Singular Value Decomposition

Kaarel Hänni

August 2023

1 Introducing the SVD

Here's the main thing:

Theorem 1.1. *An $m \times n$ real matrix M can be factored as $M = U\Sigma V^T$, where:*

U is an $m \times m$ orthogonal matrix.

Σ is an $m \times n$ diagonal matrix with $\Sigma_{11} \geq \Sigma_{22} \geq \cdots \geq \Sigma_{\min(m,n),\min(m,n)} \geq 0$.

V is in an $n \times n$ orthogonal matrix.

In words, the SVD is a way to write a matrix as a rotation given by V^T, followed by a (possibly dimension-changing) rescaling of the coordinate axes given by Σ, followed by another rotation given by U.

Here's an example:

$$\begin{bmatrix} 2 & 2 & \sqrt{2}-2 & -2-\sqrt{2} \\ \sqrt{6} & \sqrt{6} & -\sqrt{3}-\sqrt{6} & \sqrt{3}-\sqrt{6} \\ \sqrt{2} & \sqrt{2} & 1-\sqrt{2} & -1-\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} \end{bmatrix} \begin{bmatrix} 4\sqrt{3} & 0 & 0 & 0 \\ 0 & 2\sqrt{3} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Here's some more key terminology:

Terminology 1.1.

- The columns $u_1, \ldots, u_m \in \mathbb{R}^m$ of U are called the *left singular vectors* of A.
- The entries $\Sigma_{ii} =: \sigma_i$ for $i = 1, \ldots, \min(m, n)$ are called the *singular values* of A.
- The columns $v_1, \ldots, v_n \in \mathbb{R}^n$ of V are called the *right singular vectors* of A.

And here's a very useful version of the SVD in terms of the above:

Proposition 1.1.1. $M = U\Sigma V^T = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T$.

Note that $Mv_j = \sigma_i u_i v_i^T v_j = \sigma_j u_j$. So M sends the orthogonal basis v_j to the orthogonal basis u_i, with rescalings σ_i. In other words, every matrix is diagonal if one uses the right bases for the domain and codomain.

2 The SVD makes a lot about the matrix M explicit

- The **rank** of M is the number of nonzero singular values; let's call it k.
- The first k right singular vectors are an orthonormal basis of the **row space** of M.
- The last $n - k$ right singular vectors are an orthonormal basis of the **nullspace** of M. (Thus, the **nullity** of M is $n - k$.)
- The first k left singular vectors are an orthonormal basis of the **column space** of M.
- The last $m - k$ left singular vectors are an orthonormal basis of the **left nullspace** of M.

3 Also, the SVD makes a lot about M as a data matrix explicit

3.1 Approximating a data set with a lower-dimensional one

The SVD provides the best way to approximate a data set with a lower-dimensional data set. Suppose that M is a data matrix, i.e., that a vector data set $w_1, \ldots, w_m \in \mathbb{R}^n$ is placed as the rows of the $m \times n$ matrix M. Then, the top right singular vector v_1 is the direction in which the data set has the highest L^2 norm; i.e., it's a solution to the following optimization problem:

$$\max_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \left\| proj_x w_i \right\|^2 = \max_{x \in \mathbb{R}^n \text{ s.t. } \|x\|=1} x^T M^T M x.$$

More generally, the subspace spanned by the first k right singular vectors has the largest possible sum of L^2 norms of projections of the data set among all subspaces of dimension $\leq k$:

$$\text{span}(v_1, \ldots, v_k) \in \operatorname*{arg\,max}_{A \subseteq \mathbb{R}^n \text{ s.t. } \dim A = k} \sum_{i=1}^{m} \left\| proj_A w_i \right\|^2$$

Equivalently, since

$$\sum_{i=1}^{m} \left\| proj_A w_i \right\|^2 + \sum_{i=1}^{m} \left\| proj_{A^\perp} w_i \right\|^2 = \sum_{i=1}^{m} \left\| w_i \right\|^2 = \text{const},$$

this optimization problem is equivalent to minimizing the *reconstruction error*

$$\sum_{i=1}^{m} \left\| proj_{A^\perp} w_i \right\|^2 = \sum_{i=1}^{m} \left\| w_i - proj_A w_i \right\|^2.$$

Since $\text{proj}_A w_i = \arg\min_{a \in A} \|w_i - a\|^2$, we have $\|w_i - proj_A w_i\| = \min_{a \in A} \|w_i - a\|^2$, so the minimization problem can be rewritten as

$$\min_{A \text{ of dimension } k} \sum_{i=1}^{m} \min_{a \in A} \left\| w_i - a \right\|^2.$$

Restating the minimization problem in terms of a basis $a_1, \ldots, a_k$ of A and coefficients $c_{i1}, \ldots, c_{ik}$ in terms of which to approximate w_i, this optimization problem becomes

$$\min_{a_1,\ldots,a_k \in \mathbb{R}^k} \sum_{i=1}^{m} \min_{c_{i1},\ldots,c_{ik}} \left\| w_i - \sum_{j=1}^{k} c_{ij} a_j \right\|^2 = \min_{a_1,\ldots,a_k \in \mathbb{R}^k} \min_{(c_{ij})_{(i,j) \in [m] \times [k]}} \sum_{i=1}^{m} \left\| w_i - \sum_{j=1}^{k} c_{ij} a_j \right\|^2$$

$$= \min_{a_1,\ldots,a_k \in \mathbb{R}^k} \min_{(c_{ij})_{(i,j) \in [m] \times [k]}} \sum_{i=1}^{m} \sum_{k=1}^{n} \left(w_{ik} - \sum_{j=1}^{k} c_{ij} a_{jk} \right)^2 = \min_{A \text{ a } k \times n \text{ matrix}} \quad \min_{C \text{ an } m \times k \text{ matrix}} \left\| M - CA \right\|_F^2,$$

where $\|M\|_F$ denotes the *Frobenius norm* of M, i.e., the square root of the sum of squares of all entries of M. This is now the problem of finding a low-rank factorization of M that approximates it. There is an analogous optimization problem, important in data science and in interpretability, called *Non-Negative Matrix Factorization (NMF)*,[1] which is exactly the same but with the additional requirement that the entries of each matrix are non-negative,[2] which turns out to be difficult to solve exactly. However, this version of the problem turns out to be easy to solve — an optimum is given in terms of the SVD by A being V^T truncated to just the first k rows, and C being $U\Sigma$ truncated to just the first k columns.

The minimum of the problem above is the same as the minimum of the following problem:

$$\min_{\tilde{M} \text{ of rank } k} \left\| M - \tilde{M} \right\|_F,$$

which is correspondingly solved in terms of the SVD:

[1] Here are two interpretability pieces which use NMF: `https://jalammar.github.io/explaining-transformers/` and `https://arxiv.org/abs/2111.09259`.

[2] For instance, because the data set might be the non-negative activation vectors from a Relu neural net, and because we might want each basis vector to be a possible activation vector (so positive) to be individually interpretable, or because we might want coefficients to plausibly correspond to the presence of some quality in the data which does not have a reasonable 'negative presence'.

Theorem 3.1 (Eckart-Young-Mirsky Theorem (Frobenius norm version))**.** *Among matrices of rank* $\leq k$*, the* $\left\|M - \tilde{M}\right\|_F$ *is minimized at* $\tilde{M} = \sum_{i=1}^{k} \sigma_i u_i v_i^T$*.*

Finally, there is one remaining important variant of SVD for capturing directions in which the data has highest variance: *Principal Component Analysis (PCA)*. PCA solves the analogous reconstruction error minimization problem in case we get to approximate all vectors in our data set by their mean $\mu = \frac{1}{m}\sum_{i=1}^{m} w_i$ by default. That is, the optimization problem solved by PCA is

$$\min_{a_1,\ldots,a_k \in \mathbb{R}^k} \min_{(c_{ij})_{(i,j)\in[m]\times[k]}} \sum_{i=1}^{m}\sum_{k=1}^{n} \left(w_{ik} - \left(\mu + \sum_{j=1}^{k} c_{ij} a_{jk} \right) \right)^2.$$

Mathematically, this is the same as finding the SVD of the dataset $w_i' = w_i - \mu$. So PCA is just de-meaned SVD. After de-meaning, L^2 norm in a direction is variance in that direction, so PCA picks out directions of highest variance in the data.

To conclude: SVD gives the best k-dimensional approximation of a data set, and it captures the directions in which the data has highest L^2 norm. NMF is a variant sometimes used in data science and interpretabilityfootnoteSee `https://kaarelh.github.io/doc/decomposition.pdf` that seeks a subspace of high L^2 norm subject to a non-negativity constraint. for more on how all this connects to interpretability. PCA is another variant often used in data science that seeks directions of high variance in the data.

4 Also also, the SVD makes a lot about M as a linear operator explicit

4.1 Approximating a linear map with a low-rank map

Here's a natural way to measure the oomph of a matrix, called the operator norm:

$\|M\| = \max_{w\in\mathbb{R}^n \text{ s.t. } \|w\|=1} = \|Mw\|$. It turns out that this is just equal to σ_1. What's more, as norms do, this gives us another sense in which to ask for a rank k approximation to M:

$$\min_{\tilde{M} \text{ of rank } k} \left\|M - \tilde{M}\right\|.$$

It turns out that this is also solved by the SVD:

Theorem 4.1 (Eckart-Young-Mirsky Theorem (operator norm version))**.** *Among matrices of rank* $\leq k$*, we have that* $\left\|M - \tilde{M}\right\|$ *is minimized at* $\tilde{M} = \sum_{i=1}^{k} \sigma_i u_i v_i^T$*.*

Proof. We will first show that the operator norm of a matrix is its top singular value:

$$\|M\|^2 = \max_{x \text{ s.t. } \|x\|=1} \|Mx\|^2 = \max_{x \text{ s.t. } \|x\|=1} x^T M^T M x = \max_{x \text{ s.t. } \|x\|=1} x^T V \Sigma^T U^T U \Sigma V^T x$$

$$= \max_{x \text{ s.t. } \|x\|=1} (V^T x)^T \Sigma^T \Sigma (V^T x) = \max_{y \text{ s.t. } \|y\|=1} y^T (\Sigma^T \Sigma) y = \max_{y \text{ s.t. } \|y\|=1} \sum_{i=1}^{n} \sigma_i^2 y_i^2 = \sigma_1^2.$$

Since $M - \tilde{M}$ has SVD $\sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T - \sum_{i=1}^{k} \sigma_i u_i v_i^T = \sum_{i=k+1}^{\min(m,n)} \sigma_i u_i v_i^T$, its top singular value is σ_{k+1}, so its operator norm is σ_{k+1}.

It remains to show that operator norm σ_{k+1} is best possible. Since $\tilde{M}$ has rank k, there must be a linear combination the $k+1$ independent vectors $v_1, \ldots, v_{k+1}$ which is in the kernel of M; let's take it to be $v = \sum_{i=1}^{k+1} c_i v_i$, WLOG with unit norm, so $\sum_{i=1}^{k+1} c_i^2 = 1$. Then

$$\left\|M - \tilde{M}\right\|^2 = \left\|\left(M - \tilde{M}\right) v\right\|^2 = \|Mv\|^2 = \left\|\sum_{i=1}^{k+1} c_i \sigma_i u_i\right\|^2 = \sum_{i=1}^{k+1} c_i^2 \sigma_i^2 \geq \sum_{i=1}^{k+1} c_i^2 \sigma_{k+1}^2 = \sigma_{k+1}^2,$$

which is what we wanted to show. □

So, in at least two very reasonable senses (plausibly in the two most reasonable senses) — the Frobenius norm and the operator norm — the SVD gives the best low-rank approximation to M.

4.2 The pseudoinverse

The pseudoinverse M^+ of a matrix M, defined via SVD, is the best generalization of the inverse to non-invertible matrices. (The only bad thing about it is the unfortunate notation.) For a matrix with SVD $M = U\Sigma V^T$, we have

$$M^+ = V\Sigma^+ U^T,$$

where (the even more unfortunately-denoted, given that it makes the whole thing look circular) Σ^+ stands for Σ^T with each nonzero entry σ_i replaced by $\frac{1}{\sigma_i}$.

If M is an invertible square matrix, then indeed $M^+M = V\Sigma^+U^TU\Sigma V^T = V\Sigma^+\Sigma V^T = VV^T = I$, and, similarly, $MM^+ = I$. You will see on the problem set that the pseudoinverse also generalizes left-inverses and right-inverses.

5 Computing the SVD; or, an existence proof

For some motivation for the construction, note that if we had the SVD $M = U\Sigma V^T$, then $M^TM = V\Sigma^TU^TU\Sigma V^T = V(\Sigma^T\Sigma)V^T = V(\Sigma^T\Sigma)V^{-1}$, and note that $\Sigma^T\Sigma$ is a diagonal matrix. This is expressly an eigendecomposition of the symmetric matrix M^TM. Thus, if the SVD exists, the right singular vectors are eigenvectors of M^TM, and the corresponding singular values are square roots of the corresponding eigenvalues. (This will be the starting point of the existence proof and method of computation later.) And if we have the right singular vectors v_i, then for each nonzero singular value σ_i, the left singular vector u_i can be recovered with $\frac{Av_i}{\sigma_i} = \frac{1}{\sigma_j}\sum_j \sigma_j u_j v_j^T v_i = \frac{\sigma_i u_i}{\sigma_i} = u_i$. This paragraph inspires the following existence proof of the SVD:

Proof of Theorem 1.1.1. Note that M^TM is an $n \times n$ positive-semi-definite matrix, so it has a basis of orthonormal eigenvectors $v_1, \ldots, v_n$ with non-negative eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k > 0 = \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_n,$$

(possibly with $k = 0$ or $k = n$), and thus we can write:

$$M^TM = \sum_{i=1}^{n} \sigma_i^2 v_i v_i^T,$$

where we have defined $\sigma_i = \sqrt{\lambda_i}$. For $1 \leq i \leq k$, we define $u_i = \frac{Mv_i}{\sigma_i}$ (as heuristically motivated above), which implies $u_i\sigma_i = Mv_i$. Note that with $1 \leq i < j \leq k$, we have that u_i and u_j must be orthogonal, because $u_i \cdot u_j = \frac{1}{\sigma_i\sigma_j} v_i M^T M v_j = \frac{\lambda_i}{\sigma_i\sigma_j} v_i \cdot v_j = 0$. The same calculation with $i = j$ implies that $\|u_i\| = 1$, so the vectors $u_1, \ldots, u_m$ are an orthonormal set. Let U be the $m \times k$ matrix with columns u_i, V be the $k \times n$ matrix with rows v_i, and Σ be the $k \times k$ diagonal matrix with entries $\Sigma_{ii} = \sigma_i$. We can write the equation $u_i\sigma_i = Mv_i$ in terms of these matrices as $U\Sigma = MV$. Because the kernel of M^TM is the same as the kernel of M, padding V with $n-k$ more columns $v_{k+1}, \ldots, v_n$ just pads MV with $n-k$ columns of zeros. We also arbitrarily complete U with orthonormal columns into a $m \times m$ matrix, and simultaneously pad Σ with $m - k$ rows of zeros into a $m \times k$ matrix — this leaves $U\Sigma$ unchanged; and then, we pad Σ with $n - k$ columns of zeros to a $m \times k$ matrix —- this pads $U\Sigma$ with $n - k$ columns of zeros as well. Since all this has just padded both $U\Sigma$ and MV with $n - k$ columns of zeros, we still have $U\Sigma = MV$ after this. Multiplying both sides from the right with V^T gives us $U\Sigma V^T = M$, completing the proof. □

Note that this also gives a way to compute the SVD, as long as one knows how to compute eigenvectors.[3]

[3]So by now, one ought to know how to compute the SVD, because one ought to know how to compute eigenvectors :)