A Course in Graduate Algorithms **Updated: 2022**

The Singular Value Decomposition

Bo Waggoner, University of Colorado-Boulder *Lecture 13*

The Singular Value Decomposition (SVD) takes a matrix of data points and breaks it down into components. The SVD can be viewed as a form of dimensionality reduction, as in particular it allows us to approximate the original matrix by a simpler matrix of low rank. It has a number of related applications as well.[1]

Objectives:

- Be comfortable with linear algebra required to state and understand the Singular Value Decomposition.
- Understand the technical definition of the SVD of a matrix.
- Understand the intuition for singular vectors and singular values in terms of both linear algebra and applications.
- Be able to explain how SVD can be applied for principal component analysis, low-rank approximation, and collaborative filtering.

1 Introduction

We have a collection of n vectors, each of which is a data point in $\mathbb{R}^d$. As a running example, suppose that each vector represents one person's movie preferences, where the jth entry in the vector is a number representing how much they like movie j, say, between 1 and 10.

We will collect these vectors as rows of a matrix $A \in \mathbb{R}^{n \times d}$. That is,

$$A_{ij} = \text{how much person } i \text{ likes movie } j.$$

If there are millions of people and thousands of movies, this matrix is extremely large. It would be nice to have a smaller, simpler version of this matrix that still captures its essentials. Having such a simplification or approximation could help with other tasks too, such as filling in missing entries (*matrix completion*) – that is, estimating how much someone will like a movie that they haven't yet seen.

1.1 A simple model

For intuition, imagine a very simple model where there are two main genres of movie: *action* and *comedy*. Each movie $j = 1, \ldots, d$ can be described by two numbers,

- $v_1(j)$ =amount of action,
- $v_2(j)$ =amount of comedy.

Meanwhile, each person $i = 1, \ldots, n$ can be described by two numbers,

- $u_1(i)$ =how much they like action,
- $u_2(i)$ =how much they like comedy.

[1] Presentation inspired by [1], Chapter 3.

Now, suppose that how much person i likes movie j is simply the combination of these factors, i.e.

$$A_{ij} = u_1(i)v_1(j) + u_2(i)v_2(j). \tag{1}$$

We can collect all of the people's summaries as a matrix $U \in \mathbb{R}^{n\times 2}$, where column 1 is u_1 and column 2 is u_2. And we can collect all of the movie summaries as a matrix $V \in \mathbb{R}^{d\times 2}$, where the columns are v_1, v_2. So we can also write (1) as a product of two matrices:

$$A = UV^\intercal, \tag{2}$$

where $V^\intercal$ is the transpose of V, i.e. the matrix whose *rows* are the *columns* of V. This gives us (1), see Exercise 1.

We can interpret the column v_1 as the ratings of a hypothetical person who only likes action, and v_2 the ratings of someone who only likes comedy. Then a given person i's ratings, which are row i of the matrix A, are a linear combination of these, weighted by how much person i likes each genre, $(u_1(j), u_2(j))$. Similarly, we can interpret the column u_1 as the ratings of a hypothetical movie that only consists of action, and u_2 a movie that only contains comedy. A given movie j's ratings, column j of A, are a linear combination of these weighted how much that movie contains each genre, $(v_1(j), v_2(j))$.

Notice that we've taken A, a matrix with billions of entries, and fully described it as a product of matrices with much fewer entries, U and V. This is one of the key properties of SVD.

Exercise 1. Check that, if A is defined by Equation (2), then Equation (1) holds.

1.2 From the simple model to the SVD

It turns out that we can make three assumptions without loss of generality about the columns of U and V: they are unit vectors, they are orthogonal, and they are sorted in a certain order.

First, as long as neither column of U is zero, we can always renormalize them so that they are unit vectors. If $a_1 = \|u_1\|_2$ and $a_2\|u_2\|_2$, then we can let $u_1' = \frac{1}{a_1}u_1$ and $u_2' = \frac{1}{a_2}u_2$.

$$\begin{bmatrix} & U & \end{bmatrix} = \begin{bmatrix} | & | \\ u_1' & u_2' \\ | & | \end{bmatrix} \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}$$

Similarly, we can renormalize the columns of V to be unit vectors as well, i.e. if $b_1 = \|v_1\|_2$ and $b_2 = \|v_2\|_2$, then

$$\begin{bmatrix} & V^\intercal & \end{bmatrix} = \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \end{bmatrix} \begin{bmatrix} — & v_1' & — \\ — & v_2' & — \end{bmatrix}$$

Putting these together, (2) becomes

$$A = U'D(V')^\intercal$$

$$\begin{bmatrix} & A & \end{bmatrix} = \begin{bmatrix} | & | \\ u_1' & u_2' \\ | & | \end{bmatrix} \begin{bmatrix} a_1b_1 & 0 \\ 0 & a_2b_2 \end{bmatrix} \begin{bmatrix} — & v_1' & — \\ — & v_2' & — \end{bmatrix}.$$

Here the diagonal components of D, which are a_1b_1 and a_2b_2, represent how important action and comedy respectively are to the average person's preferences.

Second, we can generally assume that u_1 and u_2 are orthogonal, and similarly with v_1 and v_2. The idea behind this assumption is, if we chose the genre categories correctly, then they are completely independent, e.g. a person's liking for comedy cannot be explained on average by how much they like

action, or a movie's amount of comedy cannot be predicted from its amount of action. If this were not true, we could re-define the genres so that the first genre captures all of the variation in preferences in one "direction", while the second genre captures the variation in an orthogonal "direction".

Third and finally, we can re-sort the genres so that $a_1 b_1 > a_2 b_2$. We re-shuffle the columns of U and V accordingly, e.g. (comedy, action).

Takeaway. We will see that any matrix A can be written in the format above using a number of "genres" equal to the rank[2] of the matrix A. That is, we can always write $A = UDV^\intercal$ where the columns of U are orthonormal[3], the columns of V are orthonormal, and D is a diagonal matrix with positive entries $\sigma_1 = D_{11} > \sigma_2 = D_{22} > \cdots$, etc.. Here (U, D, V) is the **singular value decomposition (SVD)** of A. The values $\sigma_1, \sigma_2, \ldots$ are the **singular values**. The columns of U are the **left singular vectors** and the columns of V are the **right singular vectors**. Next, we will see how to construct the SVD from any given input matrix.

2 Constructing the SVD

Given a matrix $A \in \mathbb{R}^{n \times d}$, not all zeros, we now construct its singular value decomposition. Define

$$v_1 = \underset{\|v\|_2 = 1}{\operatorname{argmax}} \|Av\|_2.$$

In other words, we pick the unit vector that maximizes the total projection of A onto v. Now, let

$$\sigma_1 = \|Av_1\|_2,$$

a measure of how well aligned v_1 is with the rows of A on average. Finally, let

$$u_1 = \frac{1}{\sigma_1} A v_1,$$

i.e. the vector $u_1 \in \mathbb{R}^n$ is a renormalized version of the projection. Note that $u_1(i) = \frac{1}{\sigma_1} \sum_{j=1}^{d} A_{ij} v_1(j)$, or in other words, $u_1(i)$ is a representation of how well row i aligns with v_1.

Next, let

$$v_2 = \underset{\|v\|_2 = 1, v \perp v_1}{\operatorname{argmax}} \|Av\|_2,$$

where $v \perp v_1$ means that v is orthogonal to v_1, i.e. $v \cdot v_1 = 0$. Similarly, let

$$\sigma_2 = \|Av_2\|_2,$$

$$u_2 = \frac{1}{\sigma_2} A v_2.$$

Repeat, where at each step ℓ we pick v_ℓ from the space of unit vectors orthogonal to all of $v_1, \ldots, v_{\ell-1}$. Stop after the step r where we have

$$A = UDV^\intercal,$$

where the columns of U are $u_1, \ldots, u_r$, the columns of V are $v_1, \ldots, v_r$, and D is a diagonal matrix with diagonal entries $D_{ii} = \sigma_i$, for $i = 1, \ldots, r$. Then (U, D, V) is the **singular value decomposition** of A, the vectors $u_1, \ldots, u_r$ are the **left singular vectors**, the vectors $v_1, \ldots, v_r$ are the **right singular vectors**, and the values $\sigma_1, \ldots, \sigma_r$ are the **singular values**.

[2]The rank of a matrix is the maximum number of columns that are linearly independent, i.e. the size of the largest subset of columns where none of them can be written as a linear combination of the others.

[3]A set of vectors are orthonormal if they are all orthogonal and all unit vectors.

Remarks. An equivalent stopping condition is that we stop after r steps if, when we try to take step $r+1$, we find $\sigma_{r+1} = 0$. In other words, the maximum possible projection is zero (or the v_ℓ vectors span all of $\mathbb{R}^d$, i.e. $r = d$, and there is no way to pick an additional vector). We can show that r, the number of singular values, is the *rank* of the matrix A.

It is possible for there to be a tie in choosing some v_ℓ, i.e. there are multiple unit vectors that are equally good. Because of this, there is not necessarily a unique "the" singular value decomposition; there could be many.

Movie-model interpretation. If A_{ij} = how much person i likes movie j, then we can think of a unit vector $v \in \mathbb{R}^d$ as a hypothetical person's rating of all the movies. Let a_i be the ith row of A, i.e. person i's ratings. The dot-product of v with a_i represents represents the similarity. So $\|Av\|_2^2 = \sum_i a_i \cdot v$ is a measure of total similarity of all people in the dataset with v. So v_1, the first right singular vector, defines the "most-representative" hypothetical person.

Meanwhile, $\sigma_1 = \|Av_1\|_2$ is a measure of how well that hypothetical person aligns with everyone's preferences. And $u_1 = Av_1/\|Av_1\|_2$ is a measure of how similar each person in the data set is to v_1, i.e. $u_1(i) = \frac{1}{\sigma_1} a_i \cdot v_1$.

We can interpret this process as defining a "stereotype" or a "genre" of movie, where v_1 captures the ratings of a hypothetical person who only cares about that genre. Then u_1 captures how much each person likes that genre, and σ_1 captures the overall importance of that genre to people's final ratings. A very rough prediction of i's rating for movie j is $A_{ij} \approx \sigma_1 u_1(i) v_1(j)$, i.e. importance multiplied by i's liking for the genre multiplied by how much j contains of the genre.

Similarly, each additional iteration ℓ defines a new "most important genre remaining." Out of the space orthogonal to the preferences already explained by the first $\ell - 1$ genres, it finds a "most-representative" hypothetical set of preferences that align with the ratings in a direction not already explained.

Exercise 2. Show that r, the number of singular values of A, is equal to the rank of A.

3 Low-rank approximation

The *rank* $\text{rank}(A)$ of a matrix A is the size of the smallest set of orthonormal vectors such that every row in the matrix is some linear combination of the vectors. In the movie-rating example, we can picture a set of $r = \text{rank}(A)$ idealized "basis" people who have some stereotypical preferences $v_1, \ldots, v_r$, such that every real person's preferences are a linear combination of some of the basis people's. If the rank of a matrix is small, then every row is easy to explain or generate. Everything in this paragraph also holds true if we replace "row" with "column"; in this case, we can picture idealized "basis" movies $u_1, \ldots, u_r$ such that every movie's ratings are a linear combination of some of the basis movies'.

In reality, however, we don't expect A to always have low rank. There might be some noise or idiosyncracies in preferences so that, even if A has a lot of low rank structure, its actual rank is very high. In this case, we might like to uncover a simplified or idealized version of A with low rank. This version might also be more tractable for performing computations, if it takes much fewer numbers to describe.

Observe (see Exercise 3) that for any matrix A with rank r and SVD (U, D, V), we have

$$A = \sum_{\ell=1}^{r} \sigma_\ell u_\ell v_\ell^\intercal, \tag{3}$$

where $u_\ell v_\ell^\intercal$ is the outer product of the two vectors, i.e. the $n \times d$ matrix whose (i, j) entry is $u_\ell(i) v_\ell(j)$.

In other words, A is the sum of r matrices, each of rank one:

$$A = \sigma_1 u_1 v_1^\intercal + \cdots + \sigma_r u_r v_r^\intercal.$$

To see that each matrix has rank one, note that e.g. the jth column of the matrix $\sigma_1 u_1 v_1^\intercal$ is equal to the vector u_1 multiplied by σ_1 and $v_1(j)$. So each column is a multiple of u_1 (and similarly each row is a multiple of v_1), so the rank of the matrix is one.

Now, the first matrix, $\sigma_1 u_1 v_1^\intercal$ can be viewed as a rank-one approximation of the original matrix A. In fact, it is in a sense the *best possible* rank-one approximation. Similarly, the sum of the first two matrices, $\sigma_1 u_1 v_1^\intercal + \sigma_2 u_2 v_2^\intercal$, is in a sense the best possible rank-two approximation, and so on. The "sense" is the following.

Fact 1. *Let $A = UDV^\intercal$ be the SVD; then for any $k \geq 1$, the matrix*

$$A_k := \sum_{\ell=1}^{k} \sigma_\ell u_\ell v_\ell^\intercal$$

is a solution to

$$\operatorname*{argmin}_{A' : rank(A')=k} \|A - A'\|_F,$$

where $\|B\|_F := \sum_{i,j} B_{ij}^2$ is the Frobenius norm.

Proof. If A' has rank k, then there exists $V \in \mathbb{R}^{d \times k}$ with orthonormal columns $v_1, \ldots, v_k$ such that, for some $W \in \mathbb{R}^{n \times k}$, $A' = WV^\intercal$. To see this, note that the ith row a_i' of A' can be written $a_i' = \sum_{\ell=1}^{k} W_{\ell,i} v_\ell$, so it is a linear combination of the k orthonormal basis vectors. We will continue the proof using several lemmas.

Lemma: Suppose $A' = WV^\intercal$ where the columns of V are orthonormal. For any fixed V, the optimal choice of W is $W = AV$.
Proof of lemma: First of all, using that $\|x\|_2^2 = x \cdot x$,

$$\begin{aligned}\|A - A'\|_F &= \sum_i \|a_i - a_i'\|_2^2 \\ &= \sum_i \left(a_i \cdot a_i - 2a_i \cdot a_i' + a_i' \cdot a_i'\right).\end{aligned}$$

Recalling that the v_ℓ are orthonormal and that $a_i' = \sum_{\ell=1}^{k} W_{i\ell} v_\ell$,

$$\begin{aligned}a_i \cdot a_i - 2a_i \cdot a_i' + a_i' \cdot a_i' &= \|a_i\|_2^2 - 2\sum_\ell W_{i\ell} a_i \cdot v_\ell + \left(\sum_\ell W_{i\ell} v_\ell\right) \cdot \left(\sum_\ell W_{i\ell} v_\ell\right) \\ &= \|a_i\|_2^2 - 2\sum_\ell W_{i\ell} a_i \cdot v_\ell + \sum_\ell W_{i\ell}^2 v_\ell \cdot v_\ell \\ &= \|a_i\|_2^2 - 2\sum_\ell W_{i\ell} a_i \cdot v_\ell + \sum_\ell W_{i\ell}^2.\end{aligned}$$

By taking the derivative with respect to $W_{i\ell}$, we find the optimal choice is $W_{i\ell} = a_i \cdot v_\ell$, so $W = AV$.

Lemma: If we set $W = AV$, then the problem is equivalent to selecting V to maximize $\sum_{\ell=1}^{k} \|Av_\ell\|_2^2$.
Proof of lemma: Fix a row a_i. Let the vector x_ℓ^i be the projection of row i onto v_ℓ, i.e. $x_\ell^i = a_i \cdot v_\ell \frac{v_\ell}{\|v_\ell\|_2} = W_{i\ell} v_\ell$. Notice that

$$\|A - A'\|_F = \sum_i \|a_i - a_i'\|_2^2 = \sum_i \|a_i - \sum_\ell x_\ell^i\|_2^2.$$

We have $a_i = \sum_\ell x_\ell^i + \left(a_i - \sum_\ell x_\ell^i\right)$, and using orthogonality of the x_ℓ, we have (i.e. a Pythagorean Theorem)

$$\|a_i\|_2^2 = \sum_\ell \|x_\ell^i\|_2^2 + \|a_i - \sum_\ell x_\ell^i\|_2^2.$$

The left side doesn't depend on our choice of V, so minimizing $\|a_i - \sum_\ell x_\ell^i\|_2^2$ is equivalent to maximizing $\sum_\ell \|x_\ell^i\|_2^2$. In other words, our problem is to pick orthonormal $v_1, \ldots, v_k$ to maximize

$$\sum_i \sum_\ell \|x_\ell^i\|_2^2 = \sum_{i,\ell} W_{i,\ell}^2 \|v_\ell\|_2^2 = \sum_{i,\ell} W_{i,\ell}^2 = \sum_{i,\ell} (a_i \cdot v_\ell)^2 = \sum_\ell \|Av_\ell\|_2^2.$$

Completing the proof. By the above lemmas, all we have to show is that, over all choices of orthonormal vectors, the SVD maximizes $\sum_{\ell=1}^k \|Av_\ell\|_2^2$. Let $v_1, \ldots, v_k$ be any set of orthonormal vectors, and let $v_1^*, \ldots, v_r^*$ be all of the right singular vectors. Because they form a basis, we can write $v_\ell = \sum_{\ell'} \alpha_{\ell\ell'} v_{\ell'}^*$, where $\alpha_{\ell\ell'} = v_\ell \cdot v_{\ell'}^*$. Observe that $1 = v_\ell \cdot v_\ell = \sum_{\ell'} \alpha_{\ell\ell'}^2$, using orthonormality. Similarly, by extending V to an orthonormal basis $v_1, \ldots, v_r$, we have $v_{\ell'}^* = \sum_\ell \alpha_{\ell\ell'} v_\ell$, and $1 = \sum_{\ell=1}^r \alpha_{\ell\ell'}^2 \geq \sum_{\ell=1}^k \alpha_{\ell\ell'}^2$.

Then, using that the vectors $Av_1^*, \ldots, Av_r^*$ are orthogonal (we claim this without proof; a good exercise):

$$\begin{aligned}
\sum_{\ell=1}^k \|Av_\ell\|_2^2 &= \sum_{\ell=1}^k \|A \sum_{\ell'} \alpha_{\ell\ell'} v_{\ell'}^*\|_2^2 \\
&= \sum_{\ell=1}^k \| \sum_{\ell'} \alpha_{\ell\ell'} Av_{\ell'}^*\|_2^2 \\
&= \sum_{\ell=1}^k \sum_{\ell'} \alpha_{\ell\ell'}^2 \|Av_{\ell'}^*\|_2^2 \\
&= \sum_{\ell'} \left(\sum_{\ell=1}^k \alpha_{\ell\ell'}^2 \right) \|Av_{\ell'}^*\|_2^2.
\end{aligned}$$

Let $\beta_{\ell'} = \sum_{\ell=1}^k \alpha_{\ell\ell'}^2 \leq 1$. We have $\sum_{\ell'} \beta_{\ell'} \leq k$. What choices of $\beta_1, \ldots, \beta_r$ maximize the following sum?

$$\max \sum_{\ell'} \beta_{\ell'} \|Av_{\ell'}^*\|_2^2$$

By definition, $\|Av_1^*\|_2 \geq \cdots \geq \|Av_r^*\|_2$. So the solution is $\beta_1 = \cdots = \beta_k = 1$, $\beta_{k+1} = \cdots = \beta_r = 0$. We conclude that the optimal solution has value $\sum_{\ell'=1}^k \|Av_{\ell'}^*\|_2^2$, which is the value given by the SVD. □

Now that we have proven Fact 1, we know how to produce an optimal rank-k approximation to A:

- Compute the SVD (U, D, V).
- Take the first k singular values to get $U_k \in \mathbb{R}^{n \times k}$, $V_k \in \mathbb{R}^{d \times k}$, and $D_k \in \mathbb{R}^{k \times k}$.
- Compute $A_k = U_k D_k V_k$.

$$A_k = U_k D_k V_k$$

$$\begin{bmatrix} & A_k & \end{bmatrix} = \begin{bmatrix} | & & | \\ | & & | \\ u_1 & \cdots & u_k \\ | & & | \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{bmatrix} \begin{bmatrix} — & — & v_1 & — & — \\ & & \vdots & & \\ — & — & v_k & — & — \end{bmatrix}$$

Exercise 3. Prove that (3) holds.

Hint: Consider, for each i, j, the entry $A_{ij} = (UDV^\intercal)_{ij}$, and show it is equal to the (i, j) entry of the right-hand side of (3).

4 Other Applications

4.1 Principal component analysis (PCA)

The idea of PCA is to capture the key features or dimensions of a data set. It has many applications and variants, beyond the scope of this lecture. Sticking with the movie dataset example, suppose we wish to analyze the dataset to find out key trends that explain people's movie preferences. We compute the SVD (U, D, V) and learn several things:

- The first right singular vector, v_1, represents an idealized "genre" that best describes preferences as a one-dimensional model. Looking at which movies j have larger values of $v_1(j)$ may help us understand qualitatively what members of the "genre" have in common.
- The first left singular vector, u_1, represents people's preferences over this "genre". This may help classify people with a "viewer profile" as those who especially like or dislike it.
- The first singular value, σ_1, measures how well v_1 and u_1 explain the data. The larger it is, the better the explanation, generally.
- We can take additional singular vectors and values to obtain the next-most-important "genres" and "viewer profiles".
- By looking at the singular values $\sigma_2, \ldots$, we can estimate how important each genre is and how well the rank-k approximation explains the data.

4.2 Collaborative filtering (matrix completion)

Suppose we are given a matrix A with some missing entries. In fact, this would typically be the case in a movie ratings dataset, as not every person has seen and rated every movie. We would like to predict the missing entries, which corresponds to guessing how a person would rate a movie if they watched it. This could be the basis for a recommendation algorithm. In fact, the "Netflix Prize" was a million-dollar competition held by Netflix from 2006 to 2009 for the best such algorithm on their data set. The winners used an approach based on the Singular Value Decomposition, though much more sophisticated.

How does SVD help with these matrix completion problems, also known as "collaborative filtering"?

1. Given the matrix A, we first use a simple method to estimate the missing entries. For example, use the average value of the rest of the row, or a simple linear predictor.
2. Call the full matrix with all entries estimated A'.
3. Compute the SVD $A' = UDV^\intercal$.
4. Compute the low-rank approximation $A'_k = U_k D_k V_k^\intercal$.
5. Use the value of $A'_k(i, j)$ to estimate the missing entry i, j.

The idea is that A'_k captures the hidden structure of A. For example, suppose that many people who like movies j and j' also like movie j''. There may be a singular vector u_ℓ that has high values in all three of these entries. Meanwhile, suppose person i likes movie j and j', but hasn't seen movie j''. The low-rank approximation may find that the best estimate for person i comes from a high weight on the singular vector u_ℓ, which will result in a large entry in $A'_k(i, j'')$.

References

[1] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science.* Cambridge University Press, 2020.