Latent semantic analysis (LSA)

- Also known as Latent semantic indexing (LSI)

Descending order —>

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{bmatrix} u_1^{(1)} & \dots & u_1^{(M)} \\ \vdots & \vdots & \vdots \\ u_M^{(1)} & \dots & u_M^{(M)} \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & 0 \\ 0 & \dots & \sigma_M & \dots & 0 \end{bmatrix} \begin{bmatrix} v_1^{(1)} & \dots & v_N^{(1)} \\ \vdots & \vdots & \vdots \\ v_1^{(N)} & \dots & v_N^{(N)} \end{bmatrix}$$

- Used extensively in search engines

- Factors the document-term matrix, $\mathbf{X} \in \mathbb{R}^{M \times N}$, by computing its SVD

where $\mathbf{X} \in \mathbb{R}^{M \times N}$ term-document matrix

$\mathbf{U} \in \mathbb{R}^{M \times M}$ left singular vectors

$\mathbf{\Sigma} \in \mathbb{R}^{M \times N}$ diagonal matrix of singular values

$\mathbf{V} \in \mathbb{R}^{N \times N}$ right singular vectors

N = number of words

M = number of documents

Truncated SVD

- Because $\mathbf{X}$ is not full rank, there is some lower dimensional representation, $\mathbf{U}^* \in \mathbb{R}^{M \times K}$, which preserves most of its information. Because $\mathbf{U}^*$ is dense, standard Lp-based distance metrics are no longer meaningless

- Typically we evaluate singular values and do one of the following:
  - Set $K = r$, where r is the dimensionality of largest full rank approximation of $\mathbf{X}$
  - Set K using the elbow method [1]
  - Set $K \ll N$

- LSA uses the latter, which yields a truncated SVD form:

$$\mathbf{X} = \mathbf{U}^* \mathbf{\Sigma}^* \mathbf{V}^{*T} = \begin{bmatrix} u_1^{(1)} & \dots & u_1^{(K)} \\ \vdots & \vdots & \vdots \\ u_M^{(1)} & \dots & u_M^{(K)} \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_K \end{bmatrix} \begin{bmatrix} v_1^{(1)} & \dots & v_N^{(1)} \\ \vdots & \vdots & \vdots \\ v_1^{(K)} & \dots & v_N^{(K)} \end{bmatrix}$$

[1] Elbow Method