SVD Preliminaries

- \( \mathbb{R}^n = \left\{ \left( \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right) \mid x_i \text{ real} \right\} \)

- Inner product of \( x, y \in \mathbb{R}^n \): \( (x, y) = \sum_{i=1}^{n} x_i y_i = x^T y \). Note: \( \|x\|_2 = \sqrt{(x, x)} \).

- \( x, y \in \mathbb{R}^n \) orthogonal if \( (x, y) = 0 \).

- \( \{q_1, ..., q_k\} \subset \mathbb{R}^n \) orthonormal if \( (q_i, q_j) = \delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \)

- A matrix \( Q_{n \times n} \) is orthogonal if \( Q^T = Q^{-1} \).

- \( Q_{n \times n} \) is orthogonal \( \iff \) its columns \( \{q_1, ..., q_n\} \) are an orthonormal set.

  Proof: \( (Q^T Q)_{i,j} = (\text{row } i \text{ of } Q^T, \text{ col } j \text{ of } Q) = (q_i, q_j) = \delta_{i,j} \iff \{q_1, ..., q_n\} \) orthonormal.

- A set of orthonormal vectors \( \{q_1, ..., q_n\} \subset \mathbb{R}^n \) can be used as a basis for \( \mathbb{R}^n \). In fact, it is particularly simple to represent an arbitrary vector \( x \in \mathbb{R}^n \) as a linear combination \( x = \sum_{i=1}^{n} c_i q_i \), for this is equivalent to \( x = \underbrace{[q_1, ..., q_n]}_{Q} \underbrace{\begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}}_{c} \). Using the orthogonality of \( Q \), we have \( c = Q^T x \).

- \( Q \) orthogonal \( \implies \|Qx\|_2 = \|x\|_2 \) for all \( x \in \mathbb{R}^n \).

  Proof: \( \|Qx\|_2 = \sqrt{(Qx, Qx)} = \sqrt{x^T Q^T Q x} = \sqrt{x^T x} = \|x\|_2 \).

- \( Q_{n \times n} \) orthogonal \( \implies \|AQ\|_2 = \|A\| \) for arbitrary \( A_{m \times n} \), \( \|QB\|_2 = \|B\|_2 \) for arbitrary \( B_{n \times p} \).

  (These follow easily from the preceding result.)

- \( Q \) orthogonal \( \implies \kappa_2(Q) = 1 \). (Orthogonal matrices are perfectly conditioned in the 2-norm sense.)

- \( A_{n \times n} \) a real symmetric matrix \( \implies \) \( A \) has real eigenvalues \( \lambda_1, ..., \lambda_n \) and a corresponding set of orthonormal eigenvectors \( q_1, ..., q_n \)

\[
A q_i = \lambda_i q_i, \quad i = 1, ..., n
\]

\[
(q_i, q_j) = \delta_{i,j}
\]

Equivalently, in matrix form:

\[
A \underbrace{[q_1, ..., q_n]}_{Q} = \underbrace{[q_1, ..., q_n]}_{Q} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}
\]

\[
A = QAQ^T, \quad \Lambda = Q^T A Q
\]
Singular Value Decomposition (SVD)

**Basic result:** Given a real matrix $A_{m \times n}$, there exist orthogonal matrices

$$V_{n \times n} = [v_1, \ldots, v_n], \quad U_{m \times m} = [u_1, \ldots, u_m]$$

and a diagonal matrix

$$\Sigma_{m \times n} = \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \end{bmatrix}$$

with nonnegative diagonal entries $\sigma_i$ such that

$$A = U \Sigma V^T$$  \hspace{1cm} (1)

For proof, see Golub & Van Loan, Matrix Computations (Johns Hopkins).

**Picture of $\Sigma$**

$m > n$:

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}_{m \times n}$$

$m = n$:

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}_{n \times n}$$

$m < n$:

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}_{m \times n}$$

**terminology**

$\sigma_1, \ldots, \sigma_{\min\{m,n\}}$: *singular values of $A$* (nonnegative)

$v_1, \ldots, v_n$: *right singular vectors of $A$* (orthonormal)

$u_1, \ldots, u_m$: *left singular vectors of $A$* (orthonormal)

**Assumption:** the singular values of $A_{m \times n}$ are indexed in nonincreasing order:

$$\sigma_1 \geq \cdots \geq \sigma_r > 0, \quad \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0 \quad \text{if } r < \min\{m,n\}.$$

**Another view of the SVD**

Multiplying (1) on the right by $V$, we get

$$AV = U \Sigma$$
i.e.,
\[
A [v_1, ..., v_r | v_{r+1}, ..., v_n] = [u_1, ..., u_r | u_{r+1}, ..., u_m] = [\sigma_1 u_1, ..., \sigma_r u_r | 0, ..., 0],
\]
which is equivalent to
\[
Av_1 = \sigma_1 u_1 \\
\vdots \\
Av_r = \sigma_r u_r \\
Av_{r+1} = 0 \\
\vdots \\
Av_n = 0
\]
So what the SVD gives us is a pair of orthonormal bases
\[
\{v_1, ..., v_n\} \text{ for } \mathbb{R}^n \\
\{u_1, ..., u_m\} \text{ for } \mathbb{R}^m
\]
in terms of which \(A_{m \times n}\) has a particularly simple (decoupled) description.

- \(Av_i = \sigma_i u_i, \ i = 1, ..., r\)
- \(\text{range}(A) \equiv \{y | y = Ax \text{ for some } x\} = \text{span}\{u_1, ..., u_r\}\)
- \(\text{rank}(A) \equiv \text{dim}(\text{range}(A)) = r\)
- \(\text{null}(A) \equiv \{x | Ax = 0\} = \text{span}\{v_{r+1}, ..., v_n\} \ (= \emptyset \text{ if } r = n)\)
- \(\text{dim}(\text{null}(A)) = n - r\)

Additional facts:

- 'outer product' form of SVD
\[
A = \sum_{i=1}^{\min\{m, n\}} \sigma_i u_i v_i^T
\]

- \(A^T = V \Sigma^T U^T \) \(\Rightarrow\) the singular values of \(A^T\) are the same as those of \(A\) and the left/right singular vectors of \(A^T\) are the right/left singular vectors of \(A\)
- \(A^T A = V (\Sigma^T \Sigma) V^T \) \(\Rightarrow\) the eigenvalues and a corresponding set of orthonormal eigenvectors for \(A^T A\) are \(\{\sigma_1^2, ..., \sigma_r^2, 0, ..., 0\}\) and \(\{v_1, ..., v_n\}\)
- \(AA^T = U (\Sigma \Sigma^T) U^T \) \(\Rightarrow\) the eigenvalues and a corresponding set of orthonormal eigenvectors for \(AA^T\) are \(\{\sigma_1^2, ..., \sigma_r^2, 0, ..., 0\}\) and \(\{u_1, ..., u_m\}\).
The SVD “layers” $\mathbb{R}^n$, the domain space of $\mathbf{A}$, according to sensitivity to $\mathbf{A}$, as described below:

\[
\max_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\mathbf{v}_1\|_2 = \|\sigma_1 \mathbf{u}_1\|_2 = \sigma_1 \\
\max_{\mathbf{x} \in \mathbb{R}^n, (\mathbf{x}, \mathbf{v}_1)\neq(0), \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\mathbf{v}_2\|_2 = \|\sigma_2 \mathbf{u}_2\|_2 = \sigma_2 \\
\vdots \\
\max_{\mathbf{x} \in \mathbb{R}^n, (\mathbf{x}, \mathbf{v}_1)\neq\cdots=(\mathbf{x}, \mathbf{v}_{i-1})\neq(0), \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\mathbf{v}_i\|_2 = \|\sigma_i \mathbf{u}_i\|_2 = \sigma_i
\]

Thus the dominant “mode” or effect of $\mathbf{A}$ is $\mathbf{v}_1 \rightarrow \sigma_1 \mathbf{u}_1$, followed by $\mathbf{v}_2 \rightarrow \sigma_2 \mathbf{u}_2$, etc.

- $\mathbf{A}_{n \times n}$ is nonsingular if and only if $\kappa_2(\mathbf{A}_{n \times n}) = \sigma_1/\sigma_n$.

Applications of SVD

1. Solution of ill-conditioned system $\mathbf{A}_{n \times n}\mathbf{x} = \mathbf{b}$

\[
m = n \Rightarrow \Sigma = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \end{pmatrix}_{n \times n}
\]

\[
\mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \Rightarrow \Sigma\mathbf{V}^T\mathbf{x} = \mathbf{U}^T\mathbf{b}
\]

Defining $\xi = \mathbf{V}^T\mathbf{x}$, $\beta = \mathbf{U}^T\mathbf{b}$, we thus obtain an equivalent system with a diagonal coefficient matrix:

\[
\Sigma\xi = \beta.
\]

Note that since $\mathbf{V}$ and $\mathbf{U}$ are orthogonal, $\mathbf{x} = \mathbf{V}\xi = \sum_{i=1}^{n} \xi_i \mathbf{v}_i$ and $\mathbf{b} = \mathbf{U}\beta = \sum_{i+1}^{m} \beta_i \mathbf{u}_i$. (Thus we’re expressing $\mathbf{x}$ and $\mathbf{b}$ in terms of our orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively.)

Solution of (2):

\[
\xi_i = \frac{\beta_i}{\sigma_i}, \quad i = 1, \ldots, n \quad (\Rightarrow \mathbf{x} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i} \mathbf{v}_i)
\]

With SVD, we can now give a more complete answer to the stability question: If $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$ where $\|\delta\mathbf{b}\|_2 = \epsilon$, how large is $\delta\mathbf{x}$?

The corresponding transformed systems are $\Sigma\xi = \beta$, $\Sigma\delta\xi = \delta\beta$ where $\|\delta\beta\|_2 = \epsilon$. Thus

\[
|\delta\xi_i| = \left| \frac{(\delta\beta)_i}{\sigma_i} \right| \leq \frac{\epsilon}{\sigma_i}.
\]

Hence if $\sigma_n$ is very small in comparison to the other $\sigma_i$’s, we expect a large $\delta\xi_n$, in which case the error in $\mathbf{x} + \delta\mathbf{x}$ will be concentrated in the direction of $\mathbf{v}_n$; the remaining portion of $\mathbf{x} + \delta\mathbf{x}$ may be quite accurate.
2. Least squares solution of an overdetermined linear system

**Problem:** Choose \( x \) to minimize 
\[
Q(x) = \| A_{m \times n} x - b \|_2^2, \quad m > n
\]

**Solution via normal equations:** \( A^T A x = A^T b \)
(Denoting solution of normal equations by \( \bar{x} \), we have 
\[
Q(\bar{x} + \delta x) = Q(\bar{x}) + \| A \delta x \|_2^2 \geq Q(\bar{x}).
\]

**Solution via SVD** (more stable)... Since \( m > n \), \( \Sigma \) has the following configuration:

\[
\Sigma = \begin{bmatrix}
\sigma_1 & & \\
& \ddots & \\
& & \sigma_n \\
0 & & \\
\end{bmatrix}_{m \times n}
\]

Since \( U \) is orthogonal,

\[
\| A x - b \|_2^2 = \| U \Sigma V^T x - b \|_2^2 = \| U(\Sigma V^T x - U^T b) \|_2^2 = \| \Sigma V^T x - U^T b \|_2^2 = \| \Sigma \xi - \beta \|_2^2
\]

where \( \xi = V^T x, \beta = U^T b \). Now

\[
\Sigma \xi - \beta = \begin{bmatrix}
\sigma_1 \xi_1 - \beta_1 \\
\vdots \\
\sigma_n \xi_n - \beta_n \\
-\beta_{n+1} \\
\vdots \\
-\beta_m
\end{bmatrix}
\]

Thus

\[
\min_{x \in \mathbb{R}^n} \| A x - b \|_2 = \sqrt{\sum_{i=n+1}^{m} \beta_i^2} = \sqrt{\sum_{i=n+1}^{m} (u_i^T b)^2},
\]

which is achieved for

\[
\xi = \begin{bmatrix}
\beta_1 / \sigma_1 \\
\vdots \\
\beta_n / \sigma_n
\end{bmatrix} \quad \text{equivalently,} \quad x = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i.
\]

3. Lower rank approximation to \( A_{m \times n} \)

Define

\[
A_k = U \begin{bmatrix}
\Sigma_k & 0 \\
0 & 0
\end{bmatrix} V^T
\]
where
\[ \Sigma_k = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}. \]

Note that \( A_k \) is a rank \( k \) approximation to \( A \) (assuming \( k < r \)), and it should have good accuracy if the omitted singular values of \( A \) (i.e., \( \sigma_{k+1}, \ldots, \sigma_r \)) are small.

Observe that \( A_k \) has an alternative representation:
\[ A_k = U_k \Sigma_k V_k^T \]
where
\[ U_k = [u_1, \ldots, u_k]_{m \times k} \]
\[ V_k = [v_1, \ldots, v_k]_{n \times k} \]
and that the total storage requirement for \( U_k, \Sigma_k, V_k \) is \((m + n + 1)k\) vs. \( mn \) for \( A \) (or \( A_k \) in non-factored form). If \( k \ll \min\{m, n\} \) this can be an important saving. One application is image compression/coarsening (e.g., \( A = \) color-coded array of pixels).

Accuracy of \( A_k \) in 2-norm and Frobenius norm:
\[ \| A - A_k \|_2 = \| U \left( \Sigma - \begin{bmatrix} \Sigma_k \\ 0 \\ 0 \end{bmatrix} \right) V^T \|_2 \]
\[ = \| \Sigma - \begin{bmatrix} \Sigma_k \\ 0 \\ 0 \end{bmatrix} \|_2 \]
\[ = \sigma_{k+1} \]
\[ \| A - A_k \|_F = \| U \left( \Sigma - \begin{bmatrix} \Sigma_k \\ 0 \\ 0 \end{bmatrix} \right) V^T \|_F \]
\[ = \| \Sigma - \begin{bmatrix} \Sigma_k \\ 0 \\ 0 \end{bmatrix} \|_F \]
\[ = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2} \]

Claim: \( A_k \) minimizes both \( \| A - B \|_2, \| A - B \|_F \) over all rank \( k \) matrices \( B \). (See Golub & Van Loan for proof.)

4. Orthogonal regression (also known as principle component analysis)
Suppose we have \( m \) data points \( \{x_i\}_{i=1}^m \) in \( \mathbb{R}^n \), \( m > n \), and we wish to fit them by a \( k \)-dimensional hyperplane \( H_k \):
\[ H_k = x^* + S_k. \]
Here \( x^* \) is a point in \( \mathbb{R}^n \) and \( S_k \) denotes a \( k \)-dimensional subspace of \( \mathbb{R}^n \). The goal is to choose \( x^* \) and \( S_k \) to minimize \( \sum_{i=1}^m d_i^2 \) where \( d_i \) is the orthogonal distance from \( x_i \) to \( H_k \).

Example: \( n = 2 \), \( k = 1 \). In this case, \( H_k \) becomes
\[ H_1 = \left\{ \left( \begin{array}{c} x_1^* \\ x_2^* \end{array} \right) + c_1 q_1 \right\} \]
where $q_1$ is a unit vector in $\mathbb{R}^2$. Thus we are fitting $m$ points in 2-space by a straight line, with distances measured orthogonally (as opposed to vertically as in the case of least squares approximation).

Solution, in general:

$$x^* = \frac{1}{m} \sum_{i=1}^{m} x_i \quad \text{(the mean of the data points)}.$$

To determine $S_k$, form $A_{m \times n}$ with rows representing the data points $x_i - x_i^*$, and compute its SVD.

Then $S_k = \text{span}\{v_1, ..., v_k\}$.

5. Web search

Suppose, as result of a keyword search, we have $n$ potentially relevant web pages $P_1, ..., P_n$. Which ones to report back to user? How to rank them? Useful technique: analyze link structure via SVD...

$$G = (V, E) \quad \text{(graph representing links)}$$

$$V = \{1, ..., n\} \quad \text{(vertex $i$ represents $P_i$)}$$

$$E = \{ij \mid P_i \text{ points to } P_j\} \quad \text{(edges)}$$

The corresponding adjacency matrix, denoted by $M_{n \times n}$, has as its $i,j$th element

$$m_{i,j} = \begin{cases} 1, & \text{if } ij \in E, \\ 0, & \text{if } ij \notin E. \end{cases}$$

For any given subset of vertices $S \subset V$, there is a corresponding set of links $T$ to other vertices.

We describe sets $S, T$ in terms of vectors $x, y \in \mathbb{R}^n$ defined by

$$x_i = \begin{cases} 1, & i \in S, \\ 0, & \text{otherwise}, \end{cases}$$

$$y_j = \text{number of edges from } S \text{ to vertex } j.$$ 

Note that $y_j = \sum_{i \in S} a_{i,j} = \sum_{i=1}^{n} x_i a_{i,j}$. Thus $y^T = x^T A$; equivalently, $y = A^T x$. This suggests the possibility of using an SVD of $A^T$ in order to summarize the primary information content of $G$ in a condensed form.

Let $A = U \Sigma V^T$ be the SVD of $A$, in which case $A^T = V \Sigma U^T$ is the SVD of $A^T$. The dominant modes in the transformation $A^T x = y$ are then given by:

$$A^T u_1 = \sigma_1 v_1$$

$$A^T u_2 = \sigma_2 v_2$$

$$\ldots$$

What we are looking for is a rapid decay in the $\sigma_i$'s so all but the first few are negligible.. The entries of $v_1, v_2, ...$ yield “authority weights” for web pages $P_1, P_2, ...$ with respect to the given keyword. The vectors $u_1, u_2, ...$ furnish a corresponding set of “hub weights”.

7