

Singular Value Decomposition (SVD) tutorial

BE.400 / 7.548

Singular value decomposition takes a rectangular matrix of gene expression data (defined as A , where A is a $n \times p$ matrix) in which the n rows represents the genes, and the p columns represents the experimental conditions. The SVD theorem states:

$$A_{n \times p} = U_{n \times n} S_{n \times p} V_{p \times p}^T$$

Where

$$U^T U = I_{n \times n}$$

$$V^T V = I_{p \times p} \text{ (i.e. } U \text{ and } V \text{ are orthogonal)}$$

Where the columns of U are the left singular vectors (*gene coefficient vectors*); S (the same dimensions as A) has singular values and is diagonal (*mode amplitudes*); and V^T has rows that are the right singular vectors (*expression level vectors*). The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal.

Calculating the SVD consists of finding the eigenvalues and eigenvectors of AA^T and $A^T A$. The eigenvectors of $A^T A$ make up the columns of V , the eigenvectors of AA^T make up the columns of U . Also, the singular values in S are square roots of eigenvalues from AA^T or $A^T A$. The singular values are the diagonal entries of the S matrix and are arranged in descending order. The singular values are always real numbers. If the matrix A is a real matrix, then U and V are also real.

To understand how to solve for SVD, let's take the example of the matrix that was provided in Kuruvilla *et al*:

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

In this example the matrix is a 4x2 matrix. We know that for an $n \times n$ matrix W , then a nonzero vector \mathbf{x} is the eigenvector of W if:

$$W \mathbf{x} = l \mathbf{x}$$

For some scalar l . Then the scalar l is called an eigenvalue of A , and \mathbf{x} is said to be an eigenvector of A corresponding to l .

So to find the eigenvalues of the above entity we compute matrices AA^T and $A^T A$. As previously stated, the eigenvectors of AA^T make up the columns of U so we can do the

following analysis to find U.

$$AA^T = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 20 & 14 & 0 & 0 \\ 14 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = W$$

Now that we have a $n \times n$ matrix we can determine the eigenvalues of the matrix W.

Since $W \mathbf{x} = \lambda \mathbf{x}$ then $(W - \lambda I) \mathbf{x} = 0$

$$\begin{bmatrix} 20 - \lambda & 14 & 0 & 0 \\ 14 & 10 - \lambda & 0 & 0 \\ 0 & 0 & -\lambda & 0 \\ 0 & 0 & 0 & -\lambda \end{bmatrix} \mathbf{x} = (W - \lambda I) \mathbf{x} = 0$$

For a unique set of eigenvalues to determinant of the matrix $(W - \lambda I)$ must be equal to zero. Thus from the solution of the characteristic equation, $|W - \lambda I| = 0$ we obtain:

$\lambda = 0, \lambda = 0; \lambda = 15 + \sqrt{221.5} \sim 29.883; \lambda = 15 - \sqrt{221.5} \sim 0.117$ (four eigenvalues since it is a fourth degree polynomial). This value can be used to determine the eigenvector that can be placed in the columns of U. Thus we obtain the following equations:

$$19.883 x_1 + 14 x_2 = 0$$

$$14 x_1 + 9.883 x_2 = 0$$

$$x_3 = 0$$

$$x_4 = 0$$

Upon simplifying the first two equations we obtain a ratio which relates the value of x_1 to x_2 . The values of x_1 and x_2 are chosen such that the elements of the S are the square roots of the eigenvalues. Thus a solution that satisfies the above equation $x_1 = -0.58$ and $x_2 = 0.82$ and $x_3 = x_4 = 0$ (this is the second column of the U matrix).

Substituting the other eigenvalue we obtain:

$$-9.883 x_1 + 14 x_2 = 0$$

$$14 x_1 - 19.883 x_2 = 0$$

$$x_3 = 0$$

$$x_4 = 0$$

Thus a solution that satisfies this set of equations is $x_1 = 0.82$ and $x_2 = -0.58$ and $x_3 = x_4 = 0$ (this is the first column of the U matrix). Combining these we obtain:

$$U = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Similarly $A^T A$ makes up the columns of V so we can do a similar analysis to find the value of V .

$$A^T A = \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and similarly we obtain the expression:

$$V = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

Finally as mentioned previously the S is the square root of the eigenvalues from AA^T or $A^T A$, and can be obtained directly giving us:

$$S = \begin{bmatrix} 5.47 & 0 \\ 0 & 0.37 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Note that: $s_1 > s_2 > s_3 > \dots$ which is what the paper was indicating by the figure 4 of the Kuruvilla paper. In that paper the values were computed and normalized such that the highest singular value was equal to 1.

Proof:

$$A = USV^T \text{ and } A^T = VSU^T$$

$$A^T A = VSU^T USV^T$$

$$A^T A = VS^2 V^T$$

$$A^T A V = VS^2$$

References

- Alter O, Brown PO, Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, **97**, 10101-6.
- Golub, G.H., and Van Loan, C.F. (1989) *Matrix Computations*, 2nd ed. (Baltimore: Johns Hopkins University Press).

- Greenberg, M. (2001) Differential equations & Linear algebra (Upper Saddle River, N.J. : Prentice Hall).
- Strang, G. (1998) Introduction to linear algebra (Wellesley, MA : Wellesley-Cambridge Press).