Singular Value Decomposition – A Primer

Sonia Leach
Department of Computer Science
Brown University
Providence, RI 02912
DRAFT VERSION

1 Introduction

The singular value decomposition (SVD) is a powerful technique in many matrix computations and analyses. Using the SVD of a matrix in computations, rather than the original matrix, has the advantage of being more robust to numerical error. Additionally, the SVD exposes the geometric structure of a matrix, an important aspect of many matrix calculations. A matrix can be described as a tranformation from one vector space to another. The components of the SVD quantify the resulting change between the underlying geometry of those vector spaces.

The SVD is employed in a variety of applications, from least-squares problems to solving systems of linear equations. Each of these applications exploit key properties of the SVD – its relation to the rank of a matrix and its ability to approximate matrices of a given rank. Many fundamental aspects of linear algebra rely on determining the rank of a matrix, making the SVD an important and widely-used technique.

This primer serves as a short introduction to the SVD and its applications. More comprehensive coverage can be found in numerous references, such as [GVL83, Dep88, Vac91].

Organization of the paper is as follows. Section 2 introduces the definition of the SVD, followed by a discussion of the properties of the components of the SVD. Section 3 explores further properties of the SVD and provides a geometric interpretation of the singular values. Section 4 lists a number of interesting applications and Section 5 concludes the paper with a discussion of the advantages and disadvantages of using the SVD.

2 Definition of the SVD

In this section, we assume a familiarity with the basic terminology of linear algebra, and refer the reader to [And86] for a more complete coverage. We restrict our attention to matrices of real numbers and refer the reader to [DD88] for a discussion of the SVD using complex numbers. This presentation is largely adapted from [FMM77].

Using the superscript T to denote the transpose of a vector or matrix, we say two vectors x and y are **orthogonal** if $x^T y = 0$. In two or three dimensional space, this simply means that the vectors are perpendicular. Let A be a square matrix such that its columns are

mutually orthogonal vectors of length 1, i.e. $x^T x = 1$. Then A is an **orthogonal matrix** and $A^T A = I$, the identity matrix. To simplify the notation, assume that a matrix A has at least as many rows as columns ($M \geq N$).

A **singular value decomposition** of an $M \times N$ matrix A is any factorization of the form

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

where U is an $M \times M$ orthogonal matrix, V is an $N \times N$ orthogonal matrix, and Σ is an $M \times N$ diagonal matrix with $s_{ij} = 0$ if $i \neq j$ and $s_{ii} = s_i \geq 0$. Furthermore, it can be shown that there exist (non-unique) matrices U and V such that $s_1 \geq s_2 \geq \cdots \geq s_N \geq 0$ [GVL83]. Henceforth we will assume the SVD has such a property. The quantities s_i are called the **singular values** of A, and the columns of U and V are called the left and right **singular vectors**, respectively.

For example, the matrix $A = \begin{pmatrix} 0.96 & 1.72 \\ 2.28 & 0.96 \end{pmatrix}$ has the SVD

$$A = U\Sigma V^T = \begin{pmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{pmatrix}^T$$

We see that the columns of U and V are unit length since $(0.6)^2 + (0.8)^2 = 1$, and a simple calculation of dot products will show them to be mutually orthogonal.

From the components of the SVD, we can determine many properties of the original matrix. The **null space** of a matrix A is the set of x for which $Ax = 0$, and the **range** of A is the set of b for which $Ax = b$ has a solution for x. Let u_j and v_j be the columns of U and V respectively. Then the decomposition of $A = U\Sigma V^T$ can be written as

$$Av_j = s_j u_j, \quad j = 1, 2, \ldots, N.$$

If $s_j = 0$, then $Av_j = 0$ and v_j is in the null space of A, whereas if $s_j \neq 0$, then u_j is in the range of A. Consequently, we can construct bases for various vector subspaces defined by A. A set of vectors $v_1, v_2, \ldots, v_k$ in a vector space V is said to form a **basis** for V if every vector x in V can be expressed as a linear combination of them in exactly one way. Let V_0 be the set of columns v_j for which $s_j = 0$, and let V_1 be the remaining columns v_j. Similarly, let U_1 be the set of columns u_j for which $s_j \neq 0$, and let U_0 be the remaining columns u_j, including those with $j > n$. Thus, if k is the number of non-zero singular values, there are k columns in V_0, $N - k$ columns in V_1 and U_1, and $M - N + k$ columns in U_0. Each of these sets forms a basis for the vector subspaces of A.

1. V_0 is an orthonormal basis for $Nullspace(A)$.
2. V_1 is an orthonormal basis for the orthogonal complement of $Nullspace(A)$.
3. U_1 is an orthonormal basis for $Range(A)$.
4. U_0 is an orthonormal basis for the orthogonal complement of $Range(A)$.

As we shall see in the next two sections, the singular values of A can be used in many other ways to determine properties of A, as well as to partition the M-dimensional vector space (of the mapping defined by A) into dominant and sub-dominant subspaces.

3 Properties of the SVD

3.1 SVD and Matrix Norms

Often when speaking about vectors and matrices, we are interested in the lengths of the vectors and the resulting length of a vector when multiplied by a matrix. A familiar concept of length in two dimensions is the Euclidean distance from the origin to the point specified by the coordinates of the vector $\{x_1, x_2\}$. This distance is calculated by the formula $(x_1^2 + x_2^2)^{\frac{1}{2}}$. In the general case of N dimensions, the length (or **norm**) of a vector x is defined by

$$\| x \| = (x_1^2 + x_2^2 + \ldots + x_N^2)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}}.$$

When a vector x is multiplied by a matrix A, the length of the resulting vector Ax changes according to the matrix A. If A is orthogonal, the length is preserved. Otherwise, the quantity $\frac{\|Ax\|}{\|x\|}$ measures how much A *stretches* x. Thus, calculating the norm of a *matrix* intuitively means finding the maximum stretch factor. If the SVD of a matrix is given, this computation is simplified.

The Euclidean norm of a matrix, sometimes referred to as the L_2 norm, is defined as follows. Let x be an N dimensional vector, and A be an $M \times N$ matrix, then

$$\| A \|_E = \max_{\|x\|=1} \left\{ \frac{\| Ax \|}{\| x \|} \right\}$$

An alternative norm for A is the Frobenius norm, which is the Euclidean norm of a vector constructed by stacking the columns of A in one $M * N$ vector. The Frobenius norm is then

$$\| A \|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^2 \right)^{\frac{1}{2}}.$$

Given the SVD of a matrix A, these norms can easily be computed. Proofs of the following facts are given in [GVL83, DD88]. Let $U\Sigma V^T$ be the SVD of $M \times N$ matrix A, where $\{s_1, s_2, ..., s_k\}, k \leq N$ are the non-zero singular values in Σ. Then

$$\| A \|_E = s_1$$

$$\| A \|_F = \left(\sum_{l=1}^{k} s_k^2 \right)^{\frac{1}{2}}$$

To return to an earlier notion, we mentioned that multiplying a vector x by a matrix A effectively stretches the vector. This geometric interpretation can be viewed more clearly in terms of the singular values of A. The set of vectors x of length N for which $\| x \| = 1$ defines a unit circle. Multiplication of these vectors by the $M \times N$ matrix A results in a set of M-dimensional vectors $b = Ax$ with varying lengths. Geometrically, this set defines a k-dimensional ellipsoid embedded in an M-dimensional space, where k is the number of non-zero singular values. Figure 1 depicts the situation when $M = N = k = 2$ [FMM77]. The lengths of the axes of the ellipsoid are the singular values of A, and in the general case, the major and minor axes are given by s_{max} and s_{min} respectively. Intuitively, the singular values of a matrix describe the extent to which multiplication by the matrix distorts the original vector. The magnitude of the singular values can be used to highlight which dimensions of the vector are most affected, and in some sense more important, as we shall see next in the discussion of SVD applications.

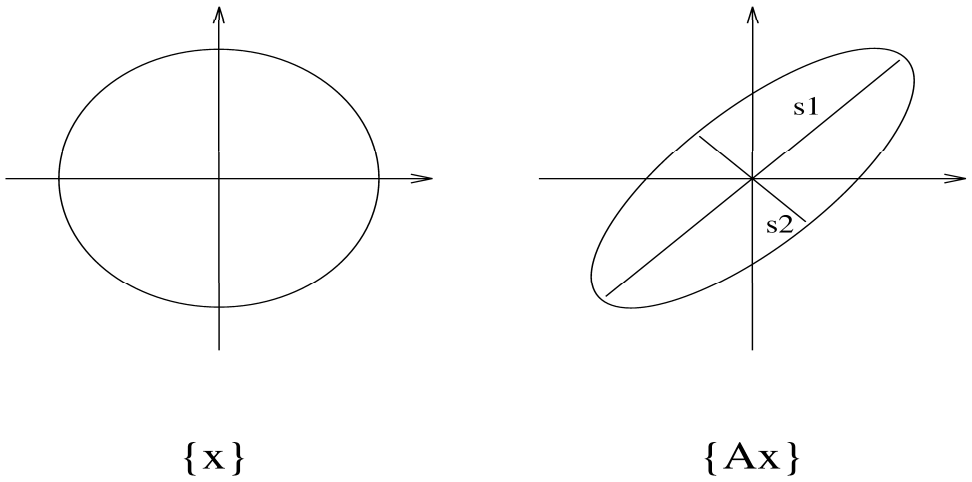

Figure 1: Mapping by A of Unit Sphere when $M = N = k = 2$

3.2 SVD and Matrix Rank

Fundamental to linear algebra is the notion of rank. Numerous theorems begin with the condition "If matrix A is of full rank, then the following property holds". However, if the matrix is rank deficient (or nearly so), then small perturbations of the matrix values (from round-off errors or fuzzy data) will yield a matrix which is of full rank. Hence, determining the rank of a matrix is non-trivial. The SVD lends us a practical definition of rank, as well as allows us to quantify the notion of near rank deficiency.

The familiar definition of rank is the number of linearly independent columns of a matrix. Let the matrix A have the SVD $U\Sigma V^T$. Since multiplication by orthogonal matrices preserves linear independence, the rank of A is precisely the rank of the diagonal matrix Σ, or equivalently, the number of non-zero singular values. If A is nearly rank deficient (singular), then the singular values will be small. Moreover, suppose that $Rank(A) = m$ and we wish to approximate A by a matrix B of lower rank k. Then we can use the singular values of A to compute a matrix with the best approximation, and to determine if the approximation is unique. Let s_i be the diagonal entries of Σ and let u_i and v_i be the column vectors of U and V respectively. Then

$$\min_{Rank(B)=k} \| A - B \|_E = \| A - A_k \|_E = s_{k+1}$$

where $A_k = \sum_{i=1}^{k} s_i u_i v_i^T$. The solution $B = A_k$ will be unique when $s_{k+1} < s_k$. Proofs of these facts can be found in [DD88, GVL83].

We see then that the SVD of A produces a sequence of approximations to A of successive ranks $A_i = U\Sigma_i V^T$, where Σ_i is the rank i version of Σ obtained by setting the last $m - i$ singular values to zero. Also, A_i is the best rank i approximation to A in the sense of Euclidean distance.

The use of SVD for matrix approximation has a number of practical advantages. First, applications which encounter round-off errors or fuzzy data typically use the *effective rank* of a matrix, i.e. the number of singular values greater than some ϵ, where ϵ reflects the accuracy of the data. Hence, decisions are made only about the negligibility of a few singular values, rather than vectors or sets of vectors. Second, storing the approximation of a matrix often

results in a significant savings over storing the whole matrix. Note that we can express a matrix A as

$$A = s_1 u_1 v_1^T + s_2 u_2 v_2^T + \cdots + s_m u_m v_m^T$$

Each outer product $u_i v_i^T$ is a simple matrix of rank 1, and can be stored in $M + N$ numbers, versus $M * N$ of the original matrix. Additionally, multiplication of $u_i v_i^T$ with a vector x requires only $M + N$ operations, instead of $M * N$ [FMM77]

3.3 SVD and Linear Independence

Another use of the SVD provides a measure, called a condition number, which is related to the measure of linear independence between the column vectors of the matrix.

The condition number (with respect to the Euclidean norm) of a matrix A is

$$cond(A) = \frac{s_{max}}{s_{min}}$$

where s_{min} and s_{max} are the largest and smallest singular values of A. If A is rank deficient, then $s_i = 0$ and we consider $cond(A) = \infty$.

Using the condition number, we can quantify the independence of the columns of A. Note that $cond(A) \geq 1$. If $cond(A)$ is close to 1, then the columns of A are very independent. When the condition number is large, the columns of A are nearly dependent.

Returning to the geometric interpretation of singular values, we see that the condition number is related to the axes of the hyperellipsoid associated with the matrix. Since $cond(A)$ is defined by the extreme singular values and these values are the lengths of the major and minor axes, the condition number describes the eccentricity of the hyperellipsoid.

As we will see in the next section, the notion of a condition number becomes important in solving linear systems, where $cond(A)$ in some sense measures the sensitivity of the system to noise in the data.

4 Applications of SVD

4.1 Solutions to Linear Equations

Numerous practical problems can be expressed in the language of linear algebra. A linear system involves a set of equations in N variables. For example, consider the following linear system.

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & x_3 & = & 8 \\ 10x_1 & + & 18x_2 & + & 12x_3 & = & 78 \\ 20x_1 & + & 22x_2 & + & 40x_3 & = & 144 \end{array}$$

This problem can be expressed in terms of a coefficient matrix A, a vector x of variables, and a vector b, such that a solution to the linear system $Ax = b$ is an assignment to the values of the vector x. For the above example, A, x, and b are

$$Ax = \begin{pmatrix} 1 & 2 & 1 \\ 10 & 18 & 12 \\ 20 & 22 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 8 \\ 78 \\ 144 \end{pmatrix} = b$$

Using the SVD of A, we can determine if a solution exists, as well as the general form of the possible solutions x. If $U\Sigma V^T$ is the SVD of the $M \times N$ matrix A ($M \geq N$), then the system $Ax = b$ becomes

$$U\Sigma V^T x = b.$$

Substituting $z = V^T x$ and $d = U^T b$, we have

$$\Sigma z = d.$$

Let $Rank(A) = k =$ the number of non-zero singular values s_i. Studying the linear equations of the diagonal system $\Sigma z = d$, we can determine whether or not there is a solution. A solution exists if and only if $d_j = 0$ whenever $s_j = 0$ or $j > N$. If $k < N$, then the z_j associated with a zero s_j can be set to any value and still yield a solution. A general form of the possible solutions can then be expressed in terms of these arbitrary components of z when transformed back to the original coordinates by $x = Vz$.

The condition number of a matrix can also describe the sensitivity of solutions of linear systems to inaccuracies in the data. Suppose we want to measure the maximal increase in relative inaccuracy for the worst position of b and error db, when solving for x in the system $Ax = b$. The answer is precisely the condition number [VDM88].

$$E = cond(A) = \max_{b,db} \frac{\| dx \| / \| x \|}{\| dy \| / \| y \|} = \frac{s_{max}}{s_{min}}$$

For the above reason, matrices with large condition numbers are said to be ill-conditioned.

As an extension of solving linear systems, suppose we wish to find a solution where Ax is approximately equal to b. By this we mean the **least-squares** solution x to minimize

$$\| Ax - b \|^2,$$

or equivalently to minimize the length $\| Ax - b \|$. The advantage of using the SVD for this problem is that it can reliably handle the rank deficient case as well as the full rank case. Since orthogonal matrices preserve norm,

$$\| U^T(AVV^T x - b) \| = \| \Sigma z = d \| .$$

Using the SVD, the least squares problem is now in terms of a diagonal matrix, where the vector z that minimizes the length $\| Ax - b \|$ is given by

$$\begin{aligned} z_j &= \tfrac{d_j}{s_j} && \text{if } s_j \neq 0 \\ z_j &= \text{anything} && \text{if } s_j = 0 \end{aligned}$$

Hence, k of the equations have exact solutions and the remaining ones yield a possibly non-zero residual vector of length $(\sum d_i^2)^{\frac{1}{2}}$, where the sum is over all i for which $s_i = 0$ or $i > N$. The solution to the original problem is then $x = Vz$ [FMM77].

4.2 Noisy Signal Filtering

Problems in signal processing often use linear models for signals. In ideal (noise-free) conditions, the measurment data can be arranged in a matrix, where the matrix is known to be

rank-deficient. By this, we mean that the signal is assumed to lie in a proper subspace of Euclidean space. However, the presence of noise, either from rounding error or instrument error, results in a measurement matrix that is often of full rank. Usually, the models assume that the error can be separated from the data, in that the noise component is that which lies in a subspace orthogonal to the signal subspace. For this reason, the SVD is used to approximate the matrix, decomposing the data into an optimal estimate of the signal and the noise components.

Suppose A is the measurement matrix, where each column consists of a signal component x and a noise component n.

$$A = (C_1 \mid C_2 \mid \cdots \mid C_N)$$

where each $C_i = x_i + n_i$. The vector x representing the signal is known to lie in a rank k subspace, though the precise subspace is not known. Therefore, let $x = Hc$ for a coefficient vector c and a matrix H whose columns are the basis vectors of some rank k subspace. The (least-squared) error between A and Hc is minimized by choosing H to be the optimal k rank approximation A_k to A. Then the k columns of U, corresponding to the k largest singular values, span the rank k subspace H. The resulting error is $e^2 = \sum_{k+1}^{N} s_i^2$. Using the SVD as above, we see that the original data matrix A is decomposed into the orthogonal components $U\Sigma_k V^T$, which is the rank k subspace corresponding to the signal subspace, and $U\Sigma_{n-k}V^T$, which corresponds to the orthogonal subspace defining the noise components [Sch91].

4.3 Time Series Analysis

The technique of delay coordinate embedding, used by [Sau94] for time series analysis, also uses the SVD. The algorithm constructs a multidimensional model of the data from a sequence of one dimensional observations. An M dimensional vector is constructed by sliding a window of length M over consecutive observations in the data sequence. The vectors are then filtered using the Discrete Fourier Transform to remove signal noise. Each vector $b = \{b_1, b_2, \ldots, b_M\}$ represents a state of the underlying dynamical system. The object is to find the best (least-squares distance) L-dimensional linear space ($L \leq M$) that passes through the center of mass c of the K nearest neighbors of b. We construct a matrix A whose rows consist of the vectors $b_1 - c, b_2 - c, \ldots, b_k - c$, and calculate the SVD $U\Sigma V^T$. Taking the first L columns of the orthogonal matrix V gives us the desired basis for the L-dimensional subspace.[1]

5 Discussion and Conclusion

By providing an approximation to rank deficient matrices, and exposing the geometric properties of the matrix, the singular value decomposition of a matrix is a powerful technique in matrix computations. Despite its usefulness, however, there are a number of drawbacks, as mentioned by [Vac91]. For problems that can be solved by simpler techniques, such as the Fourier Transform, or QR decomposition, use of the SVD may be unduly expensive computationally. Secondly, the SVD operates on a fixed matrix, hence it is not amenable to

[1]The first L columns of V provide a basis for $NullSPace(A) = Range(A^T)$ which is consistent with our previous discussion since the vectors $b_i - c$ constitute the *rows* of A, rather than the *columns* as in the other examples.

problems that require adaptive algorithms. A host of active research efforts address these problems. Further examples of the use of SVD in the field of Signal Processing, as well as discussions of implementation algorithms and architectures, can be found in [Vac91, Dep88].

References

[And86] Fobert F. V. Anderson. *Introduction to Linear Algebra.* Holt, Rinehart, and Winston, New York, 1986.

[DD88] P. Dewilde and Ed. F. Deprettere. Singular value decomposition: An introduction. In Ed. F. Deprettere, editor, *SVD and Signal Processing: Algorithms, Applications, and Architectures*, pages 3–41. Elsevier Science Publishers, North Holland, 1988.

[Dep88] Ed. F. Deprettere, editor. *SVD and Signal Processing: Algorithms, Analysis, and Applications.* Elsevier Science Publishers, North Holland, 1988.

[FMM77] George E. Forsythe, Michael A. Malcolm, and Cleve B. Moler. *Computer Methods for Mathematical Computations*, pages 201–235. Prentice Hall, Englewood Cliffs, 1977.

[GVL83] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*, pages 16–21, 293. Johns Hopkins University Press, Baltimore, Maryland, 1983.

[Sau94] Tim Sauer. Time series prediction by using delayed coordinate embedding. In Andreas S. Weigend and Neil A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past.* Adgdison-Wesley, 1994.

[Sch91] Louis L. Scharf. The SVD and reduced-rank signal processing. In R. Vaccaro, editor, *SVD and Signal Processing II: Algorithms, Applications, and Architectures*, pages 3–31. Elsevier Science Publishers, North Holland, 1991.

[Vac91] R. Vaccaro, editor. *SVD and Signal Processing II: Algorithms, Analysis, and Applications.* Elsevier Science Publishers, North Holland, 1991.

[VDM88] Joos Vandewalle and Bart De Moor. A variety of applications of singular value decomposition in identification and signal processing. In Ed. F. Deprettere, editor, *SVD and Signal Processing: Algorithms, Applications, and Architectures*, pages 43–91. Elsevier Science Publishers, North Holland, 1988.

[WG94] Andreas S. Weigend and Neil A. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, Reading, Massachusetts, 1994.