An almost self-contained introduction to the Singular Value Decomposition

Disclaimer. This is a first draft needing revision. Some parts are missing, references are missing, mistakes might be there. You are welcome to point me to mistakes or inconsistencies.

These notes are intended to provide a short, guided tour of the Singular Value Decomposition (henceforth, SVD), showing how it directly follows from general properties of square, symmetric matrices. Proofs are intended to provide an understanding of key aspects, such as the existence of an SVD. This presentation is (almost completely) self-contained and basically builds on the definitions of eigenvalue and eigenvector of a matrix.

Key notions in matrix linear algebra

We begin by considering a generic, square matrix $M \in \mathbb{R}^{n\times n}$. For such a matrix an (eigenvalue, eigenvector) pair are a scalar λ and a vector $\mathbf{x} \in \mathbb{R}^n$, such that the following holds:

$$M\mathbf{x} = \lambda\mathbf{x}.$$

In this case, $\mathbf{x}$ is a *right* eigenvector of M. Analogously, a *left* eigenvector is a vector $\mathbf{x}$, such that $\mathbf{y}^T M = \lambda\mathbf{y}^T$. The following should be noted: the equation $M\mathbf{x} = \lambda\mathbf{x}$ is equivalent to the following homogeneous system:

$$(M - \lambda I)\mathbf{x} = \mathbf{0}.$$

If you go back to your Math classes, you will remember that the system above admits a non-trivial solution (i.e., a solution $\mathbf{x} \neq \mathbf{0}$) if and only if $M - \lambda I$ is singular. In turn, singularity implies $p(\lambda) = det(M - \lambda I) = 0$. This determinant provides a polynomial in the unknown λ (the so-called *characteristic polynomial*) and, not surprisingly, the eigenvalues of M are the roots of $p(\lambda)$. Of course, eigenvalues are the same, whether we are looking at left or right eigenvectors. On the other hand, for a generic square matrix M, left and right eigenvectors associated to the same eigenvalue will differ in general.

The wonderful world of symmetric matrices

Next, assume M is symmetric. A number of marvelous things happen, most of which are very easy to see, though we begin with a property we are not going to prove:

Claim 1. All eigenvalues of a symmetric matrix are real.

Next, we show that left and right eigenvectors are indeed the same.

Claim 2. If $\mathbf{x}$ is a right eigenvector of M with eigenvalue λ, $\mathbf{x}$ is also a left eigenvector for the same eigenvalue.

Proof. We simply transpose the equation $M\mathbf{x} = \lambda\mathbf{x}$ and remember that M is symmetric:

$$M\mathbf{x} = \lambda\mathbf{x} \Rightarrow \mathbf{x}^T M^T = \mathbf{x}^T M = \lambda\mathbf{x}^T.$$

We next prove a very important property, key to the diagonalization of symmetric matrices.

Claim 3. If M is symmetric, eigenvectors associated to different eigenvalues are mutually orthogonal.

Proof. Assume $M\mathbf{x} = \lambda_1 \mathbf{x}$ and $M\mathbf{y} = \lambda_2 \mathbf{y}$, with $\lambda_1 \neq \lambda_2$. Consider $\mathbf{y}^T M\mathbf{x}$. We can expand this expression in two ways. The first time we use $M\mathbf{x} = \lambda_1 \mathbf{x}$, the second time we use **Claim 2**. Namely, symmetry of M implies that, if $M\mathbf{y} = \lambda_2 \mathbf{y}$, $\mathbf{y}^T M = \lambda_2 \mathbf{y}^T$ also holds. In the first case we have:

$$\mathbf{y}^T M\mathbf{x} = \lambda_1 \mathbf{y}^T \mathbf{x}.$$

In the second case we have:

$$\mathbf{y}^T M\mathbf{x} = \lambda_2 \mathbf{y}^T \mathbf{x}.$$

Subtracting this equation from the former we get:

$$0 = (\lambda_1 - \lambda_2)\mathbf{y}^T \mathbf{x}.$$

Since $\lambda_1 \neq \lambda_2$, $\mathbf{y}^T \mathbf{x} = 0$ has to hold, which proves **Claim 3**.

Remark. In the remainder, we assume eigenvectors are normalized, i.e., we assume they have unit 2-norm (this is the Euclidean norm). A basis of unit norm, mutually orthogonal vectors is called an *orthonormal basis*.

Next, define by V the $n \times n$ matrix, whose j-th column is $\mathbf{v}_j$ (assumed normalized). Then, Claim 3 implies $V^T V = I$. This means that V is invertible and $V^{-1} = V^T$, which in turn implies $VV^T = I$. Hence, we have Claim 4:

Claim 4. Assume V is an orthonormal eigenvector basis for a symmetric matrix M. Then,

$$V^T V = VV^T = I.$$

Note that Claim 4 also implies that i) V is invertible and ii) its inverse is V^T.

Remark. Claim 4 immediately follows from Claim 3 if we assume $\lambda_1 \neq \cdots \neq \lambda_n$. If the algebraic of λ_i is $k \geq 2$, we have exactly k linearly independent eigenvectors associated to λ_i, and we can always consider and orthonormal base of their span (using Gram-Schmidt orthogonalization procedure). While we do not prove this case (the proof follows from the fact that the null-space of $M - \lambda_i I$ must have dimension k and from Claim 3 for pairs of distinct eigenvalues). Note that this also applies to the null-space of M itself, i.e., if the eigenvalue 0 has algebraic multiplicity k ($rank(M) = n - k$), we can identify k mutually orthonormal eigenvectors associated to the eigenvalue 0.

Now, using the claims above, we can show that any symmetric matrix M can be diagonalized. This is **Claim 5** below.

Claim 5. Any symmetric matrix M can be written as:

$$M = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and where $\mathbf{v}_i$ denotes the i-th eigenvector (the one associated to λ_i).

Proof. For simplicity, we give the proof for the case $\lambda_1 > \lambda_2 > \cdots > \lambda_n$, i.e., when all eigenvalues are *simple*, but the proof can be easily extended to the case in which one or more eigenvalues have multiplicities larger than 1.

For every $i = 1, \ldots, n$, from Claim 2 we have $\mathbf{v}_i^T M\mathbf{v}_i = \lambda_i$. Define the diagonal matrix Λ, such that $\Lambda_{ii} = \lambda_i$. Further, define by V the $n \times n$ matrix, whose j-th column is $\mathbf{v}_j$. Then, the equations above can be written in matrix form as:

$$V^T MV = \Lambda,$$

If we left-multiply both sides by V and right-multiply by V^T we obtain $M = V\Lambda V^T$, which can be written as (check!):

$$M = V\Lambda V^T = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T.$$

This completes the proof of Claim 5.

Note that if we have k eigenvalues equal to 0 (i.e., $rank(M) = n - k$), we only have $n - k$ terms in the sum above, reflecting M's rank.

Decomposing a rectangular matrix - the SVD

We now possess the necessary tools to show a beautiful decomposition, which applies to any (generally rectangular) matrix $A \in \mathbb{R}^{n \times m}$.

To this purpose, we begin by considering the ($n \times n$-dimensional) matrix AA^T. This matrix is clearly symmetric (you can check this directly, by writing down $(AA^T)_{ij}$ and verifying that $(AA^T)_{ij} = (AA^T)_{ji}$) and, therefore, it admits an orthonormal eigenvector basis. Denote it by $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the corresponding eigenvalues. We next prove a number of claims:

Claim 6. For every i, $\lambda_i \geq 0$.

Proof. We have $AA^T\mathbf{u}_i = \lambda_i \mathbf{u}_i$ by definition, which implies:

$$0 \leq \mathbf{u}_i^T AA^T \mathbf{u}_i = \lambda_i \mathbf{u}_i^T \mathbf{u}_i = \lambda_i \|\mathbf{u}_i\|^2 = \lambda_i.$$

Here, the first inequality follows since $\mathbf{u}_i^T AA^T \mathbf{u}_i$ is simply the (squared) 2-norm of the vector $A\mathbf{u}_i$ and is thus non-negative, while the last equality holds because the $\mathbf{u}_i$'s are unit norm vectors.

Since $\lambda_i \geq 0$ for every i, we set $\lambda_i = \sigma_i^2$ in the remainder, which in turn implies $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2$. We next prove the following

Claim 7. For every i, consider the m-dimensional vector $\mathbf{v}_i = \frac{1}{\sigma_i} A^T \mathbf{u}_i$. Then, $\mathbf{v}_i$ is a unit 2-norm eigenvector of $A^T A$, with eigenvalue $\lambda_i = \sigma_i^2$. Moreover, the $\mathbf{v}_i$'s form an orthonormal basis.

Proof. We begin by showing that $\|\mathbf{v}_i\| = 1$. To this purpose we write:

$$\|\mathbf{v}_i\|^2 = \mathbf{v}_i^T \mathbf{v}_i = \frac{1}{\sigma_i^2} \mathbf{u}_i^T AA^T \mathbf{u}_i = \frac{1}{\sigma_i^2} \mathbf{u}_i^T \mathbf{u}_i \cdot \sigma_i^2 = 1,$$

where to derive the third equality we recall that $\mathbf{u}_i$ is the i-th eigenvector of AA^T.

Next, we have:

$$A^T A \mathbf{v}_i = \frac{1}{\sigma_i} A^T AA^T \mathbf{u}_i = \left(\frac{1}{\sigma_i} A^T \mathbf{u}_i \right) \sigma_i^2 = \sigma_i^2 \mathbf{v}_i,$$

where the first equality follows from the definition of $\mathbf{v}_i$, the second follows since $\mathbf{u}_i$ is an eigenvector of AA^T with eigenvalue σ_i^2, while the third again follows from the definition of $\mathbf{v}_i$. Hence, $\mathbf{v}_i$ is an eigenvector of AA^T, with eigenvalue σ_i^2.

Finally, consider $\mathbf{v}_i^T \mathbf{v}_j$, for $i \neq j$. We have:

$$\mathbf{v}_i^T \mathbf{v}_j = \frac{1}{\sigma_i \sigma_j} \mathbf{u}_i^T AA^T \mathbf{u}_j = \frac{1}{\sigma_i \sigma_j} \sigma_j^2 \mathbf{u}_i^T \mathbf{u}_j = 0.$$

Here, the first equality follows from the definition of the $\mathbf{v}_i$'s, the second follows since $\mathbf{u}_j$ is an eigenvector of AA^T with eigenvalues σ_j^2, while the last follows since the $\mathbf{u}_i$'s form an orthonormal basis. This completes the proof of Claim 7.

Claim 8 (existence of the SVD). Consider any real matrix $A \in \mathbb{R}^{n \times m}$. There exist matrices $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{m \times m}$ such that:

1. U is an orthonormal basis in $\mathbb{R}^n$, V is an orthonormal basis in $\mathbb{R}^m$.
2. $A = U\Sigma V^T$, where Σ is an $r \times r$ diagonal matrix and $\Sigma_{ii} \geq 0$, for every $i = 1, \ldots, r$, with $r \leq \min\{n, m\}$.

Proof. Take U, V to be the matrices whose columns are the left and right singular vectors, respectively. Let Σ be the diagonal matrix, such that $\Sigma_{ii} = \sigma_i$. From Claim 7, we know that U and V are related as follows:

$$\mathbf{v}_i = \frac{1}{\sigma_i} A^T \mathbf{u}_i \Leftrightarrow \sigma_i \mathbf{v}_i = A^T \mathbf{u}_i.$$

We can write the latter equality in compact matrix form as:

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

If we right-multiply by U^T we get:

$$V\Sigma U^T = A^T \Leftrightarrow A = U\Sigma V^T.$$

This completes the proof of Claim 8. You should convince yourself (by checking) that the formula above is equivalent to:

$$A = \sum_{i=1}^{n} \sigma_i \mathrm{u}_i \mathrm{v}_i^T$$

SVD and dimensionality reduction

In this section, we discuss interesting properties of the SVD. In particular, its ability to provide a low rank approximation of a given matrix.

Assume again we have a matrix $A \in \mathbb{R}^{n \times m}$. A will have some rank r, which is equal to the number of non-zero singular values (you may want to figure out why). Say we are interested in the matrix B of rank at most $k \leq r$ (possibly $k << r$) that best approximates A in some sense. This means that B will be an approximation of A in a lower dimensional space (at most k), i.e., a *low rank approximation*. In order to formally define this problem, we need a notion of "distance" between A and B. In the remainder, we consider the *Frobenius norm*. The Frobenius norm of an $n \times m$ matrix M is defined as:

$$\|M\|_F = \sqrt{\sum_i \sum_j M_{ij}^2}$$

At this point, identifying a low rank approximation of A can be formulated as a constrained optimization problem as follows:

$$\begin{gathered}\text{Low rank approximation problem}\\ \min_B \|A - B\|_F^2\\ \text{such that } rank(B) \leq k\end{gathered}$$

Define by A_k the *truncated SVD* of A, namely, the matrix obtained by considering the first k terms in the SVD decomposition of A:

$$A(k) = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

The following Theorem establishes a key property of the SVD:

Theorem 9 (Eckart–Young–Mirsky). $A(k)$ is an optimal solution for the low rank approximation problem.

While we are not proving this theorem here, it is interesting to provide some insight into the error guarantees afforded by choosing $A(k)$. In particular, we next prove the following

Claim 10.

$$\|A - A(k)\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.$$

Proof. We begin by observing that Claim 8 and the definition of $A(k)$ imply $A - A(k) = \sum_{i=k+1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. For the remainder of this proof, for any matrix M, we denote by M_s the s-th row of M. Then, the definitions of Frobenius norm of a matrix and of 2-norm of a vector immediately imply:

$$\|M\|_F^2 = \sum_s \|M_s\|_2^2.$$

We next leverage this observation, together with the fact that, if $\mathbf{w}$ is a *row vector*, $\|\mathbf{w}\|_2^2 = \mathbf{w}\mathbf{w}^T$. If we apply these observations to $\|A - A(k)\|_F^2$ we obtain:

$$\|A - A(k)\|_F^2 = \sum_{s=1}^{n} (A - A(k))_s (A - A(k))_s^T$$

In our case, we have $(A - A(k))_s = \sum_{i=k+1}^{n} \sigma_i \mathbf{u}_i(s) \mathbf{v}_i^T$, hence, for every $s = 1, \ldots, n$:

$$(A - A(k))_s (A - A(k))_s^T = \sum_{i=k+1}^{n} \sum_{j=k+1}^{n} \sigma_i \sigma_j \mathbf{u}_i(s) \mathbf{u}_j(s) \mathbf{v}_i^T \mathbf{v}_j = \sum_{i=k+1}^{n} \sigma_i^2 \mathbf{u}_i(s)^2 \|\mathbf{v}_i\|^2 = \sum_{i=k+1}^{n} \sigma_i^2 \mathbf{u}_i(s)^2,$$

where the second and third inequalities follow because the $\mathbf{v}_i$'s form an orthonormal basis. As a consequence we have:

$$\|A - A(k)\|_F^2 = \sum_{s=1}^{n} (A - A(k))_s (A - A(k))_s^T = \sum_{s=1}^{n} \sum_{i=k+1}^{n} \sigma_i^2 \mathbf{u}_i(s)^2 = \sum_{i=k+1}^{n} \sigma_i^2 \sum_{s=1}^{n} \mathbf{u}_i(s)^2 = \sum_{i=k+1}^{n} \sigma_i^2,$$

where the third equality follows by exchanging sums, while the fourth follows since the $\mathbf{u}_i$'s are unit norm vectors. This completes the proof of Claim 10.

Together, Theorem 9 and Claim 10 tell us that, in some sense, singular values "measure" the strength of the signal carried by each term in the SVD of A. In this respect, choosing $A(k)$ amounts to removing the weakest components in terms of strength. In light of Claim 10, if one asked us to pick a subset $I \subset \{1, \ldots, \min\{n, m\}\}$ of size k, so as to minimize $\|A - \sum_{i \in I} \sigma_i \mathbf{u}_i \mathbf{v}_i^T\|_F^2$, the obvious choice would be $I = \{1, \ldots, k\}$. Consider this as a weaker, yet pretty intuitive version of Theorem 9.

Singular values and explained variance

Very often, people (and `sklearn` documentation) speak of *explained variance*. Let us try to make this notion clearer. Recall that:

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

We next consider the sum of the squared lengths of A's rows onto the j-th right singular vector V_j , i.e., we are interested in $\|AV_j\|^2$ (note that AV_j is a vector with a number of components equal to the number of rows, with the i-th component the projection of the i-th row A_{i*} of A onto V_j). Since the above equation implies $AV = U\Sigma$ we have:

$$AV_j = \sigma_j U_j$$

Hence, if $U_j(i)$ denotes the generic, $i-th$ component of U_j:

$$\|AV_j\|^2 = \sum_i (\sigma_j U_j(i))^2 = \sigma_j^2 \sum_i U_j(i)^2 = \sigma_j^2,$$

where the last equality follows since U_j has unit 2-norm. The left-hand side of the above chain of equalities is called *spread*. It basically measures how much of total length of A's rows is spread along the direction of V_j. If the the data were centered (i.e., if the sum of the entries of each row of A were 0), $\|AV_j\|^2$ would be the statistical variance of the data points (provided these are the rows of A) along direction V_j. This is what happens with PCA, where data are first centered and then SVD is applied, hence the term explained variance.

In this perspective, the first right (left) singular vectors are the directions that maximize the overall statistical variance (or spread, when data are not centered).