Singular value decomposition

Victor Kitov

Definitions

Consider matrix $X \in \mathbb{R}^{NxD}$. For this matrix:

- square roots of eigenvalues of $X^T X$ are called *singular values*.
- orthonormal eigenvectors of $X^T X$ are called *right singular vectors*.
- orthonormal eigenvectors of XX^T are called *left singular vectors*.

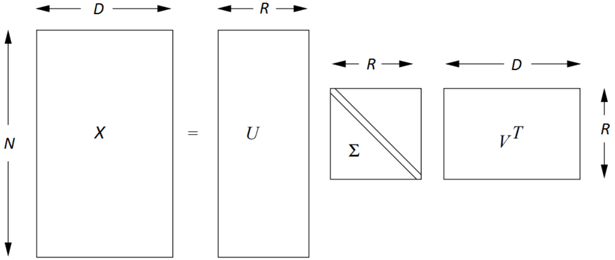
Principal component a_i is the i-th right singular vector of X, corresponding to i-th largest singular value λ_i.

SVD decomosition

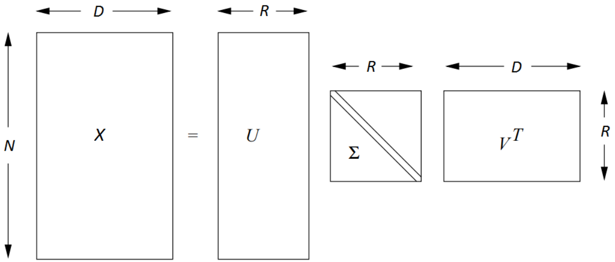
Every matrix $X \in \mathbb{R}^{NxD}$, rank $X = R$, can be decomposed into the product of three matrices:

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

where $U \in \mathbb{R}^{NxR}$, $\Sigma \in \mathbb{R}^{RxR}$, $V^T \in \mathbb{R}^{RxD}$, and $\Sigma = diag\{\sigma_1, \sigma_2, ...\sigma_R\}$, $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_R \geq 0$, $U^T U = I$, $V^T V = I$. $I \in \mathbb{R}^{DxD}$ denotes identity matrix.



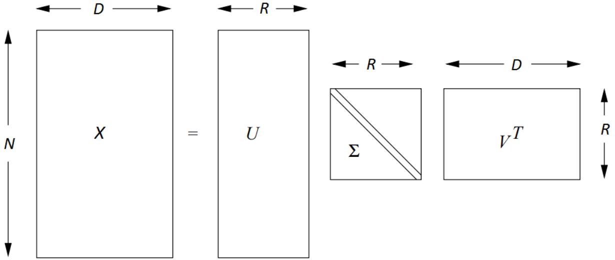
Interpretation of SVD



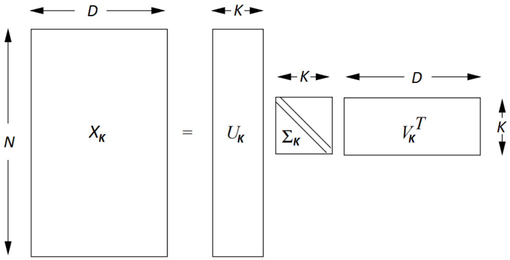
For X_{ij} let i denote objects and j denote properties.

- U represents standardized coordinates of concepts
- V^T represents standardized concepts representations
- Σ shows the magnitudes of presence of standardized concepts in X.

Original SVD decomposition



Reduced SVD decomposition



Simplification to rank $K \le R$:

$$X_K = U_K \Sigma_K V_K$$

$$\Sigma = diag\{\sigma_1, \sigma_2, ...\sigma_K, \sigma_{K+1}, ...\sigma_R\} \longrightarrow diag\{\sigma_1, \sigma_2, ...\sigma_K\} = \Sigma_K$$
$$U = [u_1, u_2, ...u_K, u_{K+1}, ...u_R] \longrightarrow [u_1, u_2, ...u_K] = U_K$$
$$V = [v_1, v_2, ...v_K, v_{K+1}, ...v_R] \longrightarrow [v_1, v_2, ...v_K] = V_K$$

Properties of reduced SVD decomposition

- Suppose $X \in \mathbb{R}^{NxD}$, $\operatorname{rank} X = R$, is approximated with $X_K = U_K \Sigma_K V_K$. Then:
  - $\operatorname{rank} X_K = K$.
  - $X_K = \arg\min_{B:\operatorname{rank} B \le K} \|X - B\|$
- Which K to choose?
  - Define Frobenius norm $\|X\|_F^2 = \sum_{n=1}^{N} \sum_{d=1}^{D} x_{nd}^2$
  - $\|X\|_F^2 = \sum_{i=1}^{R} \sigma_i^2$
  - $\|X_K\|_F^2 = \sum_{i=1}^{K} \sigma_i^2$
  - Choose $K = \arg\min_K \left\{ \frac{\|X_K\|_F^2}{\|X\|_F^2} \ge t \right\}$, where t is some threshold, say $t = 0.95$.

Table of Contents

Memory efficiency

Storage costs of $X \in \mathbb{R}^{NxD}$, assuming $N \geq D$ and each element taking 1 byte:

Memory storage costs

representation of X	memory requirements
original X	$O(ND) = O(\min\{N, D\}\max\{N, D\})$
fully SVD decomposed	$NR + R^2 + RD = O(R\max\{N, D\})$
reduced SVD to rank K	$NK + K^2 + KD = O(K\max\{N, D\})$

Performance efficiency

Suppose we have N documents, vocabulary size is D, typically $D \geq N$.

- $X \in \mathbb{R}^{NxD}$ represents normalized vector representation of documents
- $q \in \mathbb{R}^{D}$ represents normalized vector representation of search query
- $$X \approx X_K = \underbrace{U_K \Sigma_K}_{B} V_K^T = BV_K^T, \quad B \in \mathbb{R}^{NxK}$$

Document x_i relevance is proportional to $\langle x_i, q \rangle$, so to find matching documents we need to calculate $Xq = [\langle x_1, q \rangle, ... \langle x_N, q \rangle]^T$.
Direct multiplication Xq takes $O(ND) = O(max\{N, D\} \min\{N, D\})$ operations.
$X_K q = U_K \Sigma_K V_K^T q = BV_K^T q$. $V_K^T q$ takes $O(DK)$ multiplications and $BV_K^T q$ takes $O(NK)$, so total complexity is $O(K \max\{N, D\})$.

SVD for square non-degenerate matrix

For square non-degenerate matrix X:

- $X \in \mathbb{R}^{DxD}$, rg X=D, so $U \in \mathbb{R}^{DxD}$, $V \in \mathbb{R}^{DxD}$, $U^{-1} = U^T$, $V^{-1} = V^T$.
- U, V^T represent rotations, Σ represents scaling, every square matrix may be represented as superposition of rotation, scaling and another rotation.
- For full rank X:
$$X^{-1} = V\Sigma^{-1}U^T,$$
since $XX^{-1} = U\Sigma V^T V\Sigma^{-1}U^T = I$.

Table of Contents

Example

	Terminator	Gladiator	Rambo	Titanic	Love story	A walk to remember
Andrew	4	5	5	0	0	0
John	4	4	5	0	0	0
Matthew	5	5	4	0	0	0
Anna	0	0	0	5	5	5
Maria	0	0	0	5	5	4
Jessika	0	0	0	4	5	4

Example

$$U=\begin{pmatrix} 0. & 0.6 & -0.3 & 0. & 0. & -0.8 \\ 0. & 0.5 & -0.5 & 0. & 0. & 0.6 \\ 0. & 0.6 & 0.8 & 0. & 0. & 0.2 \\ 0.6 & 0. & 0. & -0.8 & -0.2 & 0. \\ 0.6 & 0. & 0. & 0.2 & 0.8 & 0. \\ 0.5 & 0. & 0. & 0.6 & -0.6 & 0. \end{pmatrix}$$

$$\Sigma = \text{diag}\{\begin{pmatrix}14. & 13.7 & 1.2 & 0.6 & 0.6 & 0.5\end{pmatrix}\}$$

$$V^T=\begin{pmatrix} 0. & 0. & 0. & 0.6 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.6 & 0. & 0. & 0. \\ 0.5 & 0.3 & -0.8 & 0. & 0. & 0. \\ 0. & 0. & 0. & -0.2 & 0.8 & -0.6 \\ -0. & -0. & -0. & 0.8 & -0.2 & -0.6 \\ 0.6 & -0.8 & 0.2 & 0. & 0. & 0. \end{pmatrix}$$

Example (excluded insignificant concepts)

$$U_2 = \begin{pmatrix} 0. & 0.6 \\ 0. & 0.5 \\ 0. & 0.6 \\ 0.6 & 0. \\ 0.6 & 0. \\ 0.5 & 0. \end{pmatrix}$$

$$\Sigma_2 = \text{diag}\{\begin{pmatrix} 14. & 13.7 \end{pmatrix}\}$$

$$V_2^T = \begin{pmatrix} 0. & 0. & 0. & 0.6 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.6 & 0. & 0. & 0. \end{pmatrix}$$

Concepts may be

- patterns among movies (along j) - action movie / romantic movie
- patterns among people (along i) - boys / girls

Dimensionality reduction case: patterns along j axis.

Applications

- Example: new movie rating by new person

$$x = \begin{pmatrix} 5 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- **Dimensionality reduction:** map x into concept space:

$$y = V_2^T x = \begin{pmatrix} 0 & 2.7 \end{pmatrix}$$

- **Recommendation system:** map y back to original movies space:

$$\widehat{x} = y V_2^T = \begin{pmatrix} 1.5 & 1.6 & 1.6 & 0 & 0 & 0 \end{pmatrix}$$

Fronebius norm

- Fronebius norm of matrix X is $\|X\|_F \stackrel{df}{=} \sqrt{\sum_{n=1}^{N}\sum_{d=1}^{D} x_{nd}^2}$
- Using properties $||X||_F^2 = \text{tr}\, XX^T$ and $\text{tr}\, AB = \text{tr}\, BA$, we obtain:

$$
\begin{aligned}
\|X\|_F^2 &= \text{tr}[U\Sigma V^T V\Sigma U^T] = \text{tr}[U\Sigma^2 U^T] = \\
&= \text{tr}[\Sigma^2 U^T U] = \text{tr}[\Sigma^2] = \sum_{r=1}^{R} \sigma_r^2
\end{aligned}
\tag{1}
$$

Matrix approximation

Consider approximation $X_k = U\Sigma_k V^T$, where
$\Sigma_k = \text{diag}\{\sigma_1, \sigma_2, ...\sigma_k, 0, 0, ..., 0\} \in \mathbb{R}^{RxR}$.

Theorem 1

X_k is the best approximation of X retaining k concepts.

Proof: consider matrix $Y_k = U\Sigma' V^T$, where Σ' is equal to Σ except some $R-k$ elements set to zero:
$\sigma'_{i_1} = \sigma'_{i_2} = ... = \sigma'_{i_{R-k}} = 0$. Then, using (1)

$$\|X - Y_k\|_F^2 = \left\|U(\Sigma - \Sigma')V^T\right\|_F^2 = \sum_{p=1}^{R-k} \sigma_{i_p}^2 \le \sum_{p=1}^{R-k} \sigma_p^2 = \|X - X_k\|_F^2$$

since $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_R \ge 0$.

Matrix approximation

How many components to retain?

General case: Since

$$\|X - X_k\|_F^2 = \left\| U(\Sigma - \Sigma_k)V^T \right\|_F^2 = \sum_{i=k+1}^{R} \sigma_i^2$$

a reasonable choice is k^* such that

$$\frac{\|X - X_{k^*}\|_F^2}{\|X\|_F^2} = \frac{\sum_{i=k^*+1}^{R} \sigma_i^2}{\sum_{i=1}^{R} \sigma_i^2} \geq threshold$$

Visualization: 2 or 3 components.

Theorem 2

For any matrix Y_k with rank $Y_k = k$: $\|X - X_k\|_F \leq \|X - Y_k\|_F$

Finding U and V

- **Finding** V
  $X^TX=\left(U\Sigma V^T\right)^T U\Sigma V^T=(V\Sigma U^T)U\Sigma V^T=V\Sigma^2V^T$. It follows that
  $$X^TXV=V\Sigma^2V^TV=V\Sigma^2$$
  So V consists of eigenvectors of X^TX with corresponding eignvalues $\sigma_1^2,\sigma_2^2,...\sigma_R^2$.

- **Finding** U**:**
  $XX^T=U\Sigma V^T\left(U\Sigma V^T\right)^T=U\Sigma V^TV\Sigma U^T=U\Sigma^2U^T$. So
  $$XX^TU=U\Sigma^2U^TU=U\Sigma^2.$$
  So U consists of eigenvectors of $XX^T$with corresponding eigenvalues $\sigma_1^2,\sigma_2^2,...\sigma_R^2$.

Comments

- Denote the average $\bar{X} \in \mathbb{R}^D : \bar{X}_j = \sum_{i=1}^N x_{ij}$
- Denote the n-th row of X be $X_n \in \mathbb{R}^D : X_{nj} = x_{nj}$
- For centered X sample covariance matrix $\widehat{\Sigma}$ equals:

$$\begin{aligned} \widehat{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})(X_n - \bar{X})^T = \frac{1}{N}\sum_{n=1}^{N} X_n X_n^T \\ &= \frac{1}{N} X^T X \end{aligned}$$

- V **consists of principal components** since
  - V consists of eigenvectors of $X^T X$,
  - principal components are eignevectors of $\widehat{\Sigma}$ and
  - $\widehat{\Sigma} \propto X^T X$.