Announcements

$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$ $\quad \frac{1}{n} \underset{1\times n}{\mathbf{1}^T} \underset{n\times d}{X} = \bar{x}^T \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

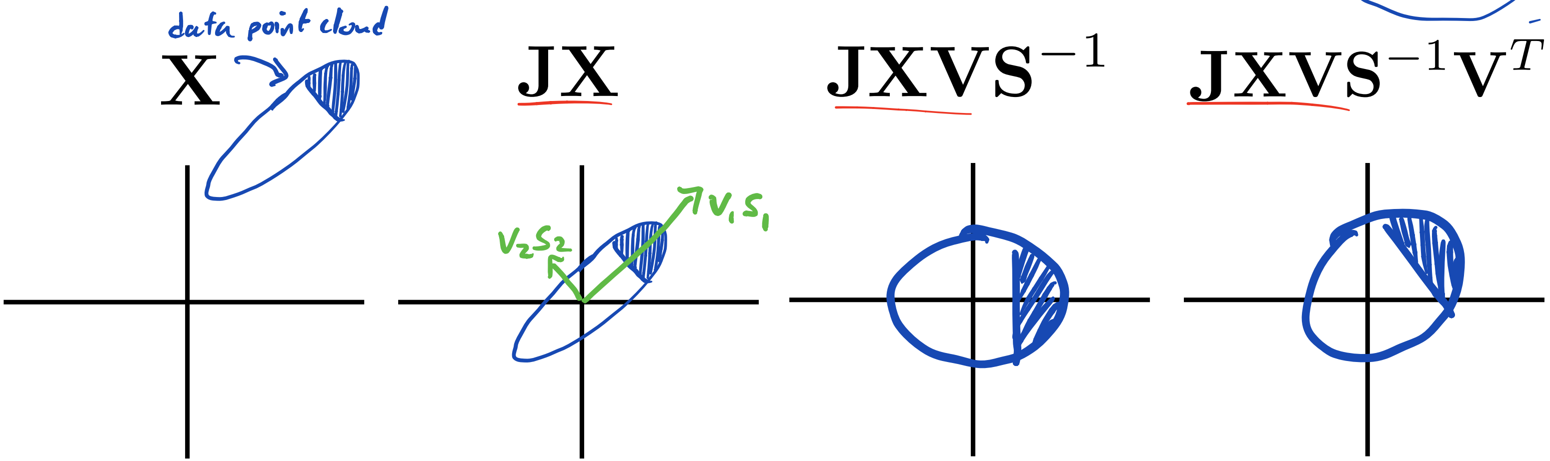
$JX = X - \mathbf{1}\bar{x}^T$

- Milestone due tonight
- Fill in the missing plots:

$$\mathbf{U}, \mathbf{S}, \mathbf{V} = \text{svd}(\overset{JX}{\cancel{\mathbf{X}}}) \Rightarrow U^TU = I,\ V^TV = I$$

$$JX = \sum_{k=1}^{d} u_k v_k^T s_k$$

$$J\mathbf{X} = \mathbf{USV}^T \qquad \mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n$$



$\mathbf{X}$ $\quad \mathbf{JX}$ $\quad \mathbf{JXVS}^{-1}$ $\quad \mathbf{JXVS}^{-1}\mathbf{V}^T$

$$JXVS^{-1} = USV^TVS^{-1}$$
$$= U$$

©2017 Kevin Jamieson

Principal Component Analysis (continued)

Machine Learning – CSE546
Kevin Jamieson
University of Washington

November 13, 2017

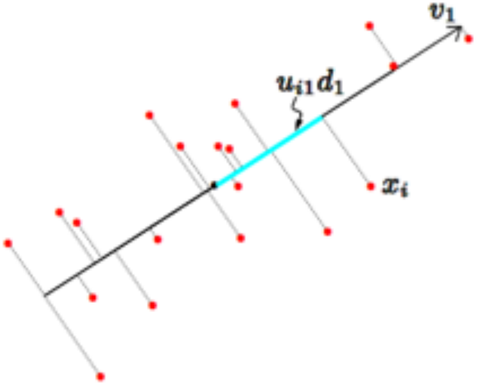
Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

where $\mathbf{V}_q = [v_1, v_2, \dots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$



$\mathbf{V}_q$ are the first q eigenvectors of Σ

$\mathbf{V}_q$ are the first q *principal components*

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

Principal Component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x}^T)$ down onto $\mathbf{V}_q$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \text{diag}(d_1, \dots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

Singular Value Decomposition defined as

$$\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

Linear projections

X is centered $(JX = X)$

$USV^T = X$

$$\underbrace{XX^T}_{n\times n} = US^2U^T$$

$\mathrm{eig}_{vec}(XX^T) = U$

$$\underbrace{X^TX}_{d\times d} = VS^2V^T$$

$\mathrm{eig}_{vec}(X^TX) = V$

$$A = XV = US\cancel{V^TV}$$

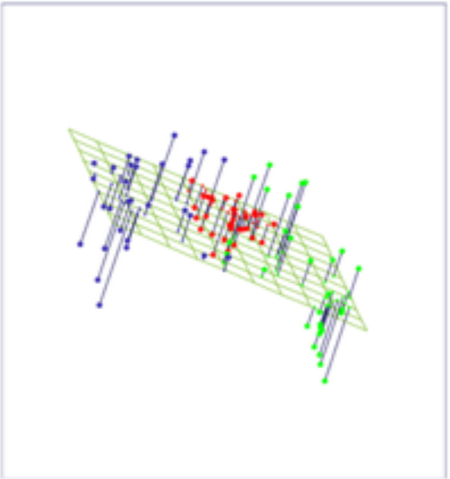
$$u_i = \frac{Ae_i}{\|Ae_i\|_2} = \frac{u_i s_i}{s_i} = u_i$$

$$Ae_i = u_i s_i$$

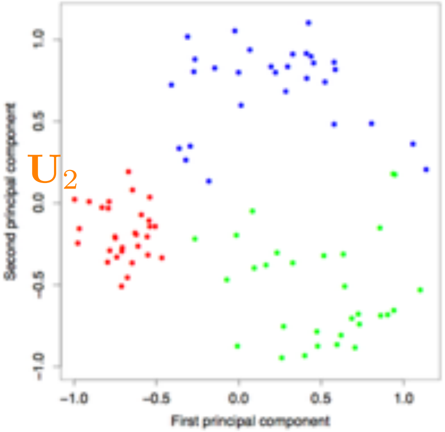
$$\|Ae_i\|_2^2 = s_i^2 \cancel{\|u_i\|_2^2} = s_i^2$$

Dimensionality reduction

$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$



$\mathbf{X} - \mathbf{1}\bar{x}^T$

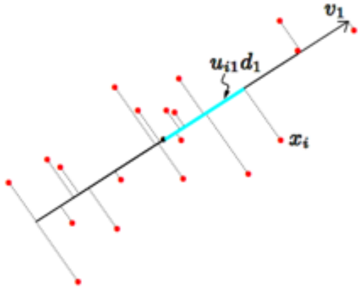


$\mathbf{U}_2$

$\mathbf{U}_1$

Power method - one at a time

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T \qquad v_* = \arg\max_v \, v^T \Sigma v$$



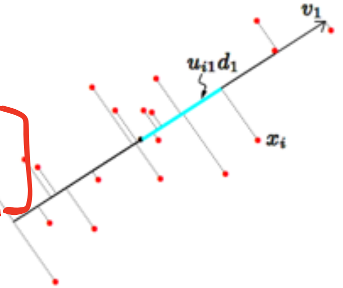
$v_1 \leftarrow \text{Power Method}(\Sigma)$ $s_i = v_i^T \Sigma v_i$ $v_2 \leftarrow \text{Power Method}(\Sigma - s_1 v_1 v_1^T)$

Power method - one at a time

$(VSV^T)^2 = VS\underbrace{V^TV}_{I}SV^T = VS^2V^T$

$w = \sum \alpha_i v_i$

$$\Sigma := \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T \qquad v_* = \arg\max_v v^T \Sigma v$$



$$v_k = \frac{\Sigma v_{k-1}}{||\Sigma v_{k-1}||_2}$$

$V^T w = \begin{bmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{bmatrix}$

$$w_{k+1} = \frac{\Sigma w_k}{||\Sigma w_k||}$$

$$w \sim \mathcal{N}(0, I)$$

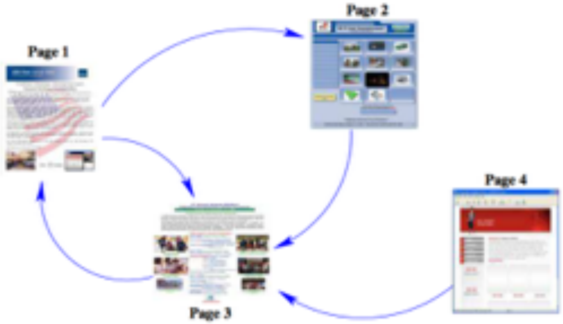
$V = [v_1, \ldots, v_n]$

$$\frac{\Sigma v_k}{||\Sigma v_k||} = \frac{\Sigma^2 v_{k-1}}{||\Sigma v_k|| \, ||\Sigma \cancel{v_{k-1}}||} \frac{\cancel{||\Sigma v_{k-1}||}}{||\Sigma^2 \cancel{v_{k-1}}||} = \frac{\Sigma^k v_0}{||\Sigma^k v_0||_2} \quad (*)$$

$$\Sigma^k = (VSV^T)^k = VS^kV^T$$

$$(*) \quad \frac{VS^kV^Tw}{||VS^kV^Tw||_2} = \frac{s_1^k V \begin{bmatrix}(s_1/s_1)^k & & 0 \\ 0 & (s_2/s_1)^k & \\ & & \ddots (s_n/s_1)^k\end{bmatrix}\alpha}{s_1^k \left\|V \begin{bmatrix}(s_1/s_1)^2 & & 0 \\ 0 & (s_2/s_1)^k & \\ & & \ddots (s_n/s_1)^k\end{bmatrix}\alpha\right\|_2} \xrightarrow{k \to \infty} \frac{s_1^k V \begin{bmatrix}\alpha_1 \\ 0 \\ \vdots \\ 0\end{bmatrix}}{s_1^k \left\|V \begin{bmatrix}\alpha_1 \\ \vdots\end{bmatrix}\right\|} = v_1$$

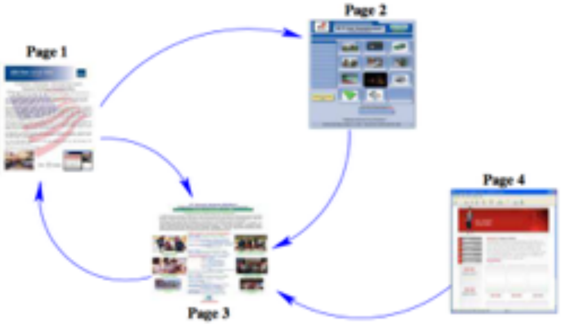
Markov chains - PageRank



Markov chains - PageRank

$$L_{i,j} = \mathbf{1}\{\text{page } j \text{ points to page } i\}$$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$



Google PageRank of page i:

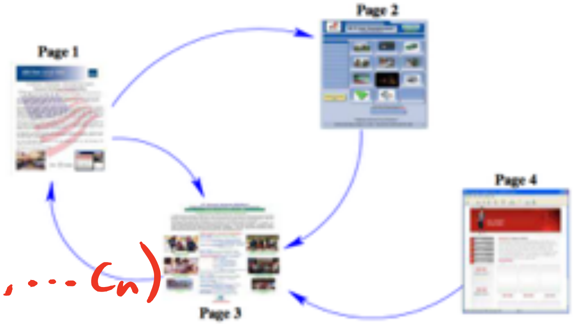
$$p_i = (1 - \lambda) + \lambda \sum_{j=1}^{n} \frac{L_{i,j}}{c_j} p_j$$

$$\underline{c_j = \sum_{k=1}^{n} L_{j,k}}$$

Markov chains - PageRank

$$L_{i,j} = \mathbf{1}\{\text{page } j \text{ points to page } i\}$$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$



Google PageRank of pages given by:

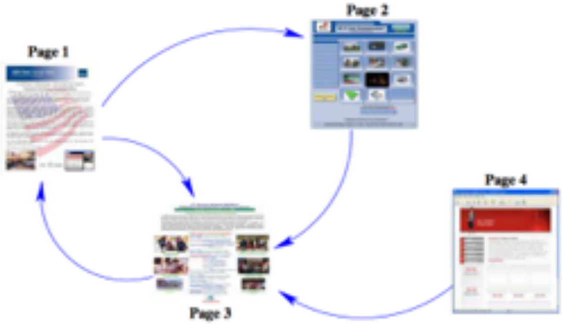
$$\underline{\mathbf{p}} = (1-\lambda)\mathbf{1} + \lambda \mathbf{L}\mathbf{D}_c^{-1}\mathbf{p}$$

$D_c = \mathrm{diag}(c_1, \ldots c_n)$

Markov chains - PageRank

$$L_{i,j} = \mathbf{1}\{\text{page } j \text{ points to page } i\}$$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$



Google PageRank of pages given by:

$$\mathbf{p} = (1-\lambda)\mathbf{1} + \lambda \mathbf{L}\mathbf{D}_c^{-1}\mathbf{p}$$

Set arbitrary normalization: $\mathbf{1}^T\mathbf{p} = n$ so that

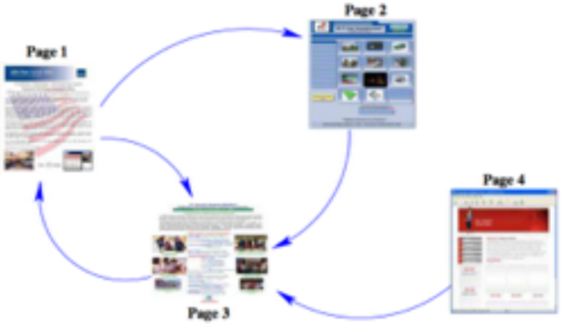
$$\mathbf{p} = \left((1-\lambda)\mathbf{1}\mathbf{1}^T/n + \lambda \mathbf{L}\mathbf{D}_c^{-1}\right)\mathbf{p}$$

$$=: \mathbf{A}\mathbf{p}$$

Markov chains - PageRank

$$L_{i,j} = \mathbf{1}\{\text{page } j \text{ points to page } i\}$$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$



Google PageRank of pages given by:

$$\mathbf{p} = (1-\lambda)\mathbf{1} + \lambda \mathbf{L}\mathbf{D}_c^{-1}\mathbf{p}$$

Set arbitrary normalization: $\mathbf{1}^T\mathbf{p} = n$ so that

$$\mathbf{p} = \left((1-\lambda)\mathbf{1}\mathbf{1}^T/n + \lambda \mathbf{L}\mathbf{D}_c^{-1}\right)\mathbf{p}$$

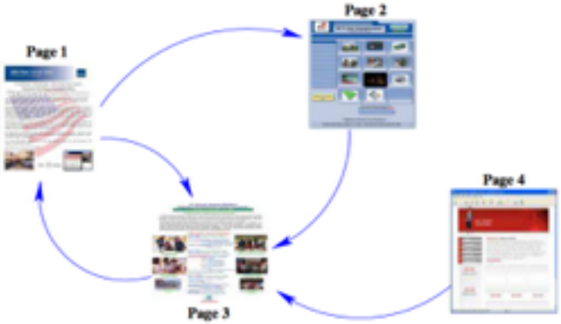
$$=: \mathbf{A}\mathbf{p}$$

$\mathbf{p}$ is an eigenvector of $\mathbf{A}$ with eigenvalue 1! And by the properties stochastic matrices, it corresponds to the *largest* eigenvalue

Markov chains - PageRank

$$L_{i,j} = \mathbf{1}\{\text{page } j \text{ points to page } i\}$$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$



Google PageRank of pages given by:

$$\mathbf{p} = (1-\lambda)\mathbf{1} + \lambda \mathbf{L}\mathbf{D}_c^{-1}\mathbf{p}$$

Set arbitrary normalization: $\mathbf{1}^T\mathbf{p} = n$ so that

$$\mathbf{p} = \left((1-\lambda)\mathbf{1}\mathbf{1}^T/n + \lambda \mathbf{L}\mathbf{D}_c^{-1}\right)\mathbf{p}$$

$$=: \mathbf{A}\mathbf{p}$$

$\mathbf{p}$ is an eigenvector of $\mathbf{A}$ with eigenvalue 1! And by the properties stochastic matrices, it corresponds to the *largest* eigenvalue

Solve using power method: $\mathbf{p}_{k+1} = \dfrac{\mathbf{A}\mathbf{p}_k}{\mathbf{1}^T\mathbf{A}\mathbf{p}_k/n}$ $\qquad \mathbf{p}_0 \sim \text{uniform}([0,1]^n)$

Matrix completion

NETFLIX

Given historical data on how users rated movies in past:

$\overset{n}{\underline{17{,}700}}$ movies, $\overset{m}{\underline{480{,}189}}$ users, 99,072,112 ratings (Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for $1 million prize)

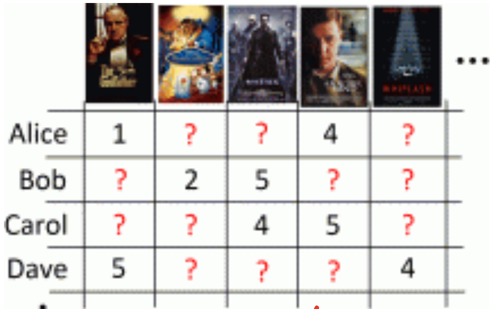
$X \approx UV^T$ $\quad \underline{U \in \mathbb{R}^{m \times d}}, \underline{V \in \mathbb{R}^{n \times d}}$

User i is assigned a vector $u_i \in \mathbb{R}^d$

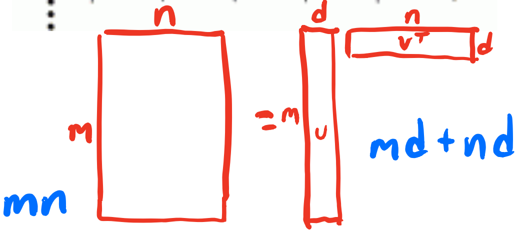
movie j is " " $v_j \in \mathbb{R}^d$

user i rates movie j as $\approx u_i^T v_j$

$X =$



						…
Alice	1	?	?	4	?	
Bob	?	2	5	?	?	
Carol	?	?	4	5	?	
Dave	5	?	?	?	4	
⋮						

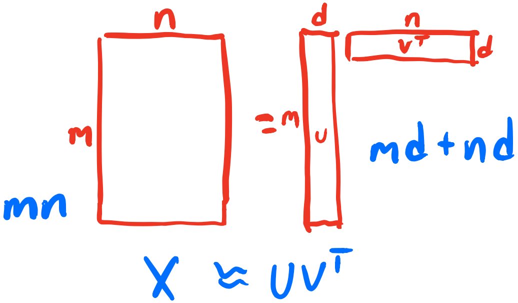


Matrix completion

n movies, m users, |*S*| ratings

$$\underset{U\in\mathbb{R}^{m\times d}, V\in\mathbb{R}^{n\times d}}{\arg\min} \sum_{(i,j,s)\in\mathcal{S}} ||(UV^T)_{i,j} - s_{i,j}||_2^2$$

How do we solve it? With full information?



Matrix completion

n movies, m users, |*S*| ratings

$$\arg\min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} \sum_{(i,j,s) \in \mathcal{S}} ||(UV^T)_{i,j} - s_{i,j}||_2^2$$

Random projections

PCA finds a low-dimensional representation that reduces population variance

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

$\mathbf{V}_q$ are the first q eigenvectors of Σ

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

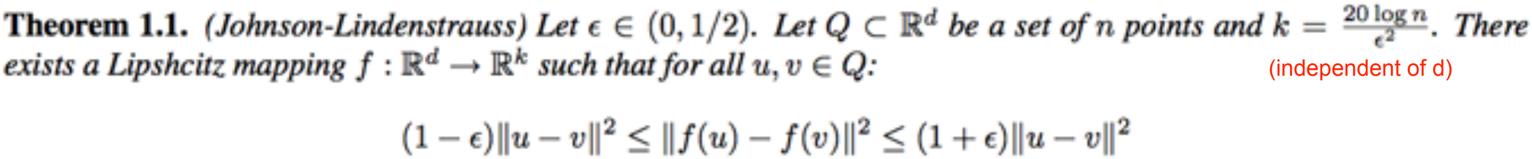
But what if I care about the reconstruction of the *individual* points?

$$\min_{\mathbf{W}_q} \max_{i=1,\ldots,n} ||(x_i - \bar{x}) - \mathbf{W}_q \mathbf{W}_q^T (x_i - \bar{x})||^2$$

Random projections

$$\min_{\mathbf{W}_q} \max_{i=1,\ldots,n} ||(x_i - \bar{x}) - \mathbf{W}_q \mathbf{W}_q^T (x_i - \bar{x})||^2$$

Johnson-Lindenstrauss (1983)



Theorem 1.1. *(Johnson-Lindenstrauss) Let $\epsilon \in (0, 1/2)$. Let $Q \subset \mathbb{R}^d$ be a set of n points and $k = \frac{20 \log n}{\epsilon^2}$. There exists a Lipshcitz mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in Q$:*

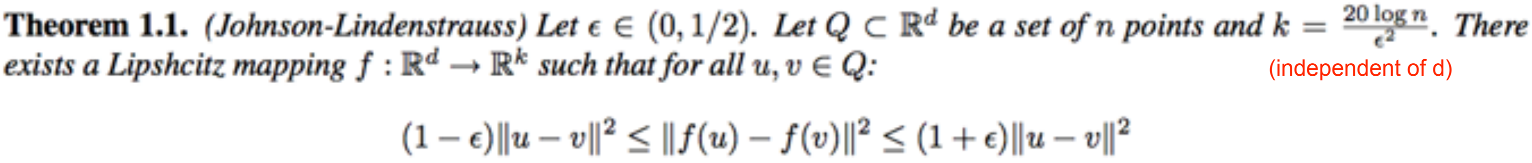
$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

(independent of d)

Random projections

$$\min_{\mathbf{W}_q} \max_{i=1,\ldots,n} ||(x_i - \bar{x}) - \mathbf{W}_q \mathbf{W}_q^T (x_i - \bar{x})||^2$$

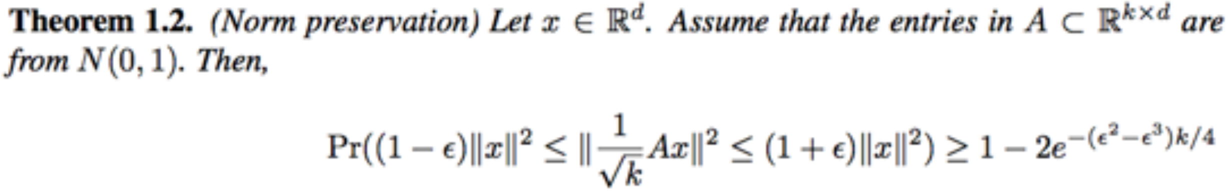
Johnson-Lindenstrauss (1983)



Theorem 1.1. *(Johnson-Lindenstrauss) Let* $\epsilon \in (0, 1/2)$. *Let* $Q \subset \mathbb{R}^d$ *be a set of* n *points and* $k = \frac{20 \log n}{\epsilon^2}$. *There exists a Lipshcitz mapping* $f : \mathbb{R}^d \to \mathbb{R}^k$ *such that for all* $u, v \in Q$:

(independent of d)

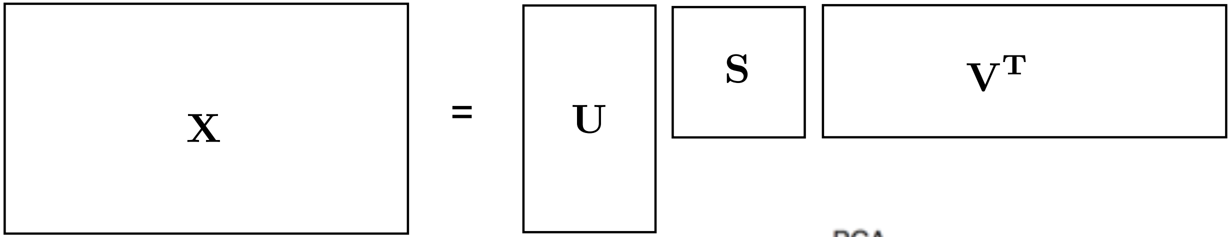
$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$



Theorem 1.2. *(Norm preservation) Let* $x \in \mathbb{R}^d$. *Assume that the entries in* $A \subset \mathbb{R}^{k \times d}$ *are sampled independently from* $N(0, 1)$. *Then,*

$$\Pr((1-\epsilon)\|x\|^2 \leq \|\frac{1}{\sqrt{k}}Ax\|^2 \leq (1+\epsilon)\|x\|^2) \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}$$

Other matrix factorizations

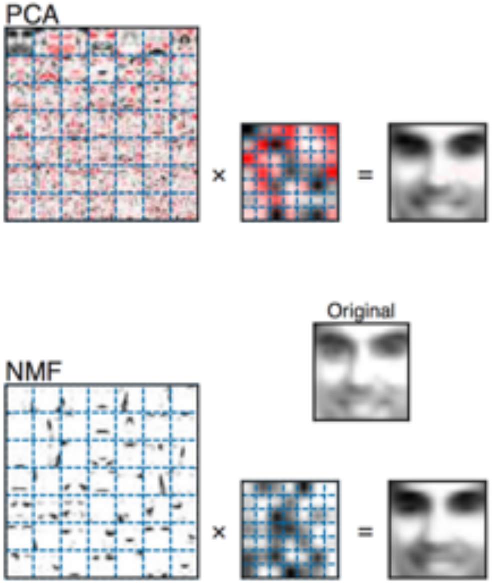


Singular value decomposition

Elements of $\mathbf{U}, \mathbf{S}, \mathbf{V}$ in $\mathbb{R}$

Nonnegative matrix factorization (NMF)

Elements of $\mathbf{U}, \mathbf{S}, \mathbf{V}$ in $\mathbb{R}_+$


PCA
×
=
Original
NMF
×
=