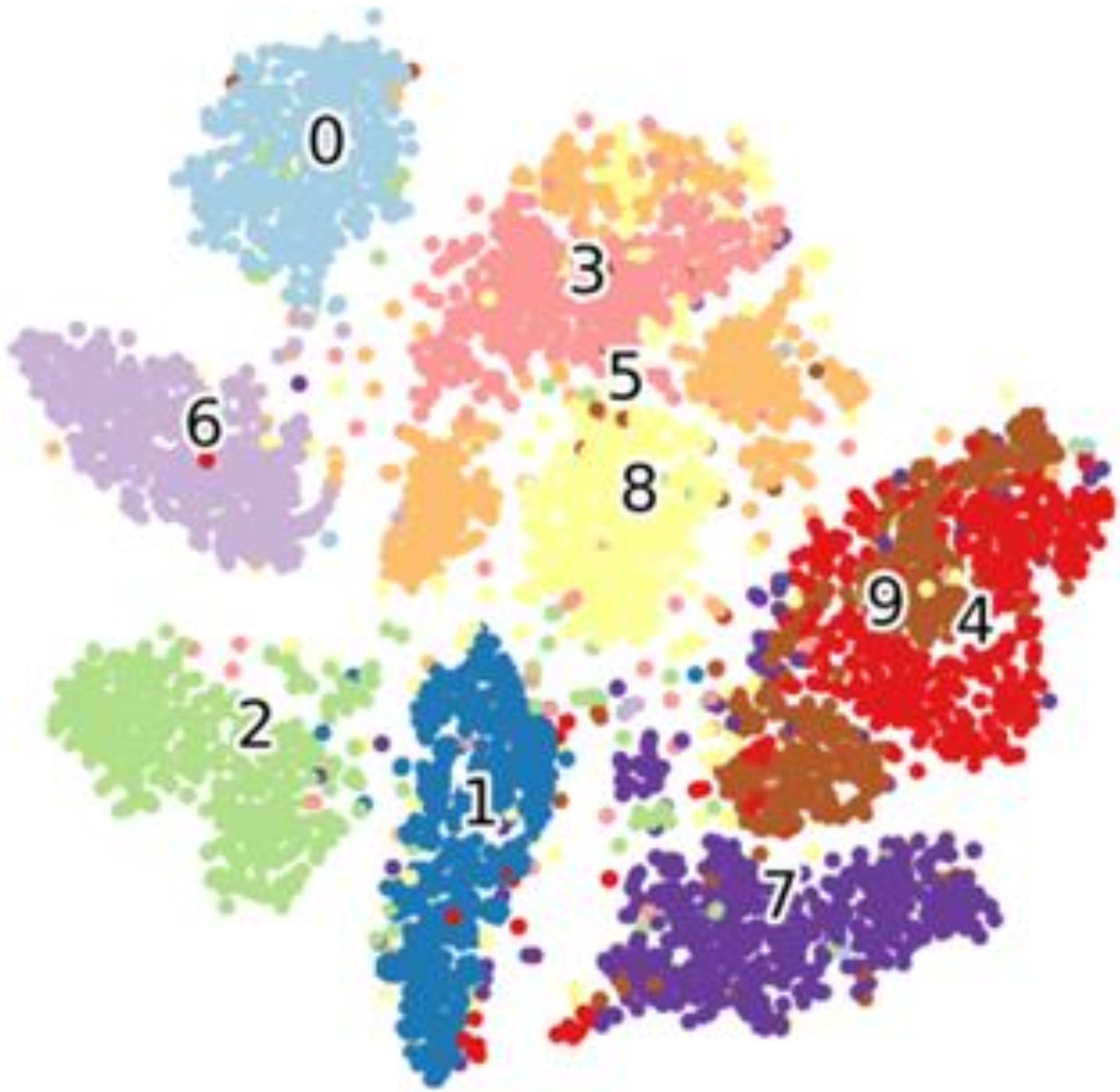
- Key idea: try to preserve nearest-neighbor relationships as much as possible
  - Penalize solution where close points (x_i, x_j) is mapped to (y_i, y_j) not close
  - Do not care (very much) about how far (x_i, x_j) is mapped
- Definition: Kullback-Leibler divergence
  - Two probability distributions P and Q
  - $KL(P||Q) = \sum_k P_k \log \frac{P_k}{Q_k}$
  - Note: asymmetric (and no triangle equality)
- Intuition:
  - KL large if P is large where Q is small (or zero)
  - The opposite is not the case: where P is zero, Q can be anything

SNE

- Map high-dimensional vectors to "probabilities of similarity":
  - $p_{j|i} = \frac{exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}$
  - σ_i^2: controls the size of the "local neighborhood"
- Make p_{ij} a symmetric joint distribution over all pairs:
  - $p_{ij} \propto p_{i|j} + p_{j|i}$
- Similar assumption in the low-dim space
- Minimize $KL(p_{ij}||q_{ij})$ → recall that this tries to avoid small q_{ij} (ie, image points far) where p_{ij} is large

MNIST dataset

t-SNE embedding of MNIST



Summary & lessons

- High-dimensional data often has structure, i.e., is exactly or approximately lower-dimensional
- Important for: visualizing; describing; modeling; compressing
- Simplest assumption: linear subspace
- SVD: exists for every matrix, describes relationships between two spaces
- PCA: projection of high-dimensional data onto "best" low-dimensional space

References

- [A. Rajaranam, J. D. Ullman: Mining of Massive Datasets (chapter 11), Cambridge, 2012]
- [J. B. Tenenbaum, V. de Silva, J. C. Langford: A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science, vol 290, 2000]