# Introduction to Machine Learning

Singular Value Decomposition

Varun Chandola

May 8, 2019

**Outline**

## Contents

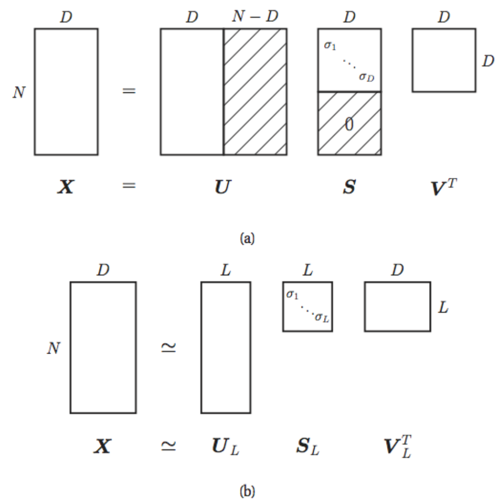


# 1 Singular Value Decompostion

- For any matrix $\mathbf{X}$ $(N \times D)$

$$\underbrace{\mathbf{X}}_{N\times D} = \underbrace{\mathbf{U}}_{N\times N}\underbrace{\mathbf{S}}_{N\times D}\underbrace{\mathbf{V}^\top}_{D\times D}$$

$\mathbf{U}$ is a $N \times N$ matrix and all columns of $\mathbf{U}$ are orthonormal, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_N$. $\mathbf{V}$ is a $D \times D$ matrix whose rows and columns are orthonormal (i.e., $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_D$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_D$). $\mathbf{S}$ is a $N \times D$ matrix containing the $r = min(N, D)$




**singular values** $\sigma_i \geq 0$ on the main diagonal and 0s in the rest of the matrix. The columns of $\mathbf{U}$ are the left singular vectors and the columns of $\mathbf{V}$ are the right singular vectors.

The lower panel above shows the truncated SVD approximation of rank $L$.

## 1.1 Economy Sized SVD

- Assume that $N > D$

$$\underbrace{\mathbf{X}}_{N\times D} = \underbrace{\tilde{\mathbf{U}}}_{N\times D}\underbrace{\tilde{\mathbf{S}}}_{D\times D}\underbrace{\tilde{\mathbf{V}}^\top}_{D\times D}$$

### 1.2 Connection between Eigenvectors and Singular Vectors

- Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$

$$\begin{aligned}\mathbf{X}^\top\mathbf{X} &= \mathbf{V}\mathbf{S}^\top\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top \\ &= \mathbf{V}(\mathbf{S}^\top\mathbf{S})\mathbf{V}^\top \\ &= \mathbf{V}\mathbf{D}\mathbf{V}^\top\end{aligned}$$

- where $\mathbf{D} = \mathbf{S}^2$ is a diagonal matrix containing squares of singular values.
- Hence,

$$(\mathbf{X}^\top\mathbf{X})\mathbf{V} = \mathbf{V}\mathbf{D}$$

- Which means that the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{D}$ contains the eigenvalues.
- Similarly one can show that the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{D}$ contains the eigenvalues.

Remember that both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices.

### 1.3 PCA Using SVD

- Assuming that $\mathbf{X}$ is centered (zero mean) the principal components are equal to the right singular vectors of $\mathbf{X}$.

### 1.4 Low Rank Approximations Using SVD

- Choose only first $L$ singular values

$$\underbrace{\mathbf{X}}_{N\times D} \approx \underbrace{\tilde{\mathbf{U}}}_{N\times L}\underbrace{\tilde{\mathbf{S}}}_{L\times L}\underbrace{\tilde{\mathbf{V}}^\top}_{L\times D}$$

- Only need $NL + LD + L$ values to represent $N \times D$ matrix
- Also known as *rank $L$ approximation* of the matrix $\mathbf{X}$ (Why?) Because the rank of the approximate matrix will be $L$.



### 1.5 The Matrix Approximation Lemma

- Among all possible rank $L$ approximations of a matrix $\mathbf{X}$, SVD gives the best approximation
  - In the sense of minimizing the *Frobenius norm*

$$\|\mathbf{X} - \mathbf{X}_L\|$$

- Also known as the `Eckart Young Mirsky` theorem

### 1.6 Equivalence Between PCA and SVD

- For data $\mathbf{X}$ (assuming it to be centered)
- Principal components are the eigenvectors of $\mathbf{X}^\top\mathbf{X}$
- Or, principal components are the columns of $\mathbf{V}$

$$\mathbf{W} = \mathbf{V}$$

- Or

$$\hat{\mathbf{W}} = \hat{\mathbf{V}}$$

- $\hat{\mathbf{W}}$ are the first $L$ principal components and $\hat{\mathbf{V}}$ are the first $L$ right singular vectors.
- For PCA, data in latent space:

$$\begin{aligned}\hat{\mathbf{Z}} &= \mathbf{X}\hat{\mathbf{W}} \\ &= \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^\top\mathbf{V} \\ &= \hat{\mathbf{U}}\hat{\mathbf{S}}\end{aligned}$$

- Optimal reconstruction for PCA:

$$\begin{aligned}\hat{\mathbf{X}} &= \hat{\mathbf{Z}}\hat{\mathbf{W}}^\top \\ &= \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^\top\end{aligned}$$

- **Optimal reconstruction is same as *truncated SVD approximation*!!**

**Singular Value Decomposition - Recap**

- What is the (column) rank of a matrix?
- Maximum number of **linearly independent** columns in the matrix.
- For $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ (SVD):
  - What is the rank of $\hat{\mathbf{X}}^{(\mathbf{1})} = \mathbf{U}_{:\mathbf{1}}\sigma_{\mathbf{1}}\mathbf{V}_{:\mathbf{1}}^\top$?
  - The rank is 1 because each column of $\hat{\mathbf{X}}^{(\mathbf{1})}$ is a scaled version of the vector $U_{:1}$.
- How much storage is needed for a rank 1 matrix?
  - $O(N)$

**Importance of the Matrix Approximation Lemma**

- There are many ways to "approximate" a matrix with a lower rank approximation
- Low rank approximation allows us to *store* the matrix using much less than $N \times D$ bits ($O(N \times L)$ bits only)
- SVD gives the *best possible* approximation

$$\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2$$

## 1.7 SVD Applications

- A faster way to do PCA (truncated SVD, sparse SVD)
- Other applications as well:
  - Image compression
  - Recommender Systems
    - There are better methods
  - Topic modeling (Latent Semantic Indexing)

# References