PCA

have data set $\vec{x}_1, \dots, \vec{x}_n$ $x_i \in \mathbb{R}^d$

$$X = \begin{pmatrix} — x_1 — \\ \vdots \\ — x_n — \end{pmatrix}$$

assume mean 0 $\left(\sum_{i=1}^{n} x_i = 0\right)$

Examples

- n images d pixels each
- n measurements d sensors
- n docs d words
- n people d movies
- n customers d products

Fix k

Find k-dimensional subspace (defined by orthonormal vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k$)

so as to minimize $\frac{1}{n} \sum_{i=1}^{n} \left[\text{distance}\left(x_i \leftrightarrow \text{subspace } S \text{ spanned by } \vec{v}_1, v_2, \dots, \vec{v}_k\right)\right]^2$

$\equiv$ maximizing $\underbrace{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} (x_i, v_j)^2}_{\text{variance of projected pts}}$

projection of $x_i \to S$

$$\sum_{j=1}^{k} (x_i, v_j)\, \vec{v}_j$$

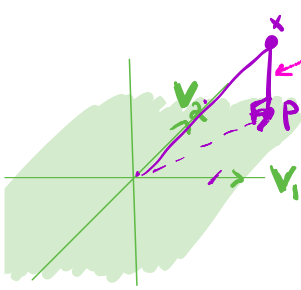
$x_i \cdot v_j$

v_1

j

find $v_1, \dots, v_k$

to max $\frac{1}{n} \sum_{i=1}^{n} \underbrace{\boxed{\sum_{j=1}^{k} (x_i, v_j)^2}}_{\|p\|^2}$



$p = a_1 v_1 + a_2 v_2$

(x, v_1)

$(x - p, v_1) = 0$

$(x, v_1) = (p, v_1) = (a_1 v_1 + a_2 v_2, v_1)$

$= a_1$

$p = (x, v_1) v_1 + (x, v_2) v_2$

maximizing $\frac{1}{n}\sum_{i=1}^{n} \sum_{j=1}^{k} (x_i v_j)^2 = \sum_{j=1}^{k} (Xv_j)^T X v_j$

$= \frac{1}{n}\sum_{j=1}^{k} v_j^T X^T X v_j$

How to find $\vec{v}_1, \dots, \vec{v}_k$

① Compute eigendecomposition of empirical covariance matrix $X^T X$

$A := X^T X = Q D Q^T$ where $Q = \begin{pmatrix} | & & | \\ q_1 & \cdots & q_d \\ | & & | \end{pmatrix}$

$q_1, \dots, q_n$ orthonormal

$Q Q^T = Q^T Q = I$

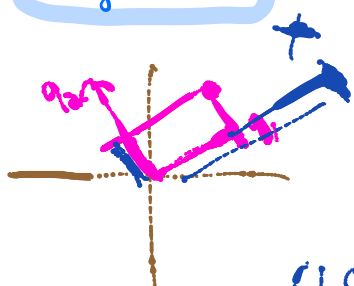
$\forall w \quad \|Qw\| = \|w\|$

$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \ \lambda_d \end{pmatrix}$

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$

all eigenvalues ≥ 0

$\boxed{A q_i = \lambda_i q_i} \quad \forall i$



$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ $\begin{pmatrix} 3 \\ 5 \end{pmatrix}$

$\begin{pmatrix} | & & | \\ q_1 & \cdots & q_d \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} \begin{pmatrix} - & q_1^T & - \\ & \vdots & \\ - & q_d^T & - \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_d \end{pmatrix}$

$\begin{pmatrix} q_1 \cdot a \\ q_2 \cdot a \\ \vdots \\ q_d \cdot a \end{pmatrix}$

$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

② Set $v_1 := q_1$

$v_2 := q_2$

$\vdots$

$v_k := q_k$

New coordinates of x_i

are $\underline{(x_i v_1), (x_i v_2), \dots, (x_i v_k)}$

low dimensional representation of $\vec{x}_i$

$\Sigma = A A^T$

$\Sigma = \underbrace{\begin{pmatrix} | & | \\ q_1 & q_2 \\ | & | \end{pmatrix}}_{Q} \underbrace{\begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} - q_1^T - \\ - q_2^T - \end{pmatrix}}_{Q^T}$

$\Phi D^{k/2} D^{k/2} \Phi^T$

$(\Phi D^{k/2})(\Phi D^{k/2})^T$

1. Find the top component, $\mathbf{v}_1$, using power iteration.
2. Project the data matrix orthogonally to $\mathbf{v}_1$:

$$\begin{bmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ & \vdots & \\ - & \mathbf{x}_m & - \end{bmatrix} \mapsto \begin{bmatrix} - & (\mathbf{x}_1 - \langle \mathbf{x}_1, \mathbf{v}_1 \rangle \mathbf{v}_1) & - \\ - & (\mathbf{x}_2 - \langle \mathbf{x}_2, \mathbf{v}_1 \rangle \mathbf{v}_1) & - \\ & \vdots & \\ - & (\mathbf{x}_m - \langle \mathbf{x}_m, \mathbf{v}_1 \rangle \mathbf{v}_1) & - \end{bmatrix}.$$

   This corresponds to subtracting out the variance of the data that is already explained by the first principal component $\mathbf{v}_1$.

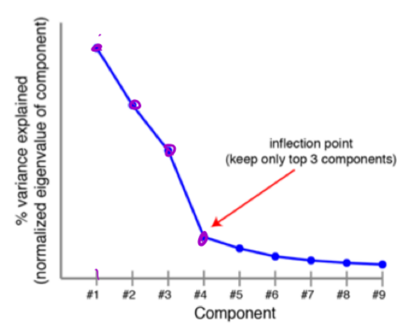
3. Recurse by finding the top $k-1$ principal components of the new data matrix.

Applications

① Visualization ② Compression ③ Learning

How to choose k?

- For data visualization: a few
- compression: Look at eigenvalues.
  As soon as small enough, happy.



k

Scree plot. Principal components are ranked by the amount of variance they capture in the original dataset, a scree plot can provide some sense of how many components are needed.

Singular Value Decomposition (SVD)

gives us best way to approximate a matrix with a "low rank" matrix

$$\begin{array}{c|c} & \text{ratings} \\ \hline \text{movies} & \begin{bmatrix} 7 & ? & ? \\ ? & 8 & ? \\ ? & 12 & 6 \\ ? & ? & 2 \\ 21 & 6 & ? \end{bmatrix}. \end{array}$$

$\boxed{\text{Motivation:}}$ can we reconstruct missing entries?

Suppose I told you that all rows are multiples of each other

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & & & \\ 2 & & & \\ 3 & & & \\ 4 & & & \\ 5 & & & \end{array} \begin{bmatrix} 7 & 2 & 1 \\ 28 & 8 & \boxed{4} \\ 42 & 12 & 6 \\ 14 & \boxed{4} & 2 \\ 21 & 6 & 3 \end{bmatrix} = \begin{array}{l} 1 \cdot (7 \;\; 2 \;\; 1) \\ 4 \cdot (7 \;\; 2 \;\; 1) \\ 6 \, (7 \;\; 2 \;\; 1) \\ 2 \, (7 \;\; 2 \;\; 1) \\ 3 \, (7 \;\; 2 \;\; 1) \end{array}$$

$$\begin{pmatrix} 1 \\ 4 \\ 6 \\ 2 \\ 3 \end{pmatrix} (7 \;\; 2 \;\; 1)$$

Rank 0 all 0 matrix

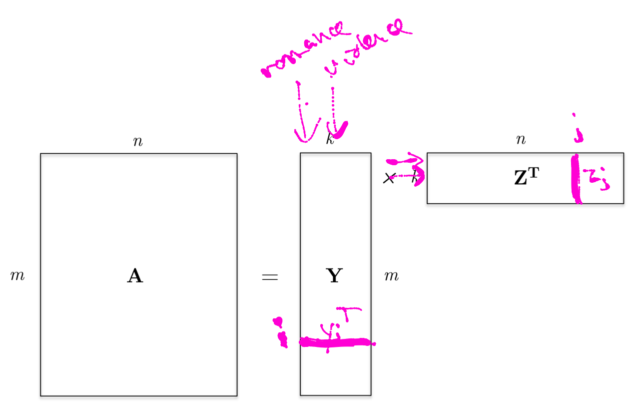
Rank 1

$$\mathbf{A} = \mathbf{u}\mathbf{v}^\top = \begin{bmatrix} — & u_1\mathbf{v}^\top & — \\ — & u_2\mathbf{v}^\top & — \\ & \vdots & \\ — & u_m\mathbf{v}^\top & — \end{bmatrix} = \begin{bmatrix} | & | & & | \\ v_1\mathbf{u} & v_2\mathbf{u} & \cdots & v_n\mathbf{u} \\ | & | & & | \end{bmatrix}$$

Rank 2

$$\mathbf{A} = \mathbf{u}\mathbf{v}^\top + \mathbf{w}\mathbf{z}^\top = \begin{bmatrix} — & u_1\mathbf{v}^\top + w_1\mathbf{z}^\top & — \\ — & u_2\mathbf{v}^\top + w_2\mathbf{z}^\top & — \\ & \vdots & \\ — & u_m\mathbf{v}^\top + w_m\mathbf{z}^\top & — \end{bmatrix} = \begin{bmatrix} | & | \\ \mathbf{u} & \mathbf{w} \\ u_i & w_i \\ | & | \end{bmatrix} \cdot \begin{bmatrix} — & \mathbf{v}^\top & v_j \\ — & \mathbf{z}^\top & z_j \end{bmatrix}$$

$u_i v_j + w_i z_j$



Find $y_1, \ldots, y_m \in \mathbb{R}^k$
$z_1, \cdots z_n \in \mathbb{R}^k$
s.t. $a_{ij} \approx y_i \cdot z_j$

Figure 1: Any matrix $\mathbf{A}$ of rank k can be decomposed into a long and skinny matrix times a short and long one.

Why might a matrix be approximately low rank?

Example: movie ratings

Suppose each movie characterized by relatively small # of attributes

e.g. romance, violence, comedy, ...

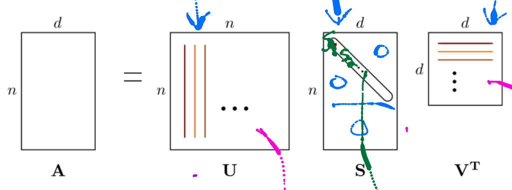
and each person characterized by their preferences on each of these

Singular Value Decomposition (SVD)

orthogonal

diagonal

orthogonal



rows are called right singular vectors

Figure 2: The singular value decomposition (SVD). Each singular value in **S** has an associated left singular vector in **U**, and right singular vector in **V**.

$UU^T = U^TU = I$

$VV^T = V^TV = I$

cols called left singular vectors

$s_1 \geq s_2 \geq \dots \geq s_{\min(n,d)} \geq 0$
called singular values.

Running time to compute

$\min\left[O(n^2d), O(d^2n)\right]$

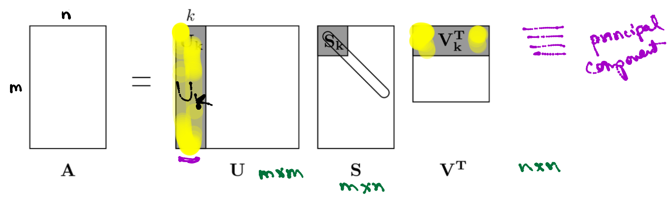


Figure 3: Low rank approximation via SVD. Recall that $\mathbf{S}$ is non-zero only on its diagonal, and the diagonal entries of S are sorted from high to low. Our low rank approximation is $\mathbf{A}_k = \mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^\top$.

is optimal rank approx

Claim: $\forall$ $m \times n$ matrix A and rank target k, and rank k matrix B

$$\|A - A_k\|_F^2 \leq \|A - B\|_F^2$$

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

Exercise: knowing that $\vec{v}_1$ (first row) is first principal component prove this theorem for $k=1$

Relationship between SVD and PCA

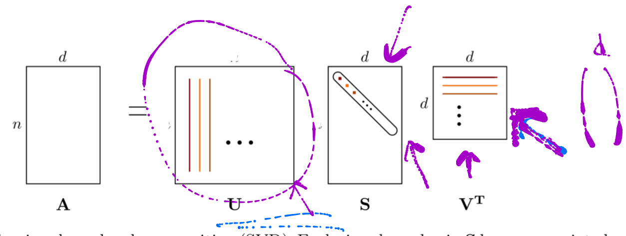


Figure 2: The singular value decomposition (SVD). Each singular value in **S** has an associated left singular vector in **U**, and right singular vector in **V**.

PCA

$$X^TX = QDQ^T$$

SVD

$$\underbrace{X}_{n\times d} = USV^T$$

$$X^TX = (USV^T)^T\, USV^T$$

$$= VS^T\underbrace{U^TU}_{I}SV^T$$

$$= VS^2V^T$$

$$\begin{pmatrix} s_1^2 & & \\ & s_2^2 & \\ & & \ddots \end{pmatrix}$$

$$Q = V$$

$$D = S^2$$

eigenvalues of X^TX are squares of singular values.

$$XX^T = \underline{US^2U^T}$$

products

customers

X^TX covariance for products

XX^T covariance for customers

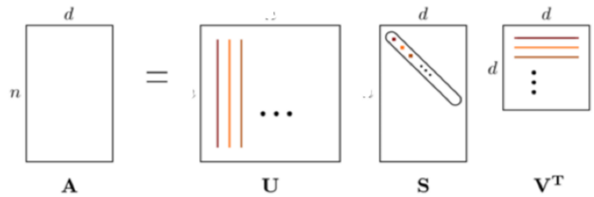


Figure 2: The singular value decomposition (SVD). Each singular value in **S** has an associated left singular vector in **U**, and right singular vector in **V**.

Application 1: Denoising

Suppose matrix A is noisy version of rank-k matrix

$$A = \underset{\substack{\uparrow \\ \text{true rank } k}}{C} + \underset{\uparrow}{N}$$

can "reconstruct C" by taking A_k

Collaborative Filtering

recomendations: which movies to see, which products to buy

Model:

$$\begin{array}{cc} & \text{movies} \\ \text{people} & \begin{bmatrix} 2.5 & 4 & 1 \\ 2.5 & 0 & 5 \\ 1 & 1 & 1 \\ 2.5 & 4 & 1 \\ 3.5 & 3 & 4 \end{bmatrix} \end{array} \longrightarrow \begin{bmatrix} \boxed{2.5} & ? & ? \\ ? & 0 & ? \\ ? & 1 & 1 \\ ? & ? & 1 \\ 3.5 & 3 & ? \end{bmatrix}$$

R $n \times d$ ground truth $\qquad\qquad R^?$

Assumptions:

R is low rank
e.g. rank k

example: humor, violence, romance

The matrix completion problem:

$$R^? \approx \begin{array}{c} k \\ n\left(\begin{array}{c} \\ \underline{\;a_i\;} \end{array} \right) \end{array} \; \begin{array}{c} d \quad j \\ k\left(\begin{array}{c} | \\ b_j \end{array} \right) \end{array}$$

Find a_i $i=1..n$
b_j $j=1..d$
s.t. if R_{ij} is present $R_{ij} \approx \vec{a}_i \cdot \vec{b}_j$

Find a_i's, b_j's to min

$$\sum_{\substack{(i,j)\in R^? \\ \text{have entry}}} (R_{ij} - a_i \cdot b_j)^2$$

Collaborative Filtering

recomendations: which movies to see, which products to buy

Model:

$$\begin{array}{cc} & \text{movies} \\ \text{people} & \begin{bmatrix} 2.5 & 4 & 1 \\ 2.5 & 0 & 5 \\ 1 & 1 & 1 \\ 2.5 & 4 & 1 \\ 3.5 & 3 & 4 \end{bmatrix} \end{array} \longrightarrow \begin{bmatrix} 2.5 & ? & ? \\ ? & 0 & ? \\ ? & 1 & 1 \\ ? & ? & 1 \\ 3.5 & 3 & ? \end{bmatrix}$$

R $n \times d$

ground truth

$R^?$

Assumptions:

1. R is low rank
   e.g. rank k

example: humor, violence, romance

The matrix completion problem:

fill in the missing entries.

Theorem $\hat{A}$ $n \times d$ matrix of indep. r.v.s, each with variance bounded by σ^2

If $A = E(\hat{A})$ is rank k

then with high probability

$$\|A - \hat{A}_k\|_F^2 \text{ is } O(k\sigma^2(n+d))$$

avg per elt error $O\left(\frac{k\sigma^2(n+d)}{nd}\right) = o(1)$

Suppose A rank k. $\longrightarrow A^?$

entry (i,j) present with prob p_{ij}

$$\begin{pmatrix} ? & 5 & - & ? \end{pmatrix}$$

Let $\hat{A}_{ij} = \begin{cases} \dfrac{A^?_{ij}}{p_{ij}} & \text{if entry present} \\ 0 & \text{o.w.} \end{cases}$

$A \longrightarrow A^?$

A_{ij} rank k

entry (i,j) survives with prob p_{ij}

$$E(\hat{A}_{ij}) = \frac{A^?_{ij}}{p_{ij}} \cdot p_{ij} + 0 = \boxed{A_{ij}}$$

$$\underset{R^?}{\begin{bmatrix} 2.5 & ? & ? \\ ? & 0 & ? \\ ? & 1 & 1 \\ ? & ? & 1 \\ 3.5 & 3 & ? \end{bmatrix}} \longrightarrow \underset{M}{\begin{bmatrix} & & \\ & & \\ & & \\ & & \\ & & \end{bmatrix}}$$