The Singular Value Decomposition (SVD)

K. Jbilou

Université du Littoral Côte d'Opale, France
2022-23

The Singular Value Decomposition

The singular value decomposition (SVD) is a factorization of a real or complex matrix that generalizes the eigen-decomposition of a square normal matrix to any $n \times n$ matrix via an extension of the polar decomposition. Specifically, the singular value decomposition of an $m \times n$ matrix A is a factorization of the form

$$A = U\Sigma V^*$$

where U and V are unitary matrices and Σ is a diagonal matrix.

History of the Singular Value Decomposition

The early five mathematiciens who contributed to the introduction of SVD are

- Eugenio Beltrami (1835-1899)
- Camille Jordan (1838-1921)
- James Joseph Sylvester (1814-1897)
- Erhard Scmidt (1876-1959)
- Hermann Weyl (1885-1955).

Algorithms for computing SVD have been proposed by

- Forsythe-Henrici (1960), Hestenes (1958)
- The SVD was introduced into modern numerical analysis by Golub and co-authors: Golub-Kahan (1965) and
- Golub-Reisch (1970) who used Householder transformations to reduce A to bidiagonal form by the (Golub-Kahan) method.
- Details on SVD, GSVD are found in the very good book of Golub and Van Loan
- For more details on history, see the good paper of Stewart: siam review 1993, vol. 35, No4.

Some Applications of SVD

- Least squares problems $min_x \parallel Ax - b \parallel_2$.
- Low-rank approximation of a matrix: $A \approx U_1 V_1$
- Image compression; Data mining
- Image restoration
- Face recognition (Principal Component Analysis PCA)
- Many applications in signal processing
- Statistics problems..

The following theorem shows that every matrix could be decomposed as a product of orthogonal matrices and a diagonal matrix.

Theorem

Let A be a real $m \times n$ matrix. Then there exist two orthogonal matrices

$$U = [u_1, u_2, \ldots, u_m] \in \mathbb{R}^{m \times m} \quad \text{and} \quad V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{n \times n}$$

and a diagonal matrix

$$\Sigma = diag(\sigma_1, \ldots, \sigma_p), \text{ with } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0, \quad p = min\{n, m\},$$

such that

$$A = U \Sigma V^T.$$

This decomposition is called the Singular Value Decomposition (SVD) of the matrix A. The σ_i's are called singular values and the u_i's and v_i's are called Left and Right singular vectors, respectively.

Proof

As the matrix $A^T A$ is symmetric and postive, it can be diagonalisable in an orthonormal basis of eigenvectors and the eigenvalues are potitive. Let the eigenvalues of $A^T A$ be

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2 > 0 = \sigma_{r+1}^2 = \ldots = \sigma_n^2.$$

If $V = [V_1, V_2]$ is an orthogonal matrix formed from the corresponding eigenvectors. Then

$$V^T A^T A V = \begin{pmatrix} \Sigma_+^2 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\Sigma_+^2 = diag(\sigma_1^2, \ldots, \sigma_r^2)$. Then we have

$$V_1^T A^T A V_1 = \Sigma_+^2, \text{ and } V_2^T A^T A V_2 = 0.$$

Proof

As the matrix $A^T A$ is symmetric and postive, it can be diagonalisable in an orthonormal basis of eigenvectors and the eigenvalues are potitive. Let the eigenvalues of $A^T A$ be

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2 > 0 = \sigma_{r+1}^2 = \ldots = \sigma_n^2.$$

If $V = [V_1, V_2]$ is an orthogonal matrix formed from the corresponding eigenvectors. Then

$$V^T A^T A V = \begin{pmatrix} \Sigma_+^2 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\Sigma_+^2 = diag(\sigma_1^2, \ldots, \sigma_r^2)$. Then we have

$$V_1^T A^T A V_1 = \Sigma_+^2, \text{ and } V_2^T A^T A V_2 = 0.$$

From the second relation, we conclude that

$$AV_2 = 0.$$

We set $U_1 = AV_1\Sigma_+^{-1}$. Then, we get $U_1^T U_1 = I$.

We choose U_2 such that the square matrix $U = [U_1, U_2]$ is orthogonal. Therefore

$$U^T AV = \Sigma = \begin{pmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{pmatrix},$$

which ends the proof

Notice that

$$Av_i = \sigma_i u_i, \quad ; \; A^T u_j = \sigma_j v_j, \tag{1}$$

and

$$A^T A v_i = \sigma_i^2 v_i, \quad AA^T u_i = \sigma_i^2 u_i, \; i = min(n, m). \tag{2}$$

- The u_i's and v_i's are called left singular vectors and right singular vectors, respectively.
- The largest singular value is denoted by $\sigma_{max}(A)$ while the smallest one is denoted by $\sigma_{min}(A)$.
- The singular value decomposition gives many important informations about the matrix A. Some of these properties are listed in the following theorem.

Example

Consider the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The eigenvalues of $A^T * A$ are $\lambda_1 = \lambda_2 = 2$ so the singular values are $\sigma_1 = \sigma_2 = \sqrt{2}$. The associated eigenvalues are $v_1 = [1\ 0]^T$ and $v_2 = [0\ 1]^T$. It is clear that the eigenvalues of $A * A^T$ are $2, 2, 0, 0$. We use the relations (1), to get an orthonormal basis$\{u_1, u_2, u_3, u_4\}$ of the proper subspace associated to $A * A^T$ to get the matrices

$$V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Theorem

Consider the SVD of $A = U\Sigma V^T$ *and define* r *by*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Then

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, *with* $r = rank(A)$.
2. $Ker(A) = span\{v_{r+1}, \ldots, v_n\}$.
3. $Range(A) = span\{u_1, \ldots, u_r\}$.
4. $\|A\|_2 = \sigma_1 = \sigma_{max}(A)$.
5. $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2$.
6. *The condition number* $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

Theorem

Consider the SVD of $A = U\Sigma V^T$ *and define* r *by*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Then

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, *with* $r = rank(A)$.
2. $Ker(A) = span\{v_{r+1}, \ldots, v_n\}$.
3. $Range(A) = span\{u_1, \ldots, u_r\}$.
4. $\|A\|_2 = \sigma_1 = \sigma_{max}(A)$.
5. $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2$.
6. *The condition number* $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

Theorem

Consider the SVD of $A = U\Sigma V^T$ *and define* r *by*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Then

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, *with* $r = rank(A)$.
2. $Ker(A) = span\{v_{r+1}, \ldots, v_n\}$.
3. $Range(A) = span\{u_1, \ldots, u_r\}$.
4. $\|A\|_2 = \sigma_1 = \sigma_{max}(A)$.
5. $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2$.
6. *The condition number* $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

Theorem

Consider the SVD of $A = U\Sigma V^T$ *and define* r *by*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Then

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, *with* $r = rank(A)$.
2. $Ker(A) = span\{v_{r+1}, \ldots, v_n\}$.
3. $Range(A) = span\{u_1, \ldots, u_r\}$.
4. $\|A\|_2 = \sigma_1 = \sigma_{max}(A)$.
5. $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2$.
6. *The condition number* $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

Theorem

Consider the SVD of $A = U\Sigma V^T$ *and define* r *by*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Then

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, *with* $r = rank(A)$.
2. $Ker(A) = span\{v_{r+1}, \ldots, v_n\}$.
3. $Range(A) = span\{u_1, \ldots, u_r\}$.
4. $\|A\|_2 = \sigma_1 = \sigma_{max}(A)$.
5. $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2$.
6. *The condition number* $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

Theorem

Consider the SVD of $A = U\Sigma V^T$ *and define* r *by*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Then

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, *with* $r = rank(A)$.
2. $Ker(A) = span\{v_{r+1}, \ldots, v_n\}$.
3. $Range(A) = span\{u_1, \ldots, u_r\}$.
4. $\|A\|_2 = \sigma_1 = \sigma_{max}(A)$.
5. $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2$.
6. *The condition number* $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

Proof

The first relation comes directly from the SVD decomposition (1). For the second relation, a vector x is in $ker(A)$ iff $Ax = 0$ which is equivalent to

$$\sum_{i=1}^{r} \sigma_i u_i v_i^T x = \sum_{i=1}^{r} \sigma_i (v_i^T . x) u_i = 0$$

Therefore, multiplying both sides by u_j^T for $j = r+1, \ldots, n$ we get $v_j^T x = 0$ which shows the result. For the relation 3, we consider a vector $y = Ax$ in $range(A)$ and this implies that $y = \sum_{i=1}^{r} (\sigma_i v_i^T x) u_i$ which shows the result.
To show the relations 4 and 5, we just use the properties of the 2 and F-norm conservations of orthogonal transformations, we get $\|A\|_2 = \|\Sigma\|_2$ and also $\|A\|_F = \|\Sigma\|_F$ and since Σ is diagonal, the results follow.

Proof

The first relation comes directly from the SVD decomposition (1). For the second relation, a vector x is in $ker(A)$ iff $Ax = 0$ which is equivalent to

$$\sum_{i=1}^{r} \sigma_i u_i v_i^T x = \sum_{i=1}^{r} \sigma_i (v_i{}^T . x) u_i = 0$$

Therefore, multiplying both sides by u_j^T for $j = r+1, \ldots, n$ we get $v_j^T x = 0$ which shows the result. For the relation 3, we consider a vector $y = Ax$ in $range(A)$ and this implies that $y = \sum_{i=1}^{r} (\sigma_i v_i{}^T x) u_i$which shows the result.

To show the relations 4 and 5, we just use the properties of the 2 and F-norm conservations of orthogonal transformations, we get $\|A\|_2 = \|\Sigma\|_2$ and also $\|A\|_F = \|\Sigma\|_F$ and since Σ is diagonal, the results follow.

Proof

The first relation comes directly from the SVD decomposition (1). For the second relation, a vector x is in $ker(A)$ iff $Ax = 0$ which is equivalent to

$$\sum_{i=1}^{r} \sigma_i u_i v_i^T x = \sum_{i=1}^{r} \sigma_i (v_i{}^T . x) u_i = 0$$

Therefore, multiplying both sides by u_j^T for $j = r+1, \ldots, n$ we get $v_j^T x = 0$ which shows the result. For the relation 3, we consider a vector $y = Ax$ in $range(A)$ and this implies that $y = \sum_{i=1}^{r} (\sigma_i v_i{}^T x) u_i$which shows the result.

To show the relations 4 and 5, we just use the properties of the 2 and F-norm conservations of orthogonal transformations, we get $\|A\|_2 = \|\Sigma\|_2$ and also $\|A\|_F = \|\Sigma\|_F$ and since Σ is diagonal, the results follow.

We can also consider the economical **Thin Singula Value Decomposition** as following:

If $A = U\Sigma V^T$ as in (1) then

$$A = U_r \Sigma_r V_r^T \tag{3}$$

where $U_r = [u_1, \ldots, u_r]$, $\Sigma_r = diag(\sigma_1, \ldots, \sigma_r)$ and $V_r = [v_1, \ldots, v_r]$

Another important result on SVD is stated in the following Eckart-Young theorem

Theorem

(Theorem of Eckart-Young) Let A be an $n \times m$ matrix of rank r. Then

$$\min_{rank(X)=k<r} \|A - X\|_2 = \sigma_{k+1}$$

where

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} \geq \ldots \geq \sigma_n$$

are the singular values of A.
A minimizer X_* is given by

$$X_* = \sigma_1 u_1 v_1^T + \ldots + \sigma_k u_k v_k^T. \tag{4}$$

We also have

$$\min_{rank(X)=k<r} \|A - X\|_F = \sum_{i=k+1}^{n} \sigma_i^2,$$

Proof

We have

$$U^T A_k V = U^T(\sum_{i=1}^{k} \sigma_i u_i v_i^T)V = (\sum_{i=1}^{k} \sigma_i U^T u_i (V^T v_i)^T$$

and since $U^T u_i = V^T v_i = e_i e_i^T$, we get
$U^T A_k V = Vdiag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. It follows that $rank(A_k) = k$ and

$U^T(A - A_k)V = diag(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$, wich implies $\|A - A_k\|_2 = \sigma_{k+1}$

Now, assume that $rank(X) = k$ for some $X \in \mathbb{R}^{m \times n}$, then we can find orthogonal vectors $w_1, \ldots, w_{n-k}$ such that $ker(X) = \{w_1, \ldots, w_{n-k}\}$. For any unit vector $z \in span\{w_1, \ldots, w_{n-k}\} \cap span\{v_1, \ldots, v_{n-k}\}$, we have

$$Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i, \text{ then}$$

$$\|A - X\|_2^2 \geq \|(A - X)z\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2,$$

Proof

We have

$$U^T A_k V = U^T(\sum_{i=1}^{k} \sigma_i u_i v_i^T)V = (\sum_{i=1}^{k} \sigma_i U^T u_i (V^T v_i)^T$$

and since $U^T u_i = V^T v_i = e_i e_i^T$, we get
$U^T A_k V = Vdiag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. It follows that $rank(A_k) = k$ and

$U^T(A - A_k)V = diag(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$, wich implies $\|A - A_k\|_2 = \sigma_{k+1}$

Now, assume that $rank(X) = k$ for some $X \in \mathbb{R}^{m\times n}$, then we can find orthogonal vectors $w_1, \ldots, w_{n-k}$ such that $ker(X) = \{w_1, \ldots, w_{n-k}\}$. For any unit vector $z \in span\{w_1, \ldots, w_{n-k}\} \cap span\{v_1, \ldots, v_{n-k}\}$, we have

$$Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i, \text{ then}$$

$$\|A - X\|_2^2 \geq \|(A - X)z\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2,$$

Proof

We have

$$U^T A_k V = U^T(\sum_{i=1}^{k} \sigma_i u_i v_i^T)V = (\sum_{i=1}^{k} \sigma_i U^T u_i (V^T v_i)^T$$

and since $U^T u_i = V^T v_i = e_i e_i^T$, we get
$U^T A_k V = Vdiag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. It follows that $rank(A_k) = k$ and

$U^T(A - A_k)V = diag(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$, wich implies $\|A - A_k\|_2 = \sigma_{k+1}$

Now, assume that $rank(X) = k$ for some $X \in \mathbb{R}^{m \times n}$, then we can find orthogonal vectors $w_1, \ldots, w_{n-k}$ such that $ker(X) = \{w_1, \ldots, w_{n-k}\}$. For any unit vector $z \in span\{w_1, \ldots, w_{n-k}\} \cap span\{v_1, \ldots, v_{n-k}\}$, we have

$$Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i, \text{ then}$$

$$\|A - X\|_2^2 \geq \|(A - X)z\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2,$$

Proof

We have

$$U^T A_k V = U^T(\sum_{i=1}^{k} \sigma_i u_i v_i^T)V = (\sum_{i=1}^{k} \sigma_i U^T u_i (V^T v_i)^T$$

and since $U^T u_i = V^T v_i = e_i e_i^T$, we get
$U^T A_k V = Vdiag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. It follows that $rank(A_k) = k$ and

$U^T(A - A_k)V = diag(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$, wich implies $\|A - A_k\|_2 = \sigma_{k+1}$

Now, assume that $rank(X) = k$ for some $X \in \mathbb{R}^{m \times n}$, then we can find orthogonal vectors $w_1, \ldots, w_{n-k}$ such that $ker(X) = \{w_1, \ldots, w_{n-k}\}$. For any unit vector $z \in span\{w_1, \ldots, w_{n-k}\} \cap span\{v_1, \ldots, v_{n-k}\}$, we have

$$Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i, \text{ then}$$

$$\|A - X\|_2^2 \geq \|(A - X)z\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2,$$

Theorem 3 can be used for applications in image processing (compression, transmission). This is shown in the following example:

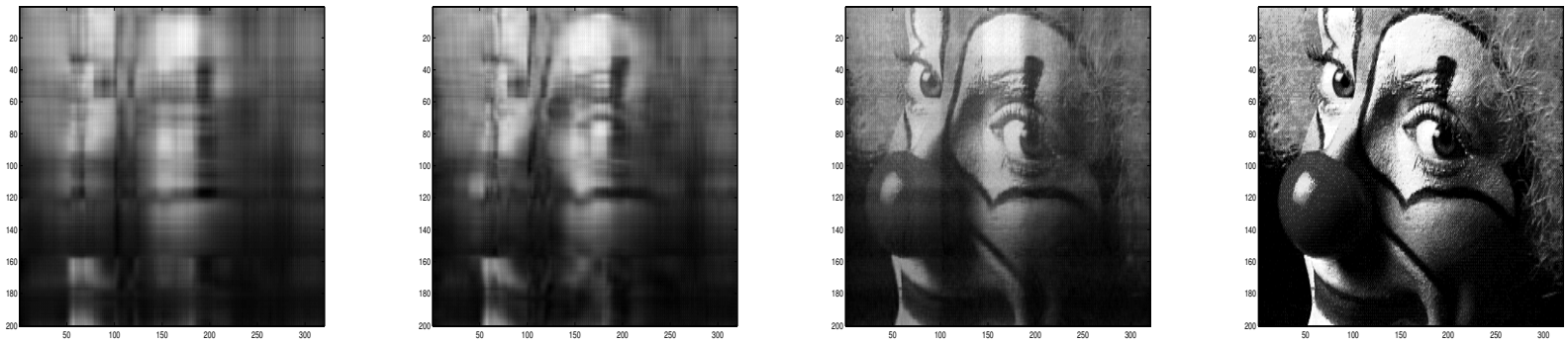


FIGURE – Low-rank TSVD approximations: $k = 5$, $k = 10$, $k = 50$ and the exact image (200×320)

Remark: For image compression FFT is usually faster.

Solving nonsingular linear systems

Theorem

Consider the nonsingular linear system of equations

$$Ax = b$$

where $A \in \mathbb{R}^{n \times n}$, *then the solution of this problem can be expressed as*

$$x = A^{-1}b = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i$$

where $A = U\Sigma V^T$.

This result shows that small changes in the entries of A or b can induce relatively large changes in the solution x if σ_n is small.

The following result enables us to compute the pseudo inverse

Theorem

Consider the SVD of A,

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

- *The $n \times m$ pseudo-inverse of the matrix A can be written as*

$$A^+ = V\Sigma^+ U^T,$$

where

$$\Sigma^+ = diag\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}\right).$$

- $A^+ = \sum_{i=1}^{r} \frac{v_i u_i^T}{\sigma_i}$.
- AA^+ *is the matrix of the orthogonal projection onto range*(A).
- A^+A *is the matrix of the orthogonal projection onto range*(A^T).

The following result enables us to compute the pseudo inverse

Theorem

Consider the SVD of A,

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

- *The $n \times m$ pseudo-inverse of the matrix A can be written as*

$$A^+ = V\Sigma^+ U^T,$$

where

$$\Sigma^+ = diag\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}\right).$$

- $A^+ = \sum_{i=1}^{r} \frac{v_i u_i^T}{\sigma_i}.$
- *AA^+ is the matrix of the orthogonal projection onto range(A).*
- *A^+A is the matrix of the orthogonal projection onto range(A^T).*

The following result enables us to compute the pseudo inverse

Theorem

Consider the SVD of A,

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

- *The $n \times m$ pseudo-inverse of the matrix A can be written as*

$$A^+ = V\Sigma^+ U^T,$$

where

$$\Sigma^+ = diag\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}\right).$$

- $A^+ = \sum_{i=1}^{r} \frac{v_i u_i^T}{\sigma_i}$.
- *AA^+ is the matrix of the orthogonal projection onto range(A).*
- *A^+A is the matrix of the orthogonal projection onto range(A^T).*

The following result enables us to compute the pseudo inverse

Theorem

Consider the SVD of A,

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

- *The $n \times m$ pseudo-inverse of the matrix A can be written as*

$$A^+ = V\Sigma^+ U^T,$$

where

$$\Sigma^+ = diag\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}\right).$$

- $A^+ = \sum_{i=1}^{r} \frac{v_i u_i^T}{\sigma_i}$.
- *AA^+ is the matrix of the orthogonal projection onto range(A).*
- *A^+A is the matrix of the orthogonal projection onto range(A^T).*

The full rank Least Squares Problem

Consider the following rectangular linear system of equations

$$Ax = b$$

where $A \in \mathbb{R}^{m \times n}$ with $m \geq n$. This linear system is overdetermined since we have more equations than unknowns. Usually, this system has no exact solution since b must be in the range of A. Then, one possibility is solve the following least squares problem

$$\min_{x \in \mathbb{R}^m} \|Ax - b\|_2. \tag{5}$$

Let us set

$$\Phi(x) = \frac{1}{2}\|Ax - b\|_2^2. \tag{6}$$

As Φ is a differentiable function, and so any minimizer of Φ must satisfy the gradient equation $\nabla\Phi(x) = 0$ which leads to solving the normal equation

$$A^T A x = A^T b. \tag{7}$$

Therefore, if A has a full rank n the normal equation (7) has a unique solution denoted by x_{LS} with the minimum residual

$$\rho_{LS} = \|Ax_{LS} - b\|_2. \tag{8}$$

Let us see now how to use SVd for solving the problem (5). Setting $A = U\Sigma V^T$ in (5), we get

$$\begin{aligned} \|Ax - b\|_2^2 &= \|U(\Sigma V^T x - U^T b)\|_2^2 && (9)\\ &= \|\Sigma z - U^T b\|_2^2 && (10)\\ &= \sum_{i=1}^{r}(\sigma_i z_i - u_i^T b)^2 + \sum_{i=r+1}^{m}(u_i^T b)^2, && (11) \end{aligned}$$

where $z = V^T x$ which is equivalent to $x = Vz$; and $r = rank(A)$.

So the minimizer of $\|Ax - b\|_2^2$ is such that $z_i = \dfrac{u_i^T b}{\sigma_i}$, $i = 1, \ldots, r$.

And then the obtained solution x_{LS} minimizing $\|Ax - b\|_2$ is given by

$$x_{LS} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i, \tag{12}$$

and the corresponding residual norm given as

$$\rho_{LS} = \|Ax_{LS} - b\|_2 = \sum_{i=r+1}^{m} (u_i^T b)^2 \tag{13}$$

Notice that an approximation of x_{LS} could be obtained by Truncating the formulae (12) for some chosen integer k to get the Truncated SVD approximation

$$x_{LS} \approx x_k = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i, \tag{14}$$

The main problem for this TSVD technique is how to chose the optimal parameter k (some methods: Tikhonov, GCV,...) will be seen later.

An example of ill-conditioned Matrices

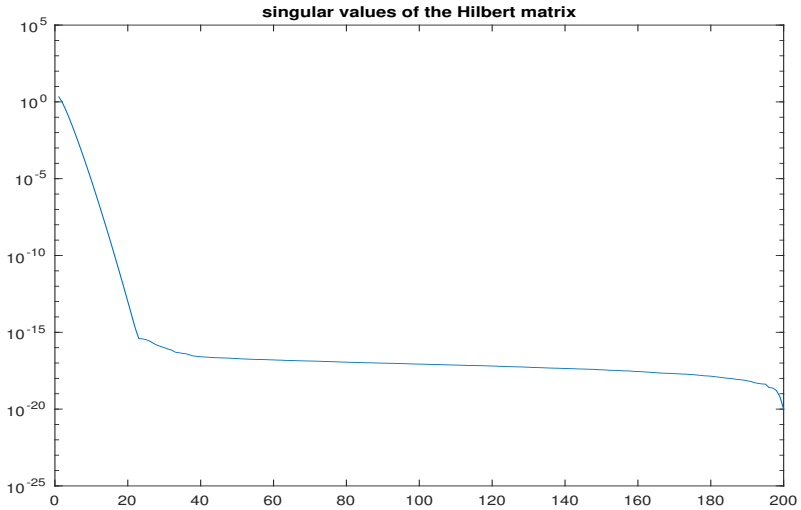


FIGURE – The singular values of the Hilbert matrix H

In this figure, we plotted the singular values of the Hilbert matrix of size $n = 200$. The plot shows that the singular values are decreasing quickly to zero which leads to a very ill-conditioned matrix. This means that when solving linear systems with the matrix H, a small change in the data will implies a large perturbation in the computed solution. The condition number is

$$cond(H) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}} = 2.3365 \times 10^{20}$$

An example of ill-conditioned Matrices

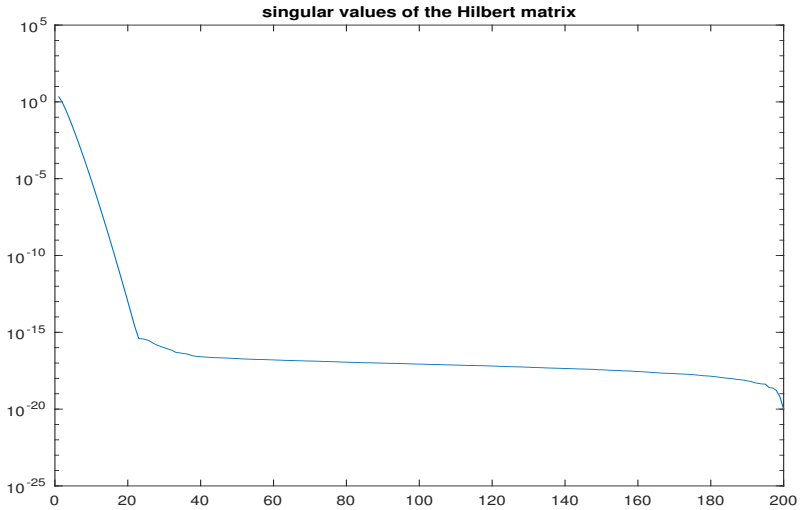


FIGURE – The singular values of the Hilbert matrix H

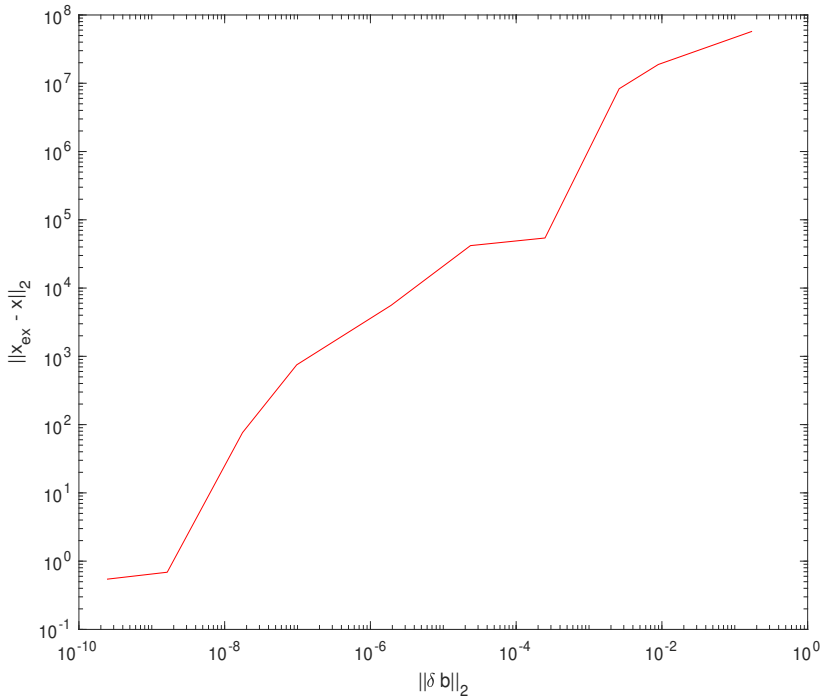
In this figure, we plotted the singular values of the Hilbert matrix of size $n = 200$. The plot shows that the singular values are decreasing quickly to zero which leads to a very ill-conditioned matrix. This means that when solving linear systems with the matrix H, a small change in the data will implies a large perturbation in the computed solution. The condition number is

$$cond(H) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}} = 2.335 \times 10^{20}$$

Example 1

$tt = (0 : -1 : -10)^T$; $t = 10^{tt}$
$A = [ones(size(t))\ t\ t^2\ t^3\ t^4\ t^5]$;
compute SVD of A: $[U, S, V] = svd(A); sigma = diag(S)$;
compute exact data: $x_{ex} = ones(6, 1); b_{ex} = A * x_{ex}$;
$fori = 1 : 10$
data perturbation: $deltab = 10^{-i} * (0.5 - rand(size(b_{ex}))). * b_{ex}$;
$b = b_{ex} + \delta b$;
solution of perturbed linear least squares problem
$w = U' * b$;
$x = V * (w(1 : 6)./sigma)$;
$err_x(i + 1) = \|x - x_{ex}\|_2; err_b(i + 1) = \|\delta b\|_2$;
end
loglog(errb,errx,*);
ylabel($||x_{ex} - x||_2$); xlabel($||\delta b||_2$).

In this example, we took a matrix A of size 11×6 with the singular values:
$\sigma_1 \approx 3.4$; $\sigma_2 \approx 2.1$; $\sigma_3 \approx 8.2 * 10^{-2}$;
$\sigma_4 \approx 7.2 * 10^{-4}$; $\sigma_5 \approx 6.6 * 10^{-7}$; $\sigma_6 \approx 5.5 * 10^{-11}$.



In this figure, we plotted the error $\|x_{ex} - x\|_2$ versus the perturbation $\|\delta b\|_2$. We see that small perturbations δb in the measurements can lead to large errors in the computed solution x.

Assume now that b is affected by some noise due to measurements, that is

$$b = b_{ex} + e$$

where e is a noise and b_{ex} is the free right-hand side . In this case the minimum norm solution of $\min\|Ax - b\|_2$ is given as

$$x_{LS} = \sum_{i=1}^{r} \left(\frac{u_i^T b_{ex}}{\sigma_i} v_i + \frac{u_i^T e}{\sigma_i} v_i \right).$$

Notice that if σ_i is small, $\frac{u_i^T e}{\sigma_i}$ could be very large. This shows that that errors e in the data can be magnified by small singular values σ_i. This is the case in image restoration where b represent the observed image that has been corrupted by some noise e. .

Assume now that b is affected by some noise due to measurements, that is

$$b = b_{ex} + e$$

where e is a noise and b_{ex} is the free right-hand side . In this case the minimum norm solution of $\min\|Ax - b\|_2$ is given as

$$x_{LS} = \sum_{i=1}^{r} \left(\frac{u_i^T b_{ex}}{\sigma_i} v_i + \frac{u_i^T e}{\sigma_i} v_i \right).$$

Notice that if σ_i is small, $\dfrac{u_i^T e}{\sigma_i}$ could be very large. This shows that that errors e in the data can be magnified by small singular values σ_i. This is the case in image restoration where b represent the observed image that has been corrupted by some noise e. .

Example 2: Image restoration: logray image 512×512

In this example, we consider an image that has been blurred with an additive Gaussian noise. The linear model is obtained as

$$Ax = b_{ex} + e = b$$

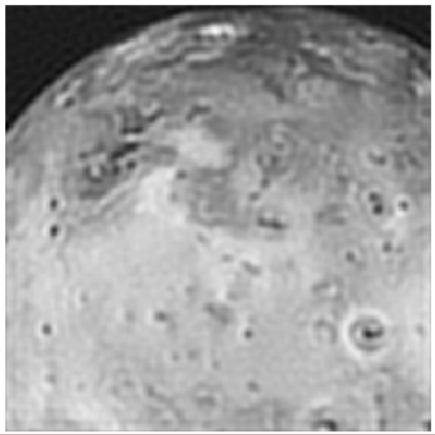
where b is the observed image and b_{ex} is the free noisy observed image. The solution x is the image to be recovered and A is the matrix corresponding to the blur and e Gaussian noise with $\|e\|_2 = 0.01$.

If X is an image of size $m \times n$, then $x = vec(X)$ is the vector obtained by puting the columns of X in a long vector x of size mn and then the blurring matrix A is large and of size $mn \times mn$. Blurred noisy image and Restored with $k = 658$.

Original Image

Blurred and Noisy Image

Restored Image



Example 2: Image restoration: logray image 512×512

In this example, we consider an image that has been blurred with an additive Gaussian noise. The linear model is obtained as

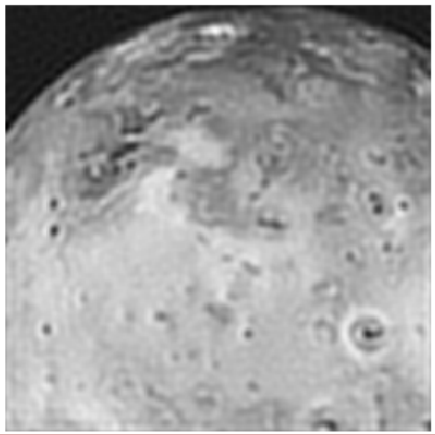
$$Ax = b_{ex} + e = b$$

where b is the observed image and b_{ex} is the free noisy observed image. The solution x is the image to be recovered and A is the matrix corresponding to the blur and e Gaussian noise with $\|e\|_2 = 0.01$.
If X is an image of size $m \times n$, then $x = vec(X)$ is the vector obtained by puting the columns of X in a long vector x of size mn and then the blurring matrix A is large and of size $mn \times mn$. Blurred noisy image and Restored with $k = 658$.

Original Image

Blurred and Noisy Image

Restored Image



Tikhonov Regularisation for Ill-posed Problems

In many situations, the $N \times N$ matrix of the least squares problem

$$\min_x \|Ax - b\|_2,$$

is ill conditioned which means that many singular values are close to zero; this occurs in image processing. As we have seen in the previous chapter, small perturbations in the data may cause large perturbations in the least squares solution. One possibility to remedy to this situation is to use the Tikhonov regularisation procedure. We replace the original problem by the following one

$$\min_x \|Ax - b\|_2^2 + \lambda^2 \|Lx\|_2^2, \tag{15}$$

for some chosen regularisation parameter λ and a regularized matrix L (we will take here $L = I$).

An example of ill-conditioned Matrices

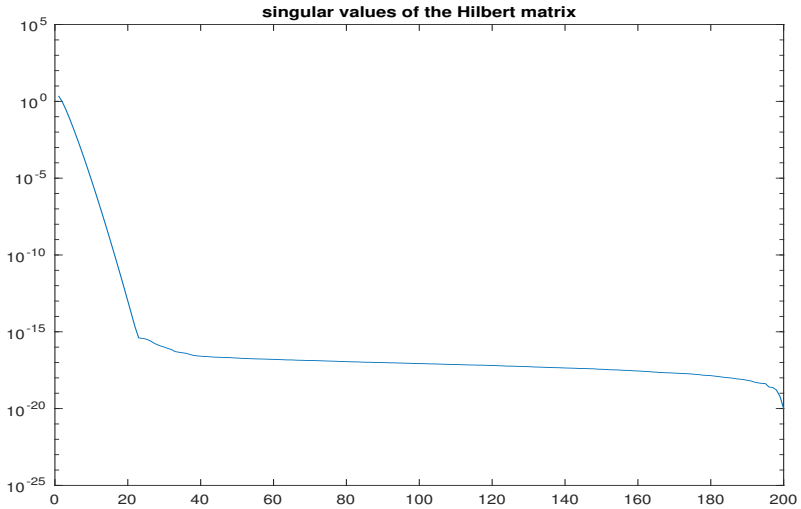


FIGURE – The singular values of the Hilbert matrix H

The minimizer of (2) is computed as the solution of the following linear system

$$A_\lambda x = A^T b, \text{ with } A_\lambda = (A^T A + \lambda^2 I). \tag{16}$$

The proof is simple since the minimizer of (15) satisfies the normal equation

$$\begin{bmatrix} A^T & \lambda I \end{bmatrix} \begin{bmatrix} A \\ \lambda I \end{bmatrix} x_\lambda = \begin{bmatrix} A^T & \lambda I \end{bmatrix} \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{17}$$

The minimizer of (2) is computed as the solution of the following linear system

$$A_\lambda x = A^T b, \text{ with } A_\lambda = (A^T A + \lambda^2 I). \tag{16}$$

The proof is simple since the minimizer of (15) satisfies the normal equation

$$\begin{bmatrix} A^T & \lambda I \end{bmatrix} \begin{bmatrix} A \\ \lambda I \end{bmatrix} x_\lambda = \begin{bmatrix} A^T & \lambda I \end{bmatrix} \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{17}$$

The parameter λ should be computed such that $\|Ax - b\|_2$ is small and $\|x\|_2$ is not very large. One of the well known method for computing an "optimal" parameter is called the Generalized Cross Validation (GCV) method due to Golub & all.

$$G(\lambda) = \frac{||Ax_\lambda - \mathrm{b}||_2^2}{[tr(I - AA_\lambda^{-1}A^T)]^2} = \frac{||(I - AA_\lambda^{-1}A^T)\mathrm{b}||_2^2}{[tr(I - AA_\lambda^{-1}A^T)]^2}. \tag{18}$$

Let $A = U\Sigma V$ be the SVD of the matrix A then the GCV function G reduces to

$$G(\lambda) = \frac{\sum_{i=1}^{N} \left(\frac{u_i^T b}{\sigma_i^2 + \lambda^2} \right)^2}{\sum_{i=1}^{N} \left(\frac{1}{\sigma_i^2 + \lambda^2} \right)^2}, \tag{19}$$

where u_i is the i-th left singular vector and σ_i is the i-th singular value of the matrix A. It was shown by Golub & all that the best parameter for the problem (15) is obtained as the minimum of the minimum of the CGV function G.

After computing the optimal parameter λ, the problem (15), is equivalent to the following one

$$\min_x \left\| \begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \lambda I \end{bmatrix} x_\lambda \right\|_2^2. \tag{20}$$

Notice that in image restoration, usually the blurring matrix A is ill conditioned and the right hand side is not known exactly but perturbed with an additive noise:

$$b = b_{exact} + e,$$

where e is some additive noise; usually it is a Gaussian noise with a zero mean. In that case, the computed solution tries to minimize the effect of the noise on the computed LS solution.

Theorem

Let $A = U\Sigma V^T$ be the singular value decomposition of the matrix A. Then the solution of the problem (17) can be expressed as follows

$$x_\lambda = x_{filt} = \sum_{i=1}^{N} \Phi_i \frac{u_i^T b}{\sigma_i} v_i \tag{21}$$

where Φ_i acts as filter and is given by

$$\Phi_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}. \tag{22}$$

proof

Let $A = U\Sigma V^T$ and then $A^T A = V\Sigma^T \Sigma V^T$. Hence the normal equation (16) can be written as

$$V(D + \lambda^2 I)V^T x = V\Sigma U^T b, \tag{23}$$

and then

$$x_\lambda = V(D + \lambda^2 I)^{-1}\Sigma U^T b, \tag{24}$$

where $D = \sigma^T \Sigma$. Therefore $D + \lambda^2 I = diag(\sigma_i + \lambda^2)$ and we get

$$x_\lambda = x_{filt} = \sum_{i=1}^{N} \Phi_i \frac{u_i^T b}{\sigma_i} v_i.$$

The solution x_{filt} can also be expressed as

$$x_{filt} = V\Phi\Sigma^{-1}U^T b = V\Sigma_{filt}^{-1}U^T b, \tag{25}$$

where $\Sigma_{filt}^{-1} = \Phi\Sigma^{-1}$ and Φ is the diagonal matrix whose elements are $\phi_1, \ldots, \phi_N$. **remark** Notice that if we choose the filter parameters as follows

$$\phi_i = \begin{cases} 1, & i = I, \ldots, k \\ 0, & i = k+1, \ldots, N, \end{cases}$$

then we get the truncated SVD and the parameter k is the optimal truncation parameter that can also be obtained by minimizing the discrete GCV function

$$G(k) = \frac{1}{(N-k)^2} \sum_{i=k+1}^{N} (u_i^T b)^2.$$

Here, the optimal truncation parameter k_{opt} is obtained by evaluating $G(k)$ for $k = 1, \ldots, N-1$ and find the index for which $G(k)$ attains its minimum.

Theorem

For the filtered solution x_{filt}, we have the following properties

1. $\|x_{filt}\|_2^2 = \sum_{i=1}^{N} \left(\phi_i \frac{u_i^T b}{\sigma_i} \right)^2 .$
2. $\|r_{filt}\|_2^2 = \|b - Ax_{filt}\|_2^2 = \sum_{i=1} \left((1 - \phi_i) u_i^T b \right)^2 .$

proof

The relation 1 comes directly from the expression (21) of the solution x_{filt} and by using the fact that the v_i's are orthonormal. For the second relation, we have

$$
\begin{aligned}
A\,x_{filt} &= U\Sigma V^T x_{filt} && (26)\\
&= U\Sigma V^T \left(\sum_{i=1}^{N} \Phi_i \frac{u_i^T b}{\sigma_i} v_i \right) && (27)\\
&= U\Sigma \left(\sum_{i=1}^{N} \Phi_i \frac{u_i^T b}{\sigma_i} V^T v_i \right) && (28)\\
&= U \left(\sum_{i=1}^{N} \Phi_i (u_i^T b)\, V^T v_i \right). && (29)
\end{aligned}
$$

Therefore

$$
b - A\,x_{filt} = \sum_{i=1}^{N} (u_i^T b) u_i - \sum_{i=1}^{N} \phi_i (u_i^T b) u_i \qquad (30)
$$

Exercices

Show that

1. $\sigma_{max} = \max_{\|x\|_2=(\|x\|_2=1} y^T A x).$
2. AA^+ is the matrix of the orthogonal projection onto $range(A)$.
3. A^+A is the matrix of the orthogonal projection onto $range(A^T)$.
4. Give an SVD of the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

For a pair of matrices, $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times m}$, there exsit two orthonormal matrices $U \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{p\times p}$ and an invertible matrix X such that

$$U^T AX = C = diag(c_1, \ldots, c_m),\ c_i \geq 0$$

and

$$V^T BX = S = diag(s_1, \ldots, s_q),\ s_i \geq 0,$$

where $q = min(p, m)$. This factorization is called the Generalized Singular Value Decomposition (GSVD).