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* x *p* matrix) in which the *n* rows represents the genes, and the *p* columns represents the experimental conditions. The SVD theorem states:

$$\mathbf{A}_{nxp}= \mathbf{U}_{nxn}\, \mathbf{S}_{nxp}\, \mathbf{V}^{\mathbf{T}}_{pxp}$$

Where

$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{\mathrm{nxn}}$

$\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{\mathrm{pxp}}$ (i.e. U and V 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^TA. The eigenvectors of A^TA 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^TA. 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 x n matrix W, then a nonzero vector **x** is the eigenvector of W if:

W **x** = l **x**

For some scalar l. Then the scalar l is called an eigenvalue of A, and **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^TA. As previously stated , the eigenvectors of AA^T make up the columns of *U* so we can do the following analysis to find U.

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

Now that we have a n x n matrix we can determine the eigenvalues of the matrix W.

Since W **x** = l **x** then (W- lI) **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-lI) must be equal to zero. Thus from the solution of the characteristic equation, |W-lI|=0 we obtain:
l=0, l=0; l = 15+Ö221.5 ~ 29.883; l = 15-Ö221.5 ~ 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 x1 + 14 x2 = 0
14 x1 + 9.883 x2 = 0
x3 = 0
x4 = 0

Upon simplifying the first two equations we obtain a ratio which relates the value of x1 to x2. The values of x1 and x2 are chosen such that the elements of the S are the square roots of the eigenvalues. Thus a solution that satisfies the above equation x1 = -0.58 and x2 = 0.82 and x3 = x4 = 0 (this is the second column of the U matrix).

Substituting the other eigenvalue we obtain:

-9.883 x1 + 14 x2 = 0
14 x1 - 19.883 x2 = 0
x3 = 0
x4 = 0

Thus a solution that satisfies this set of equations is x1 = 0.82 and x2 = -0.58 and x3 = x4 = 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^TA. and can be obtained directly giving us:

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

Note that: s1 > s2 > s3 > … 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:
$\mathbf{A}=\mathbf{USV}^T$ and $\mathbf{A}^T=\mathbf{VSU}^T$
$\mathbf{A}^T\mathbf{A} = \mathbf{VSU}^T\mathbf{USV}^T$
$\mathbf{A}^T\mathbf{A} = \mathbf{VS}^2\mathbf{V}^T$
$\mathbf{A}^T\mathbf{AV} = \mathbf{VS}^2$

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).