On the Early History of the Singular Value Decomposition

Author: G. W. Stewart

SIAM Review, Vol. 35, No. 4 (Dec. 1993)

The five mathematicians behind the theory of the SVD

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

Different areas:

"Linear algebra":

- Beltrami
- Jordan
- Sylvester

Integral equations:

- Schmidt
- Weyl

They all considered the decomposition of real, square matrices. This is implied in the remainder of the lecture.

Some Prerequisites

The Frobenius norm of a matrix ($\mathbf{A} \in \mathbb{R}^{n,n}$)*:*

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\mathrm{trace}\,(\mathbf{A}\mathbf{A}^\mathbf{T})} = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

The Frobenius norm of a vector ($\mathbf{x} \in \mathbb{R}^{n}$)*:*

$$\|\mathbf{x}\|_F = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^\mathbf{T}\mathbf{x}}$$

(**NB:** This is equal to the Euclidian norm of a vector.)

Orthogonal matrix:

$$\mathbf{A}\mathbf{A}^\mathbf{T} = \mathbf{A}^\mathbf{T}\mathbf{A} = \mathbf{I}_n$$

The rows and columns of $\mathbf{A}$ are two orthonormal bases for $\mathbb{R}^n$.

The Singular Value Decomposition

Assume: $\mathbf{A} \in \mathbb{R}^{n,n}$

Its *singular value decomposition (SVD)* is given as:

$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$

Matrix Properties:

$\mathbf{U} \in \mathbb{R}^{n,n}, \mathbf{V} \in \mathbb{R}^{n,n}$

$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$

(**i.e.** **U** and **V** are orthogonal matrices)

$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$

(**i.e.** Σ is a diagonal matrix)

Matrix interpretation:

U contains the eigenvectors of (the symmetric matrix) $\mathbf{A}\mathbf{A}^{\mathrm{T}}$.

V contains the eigenvectors of (the symmetric matrix) $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

σ_i is the square root of the eigenvalue associated with the eigenvectors $\mathbf{u}_i$ and $\mathbf{v}_i$.

Eugenio Beltrami

- Author of the first publication concerning the SVD.
- Wanted it to encourage students to become familiar with bilinear forms.
- His derivation is somewhat restricted.

Goal: Reducing the bilinear form:

$$f(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} = \sum_{j=1}^{n}\sum_{i=1}^{n} x_i a_{ij} y_j$$

to a canonical form:

$$f(\mathbf{x},\mathbf{y}) = \xi^{\mathrm{T}}\Sigma\eta = \sum_{i=1}^{n} \xi_i \sigma_i \eta_i$$

Beltrami's derivation of the SVD

The bilinear form:

$f(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y},\ \mathbf{A} \in \mathbb{R}^{n,n}$

Substitutions:

$\mathbf{x} = \mathbf{U}\xi$

$\mathbf{y} = \mathbf{V}\eta$

Rewrite:

$f(\mathbf{x},\mathbf{y}) = \xi^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{V}\eta$

Substitution:

$\mathbf{S} = \mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{V}$

Rewrite:

$f(\mathbf{x},\mathbf{y}) = \xi^{\mathrm{T}}\mathbf{S}\eta$

$\mathbf{U}$ and $\mathbf{V}$ are required to be orthogonal. This gives us $n^2 - n$ degrees of freedom in their choice.
Why? An orthogonal matrix $\mathbf{A}$ can be interpreted as a solution to the n^2 equations given by $\mathbf{A}\mathbf{A}^\mathbf{T} = \mathbf{I}$, of which only $\frac{n^2-n}{2}$ are independent. We need a pair of orthogonal matrices, meaning twice the degrees of freedom.
Use these degrees of freedom to annihilate off-diagonal elements of $\mathbf{S}$, creating the diagonal matrix $\mathbf{S} = \Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$.

U and V, being orthogonal, yield:

$$\left(\mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{V}\right)\mathbf{V}^{\mathrm{T}} = \Sigma\mathbf{V}^{\mathrm{T}} \Rightarrow \mathbf{U}^{\mathrm{T}}\mathbf{A} = \Sigma\mathbf{V}^{\mathrm{T}}$$

and also:

$$\mathbf{U}\left(\mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{V}\right) = \mathbf{U}\Sigma \Rightarrow \mathbf{A}\mathbf{V} = \mathbf{U}\Sigma$$

We multiply both sides by $\mathbf{A}^{\mathrm{T}}$ in both equations:

$$\left(\mathbf{U}^{\mathrm{T}}\mathbf{A}\right)\mathbf{A}^{\mathrm{T}} = \left(\Sigma\mathbf{V}^{\mathrm{T}}\right)\mathbf{A}^{\mathrm{T}} = \Sigma\left(\mathbf{A}\mathbf{V}\right)^{\mathrm{T}} = \Sigma\left(\mathbf{U}\Sigma\right)^{\mathrm{T}} = \Sigma^{2}\mathbf{U}^{\mathrm{T}}$$

$$\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{V}\right) = \mathbf{A}^{\mathrm{T}}\left(\mathbf{U}\Sigma\right) = \left(\mathbf{U}^{\mathrm{T}}\mathbf{A}\right)^{\mathrm{T}}\Sigma = \left(\Sigma\mathbf{V}^{\mathrm{T}}\right)^{\mathrm{T}}\Sigma = \mathbf{V}\Sigma^{2}$$

This means that the diagonal elements of Σ are the roots of the equations:

$$\det\left(\mathbf{A}\mathbf{A}^{\mathbf{T}} - \sigma^2\mathbf{I}\right) = 0$$

$$\det\left(\mathbf{A}^{\mathbf{T}}\mathbf{A} - \sigma^2\mathbf{I}\right) = 0$$

because they are the square roots of the eigenvalues associated with the eigenvectors of $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ and $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ respectively.
Assuming that $\sigma_i \neq 0$ and $\sigma_i \neq \sigma_j, i \neq j$, Beltrami argues that these two functions are identical, because:

- $\det\left(\mathbf{A}\mathbf{A}^{\mathbf{T}} - \sigma_i^2\mathbf{I}\right) = \det\left(\mathbf{A}^{\mathbf{T}}\mathbf{A} - \sigma_i^2\mathbf{I}\right), i = 1, 2, \ldots, n$
- setting $\sigma = 0$ results in the common value $\det\left(\mathbf{A}\mathbf{A}^{\mathbf{T}}\right) = \det\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right) = \det^2\left(\mathbf{A}\right)$

Beltrami states that the beforementioned roots are both real and positive. The latter is shown by:

$$0 < \left\|\mathbf{x}^{\mathbf{T}}\mathbf{A}\right\|_F^2 = \mathbf{x}^{\mathbf{T}}\left(\mathbf{A}\mathbf{A}^{\mathbf{T}}\right)\mathbf{x} = \xi^{\mathbf{T}}\Sigma^2\xi$$

The matrix $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ is clearly positive definite, meaning all its eigenvalues are positive. This argument (i.e. the inequality) is only valid when $\mathbf{A}$ is nonsingular (and we assume $\mathbf{x} \neq \mathbf{0}$).

Note: In this equation, Beltrami assumes that the vector ξ exists, although he has yet to proove this.

Beltrami's Algorithm

1. Find the roots σ_i of the beforementioned equation.
2. Find the vectors $\mathbf{u}_i$, for instance by solving $\left(\mathbf{A}\mathbf{A}^{\mathbf{T}} - \sigma_i^2\right)\mathbf{u}_i = \mathbf{c}\,\forall\,\sigma_i$.
3. Find $\mathbf{V}$ through: $\mathbf{V} = \mathbf{A}^{\mathbf{T}}\mathbf{U}\Sigma^{-1}$

Summary, Beltrami's contribution

- Dervived the SVD for a real, square, nonsingular matrix with distinct eigenvalues.
- His derivation cannot handle degeneracies.
- Intentional simplification to make the derivation more accessible to students?
- Unintentional simplification from not having thought the problem through?

Camille Jordan

- Discovered the SVD a year after Beltrami, though independently.
- The SVD was “the simplest of three problems discussed in a paper”.
- Presented as a way of reducing a bilinear form to a diagonal form by orthogonal substitutions.

Jordan's Contribution

Starts with the form:

$$P = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$$

and seeks the maximum and minimum of P subject to:

$$\|\mathbf{x}\|_F^2 = \|\mathbf{y}\|_F^2 = 1$$

The maximum is given by:

$$dP = d\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{A}d\mathbf{y} = 0$$

which must be satisfied for all:

$$d\mathbf{x}^{\mathrm{T}}\mathbf{x} = 0, d\mathbf{y}^{\mathrm{T}}\mathbf{y} = 0$$

A somewhat unclear argument from Jordan (possibly) states that $\exists \sigma, \tau$ such that the maximum can be expressed by the restrictions, resulting in the equations:

$$\mathbf{A}\mathbf{y} = \sigma \mathbf{x}$$
$$\mathbf{x}^{\mathrm{T}}\mathbf{A} = \tau \mathbf{y}^{\mathrm{T}}$$

This implies that the maximum is:

$$\mathbf{x}^{\mathrm{T}} (\mathbf{A}\mathbf{y}) = \sigma \mathbf{x}^{\mathrm{T}}\mathbf{x} = \sigma$$

but also

$$\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}\right) \mathbf{y} = \tau \mathbf{y}^{\mathrm{T}}\mathbf{y} = \tau$$

i.e. $\sigma = \tau$.

The maximum is then the value of σ where the determinant of the combined systems vanishes:

$$D = \det\left(\begin{bmatrix} -\sigma\mathbf{I} & \mathbf{A} \\ \mathbf{A}^{\mathrm{T}} & -\sigma\mathbf{I} \end{bmatrix}\right)$$

The canonical form is now found by deflation. This means that the problem is reduced to finding one set of coefficients at a time. Assume we have two vectors $\mathbf{u}$ and $\mathbf{v}$ that satisfy the equations for the largest root σ_1. By making the following substitutions:

$$\hat{\mathbf{U}} \triangleq [\mathbf{u}, \mathbf{U}_*], \hat{\mathbf{V}} \triangleq [\mathbf{v}, \mathbf{V}_*], \hat{\mathbf{U}}\hat{\mathbf{U}}^{\mathbf{T}} = \hat{\mathbf{V}}\hat{\mathbf{V}}^{\mathbf{T}} = \mathbf{I}$$

$$\mathbf{x} = \hat{\mathbf{U}}\hat{\mathbf{x}}, \mathbf{y} = \hat{\mathbf{V}}\hat{\mathbf{y}}$$

we get the modified function:

$$P = \hat{\mathbf{x}}^{\mathbf{T}}\hat{\mathbf{A}}\hat{\mathbf{y}}$$

This is clearly maximized by selecting $\hat{\mathbf{x}} = \hat{\mathbf{y}} = \mathbf{e}_1$, because this means that $\mathbf{x} = \mathbf{u}$ and $\mathbf{y} = \mathbf{v}$, and we already know these are solutions. This again implies that:

$$\hat{\mathbf{A}} = \begin{bmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{bmatrix}, \mathbf{A}_1 \in \mathbb{R}^{n-1,n-1}$$

By setting $\xi_1 = \hat{x}_1$ and $\eta_1 = \hat{y}_1$, we get:

$$P = \sigma_1 \xi_1 \eta_1 + P_1$$

The last term is now a new bilinear form that can be maximized for the next root, σ_2 in a similar way. By performing this iteration, we end up with the diagonalized (canonical) form:

$$P = \xi^{\mathrm{T}} \Sigma \eta = \sum_{i=1}^{n} \sigma_i \xi_i \eta_i$$

Summary, Jordan’s Contribution

1. Elegant solution that does not suffer under the same problems as Beltrami’s.

2. This is avoided trhough the use of deflation, a technique that was not widely recognized.

Sylvester's Contribution

Begins with the bilinear form:

$$\mathbf{B} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$$

Consider the quadratic form:

$$M = \sum_i \left(\frac{dB}{dy_i}\right)^2 = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{x}$$

Assume canoncial forms:

$$M = \sum_i \lambda_i \xi_i^2, B = \sum_i \sigma_i \xi_i \eta_i \Rightarrow \lambda_i = \sigma_i^2$$

We have that $\sum (\sigma_i \xi)^2$ is orthogonally equivalent to M, implying that $\lambda_i = \sigma_i^2$

Definitions: $M \triangleq \mathbf{AA^T}$, $\mathbf{N} \triangleq \mathbf{A^TA}$
Sylvester states that the substitutions for **x** and **y** are those who diagonalize **m** and **n**, respectively.
(This is in reality only true if all singular values of **A** are distinct.)
Finding the coefficients of the x- and y-subsitutions:

$$\mathbf{X} \triangleq \mathbf{M} - \sigma^2\mathbf{I}, \xi = [M(X)_{i1} \ldots M(X)_{in}]^{\mathbf{T}}, ||\xi||_F^2 = 1$$

$$\mathbf{Y} \triangleq \mathbf{N} - \sigma^2\mathbf{I}, \eta = [M(Y)_{i1} \ldots M(Y)_{in}]^{\mathbf{T}}, \|\eta\|_F^2 = 1$$

This is done for all σ.
NB: This only works when σ is simple.

Infinitesimal Iteration: The Problem

Assume: A problem of order $n-1$ can be (and is) solved. This gives us a problem of order n (example for n = 3), and we can assume a substitution for $\mathbf{x}$:

$$A = \begin{bmatrix} a & 0 & f \\ 0 & b & g \\ f & g & c \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 & \epsilon & \eta \\ -\epsilon & 1 & \theta \\ -\eta & -\theta & 1 \end{bmatrix} \xi$$

Perform the transformation:

$$B = \mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}$$

Goal:

- Preserve zeroes: $B_{21} = B_{12} = A_{21} = A_{12} = 0$
- Create zeroes: $f = g = 0$

Infinitesimal Iteration: The Solution

Assume: η, θ, ϵ so small that any order > 1 is approximately zero.
Step 1: Select η and θ such that:

$$\frac{1}{2}\delta\left(f^2 + g^2\right) = (a - c)f\eta + (b - c)g\theta < 0$$

Step 2: Select ϵ such that:

$$\epsilon = \frac{f\theta + g\eta}{a - b}$$

(Assuming $a \neq b$.)
Sylvester claims: This process repeted infinitely will force f or g to become zero, or result in a special case where the algrithm does not apply (which can be solved another way).

Summary, Sylvester's contribution

- Sylvester did not know of earlier, similar results by Jordan and Jacobi.
- He ignores second-order terms, possibly intentionally?

Erhard Schmidt

- ... as in “Gram-Schmidt orthogonalization”
- SVD introduced in connection with integral equations with unsymmetric kernels (*not linear algebra*).
- ... or rather the infinite dimension analogue to the SVD.
- **Application:** Used the SVD to obtain optimal, low-rank approximations to an operator.

Schmidt's Contribution

Assume a kernel $A(s,t)$ which is continous and symmetric on $[a,b] \times [a,b]$. A continous, nonvanishing function satisfying:

$$\phi(s) = \lambda \int_a^b A(s,t)\phi(t)\,dt$$

is an *eigenfunction* of $A(s,t)$ corresponding to the eigenvalue λ. (**Note:** this eigenvalue is the inverse of its "ordinary" counterpart.)

Facts:

1. $A(s,t)$ has at least one eigenfunction.
2. All eigenfunctions and eigenvalues are real.
3. An eigenvalue has a finite number of corresponding, linearily independent eigenfunctions.
4. Every eigenfunction of $A(s,t)$ can be expressed as a linear combination of a finite number of members from a set of linearily independent eigenfunctions.

The eigenvalues of $A(s,t)$ satisfy the following inequality:

$$\int_a^b \int_a^b (A(s,t))^2 \, ds \, dt \geq \sum_i \frac{1}{\lambda_i}^2$$

i.e. the sequence of eigenvalues is unbounded.

Unsymmetric kernels

Assume $A(s,t)$ to be unsymmetric. A pair of adjoint eigenfunctions is any nonzero pair $u(s)$ and $v(t)$ satisfying:

$$u(s) = \lambda \int_a^b A(s,t)v(t)\,dt$$

and:

$$v(t) = \lambda \int_a^b A(s,t)u(s)\,ds$$

where λ is the eigenvalue connected to the pair of eigenfunctions. We can create two symmetric kernels:

$$\overline{A}(s,t) = \int_a^b A(s,r)A(t,r)\,dr$$

$$\underline{A}(s,t) = \int_a^b A(r,s)A(r,t)\,dr$$

Assume the eigenfunctions and eigenvalues for $\overline{A}(s,t)$ are those who satisfy:

$$u_i(s) = \lambda_i \int_a^b A(s,t)u_i(t)\,dt$$

This gives us the eigenfunctions of $\underline{A}(s,t)$ as:

$$v_i(t) = \sqrt{\lambda_i} \int_a^b A(s,t)u_i(s)\,ds$$

The previously mentioned adjoint pairs of eigenfunctions are $u_i(s), v_i(t), t = 1, 2, \ldots$

Expanding functions in series of eigenfunctions

If:

$$g(s) = \int_a^b A(s,t)h(t)\,dt$$

Then:

$$g(s) = \sum_i \frac{u_i(s)}{\lambda_i} \int_a^b h(t)v_i(t)\,dt$$

The canonical decomposition of a bilinear form

$$\int_a^b \int_a^b A(s,t)g(s)h(t)\,ds\,dt = \sum_i \frac{1}{\lambda_i} \int_a^b g(s)u_i(s)\,ds \int_a^b h(t)v_i(t)\,dt$$

The Problem:

The problem is on the form of finding the best approximation of a matrix $A \approx \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathbf{T}}$, where "best" is defined as satisfying the following minimization:

$$\min_{\mathbf{x}_i, \mathbf{y}_i, i \in [1,k]} \left\| A - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathbf{T}} \right\|_F$$

If the approximation can be written as the sum of the first k components (i.e. column vectors of **U** and **V** plus singular values) of its SVD:

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathbf{T}}$$

then the norm can be written as:

$$\|A - A_k\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{k} \sigma_i^2 = \left\| \sum_{i=k+1}^{n} \sigma_i^2 \right\|_F$$

because the Frobenius-norm squared is equal to the trace of the quadratic diagonal matrix of singular values (i.e. the diagonal matrix of eigenvalues). More complete evaluation:

Any other choice of vectors for the approximation will yield:

$$\left\| \mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathbf{T}} \right\|_F \geq \|\mathbf{A}\|_F^2 - \sum_{i=1}^{k} \sigma_i^2$$

which means that our initial approximation was optimal. We can show this. Assume that if the set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ either is orthogonal to begin with, or can be made orthogonal using the Gram-Schmidt orthogonalization process...

The norm for this alternative choice of vectors is:

$$\left\| \mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathbf{T}} \right\|_F = \operatorname{trace}\left(\left(\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathbf{T}} \right)^{\mathbf{T}} \left(\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathbf{T}} \right) \right)$$

$$= \operatorname{trace}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A} + \sum_{i=1}^{k} \left(\mathbf{y}_i - \mathbf{A}^{\mathbf{T}}\mathbf{x}_i \right) \left(\mathbf{y}_i - \mathbf{A}^{\mathbf{T}}\mathbf{x}_i \right)^{\mathbf{T}} - \sum_{i=1}^{k} \mathbf{A}^{\mathbf{T}}\mathbf{x}_i\mathbf{x}_i^{\mathbf{T}}\mathbf{A} \right)$$

Where we recognize the first term as $\|\mathbf{A}\|_F$. The second term is clearly nonnegative, meaning we can ignore it for our purpose. The last term can be recognized as a sum over $\|\mathbf{A}\mathbf{x}_i\|_F^2$. What we need to show is now that:

$$\sum_{i=1}^{k} \|\mathbf{A}\mathbf{x}_i\|_F^2 \leq \sum_{i=1}^{k} \sigma_i^2$$

Assuming $\mathbf{V} = [\mathbf{V}_1 \mathbf{V}_2], \Sigma = \operatorname{diag}(\Sigma_1 \Sigma_2)$, we can expand one term of this sum:

$$\begin{aligned} \|\mathbf{A}\mathbf{x}_i\|_F^2 = \sigma_k^2 &+ \left(\left\| \Sigma_1 \mathbf{V}_1^{\mathbf{T}} \mathbf{x}_i \right\|_F^2 - \sigma_k^2 \left\| \mathbf{V}_1^{\mathbf{T}} \mathbf{x}_i \right\|_F^2 \right) \\ &- \left(\sigma_k^2 \left\| \mathbf{V}_2^{\mathbf{T}} \mathbf{x}_i \right\|_F^2 - \left\| \Sigma_2 \mathbf{V}_2^{\mathbf{T}} \mathbf{x}_i \right\|_F^2 \right) \\ &- \sigma_k^2 \left(1 - \left\| \mathbf{v}^{\mathbf{T}} \mathbf{x}_i \right\|_F \right) \end{aligned}$$

Hermann Weyl

- Developed a general perturbation theory.
- Gave an elegan proof of the approximation theorem.

Lemma:
If $\mathbf{B}_k = \mathbf{X}\mathbf{Y}^\mathrm{T}, \mathbf{X} \in \mathbb{R}^{a,k}, \mathbf{B} \in \mathbb{R}^{b,k} \Rightarrow \mathrm{rank}\,(\mathbf{B}_k) \le k$, then:

$$\sigma_1\,(\mathbf{A} - \mathbf{B}_k) \ge \sigma_{k+1}\,(\mathbf{A})$$

where $\sigma_i(\cdot)$ means "the ith singular value of its argument". We know that:

$$\exists \mathbf{v} = \sum_{i=1}^{k+1} \gamma_i \mathbf{v}_i \text{ s.t. } \mathbf{Y}^\mathrm{T}\mathbf{v} = \mathbf{0}_k$$

where $\{\mathbf{x}_i\}_{i=1}^{k+1}$ are the first $k+1$ column vectors of the matrix $\mathbf{V}$ from the SVD of $\mathbf{A}$. We assume that:

$$\|\mathbf{v}\|_F = 1$$

or equivalently:

$$\sum_{i=1}^{k+1} \gamma_i^2 = 1$$

Proof:

$$
\begin{aligned}
\sigma_1^2\left(\mathbf{A}-\mathbf{B}\right) &\geq \mathbf{v}^{\mathbf{T}}\left(\mathbf{A}-\mathbf{B}\right)^{\mathbf{T}}\left(\mathbf{A}-\mathbf{B}\right)\mathbf{v} \\
&= \mathbf{v}^{\mathbf{T}}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)\mathbf{v} \\
&= \sum_{i=1}^{k+1} \gamma_i^2 \sigma_i^2 \\
&\geq \sigma_{k+1}^2
\end{aligned}
$$

Two Theorems

Theorem#1:

$$\mathbf{A} = \mathbf{A}' + \mathbf{A}'' \Rightarrow \sigma_{i+j-1} \leq \sigma_i' + \sigma_j''$$

Theorem#2:

$$\sigma_i \left(\mathbf{A} - \mathbf{B}_k\right) \geq \sigma_{k+i}, i = 1, 2, \ldots$$

Proof, Theorem#1:

For the case $i = j = 1$:

$$\sigma_1 = \mathbf{u}_1^{\mathrm{T}} \mathbf{A} \mathbf{v}_1 = \mathbf{u}_1^{\mathrm{T}} \mathbf{A}' \mathbf{v}_1 + \mathbf{u}_1^{\mathrm{T}} \mathbf{A}'' \mathbf{v}_1 \leq \sigma_1' + \sigma_1''$$

For the general case:

$$\begin{aligned} \sigma_i' + \sigma_j'' &= \sigma_1 \left(\mathbf{A}' - \mathbf{A}_{i-1}' \right) + \sigma_1 \left(\mathbf{A}'' - \mathbf{A}_{j-1}'' \right) \\ &\geq \sigma_1 \left(\mathbf{A} - \mathbf{A}_{i-1}' - \mathbf{A}_{j-1}'' \right) \end{aligned}$$

(and because rank $\left(\mathbf{A}_{i-1}' + \mathbf{A}_{j-1}'' \right) \leq i + j - 2$, it follows from the lemma:)

$$\geq \sigma_{i+j-1}$$

Proof, Theorem#2:

Not so much a theorem as a corollary of theorem#1. We know that:

$$\operatorname{rank}(\mathbf{B}_k) \leq k \Rightarrow \sigma_{k+1}(\mathbf{B}_k) = 0$$

Setting $j = k + 1$ in theorem#1 yields:

$$\sigma_i(\mathbf{A} - \mathbf{B}_k) = \sigma_i(\mathbf{A}')$$

$$\geq \sigma_{i+j-1} - \sigma''_j = \sigma_{k+i} - \sigma_{k+1}(\mathbf{B}_k) = \sigma_{k+i}$$

So, we have that:

$$\sigma_i(\mathbf{A} - \mathbf{B}_k) \geq \sigma_{k+i} \Rightarrow \|\mathbf{A} - \mathbf{B}_k\|_F^2 \geq \sum_{i=k+1}^{n} \sigma_i^2$$

Discussion: Weyl's Contribution

- This is **not** Weyl's original derivation of the SVD.
- Symmetric kernels can have positive and negative eigenvalues, so Weyls wrote down three inequalities (delaing with positive and negative eigenvalues and their absolute values).

Summary

- The expression "singular value" probably from integral equation theory.
- Not used consistently until the middle of the 20th century.
- SVD closely related to *spectral decompositions* of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.
- The SVD can be generalized, this derivation involves the *CS-decomposition*.
- Can be used to derive the *polar decomposition*.
- Can be used to calculate the Moore-Penrose pseudoinverse: $\mathbf{A}^+ = \mathbf{U}\Sigma^+\mathbf{V}^T$.
- Used in deriving the solution of the *Procrustes problem*.
- Used in *PCA*.
- Stable and efficient numerical algorithm by Golub and Kahan.