LA-UR -84-2336 Revised
CONF-840872--14-Rev.

NOTICE

PORTIONS OF THIS REPORT ARE ILLEGIBLE. It has been reproduced from the best available copy to permit the broadest possible availability.

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

LA-UR--84-2336-Rev.

DE84 016488

TITLE: ALGORITHM FOR SINGULAR DECOMPOSITION

AUTHOR(S): D. C. Ross

SUBMITTED TO REAL-TIME SIGNAL PROCESSING VII, SPIE 28th Annual International Technical Symposium and Instrument Exhibition, San Diego, CA, August 1984.

MASTER

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article the publisher recognizes that the U S Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so, for U S Government purposes

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U S Department of Energy

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO 836 R4
ST NO 2629 5/81

Algorithm for Singular Value Decomposition

D. C. Ross

Electronics Division
Los Alamos National Laboratory
P.O. Box 1663, Los Alamos, New Mexico 87545

Abstract

An iterative algorithm for the singular value decomposition (SVD) of a non-zero $m \times n$ matrix $\underline{M}$ is described and illustrated numerically. Derivations of the algorithm and sufficient conditions for convergence are outlined.

SVD is one of the most important procedures in digital processing of signals and images, and in applied mathematics generally. SVD provides an effective way to find the rank of a matrix, to compress data, to find the pseudo-inverse of a matrix and, in general, to calculate with rectangular and square asymmetric matrices almost as easily as with square symmetric matrices. The eigensystem of covariance matrices and other symmetric matrices of the form $\underline{A}\,\underline{\tilde{A}}$ may be found accurately from the SVD of $\underline{A}$. The theory of the SVD is well presented in a famous 1958 paper by Cornelius Lanczos in American Mathematical Monthly. The basic facts are that any non-zero matrix $\underline{M}$ of rank r may be written as the product of three factors: (1) $m \times r$ partial isometry $\underline{\tilde{U}}$, (2) positive-definite $r \times r$ diagonal matrix $\underline{D}$, (3) $r \times n$ partial isometry $\underline{V}$.

$$\underline{M} = \underline{\tilde{U}}\,\underline{D}\,\underline{V} \qquad \underline{U}\,\underline{\tilde{U}} = \underline{V}\,\underline{\tilde{V}} = \underline{I} \qquad \underline{D} = \underline{U}\,\underline{M}\,\underline{\tilde{V}}$$

The pseudo-inverse of $\underline{M}$ is then: $\underline{M}^{\dagger} = \underline{\tilde{V}}\,\underline{D}^{-1}\underline{U}$.

Each pass of the algorithm produces one set of corresponding singular elements; that is, one diagonal element of $\underline{D}$ along with the corresponding rows of $\underline{U}$ and $\underline{V}$. Each pass starts with a trial row of $\underline{U}$ and a trial row of $\underline{V}$. A trial singular value corresponding to the starting rows is then found along with measures of error. The algorithm then finds a new pair of trial singular rows to start the next iteration. The matrix $\underline{M}$ is progressively deflated and the deflated matrix is used to start each pass, but the original matrix is used for all computations within each pass to avoid unnecessary accumulation of round-off error.

****

1. An important procedure that is often used in the digital processing of signals and images, and in applied mathematics generally, is the singular value decomposition (SVD) of a non-zero $m \times n$ matrix $\underline{M}$ of rank r.

$$\underline{M} = \underline{\tilde{U}}\,\underline{D}\,\underline{V} = \sum_{k=1}^{r} |\tilde{U}_k\rangle\, \gamma_k \,\langle V_k| \qquad \underline{U}\,\underline{\tilde{U}} = \underline{I} = \underline{V}\,\underline{\tilde{V}} \tag{1}$$

$$\underline{D} = \underline{U}\,\underline{M}\,\underline{\tilde{V}} \qquad \gamma_k = \langle U_k|\,\underline{M}\,|\tilde{V}_k\rangle \qquad \underline{U}\,\underline{M} = \underline{D}\,\underline{V} \qquad \underline{M}\,\underline{\tilde{V}} = \underline{\tilde{U}}\,\underline{D} \tag{2}$$

$\underline{D}$ is a positive-definite $r \times r$ diagonal matrix and $\underline{I}$ the $r \times r$ identity matrix (Lanczos 1958). Each set

$$\left\{ |\tilde{U}_k\rangle\,,\ \gamma_k\,,\ \langle V_k| \right\} \qquad k = 1, 2, \ldots r \tag{3}$$

is called a set of corresponding singular elements of the matrix $\underline{M}$ and their product is an $m \times n$ matrix called the corresponding dyad. The $n \times m$ matrix $\underline{M}^{\dagger} = \underline{V}\,\underline{D}^{-1}\underline{U}$ is the pseudo-inverse of $\underline{M}$. The $m \times m$ matrix $\underline{M}\,\underline{M}^{\dagger} = \underline{\tilde{U}}\,\underline{U}$ and the $n \times n$ matrix $\underline{M}^{\dagger}\underline{M} = \underline{\tilde{V}}\,\underline{V}$ represent perpendicular projectors on the column space and on the row space, respectively, of the matrix $\underline{M}$.

The main purpose of this paper is to present an iterative algorithm for the SVD. The Euclidean norm is used throughout, along with the standard inner product and Dirac notation (Muggins 1964). A vector $\langle X|$ in a primal space (a space of signals for example) is represented by a row $\langle X|$. A vector $|F\rangle$ in the dual space (for example, the set of linear measuring devices appropriate for the given signals) is represented by a column $|F\rangle$. Bases are orthonormal unless otherwise indicated. Though our results may be extended easily to complex vector spaces, we confine attention in this paper to real spaces so that $\sim$ means transposition of rows and columns.

2. It is convenient to scale $\underline{M}$ initially so that the magnitude of its largest element is near one and then to scale $\underline{D}$ appropriately at the end of the algorithm. A unit trial m-row $\langle \underline{U}$ and a unit trial n-row $\langle \underline{V}$ are needed to start the algorithm. The normalized transpose of the largest column of $\underline{M}$ is a reasonable choice for $\langle \underline{U}$ along with the normalized largest row of $\underline{M}$ for $\langle \underline{V}$.

3. A trial singular value corresponding to the trial rows is then found by

$$\gamma = \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle = \langle \underline{V}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \tag{4}$$

which must be positive, by reversing the sign of one of the trial rows if necessary. The squared norms of the residual rows

$$\langle \underline{R} \triangleq \langle \underline{V}\, \widetilde{\underline{M}} - \gamma \langle \underline{U} \qquad \langle \underline{S} \triangleq \langle \underline{U}\, \underline{M} - \gamma \langle \underline{V}$$

are then found by

$$\langle \underline{R}\, \widetilde{\underline{R}} \rangle = \nu^2 - \gamma^2 \qquad \nu^2 \triangleq \langle \underline{V}\, \widetilde{\underline{M}}\, \underline{M}\, \widetilde{\underline{V}} \rangle$$
$$\langle \underline{S}\, \widetilde{\underline{S}} \rangle = \mu^2 - \gamma^2 \qquad \mu^2 \triangleq \langle \underline{U}\, \underline{M}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \qquad \tau \triangleq \langle \underline{R}\, \widetilde{\underline{R}} \rangle + \langle \underline{S}\, \widetilde{\underline{S}} \rangle \tag{5}$$

If both residuals are sufficiently small as measured by τ, then $\langle \widetilde{\underline{U}} \rangle$, γ, $\langle \underline{V} \rangle$ closely equals one of the sets of corresponding singular elements of $\underline{M}$ as can be seen from the following theorems.

Theorem 1: $\langle R | \perp \langle U |$
Proof: $\langle R | \widetilde{U} \rangle = \langle \underline{R}\, \widetilde{\underline{U}} \rangle = \langle \underline{V}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle - \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle \langle \underline{U}\, \widetilde{\underline{U}} \rangle = \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle - \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle = 0$ □

Theorem 2: $\langle S | \perp \langle V |$
Proof: $\langle S | \widetilde{V} \rangle = \langle \underline{S}\, \widetilde{\underline{V}} \rangle = \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle - \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle \langle \underline{V}\, \widetilde{\underline{V}} \rangle = \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle - \langle \underline{U}\, \underline{M}\, \widetilde{\underline{V}} \rangle = 0$ □

Theorem 3: $\gamma \langle U |$ is the $\perp$ projection of $\langle V | \widetilde{M} |$ onto $\langle U |$.
Proof: $\gamma \langle U |$ is collinear with $\langle U | \cdot | \widetilde{U} \rangle \langle U |$ is self-adjoint and idempotent.
$\gamma \langle U | = \langle U | M | \widetilde{V} \rangle \langle U | = \langle V | \widetilde{M} | \widetilde{U} \rangle \langle U | = \left[\langle V | \widetilde{M} | \right] \left[| \widetilde{U} \rangle \langle U | \right]$ □

Theorem 4: $\gamma \langle V |$ is the $\perp$ projection of $\langle U | M |$ onto $\langle V |$.
Proof: Similar to Theorem 3. □

Theorem 5: $\left\{ \langle \underline{U} = \langle \underline{U}_k \,,\ \langle \underline{V} = \langle \underline{V}_k \,,\ \gamma = \gamma_k \right\} \Rightarrow \left\{ \langle \underline{R} = \langle \underline{0} \,,\ \langle \underline{S} = \langle \underline{0} \right\}$
Proof: Substitute premises into Equation (1). □

Theorem 6: $\left\{ \langle \widetilde{\underline{U}} \rangle \,,\ \gamma \,,\ \langle \underline{V} \rangle \right.$ is one of the sets of corresponding singular elements of $\left. \underline{M} \right\} \Rightarrow$ $\begin{cases} \langle \underline{U} \text{ is an eigenrow of } \underline{M}\, \widetilde{\underline{M}} \\ \text{belonging to eigenvalue } \gamma^2; \\ \langle \underline{V} \text{ is an eigenrow of } \widetilde{\underline{M}}\, \underline{M} \\ \text{belonging to eigenvalue } \gamma^2. \end{cases}$

Theorem 7: $\left\{ \langle \underline{R} = \langle \underline{0} \,,\ \langle \underline{S} = \langle \underline{0} \right\} \Rightarrow \left\{ \langle \underline{U} = \langle \underline{U}_k \,,\ \langle \underline{V} = \langle \underline{V}_k \,,\ \gamma = \gamma_k \right\}$ □

4. If either residual differs appreciably from zero, then the algorithm is repeated with the trial rows perturbed in a direction intended to reduce the residuals. Let $\langle \underline{U}$ and $\langle \underline{V}$ be replaced with normalized versions of $\langle \underline{U} + \langle \underline{P}$ and $\langle \underline{V} + \langle \underline{Q}$ where $\langle \underline{P}$ and $\langle \underline{Q}$ are small perturbations. The residual norms squared then become

$$\langle \underline{R}\, \widetilde{\underline{R}} \rangle = \frac{\left[\langle \underline{V} + \langle \underline{Q} \right] \widetilde{\underline{M}}\, \underline{M} \left[\widetilde{\underline{V}} \rangle + \widetilde{\underline{Q}} \rangle \right]}{\left[\langle \underline{V} + \langle \underline{Q} \right] \left[\widetilde{\underline{V}} \rangle + \widetilde{\underline{Q}} \rangle \right]} - \frac{\left\{ \left[\langle \underline{U} + \langle \underline{P} \right] \underline{M} \left[\widetilde{\underline{V}} \rangle + \widetilde{\underline{Q}} \rangle \right] \right\}^2}{\left[\langle \underline{U} + \langle \underline{P} \right] \left[\widetilde{\underline{U}} \rangle + \widetilde{\underline{P}} \rangle \right] \left[\langle \underline{V} + \langle \underline{Q} \right] \left[\widetilde{\underline{V}} \rangle + \widetilde{\underline{Q}} \rangle \right]}$$
$$\langle \underline{S}\, \widetilde{\underline{S}} \rangle = \frac{\left[\langle \underline{U} + \langle \underline{P} \right] \underline{M}\, \widetilde{\underline{M}} \left[\widetilde{\underline{U}} \rangle + \widetilde{\underline{P}} \rangle \right]}{\left[\langle \underline{U} + \langle \underline{P} \right] \left[\widetilde{\underline{U}} \rangle + \widetilde{\underline{P}} \rangle \right]} - \frac{\left\{ \left[\langle \underline{U} + \langle \underline{P} \right] \underline{M} \left[\widetilde{\underline{V}} \rangle + \widetilde{\underline{Q}} \rangle \right] \right\}^2}{\left[\langle \underline{U} + \langle \underline{P} \right] \left[\widetilde{\underline{U}} \rangle + \widetilde{\underline{P}} \rangle \right] \left[\langle \underline{V} + \langle \underline{Q} \right] \left[\widetilde{\underline{V}} \rangle + \widetilde{\underline{Q}} \rangle \right]} \tag{6}$$

Setting both residuals to zero and dropping terms in $\langle \underline{P}$ and $\langle \underline{Q}$ of second and higher order yields

$$\begin{cases} \left[\nu^2 + 2 \langle \underline{Q}\, \widetilde{\underline{M}}\, \underline{M}\, \widetilde{\underline{V}} \rangle + 2\nu^2 \langle \underline{P}\, \widetilde{\underline{U}} \rangle \right] - \left[\gamma^2 + 2\gamma \langle \underline{P}\, \underline{M}\, \widetilde{\underline{V}} \rangle + 2\gamma \langle \underline{Q}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \right] = 0 \\ \left[\mu^2 + 2 \langle \underline{P}\, \underline{M}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle + 2\nu^2 \langle \underline{Q}\, \widetilde{\underline{V}} \rangle \right] - \left[\gamma^2 + 2\gamma \langle \underline{P}\, \underline{M}\, \widetilde{\underline{V}} \rangle + 2\gamma \langle \underline{Q}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \right] = 0 \end{cases} \tag{7}$$

which can be rewritten in partitioned matrix form.

$$\begin{array}{|c|c|} \hline \overset{m}{\langle \underline{P}} & \overset{n}{\langle \underline{Q}} \\ \hline \end{array} \qquad \begin{array}{|l|l|} \hline \underline{M}\, \widetilde{\underline{V}} \rangle \gamma - \widetilde{\underline{U}} \rangle \nu^2 & \underline{M}\, \widetilde{\underline{V}} \rangle \gamma - \underline{M}\, \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \\ \hline \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \gamma - \widetilde{\underline{M}}\, \underline{M}\, \widetilde{\underline{V}} \rangle & \widetilde{\underline{M}}\, \widetilde{\underline{U}} \rangle \gamma - \widetilde{\underline{V}} \rangle \mu^2 \\ \hline \end{array} \simeq \frac{1}{2} \begin{array}{|c|c|} \hline \nu^2 - \gamma^2 & \mu^2 - \gamma^2 \\ \hline \end{array} \tag{8}$$

$$\begin{array}{c|c} \gamma\langle \underline{V}\,\underline{\tilde{M}} - \nu^2\langle \underline{U} & \gamma\langle \underline{U}\,\underline{M} - \langle \underline{V}\,\underline{\tilde{M}}\,\underline{M} \\ \hline \gamma\langle \underline{V}\,\underline{\tilde{M}} - \langle \underline{U}\,\underline{M}\,\underline{\tilde{M}} & \gamma\langle \underline{U}\,\underline{M} - \mu^2\langle \underline{V} \end{array} \triangleq \underline{W} \tag{9}$$

$$\begin{array}{c|c} \nu^2 - \gamma^2 & \mu^2 - \gamma^2 \end{array} \triangleq \langle \underline{T} \tag{10}$$

With these definitions, our problem reduces to solving Equation (11) for $\langle \underline{Z}$ and then the direct sum of $\langle \underline{P}$ and $\langle \underline{Q}$ is given approximately by $\langle \underline{Z}$.

$$\langle \underline{Z}\,\underline{\tilde{W}} = \tfrac{1}{2}\langle \underline{T} \tag{11}$$

$$\begin{array}{c|c} \overset{m}{\langle \underline{P}} & \overset{n}{\langle \underline{Q}} \end{array} \approx \langle \underline{Z} \tag{12}$$

The general solution for $\langle \underline{Z}$ in Equation (11) is $(1/2)\langle \underline{T}[\underline{W}\,\underline{\tilde{W}}]^{-1}\underline{W} + \phi\langle \underline{A}$ where $\langle \underline{A}$ is any unit row orthogonal to $\langle \underline{T}\,[\underline{W}\,\underline{\tilde{W}}]^{-1}\underline{W}$. Clearly, the solution of least norm is the one with $\phi = 0$.

The method used here to locate a minimum of τ is an example of the Newton-Raphson algorithm which is analogous to the use of Newton's Method to locate the vertex of an erect parabola starting from any point on the parabola. As shown in Figure 1, doubling the computed perturbation greatly accelerates convergence, assuming in our case that $\langle \underline{P}$ and $\langle \underline{Q}$ are small. Applying this principle to our SVD algorithm, we redefine $\langle \underline{Z}$ as follows.

$$\begin{array}{c|c} \overset{m}{\langle \underline{P}} & \overset{n}{\langle \underline{Q}} \end{array} \approx \langle \underline{Z} = \langle \underline{T}\,[\underline{W}\,\underline{\tilde{W}}]^{-1}\underline{W} \tag{13}$$

After approximate perturbation rows $\langle \underline{P}$ and $\langle \underline{Q}$ are found from Equation (13) and added respectively to $\langle \underline{U}$ and $\langle \underline{V}$, the resulting sums are normalized and used as new trial rows for the next iteration of the inner loop of the algorithm.

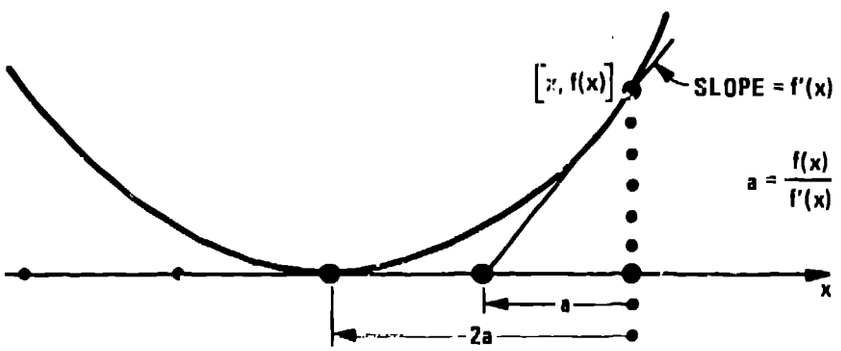

Figure 1. Acceleration of convergence.

5. The inner loop of the algorithm, defined in Sections 3 and 4, is repeated until both residual norms are sufficiently small as measured by their sum τ. The current pair of trial rows and corresponding trial singular value is then selected as a close approximation to one of the sets of corresponding singular elements of $\underline{M}$. This step ends the pass that started with the selection of initial trial singular rows and returns control to the outer loop of the algorithm for the selection of a new pair of trial singular rows to use in restarting the inner loop. To accomplish this objective, the dyad $\tilde{U}\rangle\,\gamma\,\langle \underline{V}$ is found from the results of the previous pass, subtracted from $\underline{M}$ to form a deflated matrix $\underline{N}$ which is then used to determine a new pair of starting rows $\langle \underline{U}$ and $\langle \underline{V}$. $\underline{N}$ is progressively deflated at the end of each pass. The original matrix $\underline{M}$ is used for all calculations in the inner loop of the algorithm in order to avoid unnecessary accumulation of round-off error.

A reasonable stopping rule for the outer loop of the algorithm can be based on the computation of the Euclidean norm of the deflated matrix $\underline{N}$ at the end of each pass and stopping the algorithm when the norm of $\underline{N}$ is sufficiently small. Additional measures of error in the SVD representation of $\underline{M}$ may be obtained from the Euclidean norm of any of the following matrices: $\underline{M} - \underline{\tilde{U}}\,\underline{D}\,\underline{\tilde{V}}$, $\underline{\tilde{D}} - \underline{U}\,\underline{M}\,\underline{\tilde{V}}$, $\underline{U}\,\underline{\tilde{U}} - \underline{I}$, $\underline{V}\,\underline{\tilde{V}} - \underline{I}$. These measures are relative if applied to the scaled version of $\underline{M}$ and absolute if applied to $\underline{M}$.

6. An outline of the inner loop of the SVD algorithm is given in Table 1 along with a measure of complexity of each step. For the sake of simplicity, only multiplications are counted. The total number of multiplications in each iteration in which all four tests are passed is $4mn + 12(m + n) + 10$. Tests 2-4 will be discussed later in this paper. Entry to the inner loop from the outer loop is at Step 0 and return to the outer loop follows Step 5.

Table 1 Inner Loop of Algorithm

Step	Computations	Multiplications
0	Given: m x n matrix $\underline{M}$, starting m-row $\langle \underline{U}$, starting n-row $\langle \underline{V}$	
1	$\langle \underline{U}\,\underline{M}$ and $\langle \underline{V}\,\underline{\tilde{M}}$	2mn
2	μ^2 and ν^2	m + n
3	γ, including check	m + n
4	$\langle \underline{R}\,\underline{\tilde{R}}\rangle$, $\langle \underline{S}\,\underline{\tilde{S}}\rangle$, τ	1
5	Test 1: Compare τ with stopping limit of inner loop.	
6	Test 2: Compare with previous iteration to ensure substantial reduction in τ	
7	$\gamma \langle \underline{V}\,\underline{\tilde{M}}$, $\nu^2 \langle \underline{U}$, $\langle \underline{U}\,\underline{M}\,\underline{\tilde{M}}$	2m + mn
8	$\gamma \langle \underline{U}\,\underline{M}$, $\mu^2 \langle \underline{V}$, $\langle \underline{V}\,\underline{\tilde{M}}\,\underline{M}$	2n + mn
9	Assemble $\langle \underline{T}$ and $\underline{W}$.	
10	$\underline{W}\,\underline{\tilde{W}}$ and Det $[\underline{W}\,\underline{\tilde{W}}]$	3(m + n) + -
11	Test 3: Compare determinant with lower limit.	
12	$[\underline{W}\,\underline{\tilde{W}}]^{-1}$	3
13	$\langle \underline{T}[\underline{W}\,\underline{\tilde{W}}]^{-1}$	4
14	$\langle \underline{P} \mid \langle \underline{Q}$ = $\langle \underline{T}[\underline{W}\,\underline{\tilde{W}}]^{-1}\underline{W}$	2(m + n)
15	Test 4: Compare norm of perturbation row with upper limit to prevent excessive overshoot.	m + n
16	$\langle \underline{U} + \langle \underline{P}$ and $\langle \underline{V} + \langle \underline{Q}$	
17	Normalize to find new $\langle \underline{U}$ and $\langle \underline{V}$.	2(m + n)
18	Return to Step 1.	

7. Several numerical examples of the application of the SVD algorithm to some relatively small matrices (3 x 5) have been developed and studied. In several of these examples, the starting rules normally employed with the algorithm were ignored and intentionally poor choices were made for the starting rows. In all of these examples, one iteration of the inner loop of the algorithm resulted in substantial reduction of the residual norms and of the sum of the angles between the trial vectors and the singular pair to which the trial vectors were converging. In view of the tedium of the required calculations even for small matrices, only one numerical example is presented here. In this example, the initial trial rows are determined by reasonable starting rules.

Given that $\underline{M}$ =

.640	-.640	1.000	.384	.640
.480	-.480	.016	.288	.480
-.300	.300	.240	.820	-.400

We obtain starting rows by normalizing the largest row and largest column of $\underline{M}$.

$$\langle \underline{U} = \begin{bmatrix} 1.088 & .816 & .240 \end{bmatrix} \frac{1}{\sqrt{1.9072}}$$

$$\langle \underline{V} = \begin{bmatrix} .640 & -.640 & 1.088 & .384 & .640 \end{bmatrix} \frac{1}{\sqrt{2.56}}$$

$$\langle \underline{U}\,\underline{M} = \begin{bmatrix} .735691 & -.735691 & 1.381014 & .615200 & .735691 \end{bmatrix}$$

$$\langle \underline{V}\,\underline{\tilde{M}} = \begin{bmatrix} 1.6 & 1.2 & 0.0 \end{bmatrix}$$

$\gamma = \left[\langle \underline{V}\,\underline{\tilde{M}}\right]\underline{\tilde{U}}\rangle = 1.969567$ Check: $\left[\langle \underline{U}\,\underline{M}\right]\underline{\tilde{V}}\rangle = 1.969567$

$\gamma^2 = 3.879194$ $\mu^2 = 3.909394$ $\nu^2 = 4.000000$

$\langle \underline{R}\,\underline{\tilde{R}}\rangle = \nu^2 - \gamma^2 = 0.120806$ $\langle \underline{S}\,\underline{\tilde{S}}\rangle = \mu^2 - \gamma^2 = 0.030200$

$\tau = \langle \underline{R}\,\underline{\tilde{R}}\rangle + \langle \underline{S}\,\underline{\tilde{S}}\rangle = 0.151006$ $\langle \underline{T} = \begin{bmatrix} .120806 & .030200 \end{bmatrix}$

$$\underline{W} = \begin{bmatrix} .00000 & .00000 & -.69514 & -.15101 & .15101 & .00000 & .25168 & -.15101 \\ .00000 & .00000 & -.17378 & -.11476 & .11476 & .06161 & .27342 & -.11476 \end{bmatrix}$$

$$\underline{W}\,\underline{\tilde{W}} = \begin{bmatrix} .614974 & .241605 \\ .241605 & .148263 \end{bmatrix} \qquad [\underline{W}\,\underline{\tilde{W}}]^{-1} = \frac{1}{.032805} \begin{bmatrix} .148263 & -.241605 \\ -.241605 & .614974 \end{bmatrix}$$

$$\langle \underline{T}[\underline{W}\,\underline{\tilde{W}}]^{-1}\underline{W} = \begin{bmatrix} \langle \underline{P} & \langle \underline{Q} \end{bmatrix}$$

$$= \begin{bmatrix} .00000 & .00000 & -.16869 & -.01173 & .01173 & -.01994 & -.00706 & -.01173 \end{bmatrix}$$

After adding these perturbation rows to the initial trial rows and normalizing, we have a new pair of trial rows with which to start the next iteration.

$$\langle \underline{U} = \begin{bmatrix} .79998 & .59998 & .00517 \end{bmatrix}$$

$$\langle \underline{V} = \begin{bmatrix} .40000 & -.40000 & .68000 & .23998 & .40000 \end{bmatrix}$$

In this example, the SVD of $\underline{M}$ is known exactly and we can deduce that the algorithm is converging to the following set of corresponding singular elements.

$$\langle \underline{U}_1 = \begin{bmatrix} .00 & .60 & .00 \end{bmatrix} \qquad \gamma_1 = 2.00$$

$$\langle \underline{V}_1 = \begin{bmatrix} .40 & -.40 & .68 & .24 & .40 \end{bmatrix}$$

The angle between the trial row $\langle \underline{U}$ and the singular row $\langle \underline{U}_1$ is reduced by one iteration of the inner loop of the algorithm from arccos(.984784) = 10.008° to arccos(.999972) = 0.429°, that is, by a factor of more than 20. The initial trial row $\langle \underline{V}$ was chosen equal to $\langle \underline{V}_1$ and the effect of one iteration is to change the angle between $\langle \underline{V}$ and $\langle \underline{V}_1$ from 0° to arccos(.999995) = 0.177°. Thus, even when the algorithm is converging normally, one of the two angles may increase by a small amount and this effect is more than cancelled by a large reduction in the other angle. This behavior is related to the zig-zag approach to solution called "hemstitching" which is common with algorithms of the Newton-Raphson or steepest-descent type.

In normal use of the algorithm, we must rely on τ, the total residual norm squared, to measure distance between our trial solution and the final solution. Let us carry the calculation of our numerical example through the second iteration as far as finding τ.

$$\langle \underline{U}\,\underline{M} = \begin{bmatrix} .79843 & -.79843 & 1.36120 & .48423 & .79843 \end{bmatrix}$$

$$\langle \underline{V}\,\underline{\tilde{M}} = \begin{bmatrix} 1.59999 & 1.19999 & -.00002 \end{bmatrix}$$

$\gamma = \left[\langle \underline{U}\,\underline{M}\right]\widetilde{\underline{V}}\rangle = 1.99994$ Check: $\left[\langle \underline{V}\,\widetilde{\underline{M}}\right]\underline{U}\rangle = 1.99993$

$\mu^2 = 3.99982$ $\nu^2 = 3.99994$ $\tau = .00020 + .00008 = .00028$

Note that one iteration of the inner loop has reduced both $\langle \underline{R}\,\widetilde{\underline{R}}\rangle$ and $\langle \underline{S}\,\widetilde{\underline{S}}\rangle$, and that it reduced $\sqrt{\tau}$ by a factor of more than 20.

8. There are several special cases to consider. First, we note that the null matrix $\underline{0}$ has no SVD. If the source of matrices to be factored by the algorithm can possibly produce $\underline{0}$, then the outer loop of the algorithm must first test the given matrix to ensure that it is not $\underline{0}$ before proceeding. Second, it is important that the γ found in Step 7 of the inner loop be positive, so we need to consider situations where γ could be zero. If $\langle \underline{U}$ is orthogonal to all of the singular rows $\langle \underline{U}_k$, then $\langle \underline{U}\,\underline{M} = \langle \underline{0}$ and $\gamma = 0$. Similarly, if $\langle \underline{V}$ is orthogonal to all of the singular rows $\langle \underline{V}_k$, then $\langle \underline{V}\,\widetilde{\underline{M}} = \langle \underline{0}$ and $\gamma = 0$. Another way that γ can be zero is: $\langle \underline{U} = \langle \underline{U}_i$ and $\langle \underline{V} = \langle \underline{V}_k$ where neither i nor k exceed r and $i \neq k$. Thus, the check in Step 3 must ensure that the two ways of finding γ agree within allowable round-off error and that γ exceed zero by more than allowable round-off error. If $\gamma \approx 0$, then either $\langle \underline{U}$ or $\langle \underline{V}$ must be replaced before proceeding. If $\mu^2 < \nu^2$, replace $\langle \underline{U}$; if $\mu^2 \geq \nu^2$, replace $\langle \underline{V}$. The replacement ought to be orthogonal to all previous trial rows.

Let us consider cases in which $\mathrm{Det}[\underline{W}\,\widetilde{\underline{W}}] = 0$. The cases described above that lead to $\gamma = 0$ also lead to zero for the determinant. The cure is the same as that described in the preceding paragraph. Another way that the determinant can be zero is: $\langle \underline{U} = \langle \underline{U}_k$, $\langle \underline{V} = \langle \underline{V}_k$, $k \leq r$. This case corresponds to normal convergence behavior and causes no problem, because any iteration that would lead to $\mathrm{Det}[\underline{W}\,\widetilde{\underline{W}}] = 0$ with $\gamma \neq 0$ would be terminated at Step 5 and the determinant calculation in Step 10 would not be reached. Further study may show that the check in Step 3 may eliminate the need for Test 3.

9. In order to study the behavior of the algorithm and sufficient conditions for convergence, let the trial rows be described in terms of deviations from a pair of corresponding singular rows $\langle \underline{U}_k$ and $\langle \underline{V}_k$ of the given $m \times n$ matrix $\underline{M}$ of rank r. Let $\underline{N}$ be the deflated matrix defined by: $\underline{N} = \underline{M} - \widetilde{\underline{U}_k}\rangle\,\gamma_k\,\langle \underline{V}_k$.

$$\begin{cases} \langle \underline{U} = \sqrt{1 - x^2}\,\langle \underline{U}_k + x\,\langle \underline{F} & 0 \leq x^2 < 1 \\ \langle \underline{V} = \sqrt{1 - y^2}\,\langle \underline{V}_k + y\,\langle \underline{G} & 0 \leq y^2 < 1 \end{cases}$$

The rows $\langle \underline{F}$ and $\langle \underline{G}$ may be any unit rows such that $\langle \underline{F} \perp \langle \underline{U}_k$ and $\langle \underline{G} \perp \langle \underline{V}_k$ and such that $\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle \geq 0$ for (x,y) in the first or third quadrants or $\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle \leq 0$ for (x,y) in the second or fourth quadrants.

$$\langle \underline{U}\,\underline{M} = \gamma_k\sqrt{1 - x^2}\,\langle \underline{V}_k + x\,\langle \underline{F}\,\underline{N}$$

$$\langle \underline{V}\,\widetilde{\underline{M}} = \gamma_k\sqrt{1 - y^2}\,\langle \underline{U}_k + y\,\langle \underline{G}\,\widetilde{\underline{N}}$$

$$\mu^2 = \langle \underline{U}\,\underline{M}\,\widetilde{\underline{M}}\,\widetilde{\underline{U}}\rangle = \gamma_k^2(1 - x^2) + x^2\langle \underline{F}\,\underline{N}\,\widetilde{\underline{N}}\,\widetilde{\underline{F}}\rangle$$

$$\nu^2 = \langle \underline{V}\,\widetilde{\underline{M}}\,\underline{M}\,\widetilde{\underline{V}}\rangle = \gamma_k^2(1 - y^2) + y^2\langle \underline{G}\,\widetilde{\underline{N}}\,\underline{N}\,\widetilde{\underline{G}}\rangle$$

$$\gamma = \langle \underline{U}\,\underline{M}\,\widetilde{\underline{V}}\rangle = \langle \underline{V}\,\widetilde{\underline{M}}\,\widetilde{\underline{U}}\rangle = \gamma_k\sqrt{(1 - x^2)(1 - y^2)} + xy\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle$$

At this point, we can see that the condition on the sign of $\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle$ is necessary to ensure that γ is positive for all (x,y).

$$\langle \underline{R}\,\widetilde{\underline{R}}\rangle = \nu^2 - \gamma^2 = \gamma_k^2(1 - y^2)x^2 + y^2\langle \underline{G}\,\widetilde{\underline{N}}\,\underline{N}\,\widetilde{\underline{G}}\rangle - x^2y^2\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle^2 - 2xy\sqrt{(1 - x^2)(1 - y^2)}\,\gamma_k\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle$$

$$\langle \underline{S}\,\widetilde{\underline{S}}\rangle = \mu^2 - \gamma^2 = \gamma_k^2(1 - x^2)y^2 + x^2\langle \underline{F}\,\underline{N}\,\widetilde{\underline{N}}\,\widetilde{\underline{F}}\rangle - x^2y^2\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle^2 - 2xy\sqrt{(1 - x^2)(1 - y^2)}\,\gamma_k\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle$$

$$\tau = \langle \underline{R}\,\widetilde{\underline{R}}\rangle + \langle \underline{S}\,\widetilde{\underline{S}}\rangle = \gamma_k^2(x^2 + y^2 - 2x^2y^2) + x^2\langle F\,N\,\widetilde{N}\,\widetilde{F}\rangle + y^2\langle \underline{G}\,\widetilde{\underline{N}}\,\underline{N}\,\widetilde{\underline{G}}\rangle - 2x^2y^2\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle^2 - 4xy\sqrt{(1 - x^2)(1 - y^2)}\,\gamma_k\langle \underline{F}\,\underline{N}\,\widetilde{\underline{G}}\rangle \qquad (14)$$

Note that $\tau(x,y) \geq 0$ and $\tau(0,0) = 0$.

In order to develop a geometric picture of the $\tau(x,y)$ surface, let us consider a few of its cross-sections: $x = 0$; $y = 0$; $x = y$.

$$x = 0 \Rightarrow \tau = \left[\gamma_k^2 + \langle \underline{G}\, \tilde{\underline{N}}\, \underline{N}\, \tilde{\underline{G}} \rangle \right] y^2 \tag{15}$$

$$y = 0 \Rightarrow \tau = \left[\gamma_k^2 + \langle \underline{F}\, \underline{N}\, \tilde{\underline{N}}\, \tilde{\underline{F}} \rangle \right] x^2 \tag{16}$$

$$x = y \Rightarrow \tau = 2\gamma_k^2(x^2 - x^4) + x^2 \langle \underline{F}\, \underline{N}\, \tilde{\underline{N}}\, \tilde{\underline{F}} \rangle - 2x^4 \langle \underline{F}\, \underline{N}\, \tilde{\underline{G}} \rangle^2 + x^2 \langle \underline{G}\, \tilde{\underline{N}}\, \underline{N}\, \tilde{\underline{G}} \rangle - 4x^2(1 - x^2)\gamma_k \langle \underline{F}\, \underline{N}\, \tilde{\underline{G}} \rangle$$

$$= x^2(2C^2 + D^2) - x^4(2C^2) \quad \text{where } C^2 = \left(\gamma_k - \langle \underline{F}\, \underline{N}\, \tilde{\underline{G}} \rangle\right)^2 + \langle \underline{G}\, \tilde{\underline{N}}(\underline{I} - \tilde{\underline{F}} \rangle \langle \underline{F})^2 \underline{N}\, \tilde{\underline{G}} \rangle \tag{17}$$
$$D^2 = \langle \underline{F}\, \underline{N}(\underline{I} - \tilde{\underline{G}} \rangle \langle \underline{G})^2 \underline{N}\, \tilde{\underline{F}} \rangle$$

Note that $\tilde{\underline{F}} \rangle \langle \underline{F}$, $\tilde{\underline{G}} \rangle \langle \underline{G}$, $\underline{I} - \tilde{\underline{F}} \rangle \langle \underline{F}$, and $\underline{I} - \tilde{\underline{G}} \rangle \langle \underline{G}$ represent perpendicular projectors.

We see that Equations (15,16) represent parabolas and Equation (17) represents a quartic that approaches a parabola for small x^2 as shown in Figure 2. Our algorithm was designed to work well with paraboloids, so we expect convergence for almost any y^2 when x^2 is small and for almost any x^2 when y^2 is small. To obtain some initial ideas of conditions for convergence when neither x^2 nor y^2 are small, let us consider Equation (17) and its first derivative.

$$d\tau/dx = 2x[(2C^2 + D^2) - 2x^2(2C^2)] \tag{18}$$

The stationary points of $\tau(x,x)$ are at $x = 0$ and at $x = \pm x_M$ where

$$x_M^2 = \frac{2C^2 + D^2}{4C^2} \geq \frac{1}{2} = (\sin 45°)^2 \tag{19}$$

and $\tau(x_M, x_m) = (2C^2 + D^2)^2/8C^2$ while $\tau(1,1) = D^2$. The stationary point at $x = 0$ is a global minimum where $\tau(0,0) = 0$. The other stationary points are local maxima of $\tau(x,x)$ which do not occur if $2C^2 \leq D^2$. Note that the smallest x_M^2 is 1/2 corresponding to the case where both trial vectors are 45° from one of the vectors in a pair of singular vectors.

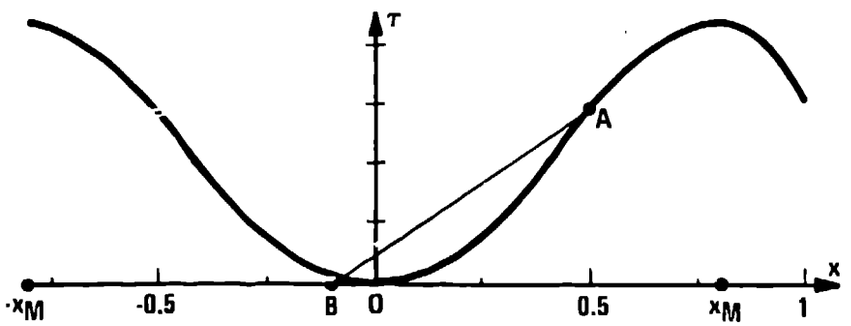

Figure 2. $\tau(x,x)$ for $D^2 < 2C^2$.

A somewhat oversimplified view of the behavior of the algorithm can be seen by studying Figure 2 where point A represents the initial choice of trial rows. One iteration of the inner loop of the algorithm would yield a new trial represented by the point B which would lie at or near the origin if $\tau(x,y)$ were paraboloidal or nearly paraboloidal. Because of the quartic nature of the $\tau(x,y)$ surface, the point B will lie beyond the point nearest to the origin along the direction of the perturbation vector. That is, the algorithm will overshoot. The amount of overshoot will be small for small x_A^2 but will become very large for x_A^2 near x_M^2. However, as long as $x_A^2 < x_M^2$, the algorithm will make changes in the correct direction. This observation suggests that the length of each perturbation vector obtained, at least in the first iteration in each pass, be checked and reduced, if necessary, to some preset maximum allowable length. Thus, Test 4 is included at Step 15 of the inner loop just before the vector addition and renormalization steps. Calculations have been made on the example of Figure 2 and on several other examples for two values of the maximum length of the perturbation vector: $1/\sqrt{2}$ and 1/2. Best results have been obtained with this limit set at $1/\sqrt{2}$.

If x_A^2 exceeds x_M^2, the algorithm may make changes in the wrong direction. Apparently, many, if not all, of such cases may be detected by noting that the new pair of trial rows

does not yield a substantial decrease in $\mathcal{T}$. It appears that such cases may be treated by reversing the direction of the calculated perturbation via Test 2 at Step 6 of the inner loop. Further study is needed to set the limiting reduction ratio used in Test 2 and to determine whether Test 2 may be eliminated by selection of good starting rules.

The choice of the limit on the norm of the perturbation vector also requires further study and computer testing. In order to speed convergence in most cases, it appears to be desirable to set this limit at about $1/\sqrt{2}$, that is, $\langle \underline{P}\,\underline{P} \rangle + \langle \underline{Q}\,\underline{Q} \rangle \leq 1/2$.

One result of the analysis of the algorithm to date is Theorem 8 which leads to the following conjecture: The algorithm converges for all starting points inside the square S of side $\sqrt{2}$ centered on the origin of the (x,y) plane, that is, for

$$x^2 < 1/2 > y^2$$

An alternative statement of the conjecture is: The algorithm converges if each trial vector is less than 45^0 from the corresponding vector in a singular pair.

Theorem 8: The only stationary point of $\mathcal{T}(x,y)$ inside the square S is the global minimum at the origin.

Proof: Equate to zero the partial derivatives of $\mathcal{T}$ with respect to x and y.
Transfer radicals to the opposite side of the equals signs.
Multiply the resulting equations together to form one equation: $f(x,y) = 0$.
Note that $f(x,y) = 0$ at the origin and at no other points inside S. □

Study of the effect of the ratio $D^2/2C^2$ on the $\mathcal{T}(x,y)$ surface leads to the further conjecture that the algorithm converges for all (x,y) if the following conditions are satisfied.

$$\langle \underline{F}\,\underline{N}\,\underline{\tilde{N}}\,\underline{\tilde{F}} \rangle > \left[\gamma_k^2 + \langle \underline{F}\,\underline{N}\,\underline{\tilde{G}} \rangle^2\right] < \langle \underline{G}\,\underline{\tilde{N}}\,\underline{N}\,\underline{\tilde{G}} \rangle \qquad (20)$$

These conditions are unlikely to be satisfied on the first pass of the algorithm but are usually satisfied on subsequent passes associated with calculation of the smaller singular values.

The algorithm appears to be particularly applicable to the updating of the SVD of a matrix of observations as each new set of observations is obtained and to the checking and refinement of results obtained from other SVD algorithms. Further study and computer testing is needed to put the algorithm in final form and to determine its rate of convergence under various conditions. In the limited number of numerical examples developed to date, the convergence of the algorithm is surprisingly fast.

References

1. Cornelius Lanczos
"Linear Systems in Self-Adjoint Form"
American Mathematical Monthly
v 65 (1958), pp 665-679

2. William H. Huggins
"An Algebra for Signal Representation"
Yearbook of the Society for General Systems Research
v VIII (1963), pp 129-143