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

Query projection and cosine similarity

- When computing relevance between a query, $\mathbf{q}$, and some document, $\mathbf{d}$, we first must project $\mathbf{q}$ onto the K-dimensional manifold that we computed from the SVD of $\mathbf{X}$:

$$\hat{\mathbf{q}} = \mathbf{q}\mathbf{V}^{*}\boldsymbol{\Sigma}^{*-1} \quad \in \mathbb{R}^{1\times K}$$

- Document retrieval then amounts to measuring a distance between our projected query, $\hat{\mathbf{q}}$, and document, $\hat{\mathbf{d}}$, in our projected K-dimensional space. In LSA, it is common to use cosine similarity:

$$\text{cosine}(\hat{\mathbf{q}}, \hat{\mathbf{d}}) = \frac{\hat{\mathbf{q}} \cdot \hat{\mathbf{d}}}{\|\hat{\mathbf{q}}\| \|\hat{\mathbf{d}}\|} = \frac{\sum_{i=1}^{K} \hat{q}_i \hat{d}_i}{\sqrt{\sum_{i=1}^{K} \hat{q}_i^2} \sqrt{\sum_{i=1}^{K} \hat{d}_i^2}}$$

- In practice, the above can be computed efficiently using matrix-vector operations between $\hat{\mathbf{u}}$ and a projected corpus of documents $\hat{\mathbf{D}}$.