The Singular Value Decomposition

The singular value decomposition (henceforth SVD) of an $m \times n$ matrix $\boldsymbol{X}$, with $m \geq n$, can be expressed in the form $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top$. Here $\boldsymbol{U}$ is $m \times n$, like $\boldsymbol{X}$, $\boldsymbol{W}$ is a diagonal $n \times n$ matrix, and $\boldsymbol{V}$ is a $n \times n$ orthogonal matrix, that is, such that $\boldsymbol{V}^\top = \boldsymbol{V}^{-1}$ or, equivalently, $\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{V}^\top\boldsymbol{V} = \mathbf{I}_n$, the $n \times n$ identity matrix. In addition, the columns of $\boldsymbol{U}$ are orthonormal, which means that $\boldsymbol{U}^\top\boldsymbol{U} = \mathbf{I}_n$.

The first question to be asked is whether the SVD exists for an arbitrary $m \times n$ matrix $\boldsymbol{X}$ with $m \geq n$. Suppose that it does. Then it follows that

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{W}^2\boldsymbol{V}^\top.$$

On postmultiplying by $\boldsymbol{V}$, we get $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{W}^2$. This shows that the elements of the diagonal matrix $\boldsymbol{W}^2$ are the eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$, which must be non-negative since $\boldsymbol{X}^\top\boldsymbol{X}$ is by construction positive semi-definite, and that the columns of $\boldsymbol{V}$ are the corresponding eigenvectors. This establishes the existence of $\boldsymbol{W}$ and $\boldsymbol{V}$, although they are by no means unique. But given some choice of $\boldsymbol{W}$ and $\boldsymbol{V}$, then, if $\boldsymbol{U}$ exists, postmultiplying the defining equation by $\boldsymbol{V}$ gives $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{W}$. Then, if there are no zero elements on the diagonal of $\boldsymbol{W}$, we see that $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{W}^{-1}$, and so in this case $\boldsymbol{U}$ exists, and does indeed satisfy the requirement that $\boldsymbol{U}^\top\boldsymbol{U} = \mathbf{I}_n$, because

$$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{W}^{-1}\boldsymbol{V}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{V}\boldsymbol{W}^{-1} = \boldsymbol{W}^{-1}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{W}^2\boldsymbol{W}^{-1},$$

since $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{W}^2$. Then, since $\boldsymbol{V}^\top\boldsymbol{V} = \mathbf{I}_n$, we see that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{W}^{-1}\boldsymbol{W}^2\boldsymbol{W}^{-1} = \mathbf{I}_n$.

In fact, the eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$ are strictly positive if (and only if) $\boldsymbol{X}$ is of full rank. If not, this means that there exists a nonzero n-vector $\boldsymbol{v}$ such that $\boldsymbol{X}\boldsymbol{v} = \mathbf{0}$. Then $\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v} = 0$, which implies that the positive semi-definite matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is not positive definite, and so it has at least one zero eigenvalue. Conversely, if there is a zero eigenvalue of $\boldsymbol{X}^\top\boldsymbol{X}$, with corresponding (nonzero) eigenvector $\boldsymbol{v}$ say, then $\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v} = \|\boldsymbol{X}\boldsymbol{v}\|^2 = 0$, which implies that $\boldsymbol{X}\boldsymbol{v} = \mathbf{0}$, so that $\boldsymbol{X}$ does not have full rank.

Let us suppose that $\boldsymbol{X}$ has full rank. We see immediately that the three matrices of the SVD exist. If one requires the elements of $\boldsymbol{W}$ to be the positive square roots of the eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$, and to be sorted in decreasing order from top left to bottom right, then the SVD is essentially unique (up to the signs of the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$), except in the non-generic case in which there is degeneracy of the eigenvalues.

The matrix $\boldsymbol{X}\boldsymbol{X}^\top$ is of dimension $m \times m$, and it given by

$$\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{W}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{W}^2\boldsymbol{U}^\top$$

Thus $\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{U} = \boldsymbol{U}\boldsymbol{W}^2$, which shows that the n diagonal elements of $\boldsymbol{W}^2$ are eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$, as well as of $\boldsymbol{X}^\top\boldsymbol{X}$, while the columns of $\boldsymbol{U}$ are the

corresponding eigenvectors. If $n < m$, there are also $m - n$ zero eigenvalues, the corresponding eigenspace being the complement in E^m of the span of the columns of $\boldsymbol{U}$, or equivalently those of $\boldsymbol{X}$. Note that $\boldsymbol{U}\boldsymbol{U}^\top$ is an orthogonal projection matrix:

$$(\boldsymbol{U}\boldsymbol{U}^\top)^2 = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}\mathbf{I}_k\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{U}^\top,$$

which establishes idempotency, symmetry being obvious. In fact, $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{P}_{\boldsymbol{X}}$, the orthogonal projection on to the span of the columns of $\boldsymbol{X}$, since the columns of $\boldsymbol{U}$ and $\boldsymbol{X}$ span the same space. The columns of $\boldsymbol{U}$ therefore provide an orthonormal basis for this span. This suggests that the SVD can provide a way to implement ordinary least squares, and indeed it is one of the most numerically stable ways to do so.

The columns of $\boldsymbol{U}$ are called the left singular vectors of $\boldsymbol{X}$, and those of $\boldsymbol{V}$ the right singular vectors of $\boldsymbol{X}$.

Generalised Inverse of a Matrix

A generalised inverse of an $m \times n$ matrix $\boldsymbol{X}$ is an $n \times m$ matrix $\boldsymbol{X}^+$ that satisfies the two relations

$$\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{X}, \quad \text{and} \quad \boldsymbol{X}^+\boldsymbol{X}\boldsymbol{X}^+ = \boldsymbol{X}^+.$$

If $\boldsymbol{X}$ is a non-singular square matrix, then we can pre- and post-multiply the first of these requirements by $\boldsymbol{X}^{-1}$ to get $\boldsymbol{X}^+ = \boldsymbol{X}^{-1}$. This shows that the notion of a generalised inverse does indeed generalise that of the inverse, when the latter exists.

If $m > n$ and $\boldsymbol{X}$ has full column rank n, a possible generalised inverse is given by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}$. However, it is not unique. If the nonzero $m \times n$ matrix $\boldsymbol{B}$ has columns that are orthogonal to those of $\boldsymbol{X}$, it is easy to check that $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}^\top$ is also a generalised inverse of $\boldsymbol{X}$. Provided that $m > n$, it is always possible to find such a matrix $\boldsymbol{B}$.

Note that, for any generalised inverse $\boldsymbol{X}^+$, both $\boldsymbol{X}\boldsymbol{X}^+$ and $\boldsymbol{X}^+\boldsymbol{X}$ are idempotent, so that, if either of them is symmetric, it is an orthogonal projection matrix. We have the following result:

Moore-Penrose Generalised Inverse

For every $m \times n$ matrix $\boldsymbol{X}$, $m \geq n$, there exists a unique matrix $\boldsymbol{X}^+$, called the Moore-Penrose generalised inverse, or the pseudo-inverse, satisfying the four conditions

$$\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{X}, \quad \boldsymbol{X}^+\boldsymbol{X}\boldsymbol{X}^+ = \boldsymbol{X}^+,$$
$$(\boldsymbol{X}\boldsymbol{X}^+)^\top = \boldsymbol{X}\boldsymbol{X}^+, \quad (\boldsymbol{X}^+\boldsymbol{X})^\top = \boldsymbol{X}^+\boldsymbol{X}.$$

I omit the full proof.

However, the existence part can be shown constructively by using the SVD. The Moore-Penrose generalised inverse of $\boldsymbol{X}$ is $\boldsymbol{X}^+ = \boldsymbol{V}\boldsymbol{W}^+\boldsymbol{U}^\top$, where $\boldsymbol{W}^+$ is a specific generalised inverse of the diagonal matrix $\boldsymbol{W}$. If $\boldsymbol{W}$ has no zero singular values, then $\boldsymbol{W}^+$ is the usual inverse, the diagonal matrix the elements of which are the inverses of the singular values in $\boldsymbol{W}$. If there are singular values equal to zero, they occur in the bottom right corner of $\boldsymbol{W}$. Let the rank of $\boldsymbol{W}$ be $p < n$, then we can write

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{D}_{p\times p} & \mathbf{O}_{p\times(n-p)} \\ \mathbf{O}_{(n-p)\times p} & \mathbf{O}_{(n-p)\times(n-p)} \end{bmatrix},$$

where the matrix $\boldsymbol{D}$ contains all the nonzero singular values, so that $\boldsymbol{D}^{-1}$ is well defined. Then define

$$\boldsymbol{W}^+ = \begin{bmatrix} \boldsymbol{D}^{-1}_{p\times p} & \mathbf{O}_{p\times(n-p)} \\ \mathbf{O}_{(n-p)\times p} & \mathbf{O}_{(n-p)\times(n-p)} \end{bmatrix}.$$

We see immediately that $\boldsymbol{W}^+$ thus defined is indeed a generalised inverse of $\boldsymbol{W}$.

Calculate as follows:

$$\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{V}\boldsymbol{W}^+\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{W}^+\boldsymbol{W}\boldsymbol{V}^\top.$$

This is clearly symmetric, and it is easy to show that $\boldsymbol{X}\boldsymbol{X}^+$ is so as well. Then

$$\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{W}^+\boldsymbol{W}\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top = \boldsymbol{X}, \text{ and}$$
$$\boldsymbol{X}^+\boldsymbol{X}\boldsymbol{X}^+ = \boldsymbol{V}\boldsymbol{W}^+\boldsymbol{W}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{W}^+\boldsymbol{U}^\top = \boldsymbol{V}\boldsymbol{W}^+\boldsymbol{U}^\top = \boldsymbol{X}^+,$$

where I have used the fact that $\boldsymbol{W}^+$ is a generalised inverse of $\boldsymbol{W}$. This show that $\boldsymbol{X}^+$ satisfies the defining properties of the Moore-Penrose inverse. Only uniqueness remains to be proved.

The *Deep Learning* book claims that $\boldsymbol{X}^+$ can also be defined by the formula

$$\boldsymbol{X}^+ = \lim_{\alpha\downarrow 0}(\boldsymbol{X}^\top\boldsymbol{X} + \alpha\mathbf{I})^{-1}\boldsymbol{X}^\top$$

This formula is harder to implement than the one based on the SVD, as very efficient algorithms exist for the computation of the SVD. The right-hand side is easy enough to compute, but the limit is not.

Principal Components Analysis

Another use of the SVD is to implement Principal Components Analysis (PCA). PCA has been around for a long time as a tool for the analysis of data. For a given full-rank $m \times n$ matrix $\boldsymbol{X}$, an orthogonal linear transformation of the columns of $\boldsymbol{X}$ is sought, such that the transformed matrix has mutually orthogonal columns, and has the property that the successive

columns have progressively decreasing variance. Before performing the transformation, the elements in each column of $\boldsymbol{X}$ must be centred by subtracting the column means.

The desired transformation is given by $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{V}$. We saw that $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{W}$, and so $\boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{W}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{W} = \boldsymbol{W}^2$, a diagonal matrix. This implies that the columns of $\boldsymbol{T}$ are indeed mutually orthogonal. The relation between the columns of $\boldsymbol{T}$ and those of $\boldsymbol{U}$ can be written as $\boldsymbol{t}_i = w_i\boldsymbol{u}_i$, $i = 1, \ldots, n$, where $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$ are the i^{th} columns of $\boldsymbol{T}$ and $\boldsymbol{U}$ respectively, and w_i is the i^{th} diagonal element of $\boldsymbol{W}$. Thus $\boldsymbol{t}_i{}^\top\boldsymbol{t}_i = w_i^2\boldsymbol{u}_i{}^\top\boldsymbol{u}_i = w_i^2$ since the columns of $\boldsymbol{U}$ have unit norm. Now we chose to sort the w_i in decreasing order, and so the norms of the columns of $\boldsymbol{T}$ are also in decreasing order, as desired.

One purpose of PCA is dimension reduction. One seeks an $m \times n$ matrix $\boldsymbol{Y}$, of rank $p < n$, that approximates the original matrix $\boldsymbol{X}$ as well as possible. In the formula $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^\top$, we can set the $n - p$ smallest diagonal elements of $\boldsymbol{W}$ equal to zero, and denote the result by $\boldsymbol{W}_p$, which has only p nonzero diagonal elements, so that the approximation $\boldsymbol{Y} \equiv \boldsymbol{U}\boldsymbol{W}_p\boldsymbol{V}^\top$ has rank p. It can be shown that $\boldsymbol{Y}$ is closer to $\boldsymbol{X}$ than any other $m \times n$ matrix of rank p, where distance is measured by the sum of the squares of the elements of the difference between $\boldsymbol{X}$ and a rank-p approximation to $\boldsymbol{X}$. This result is known as the Eckart-Young theorem. The name of Mirsky is sometimes added to the other two names, as Mirsky is responsible for an extension of the result.

Proof: The square of the **Frobenius norm** of an $m \times n$ matrix $\boldsymbol{A}$ is the sum of the squares of the elements of $\boldsymbol{A}$. It is easy to check that this is equal to $\mathrm{Tr}(\boldsymbol{A}^\top\boldsymbol{A}) = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^\top)$.

Consider an $m \times n$ rank-p matrix $\boldsymbol{Y}$, and the Frobenius norm of the difference $\boldsymbol{X} - \boldsymbol{Y}$. We have $\boldsymbol{X} - \boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Y}$, and so

$$\mathrm{Tr}(\boldsymbol{X} - \boldsymbol{Y})^\top(\boldsymbol{X} - \boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y})^\top(\boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}) + \mathrm{Tr}(\boldsymbol{Y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Y}).$$

The second term on the right-hand side is non-negative, and so, if we wish to minimise the norm of $\boldsymbol{X} - \boldsymbol{Y}$, we should choose $\boldsymbol{Y}$ such that $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Y} = \mathbf{O}$ and $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} = \boldsymbol{Y}$. Let us write $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Z}$, where the $n \times n$ matrix $\boldsymbol{Z}$ has rank p. Above, we had $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{W}_p\boldsymbol{V}^\top$, which implies that, in this case, $\boldsymbol{Z} = \boldsymbol{W}_p\boldsymbol{V}^\top$. Then $\boldsymbol{X} - \boldsymbol{Y} = \boldsymbol{U}(\boldsymbol{W} - \boldsymbol{W}_p)\boldsymbol{V}^\top$, and so, where $\|\cdot\|_F$ denotes the Frobenius norm,

$$\begin{aligned}\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 &= \mathrm{Tr}(\boldsymbol{V}(\boldsymbol{W} - \boldsymbol{W}_p)\boldsymbol{U}^\top\boldsymbol{U}(\boldsymbol{W} - \boldsymbol{W}_p)\boldsymbol{V}^\top) \\ &= \mathrm{Tr}(\boldsymbol{V}(\boldsymbol{W} - \boldsymbol{W}_p)^2\boldsymbol{V}^\top) = \mathrm{Tr}((\boldsymbol{W} - \boldsymbol{W}_p)^2), \qquad (1)\end{aligned}$$

where we have used the invariance of the trace of a matrix product under a cyclic permutation of the factors, and the fact that $\boldsymbol{V}^\top\boldsymbol{V} = \mathbf{I}$. Now the matrix $\boldsymbol{W} - \boldsymbol{W}_p$ is a diagonal matrix with only the last $n - p$ elements nonzero, these being equal to the $n - p$ smallest singular values. Thus

$$\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 = \sum_{i=n-p+1}^{n} w_i^2. \qquad (2)$$

For the theorem, we have to show that, with any other choice of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, the squared norm of $\boldsymbol{X} - \boldsymbol{Y}$ is no less than the sum above.

In the general case, then, $\boldsymbol{X} - \boldsymbol{Y} = \boldsymbol{U}(\boldsymbol{W}\boldsymbol{V}^\top - \boldsymbol{Z})$. Let $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{V}^\top$, where the $n \times n$ matrix $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{V}$. With this, $\boldsymbol{X} - \boldsymbol{Y} = \boldsymbol{U}(\boldsymbol{W} - \boldsymbol{A})\boldsymbol{V}^\top$. Similarly to the calculation in (1), we find that

$$\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 = \mathrm{Tr}\big((\boldsymbol{W}-\boldsymbol{A}^\top)(\boldsymbol{W}-\boldsymbol{A})\big) = \mathrm{Tr}\big((\boldsymbol{W}-\boldsymbol{A}^\top)\boldsymbol{P}_{\boldsymbol{A}}(\boldsymbol{W}-\boldsymbol{A})\big)+\mathrm{Tr}(\boldsymbol{W}\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{W}),$$

where $\boldsymbol{P}_{\boldsymbol{A}}$ is the orthogonal projection in the n-dimensional space spanned by the columns of $\boldsymbol{V}$, and $\boldsymbol{M}_{\boldsymbol{A}}$ is the complementary projection, so that $\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{A} = \boldsymbol{O}$. Since $\boldsymbol{W}$ is a diagonal matrix, it follows that

$$\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 \geq \mathrm{Tr}(\boldsymbol{W}^2\boldsymbol{M}_{\boldsymbol{A}}) = \sum_{i=1}^{n} w_i^2 (\boldsymbol{M}_{\boldsymbol{A}})_{ii}, \tag{3}$$

where $(\boldsymbol{A})_{ii}$ is the i^{th} diagonal element of the orthogonal projection matrix $\boldsymbol{M}_{\boldsymbol{A}}$. These diagonal elements all lie between 0 and 1, and they sum to $n - p$, which is the dimension of the image of $\boldsymbol{M}_{\boldsymbol{A}}$. Comparison of the right-hand sides of (2) and (3) shows that the latter cannot be smaller than the former, given that the w_i are in decreasing in i. ∎