Computation of the Generalized Singular Value Decomposition Using Mesh-Connected Processors

Richard P. Brent
Centre for Mathematical Analysis
Australian National University
Canberra, A.C.T. 2601
Australia

Franklin T. Luk
Department of Computer Science
Cornell University
Ithaca, New York 14853
U.S.A.

Charles Van Loan
Department of Computer Science
Cornell University
Ithaca, New York 14853
U.S.A.

ABSTRACT

This paper concerns the systolic array computation of the generalized singular value decomposition. Numerical algorithms for both one- and two-dimensional systolic architectures are discussed.

Keywords and Phrases: Systolic arrays, QR-decomposition, singular value decomposition, generalized singular value decomposition, real-time computation, VLSI.

Introduction

Two of the most important ways to decompose a given matrix $A \in \mathbb{R}^{m \times n}$ ($m \geq n$) are the Q-R factorization:

$$A = QR,$$  \hspace{1cm} (1)

where $Q \in \mathbb{R}^{m \times n}$ has orthonormal columns and $R \in \mathbb{R}^{n \times n}$ is upper triangular, and the singular value decomposition (SVD):

$$A = U \Sigma V^T,$$  \hspace{1cm} (2)

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{n \times n} = \text{diag}(\sigma_1, \ldots, \sigma_n)$, with $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$ and $r = \text{rank}(A)$. See Golub and Van Loan \(^1\) and Dongarra et al.\(^2\) for details.

The systolic array computation of these decompositions has recently attracted a great deal of attention. QR-arrays are discussed in Bojanczyk, Brent and Kung \(^3\), Gentleman and Kung \(^4\) and Heller and Ipsen \(^5\); SVD arrays in Brent and Luk \(^6\), Brent, Luk and Van Loan \(^7\), Finn, Luk and Pottle \(^8\), Heller and Ipsen \(^9\) and Schreiber \(^10\). In this paper we discuss the systolic array computation of the generalized singular value decomposition (GSVD). It has been suggested (see Speiser and Whitehouse \(^11\)) that real-time computation of this decomposition is important in modern signal processing.

The GSVD amounts to a simultaneous diagonalization of a pair of matrices $A \in \mathbb{R}^{m \times n}$ ($m \geq n$) and $B \in \mathbb{R}^{p \times n}$:

$$
\begin{bmatrix}
U^T & 0 \\
0 & V^T
\end{bmatrix}
\begin{bmatrix}
A \\
B
\end{bmatrix} =
\begin{bmatrix}
D_A \\
D_B
\end{bmatrix},
$$

\hspace{1cm} (3)

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal, $X \in \mathbb{R}^{n \times n}$ is nonsingular, $D_A = \text{diag}(\alpha_1, \ldots, \alpha_s) \geq 0$, $D_B = \text{diag}(\beta_1, \ldots, \beta_s) \geq 0$ and $q = \min\{p,n\}$. We call $(\alpha_i, \beta_i)$ a singular value pair of $A$ and $B$. Note that when $B$ is square and nonsingular, the singular values of $AB^{-1}$ are $\alpha_i/\beta_i$, for $i = 1, \cdots, n$, and when $B = I_n$ these ratios are just the singular values of $A$. For a general $B$, we may refer to $\alpha_i/\beta_i$ as the generalized singular values of $A$ with respect to $B$, although some of these values may be infinite or undefined. The use of singular
value pairs, however, avoids the distinction between $A$ and $B$. The GSVD was first introduced by Van Loan \cite{12} and further discussed in Paige and Saunders \cite{13}. The decomposition is useful for certain constrained and generalized least squares problems (see Golub and Van Loan \cite{1}).

We briefly discuss the computation of the GSVD. Suppose that the null spaces of $A$ and $B$ intersect trivially, i.e., $N(A) \cap N(B) = \{0\}$. Let

$$E = \begin{pmatrix} A \\ B \end{pmatrix},$$

and compute its QR-factorization:

$$E = QR.$$  

By assumption, the matrix $R$ is nonsingular. Partition $Q$ in the form

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix},$$

such that $Q_1 \in \mathbb{R}^{m \times n}$ and $Q_2 \in \mathbb{R}^{p \times n}$. Then we can find orthogonal matrices $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{p \times p}$ and $W \in \mathbb{R}^{n \times n}$ such that

$$\begin{pmatrix} UT & 0 \\ 0 & V^T \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} W = \begin{pmatrix} C \\ S \end{pmatrix},$$

where $C = \text{diag}(c_1, \ldots, c_\ell) \geq 0$, $S = \text{diag}(s_1, \ldots, s_\ell) \geq 0$ and $C^T C + S^T S = I_n$. The decomposition (5) is referred to as the CS-decomposition. It says that the SVD's of the blocks in a partitioned orthonormal matrix are related. The CS-decomposition first appears in Stewart \cite{14}, where it is pointed out that the result is implicit in Davis and Kahan \cite{15}. Van Loan \cite{16} shows how this decomposition can be used to analyze certain important problems involving orthogonal matrices. If we set

$$D_A = C, \ D_B = S \text{ and } X = R^{-1} W,$$

we obtain a GSVD of $A$ and $B$.

If the null spaces of $A$ and $B$ intersect nontrivially, or nearly so, then it is advisable to compute an SVD of the matrix $E$:

$$\begin{pmatrix} A \\ B \end{pmatrix} = Q \Sigma Z^T = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Z_r^T \\ Z_\omega^T \end{pmatrix}.$$
Here, $\Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$, $Q_{11} \in \mathbb{R}^{m \times r}$, $Q_{21} \in \mathbb{R}^{p \times r}$, $Z_t \in \mathbb{R}^{n \times r}$, and $r = \text{rank}(E)$.

Let

$$\begin{pmatrix} \bar{U}^T & 0 \\ 0 & \bar{V}^T \end{pmatrix} \begin{pmatrix} Q_{11} \\ Q_{21} \end{pmatrix} \bar{W} = \begin{pmatrix} \tilde{\mathcal{C}} \\ \tilde{\mathcal{S}} \end{pmatrix}$$

be a CS-decomposition of $Q_{11}$ and $Q_{21}$. Then

$$A = Q_{11} \Sigma_r Z_t^T = \bar{U}(\tilde{\mathcal{C}}, 0) \begin{pmatrix} \bar{W}^T \Sigma_r & 0 \\ 0 & I_{n-r} \end{pmatrix} Z_t^T$$

and

$$B = Q_{21} \Sigma_r Z_t^T = \bar{V}(\tilde{\mathcal{S}}, 0) \begin{pmatrix} \bar{W}^T \Sigma_r & 0 \\ 0 & I_{n-r} \end{pmatrix} Z_t^T.$$

A GSVD results by setting $D_A = (\tilde{\mathcal{C}}, 0)$, $D_B = (\tilde{\mathcal{S}}, 0)$ and $X = Z \begin{pmatrix} \Sigma_r^{-1} \bar{W} & 0 \\ 0 & I_{n-r} \end{pmatrix}$.

From the above discussion, we see that the key problem confronting us is the systolic array calculation of the CS-decomposition.
Stewart's algorithm

We desire a C3-decomposition of a partitioned orthonormal matrix

\[ Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}, \]

where \( Q_1 \in \mathbb{R}^{m \times n} (m \geq n) \) and \( Q_2 \in \mathbb{R}^{r \times n} \). First, an SVD of \( Q_1 \) may be determined via standard techniques:

\[ U^T Q_1 W = C. \]

Since

\[ Q_1^T Q_1 + Q_2^T Q_2 = I_n, \]

the nonnull columns of the matrix

\[ \overline{Q}_2 = Q_2 W \]

are orthogonal. Suppose that \( \overline{Q}_2 \) has rank \( r \) and that its first \( r \) columns are nonzero. These columns can be normalized to yield

\[ \overline{Q}_2 = (V_1, 0) \begin{pmatrix} S_1 \\ 0 \end{pmatrix}, \]

where \( V_1 \in \mathbb{R}^{r \times r} \) is orthonormal and \( S_1 = \text{diag}(s_1, \ldots, s_r) \geq 0 \). Let \( V = (V_1, V_2) \in \mathbb{R}^{r \times p} \) be an orthogonal matrix. Then we have

\[ V^T Q_2 W = \begin{pmatrix} S_1 \\ 0 \end{pmatrix} = S, \]

an SVD of \( Q_2 \).

Unfortunately, the preceding procedure is numerically unsound. Troubles may arise when some columns of \( \overline{Q}_2 \) have euclidean lengths less than \( \epsilon^{1/2} \), where \( \epsilon \) denotes the machine precision. Numerical examples are given in Stewart.\(^{17,18}\). To simplify our presentation, let us assume from here on that \( Q_2 \) has full column rank, i.e.,

\[ \text{rank}(Q_2) = n \leq p. \tag{6} \]

Stewart\(^ {17,18}\) presents the following cleanup procedure:
1. Determine an orthogonal matrix $J$ such that the columns of $Q_2 J$ can be normalized to give a matrix $V$ whose columns are then orthogonal to working accuracy.

2. Determine an orthogonal matrix $K$ such that $K^T CJ$ is diagonal.

If we replace $W$ by $WJ$ and $U$ by $UK$, and normalize the columns of $Q_2 J$ to get $V$, we obtain

$$
\begin{pmatrix}
U^T Q_1 \\
V^T Q_2
\end{pmatrix} V = \begin{pmatrix}
K^T CJ \\
V^T Q_2
\end{pmatrix}.
$$

Since $K^T CJ$ and $V^T Q_2$ are diagonal, we have computed a CS-decomposition of $Q$.

Stewart chooses $J$ and $K$ by working with the matrix

$$F = Q_1^T Q_2,$$

and using the Jacobi method, as implemented by Rutishauser, to determine $J$ such that $J^T F J$ is diagonal. Stewart then shows why we may take $K = J$, so long as certain unnecessary rotations are not performed in the Jacobi method. Specifically, a Jacobi rotation $R_{ij}$ in the $(i,j)$-plane will be suppressed if

$$c_i + c_j < r,$$

where $c_i$ and $c_j$ are the $i$-th and $j$-th diagonal elements of $C$ and $r$ is some preset tolerance. A value of $r = 0.7$ is proposed, for if

$$c_i + c_j = 0.7,$$

then the error made in accepting $R_{ij}^T CR_{ij}$ as a diagonal matrix is roughly equal to the error made in accepting the $i$-th and $j$-th columns of $Q_2$ as orthogonal. Finally, Stewart proves that, because of the suppression, the diagonal entries of $C$ are effectively unchanged in the passage to $J^T CJ$. 
Linear arrays

Brent and Luk \(^6\) present a systolic array of \(O(n)\) linearly-connected processors for computing an SVD of an \(l \times n\) matrix, say \(M\). Their array implements a one-sided orthogonalization method due to Hestenes \(^21\). The idea is to determine an orthogonal matrix \(V\) such that the non-null columns of \(MV\) are mutually orthogonal. These columns are normalized to give a matrix \(\tilde{U}\) with orthonormal columns and a nonnegative diagonal matrix \(\Sigma\). We have thus determined an SVD of \(M\):

\[
M = \tilde{U} \Sigma V^T.
\]

The orthogonal transformation \(V\) is constructed as a sequence of plane rotations; the rotations are generated to orthogonalize column pairs of \(M\). Hence the Hestenes method is mathematically equivalent to the serial Jacobi procedure for finding an eigenvalue decomposition of \(M^T M\). For the sake of parallel computing, Brent and Luk discard the classical scheme of rotating column pairs in the order:

\((1,2),(1,3), \ldots, (1,n),(2,3), \ldots, (2,n),(3,4), \ldots, (3,n), \ldots, (n-1,n)\),

in preference for a new ordering that allows \(\lceil n/2 \rceil\) simultaneous rotations. Their new ordering is amply illustrated by the \(n = 8\) case:

\[
(p,q) = (1,2), (3,4), (5,6), (7,8),
(1,4), (2,6), (3,8), (5,7),
(1,6), (4,8), (2,7), (3,5),
(1,8), (6,7), (4,5), (2,3),
(1,7), (8,5), (6,3), (4,2),
(1,5), (7,3), (8,2), (6,4),
(1,3), (5,2), (7,4), (8,6).
\]

Note that the rotation pairs associated with each "row" of the above can be calculated concurrently. Brent and Luk \(^22\) conjecture that this Jacobi approach would require \(O(\log n)\) sweeps for convergence. Their algorithm for computing an SVD of an \(l \times n\) matrix thus requires \(O(n \log n)\) time.

We may compute the GSVD using the linear systolic array of Brent and Luk \(^6\) as follows:
1. Compute an SVD of
\[
\begin{pmatrix} A \\ B \end{pmatrix} = Q \Sigma Z^T = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \Sigma Z^T
\]

2. Compute an SVD of
\[Q_1 = UCW^T,\]
and apply the appropriate transformations to get
\[\overline{Q}_2 = Q_2 W.\]

3. Initiate Stewart’s algorithm. (We note that the Jacobi procedure applied to \(F\) is equivalent to the Hestenes method applied to \(\overline{Q}_2.\))

Our procedure requires time \(O((m+p)n \log n)\).
Quadratic arrays

An array for computing an $l \times l$ matrix is proposed in Brent, Luk and Van Loan. It requires $O(l^2)$ processors and $O(l \log l)$ time to execute. The array implements a two-sided Jacobi procedure that is detailed in Forsythe and Henrici. In essence, the off-diagonal elements of the given matrix are reduced to zero by a sequence of plane rotations that are determined by solving carefully chosen two-by-two SVD's. The algorithm is very similar to the classical Jacobi algorithm for the symmetric eigenvalue problem, for which a systolic array has been proposed by Brent and Luk. Briefly, the new ordering of Brent and Luk, illustrated in the previous section, is extended in an obvious manner to allow the simultaneous computations of $\lfloor l/2 \rfloor$ two-by-two SVD's. In addition, a staggering of computations allows the execution of the equivalence transformations without requiring that the rotation parameters be broadcasted. For details see Brent et al.

If we want an SVD of a matrix $M \in \mathbb{R}^{m \times n}$, where $m, n \leq l$, we feed the matrix

$$
\hat{M} = \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{l \times l}
$$

into the array of Brent, Luk and Van Loan. An SVD:

$$
\hat{M} = \begin{pmatrix} U & 0 \\ 0 & I_{l-m} \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V & 0 \\ 0 & I_{l-n} \end{pmatrix}^T
$$

will emerge, and we see that $M = UV^T$, as desired.

Let us point out how we can compute a nonsquare CS-decomposition using a "square" $l$-by-$l$ hardware. Suppose that $Q_1 \in \mathbb{R}^{m \times l}$, $Q_2 \in \mathbb{R}^{p \times n}$, $Q_1^T Q_1 + Q_2^T Q_2 = I_n$ and $l \geq m, n, p$. If

$$
\hat{Q}_1 = \begin{pmatrix} Q_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{l \times l} \quad \text{and} \quad \hat{Q}_2 = \begin{pmatrix} Q_2 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{l \times l},
$$

then

$$
\hat{Q}_1^T \hat{Q}_1 + \hat{Q}_2^T \hat{Q}_2 = \begin{pmatrix} I_n \ 0 \\ 0 \ 0 \end{pmatrix}.
$$

It is not hard to show that there exist orthogonal matrices of the form
\[
\hat{U}_1 = \begin{pmatrix} U_1 & 0 \\ 0 & I_{n-m} \end{pmatrix}, \quad \hat{U}_2 = \begin{pmatrix} U_2 & 0 \\ 0 & I_{r-s} \end{pmatrix} \quad \text{and} \quad \hat{W} = \begin{pmatrix} W & 0 \\ 0 & I_{n-s} \end{pmatrix},
\]

such that
\[
\hat{U}_1^T \hat{Q}_1 \hat{W} = \begin{pmatrix} C \\ 0 \end{pmatrix}, \quad \hat{U}_2^T \hat{Q}_2 \hat{W} = \begin{pmatrix} S \\ 0 \end{pmatrix} \quad \text{and} \quad C^T C + S^T S = I_n.
\]

Thus, applying Stewart’s algorithm to \( \hat{Q}_1 \) and \( \hat{Q}_2 \) will produce a CS-decomposition of \( Q_1 \) and \( Q_2 \).

We now outline how we may compute a GSVD of \( A \) and \( B \) using a QR-array, a matrix-matrix multiply array (see, e.g., Kung and Leiserson\( ^{34} \)) and an SVD array:

1. Compute a QR-decomposition of
   \[
   \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} R.
   \]

2. Compute \( T = Q_2^T Q_2 \).

3. Set \( \hat{Q}_1 = \begin{pmatrix} Q_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{Q}_2 = \begin{pmatrix} Q_2 & 0 \\ 0 & 0 \end{pmatrix} \) and \( \hat{T} = \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix}, \)
   so that they are all \( n \times n \) matrices.

4. Compute an SVD of \( \hat{Q}_1 \) and apply the appropriate transformations to \( \hat{Q}_2 \) and \( \hat{T} \).

5. Initiate Stewart’s procedure.

The complete procedure requires time \( O(\log n) \).

Acknowledgements

The work of the second author was supported in part by the National Science Foundation under grant MCS-8213718, and in part by the Centre for Mathematical Analysis, Australian National University.
References


