4 Singular Value Decomposition (SVD)

The singular value decomposition of a matrix A is the factorization of A into the product of three matrices $A = UDV^T$ where the columns of U and V are orthonormal and the matrix D is diagonal with positive real entries. The SVD is useful in many tasks. Here we mention two examples. First, the rank of a matrix A can be read off from its SVD. This is useful when the elements of the matrix are real numbers that have been rounded to some finite precision. Before the entries were rounded the matrix may have been of low rank but the rounding converted the matrix to full rank. The original rank can be determined by the number of diagonal elements of D not exceedingly close to zero. Second, for a square and invertible matrix A, the inverse of A is $VD^{-1}U^T$.

To gain insight into the SVD, treat the rows of an $n \times d$ matrix A as n points in a d-dimensional space and consider the problem of finding the best k-dimensional subspace with respect to the set of points. Here best means minimize the sum of the squares of the perpendicular distances of the points to the subspace. We begin with a special case of the problem where the subspace is 1-dimensional, a line through the origin. We will see later that the best-fitting k-dimensional subspace can be found by k applications of the best fitting line algorithm. Finding the best fitting line through the origin with respect to a set of points $\{\mathbf{x_i} | 1 \le i \le n\}$ in the plane means minimizing the sum of the squared distances of the points to the line. Here distance is measured perpendicular to the line. The problem is called the *best least squares fit*.

In the best least squares fit, one is minimizing the distance to a subspace. An alternative problem is to find the function that best fits some data. Here one variable y is a function of the variables $x_1, x_2, \ldots, x_d$ and one wishes to minimize the vertical distance, i.e., distance in the y direction, to the subspace of the x_i rather than minimize the perpendicular distance to the subspace being fit to the data.

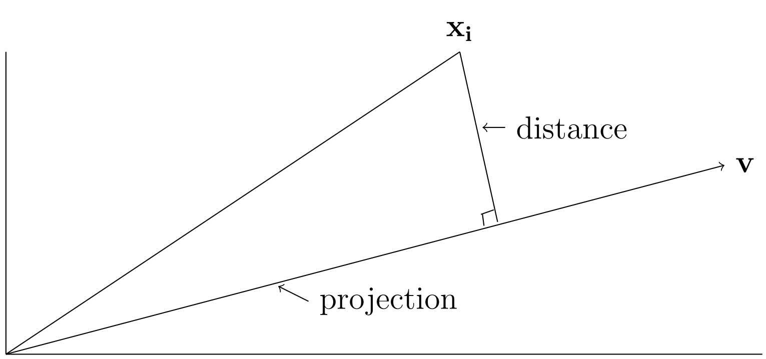

Figure 4.1: The projection of the point $\mathbf{x_i}$ onto the line through the origin in the direction of $\mathbf{v}$

Returning to the best least squares fit problem, consider projecting a point $\mathbf{x_i}$ onto a line through the origin. Then

$$x_{i1}^2 + x_{i2}^2 + \cdots +_{id}^2 = (\text{length of projection})^2 + (\text{distance of point to line})^2 .$$

See Figure 4.1. Thus

$$(\text{distance of point to line})^2 = x_{i1}^2 + x_{i2}^2 + \cdots +_{id}^2 - (\text{length of projection})^2 .$$

To minimize the sum of the squares of the distances to the line, one could minimize $\sum_{i=1}^{n} (x_{i1}^2 + x_{i2}^2 + \cdots +_{id}^2)$ minus the sum of the squares of the lengths of the projections of the points to the line. However, $\sum_{i=1}^{n} (x_{i1}^2 + x_{i2}^2 + \cdots +_{id}^2)$ is a constant (independent of the line), so minimizing the sum of the squares of the distances is equivalent to maximizing the sum of the squares of the lengths of the projections onto the line. Similarly for best-fit subspaces, we could maximize the sum of the squared lengths of the projections onto the subspace instead of minimizing the sum of squared distances to the subspace.

4.1 Singular Vectors

We now define the *singular vectors* of an $n \times d$ matrix A. Consider the rows of A as n points in a d-dimensional space. Consider the best fit line through the origin. Let $\mathbf{v}$ be a unit vector along this line. The length of the projection of $\mathbf{a_i}$, the i^{th} row of A, onto $\mathbf{v}$ is $|\mathbf{a_i} \cdot \mathbf{v}|$. From this we see that the sum of length squared of the projections is $|A\mathbf{v}|^2$. The best fit line is the one maximizing $|A\mathbf{v}|^2$ and hence minimizing the sum of the squared distances of the points to the line.

With this in mind, define the *first singular vector*, $\mathbf{v_1}$, of A, which is a column vector, as the best fit line through the origin for the n points in d-space that are the rows of A. Thus

$$\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|.$$

The value $\sigma_1(A) = |A\mathbf{v_1}|$ is called the *first singular value* of A. Note that σ_1^2 is the sum of the squares of the projections of the points to the line determined by $\mathbf{v_1}$.

The greedy approach to find the best fit 2-dimensional subspace for a matrix A, takes $\mathbf{v_1}$ as the first basis vector for the 2-dimenional subspace and finds the best 2-dimensional subspace containing $\mathbf{v_1}$. The fact that we are using the sum of squared distances helps. For every 2-dimensional subspace containing $\mathbf{v_1}$, the sum of squared lengths of the projections onto the subspace equals the sum of squared projections onto $\mathbf{v_1}$ plus the sum of squared projections along a vector perpendicular to $\mathbf{v_1}$ in the subspace. Thus, instead of looking for the best 2-dimensional subspace containing $\mathbf{v_1}$, look for a unit vector, call it $\mathbf{v_2}$, perpendicular to $\mathbf{v_1}$ that maximizes $|A\mathbf{v}|^2$ among all such unit vectors. Using the same greedy strategy to find the best three and higher dimensional subspaces, defines $\mathbf{v_3}, \mathbf{v_4}, \ldots$ in a similar manner. This is captured in the following definitions. There is no apriori guarantee that the greedy algorithm gives the best fit. But, in fact, the greedy algorithm does work and yields the best-fit subspaces of every dimension.

The *second singular vector*, $\mathbf{v_2}$, is defined by the best fit line perpendicular to $\mathbf{v_1}$

$$\mathbf{v_2} = \arg\max_{\mathbf{v}\perp\mathbf{v_1}, |\mathbf{v}|=1} |A\mathbf{v}| .$$

The value $\sigma_2(A) = |A\mathbf{v_2}|$ is called the *second singular value* of A**.** The *third singular vector* $\mathbf{v}_3$ is defined similarly by

$$\mathbf{v_3} = \arg\max_{\mathbf{v}\perp\mathbf{v_1},\mathbf{v_2}, |\mathbf{v}|=1} |A\mathbf{v}|$$

and so on. The process stops when we have found

$$\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$$

as singular vectors and

$$\arg\max_{\substack{\mathbf{v}\perp\mathbf{v_1},\mathbf{v_2},\ldots,\mathbf{v_r} \\ |\mathbf{v}|=1}} |A\mathbf{v}| = 0.$$

If instead of finding $\mathbf{v_1}$ that maximized $|A\mathbf{v}|$ and then the best fit 2-dimensional subspace containing $\mathbf{v_1}$, we had found the best fit 2-dimensional subspace, we might have done better. This is not the case. We now give a simple proof that the greedy algorithm indeed finds the best subspaces of every dimension.

Theorem 4.1 *Let A be an $n \times d$ matrix where $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are the singular vectors defined above. For $1 \le k \le r$, let V_k be the subspace spanned by $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}$. Then for each k, V_k is the best-fit k-dimensional subspace for A.*

Proof: The statement is obviously true for $k = 1$. For $k = 2$, let W be a best-fit 2-dimensional subspace for A. For any basis $\mathbf{w_1}, \mathbf{w_2}$ of W, $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2$ is the sum of squared lengths of the projections of the rows of A onto W. Now, choose a basis $\mathbf{w}_1, \mathbf{w_2}$ of W so that $\mathbf{w_2}$ is perpendicular to $\mathbf{v_1}$. If $\mathbf{v_1}$ is perpendicular to W, any unit vector in W will do as $\mathbf{w_2}$. If not, choose $\mathbf{w_2}$ to be the unit vector in W perpendicular to the projection of $\mathbf{v_1}$ onto W. Since $\mathbf{v_1}$ was chosen to maximize $|A\mathbf{v_1}|^2$, it follows that $|A\mathbf{w_1}|^2 \le |A\mathbf{v_1}|^2$. Since $\mathbf{v_2}$ was chosen to maximize $|A\mathbf{v_2}|^2$ over all $\mathbf{v}$ perpendicular to $\mathbf{v_1}$, $|A\mathbf{w_2}|^2 \le |A\mathbf{v_2}|^2$. Thus

$$|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 \le |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2.$$

Hence, V_2 is at least as good as W and so is a best-fit 2-dimensional subspace.

For general k, proceed by induction. By the induction hypothesis, V_{k-1} is a best-fit k-1 dimensional subspace. Suppose W is a best-fit k-dimensional subspace. Choose a basis $\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}$ of W so that $\mathbf{w_k}$ is perpendicular to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$. Then

$$|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_k}|^2 \le |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2 + |A\mathbf{w_k}|^2$$

since V_{k-1} is an optimal k -1 dimensional subspace. Since $\mathbf{w_k}$ is perpendicular to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$, by the definition of $\mathbf{v_k}$, $|A\mathbf{w_k}|^2 \leq |A\mathbf{v_k}|^2$. Thus

$$|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_{k-1}}|^2 + |A\mathbf{w_k}|^2 \leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2 + |A\mathbf{v_k}|^2,$$

proving that V_k is at least as good as W and hence is optimal. ■

Note that the n-vector $A\mathbf{v_i}$ is really a list of lengths (with signs) of the projections of the rows of A onto $\mathbf{v_i}$. Think of $|A\mathbf{v_i}| = \sigma_i(A)$ as the "component" of the matrix A along $\mathbf{v_i}$. For this interpretation to make sense, it should be true that adding up the squares of the components of A along each of the $\mathbf{v_i}$ gives the square of the "whole content of the matrix A". This is indeed the case and is the matrix analogy of decomposing a vector into its components along orthogonal directions.

Consider one row, say $\mathbf{a_j}$, of A. Since $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ span the space of all rows of A, $\mathbf{a_j} \cdot \mathbf{v} = 0$ for all $\mathbf{v}$ perpendicular to $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$. Thus, for each row $\mathbf{a_j}$, $\sum\limits_{i=1}^{r} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = |\mathbf{a_j}|^2$. Summing over all rows j,

$$\sum_{j=1}^{n} |\mathbf{a_j}|^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} |A\mathbf{v_i}|^2 = \sum_{i=1}^{r} \sigma_i^2(A).$$

But $\sum\limits_{j=1}^{n} |\mathbf{a_j}|^2 = \sum\limits_{j=1}^{n} \sum\limits_{k=1}^{d} a_{jk}^2$, the sum of squares of all the entries of A. Thus, the sum of squares of the singular values of A is indeed the square of the "whole content of A", i.e., the sum of squares of all the entries. There is an important norm associated with this quantity, the Frobenius norm of A, denoted $||A||_F$ defined as

$$||A||_F = \sqrt{\sum_{j,k} a_{jk}^2}.$$

Lemma 4.2 *For any matrix A, the sum of squares of the singular values equals the Frobenius norm. That is, $\sum \sigma_i^2(A) = ||A||_F^2$.*

Proof: By the preceding discussion. ■

A matrix A can be described fully by how it transforms the vectors $\mathbf{v_i}$. Every vector $\mathbf{v}$ can be written as a linear combination of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ and a vector perpendicular to all the $\mathbf{v_i}$. Now, $A\mathbf{v}$ is the same linear combination of $A\mathbf{v_1}, A\mathbf{v_2}, \ldots, A\mathbf{v_r}$ as $\mathbf{v}$ is of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$. So the $A\mathbf{v_1}, A\mathbf{v_2}, \ldots, A\mathbf{v_r}$ form a fundamental set of vectors associated with A. We normalize them to length one by

$$\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}.$$

The vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_r}$ are called the *left singular vectors* of A. The $\mathbf{v_i}$ are called the *right singular vectors*. The SVD theorem (Theorem 4.5) will fully explain the reason for these terms.

Clearly, the right singular vectors are orthogonal by definition. We now show that the left singular vectors are also orthogonal and that $A = \sum\limits_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$.

Theorem 4.3 *Let A be a rank r matrix. The left singular vectors of A, $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_r}$, are orthogonal.*

Proof: The proof is by induction on r. For $r = 1$, there is only one $\mathbf{u_i}$ so the theorem is trivially true. For the inductive part consider the matrix

$$B = A - \sigma_1 \mathbf{u_1} \mathbf{v_1}^T.$$

The implied algorithm in the definition of singular value decomposition applied to B is identical to a run of the algorithm on A for its second and later singular vectors and singular values. To see this, first observe that $B\mathbf{v_1} = A\mathbf{v_1} - \sigma_1 \mathbf{u_1} \mathbf{v_1^T} \mathbf{v_1} = 0$. It then follows that the first right singular vector, call it $\mathbf{z}$, of B will be perpendicular to $\mathbf{v_1}$ since if it had a component $\mathbf{z_1}$ along $\mathbf{v_1}$, then, $\left| B \frac{\mathbf{z}-\mathbf{z_1}}{|\mathbf{z}-\mathbf{z_1}|} \right| = \frac{|B\mathbf{z}|}{|\mathbf{z}-\mathbf{z_1}|} > |B\mathbf{z}|$, contradicting the arg max definition of z. But for any $\mathbf{v}$ perpendicular to $\mathbf{v_1}$, $B\mathbf{v} = A\mathbf{v}$. Thus, the top singular vector of B is indeed a second singular vector of A. Repeating this argument shows that a run of the algorithm on B is the same as a run on A for its second and later singular vectors. This is left as an exercise.

Thus, there is a run of the algorithm that finds that B has right singular vectors $\mathbf{v_2}, \mathbf{v_3}, \ldots, \mathbf{v_r}$ and corresponding left singular vectors $\mathbf{u_2}, \mathbf{u_3}, \ldots, \mathbf{u_r}$. By the induction hypothesis, $\mathbf{u_2}, \mathbf{u_3}, \ldots, \mathbf{u_r}$ are orthogonal.

It remains to prove that $\mathbf{u_1}$ is orthogonal to the other $\mathbf{u_i}$. Suppose not and for some $i \geq 2$, $\mathbf{u_1}^T \mathbf{u_i} \neq 0$. Without loss of generality assume that $\mathbf{u_1}^T \mathbf{u_i} > 0$. The proof is symmetric for the case where $\mathbf{u_1}^T \mathbf{u_i} < 0$. Now, for infinitesimally small $\varepsilon > 0$, the vector

$$A\left(\frac{\mathbf{v}_1 + \varepsilon \mathbf{v_i}}{|\mathbf{v_1} + \varepsilon \mathbf{v_i}|}\right) = \frac{\sigma_1 \mathbf{u_1} + \varepsilon \sigma_i \mathbf{u_i}}{\sqrt{1+\varepsilon^2}}$$

has length at least as large as its component along $\mathbf{u_1}$ which is

$$\mathbf{u_1^T}\left(\frac{\sigma_1 \mathbf{u_1} + \varepsilon \sigma_i \mathbf{u_i}}{\sqrt{1+\varepsilon^2}}\right) = \left(\sigma_1 + \varepsilon \sigma_i \mathbf{u_1}^T \mathbf{u_i}\right) \left(1 - \tfrac{\varepsilon^2}{2} + O\left(\varepsilon^4\right)\right) = \sigma_1 + \varepsilon \sigma_i \mathbf{u_1}^T \mathbf{u_i} - O\left(\varepsilon^2\right) > \sigma_1$$

a contradiction. Thus, $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_r}$ are orthogonal. ■

4.2 Singular Value Decomposition (SVD)

Let A be an $n \times d$ matrix with singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ and corresponding singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$. Then $\mathbf{u_i} = \frac{1}{\sigma_i} A\mathbf{v_i}$, for $i = 1, 2, \ldots, r$, are the left singular vectors and by Theorem 4.5, A can be decomposed into a sum of rank one matrices as

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$

We first prove a simple lemma stating that two matrices A and B are identical if $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. The lemma states that in the abstract, a matrix A can be viewed as a transformation that maps vector $\mathbf{v}$ onto $A\mathbf{v}$.

Lemma 4.4 *Matrices A and B are identical if and only if for all vectors $\mathbf{v}$, $A\mathbf{v} = B\mathbf{v}$.*

Proof: Clearly, if $A = B$ then $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. For the converse, suppose that $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. Let $\mathbf{e_i}$ be the vector that is all zeros except for the i^{th}component which has value 1. Now $A\mathbf{e_i}$ is the i^{th} column of A and thus $A = B$ if for each i, $A\mathbf{e_i} = B\mathbf{e_i}$.

■

Theorem 4.5 *Let A be an $n \times d$ matrix with right singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$, left singular vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_r}$, and corresponding singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$. Then*

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$

Proof: For each singular vector $\mathbf{v_j}$, $A\mathbf{v_j} = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T \mathbf{v_j}$. Since any vector $\mathbf{v}$ can be expressed as a linear combination of the singular vectors plus a vector perpendicular to the $\mathbf{v_i}$, $A\mathbf{v} = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T \mathbf{v}$ and by Lemma 4.4, $A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$.

The decomposition is called the *singular value decomposition*, *SVD*, of A. In matrix notation $A = UDV^T$ where the columns of U and V consist of the left and right singular vectors, respectively, and D is a diagonal matrix whose diagonal entries are the singular values of A.

For any matrix A, the sequence of singular values is unique and if the singular values are all distinct, then the sequence of singular vectors is unique also. However, when some set of singular values are equal, the corresponding singular vectors span some subspace. Any set of orthonormal vectors spanning this subspace can be used as the singular vectors.

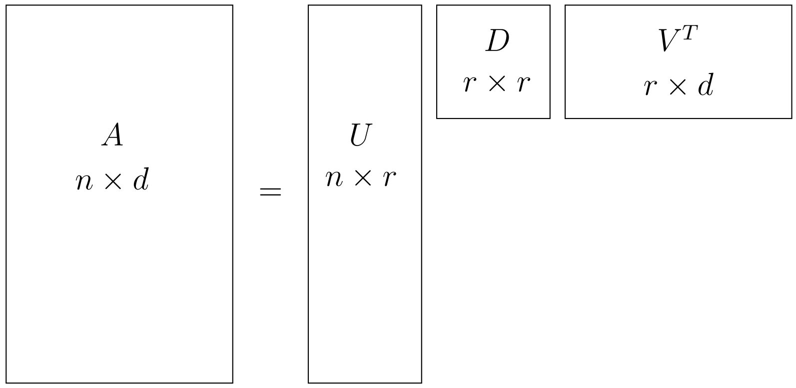

Figure 4.2: The SVD decomposition of an $n \times d$ matrix.

4.3 Best Rank k Approximations

There are two important matrix norms, the Frobenius norm denoted $||A||_F$ and the 2-norm denoted $||A||_2$. The 2-norm of the matrix A is given by

$$\max_{|\mathbf{v}|=1} |A\mathbf{v}|$$

and thus equals the largest singular value of the matrix.

Let A be an $n \times d$ matrix and think of the rows of A as n points in d-dimensional space. The Frobenius norm of A is the square root of the sum of the squared distance of the points to the origin. The 2-norm is the square root of the sum of squared distances to the origin along the direction that maximizes this quantity.

Let

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

be the SVD of A. For $k \in \{1, 2, \ldots, r\}$, let

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

be the sum truncated after k terms. It is clear that A_k has rank k. Furthermore, A_k is the best rank k approximation to A when the error is measured in either the 2-norm or the Frobenius norm.

Lemma 4.6 *The rows of A_k are the projections of the rows of A onto the subspace V_k spanned by the first k singular vectors of A.*

Proof: Let $\mathbf{a}$ be an arbitrary row vector. Since the $\mathbf{v_i}$ are orthonormal, the projection of the vector a onto V_k is given by $\sum_{i=1}^{k} (\mathbf{a} \cdot \mathbf{v_i})\mathbf{v_i}^T$. Thus, the matrix whose rows are the projections of the rows of A onto V_k is given by $\sum_{i=1}^{k} A\mathbf{v_i}\mathbf{v_i}^T$. This last expression simplifies to

$$\sum_{i=1}^{k} A\mathbf{v_i}\mathbf{v_i}^T = \sum_{i=1}^{k} \sigma_i \mathbf{u_i}\mathbf{v_i}^T = A_k.$$

■

The matrix A_k is the best rank k approximation to A in both the Frobenius and the 2-norm. First we show that the matrix A_k is the best rank k approximation to A in the Frobenius norm.

Theorem 4.7 *For any matrix B of rank at most k*

$$\|A - A_k\|_F \leq \|A - B\|_F$$

Proof: Let B minimize $\|A - B\|_F^2$ among all rank k or less matrices. Let V be the space spanned by the rows of B. The dimension of V is at most k. Since B minimizes $\|A - B\|_F^2$, it must be that each row of B is the projection of the corresponding row of A onto V, otherwise replacing the row of B with the projection of the corresponding row of A onto V does not change V and hence the rank of B but would reduce $\|A - B\|_F^2$. Since each row of B is the projection of the corresponding row of A, it follows that $\|A - B\|_F^2$ is the sum of squared distances of rows of A to V. Since A_k minimizes the sum of squared distance of rows of A to any k-dimensional subspace, it follows that $\|A - A_k\|_F \leq \|A - B\|_F$. ■

Next we tackle the 2-norm. We first show that the square of the 2-norm of $A - A_k$ is the square of the $(k+1)^{st}$ singular value of A,

Lemma 4.8 $\|A - A_k\|_2^2 = \sigma_{k+1}^2$.

Proof: Let $A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i}\mathbf{v_i}^T$ be the singular value decomposition of A. Then $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u_i}\mathbf{v_i}^T$ and $A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u_i}\mathbf{v_i}^T$. Let $\mathbf{v}$ be the top singular vector of $A - A_k$. Express $\mathbf{v}$ as a linear combination of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$. That is, write $\mathbf{v} = \sum_{i=1}^{r} \alpha_i \mathbf{v_i}$. Then

$$|(A - A_k)\mathbf{v}| = \left| \sum_{i=k+1}^{r} \sigma_i \mathbf{u_i}\mathbf{v_i}^T \sum_{j=1}^{r} \alpha_j \mathbf{v_j} \right| = \left| \sum_{i=k+1}^{r} \alpha_i \sigma_i \mathbf{u_i}\mathbf{v_i}^T \mathbf{v_i} \right|$$
$$= \left| \sum_{i=k+1}^{r} \alpha_i \sigma_i \mathbf{u_i} \right| = \sqrt{\sum_{i=k+1}^{r} \alpha_i^2 \sigma_i^2}.$$

The $\mathbf{v}$ maximizing this last quantity, subject to the constraint that $|\mathbf{v}|^2 = \sum\limits_{i=1}^{r} \alpha_i^2 = 1$, occurs when $\alpha_{k+1} = 1$ and the rest of the α_i are 0. Thus, $\|A - A_k\|_2^2 = \sigma_{k+1}^2$ proving the lemma. ∎

Finally, we prove that A_k is the best rank k 2-norm approximation to A.

Theorem 4.9 *Let A be an $n \times d$ matrix. For any matrix B of rank at most k*

$$\|A - A_k\|_2 \leq \|A - B\|_2$$

Proof: If A is of rank k or less, the theorem is obviously true since $\|A - A_k\|_2 = 0$. Thus assume that A is of rank greater than k. By Lemma 4.8, $\|A - A_k\|_2^2 = \sigma_{k+1}^2$. Now suppose there is some matrix B of rank at most k such that B is a better 2-norm approximation to A than A_k. That is, $\|A - B\|_2 < \sigma_{k+1}$. The null space of B, $\text{Null}\,(B)$, (the set of vectors $\mathbf{v}$ such that $B\mathbf{v} = 0$) has dimension at least $d - k$. Let $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k+1}}$ be the first $k+1$ singular vectors of A. By a dimension argument, it follows that there exists a $\mathbf{z} \neq 0$ in

$$\text{Null}\,(B) \cap \text{Span}\,\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k+1}}\}\,.$$

Scale $\mathbf{z}$ so that $|\mathbf{z}| = 1$. We now show that for this vector $\mathbf{z}$, which lies in the space of the first $k+1$ singular vectors of A, that $(A - B)\,\mathbf{z} \geq \sigma_{k+1}$. Hence the 2-norm of $A - B$ is at least σ_{k+1} contradicting the assumption that $\|A - B\|_2 < \sigma_{k+1}$. First

$$\|A - B\|_2^2 \geq |(A - B)\,\mathbf{z}|^2\,.$$

Since $B\mathbf{z} = 0$,

$$\|A - B\|_2^2 \geq |A\mathbf{z}|^2\,.$$

Since $\mathbf{z}$ is in the Span $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k+1}}\}$

$$|A\mathbf{z}|^2 = \left|\sum_{i=1}^{n} \sigma_i \mathbf{u_i} \mathbf{v_i}^T \mathbf{z}\right|^2 = \sum_{i=1}^{n} \sigma_i^2 \left(\mathbf{v_i}^T \mathbf{z}\right)^2 = \sum_{i=1}^{k+1} \sigma_i^2 \left(\mathbf{v_i}^T \mathbf{z}\right)^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} \left(\mathbf{v_i}^T \mathbf{z}\right)^2 = \sigma_{k+1}^2.$$

It follows that

$$\|A - B\|_2^2 \geq \sigma_{k+1}^2$$

contradicting the assumption that $||A - B||_2 < \sigma_{k+1}$. This proves the theorem. ∎

4.4 Power Method for Computing the Singular Value Decomposition

Computing the singular value decomposition is an important branch of numerical analysis in which there have been many sophisticated developments over a long period of time. Here we present an "in-principle" method to establish that the approximate SVD of a matrix A can be computed in polynomial time. The reader is referred to numerical analysis texts for more details. The method we present, called the Power Method, is conceptually simple. The word power refers to taking high powers of the matrix $B = AA^T$. If the SVD of A is $\sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, then by direct multiplication

$$\begin{aligned} B = AA^T &= \left(\sum_i \sigma_i \mathbf{u_i} \mathbf{v_i}^T \right) \left(\sum_j \sigma_j \mathbf{v_j} \mathbf{u_j}^T \right) \\ &= \sum_{i,j} \sigma_i \sigma_j \mathbf{u_i} \mathbf{v_i}^T \mathbf{v_j} \mathbf{u_j}^T = \sum_{i,j} \sigma_i \sigma_j \mathbf{u_i} (\mathbf{v}_i^T \cdot \mathbf{v_j}) \mathbf{u_j}^T \\ &= \sum_i \sigma_i^2 \mathbf{u_i} \mathbf{u_i}^T, \end{aligned}$$

since $\mathbf{v_i}^T \mathbf{v_j}$ is the dot product of the two vectors and is zero unless $i = j$. [Caution: $\mathbf{u_i} \mathbf{u_j}^T$ is a matrix and is not zero even for $i \neq j$.] Using the same kind of calculation,

$$B^k = \sum_i \sigma_i^{2k} \mathbf{u_i} \mathbf{u_i}^T.$$

As k increases, for $i > 1$, $\sigma_i^{2k} / \sigma_1^{2k}$ goes to zero and B^k is approximately equal to

$$\sigma_1^{2k} \mathbf{u_1} \mathbf{u_1}^T$$

provided that for each $i > 1$, $\sigma_i(A) < \sigma_1(A)$.

This suggests a way of finding σ_1 and $\mathbf{u_1}$, by successively powering B. But there are two issues. First, if there is a significant gap between the first and second singular values of a matrix, then the above argument applies and the power method will quickly converge to the first left singular vector. Suppose there is no significant gap. In the extreme case, there may be ties for the top singular value. Then the above argument does not work. We overcome this problem in Theorem 4.11 below which states that even with ties, the power method converges to some vector in the span of those singular vectors corresponding to the "nearly highest" singular values.

A second issue is that computing B^k costs k matrix multiplications when done in a straight-forward manner or $O(\log k)$ when done by successive squaring. Instead we compute

$$B^k \mathbf{x}$$

where $\mathbf{x}$ is a random unit length vector. Each increase in k requires a matrix-vector product which takes time proportional to the number of nonzero entries in B. Further saving may be achieved by writing

$$B^k \mathbf{x} = AA^T \left(B^{k-1} \mathbf{x} \right).$$

Now the cost is proportional to the number of nonzero entries in A. Since $B^k \mathbf{x} \approx \sigma_1^{2k} \mathbf{u_1} (\mathbf{u}_1^T \cdot \mathbf{x})$ is a scalar multiple of $\mathbf{u_1}$, $\mathbf{u_1}$ can be recovered from $B^k \mathbf{x}$ by normalization.

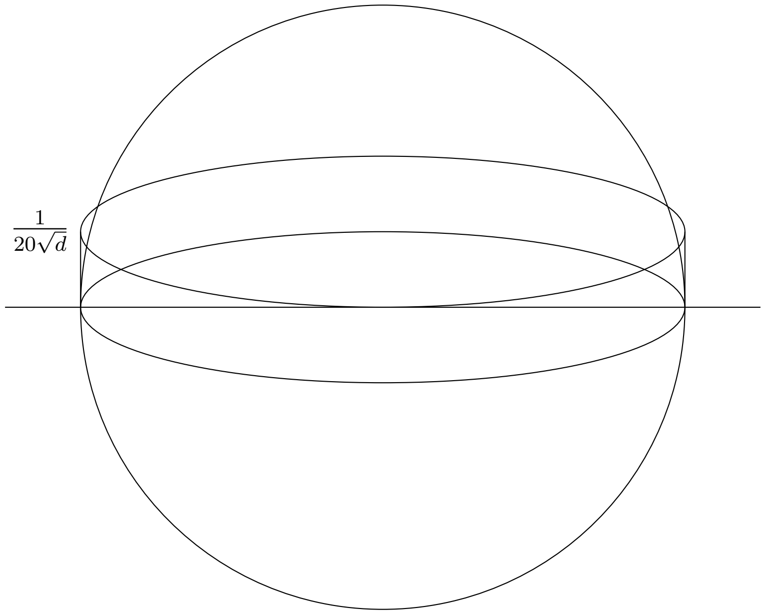

Figure 4.3: The volume of the cylinder of height $\frac{1}{20\sqrt{d}}$ is an upper bound on the volume of the hemisphere below $x_1 = \frac{1}{20\sqrt{d}}$

We start with a technical Lemma needed in the proof of the theorem.

Lemma 4.10 *Let $(x_1, x_2, \ldots, x_d)$ be a unit d-dimensional vector picked at random. The probability that $|x_1| \geq \frac{1}{20\sqrt{d}}$ is at least 9/10.*

Proof: We first show that for a vector $\mathbf{v}$ picked at random with $|\mathbf{v}| \leq 1$, the probability that $v_1 \geq \frac{1}{20\sqrt{d}}$ is at least $9/10$. Then we let $\mathbf{x} = \mathbf{v}/|\mathbf{v}|$. This can only increase the value of v_1, so the result follows.

Let $\alpha = \frac{1}{20\sqrt{d}}$. The probability that $|v_1| \geq \alpha$ equals one minus the probability that $|v_1| \leq \alpha$. The probability that $|v_1| \leq \alpha$ is equal to the fraction of the volume of the unit sphere with $|v_1| \leq \alpha$. To get an upper bound on the volume of the sphere with $|v_1| \leq \alpha$, consider twice the volume of the unit radius cylinder of height α. The volume of the portion of the sphere with $|v_1| \leq \alpha$ is less than or equal to $2\alpha A(d-1)$ and

$$\text{Prob}(|v_1| \leq \alpha) \leq \frac{2\alpha A(d-1)}{V(d)}$$

Now the volume of the unit radius sphere is at least twice the volume of the cylinder of height $\frac{1}{\sqrt{d-1}}$ and radius $\sqrt{1 - \frac{1}{d-1}}$ or

$$V(d) \geq \frac{2}{\sqrt{d-1}} V(d-1)(1 - \frac{1}{d-1})^{\frac{d-2}{2}}$$

Using $(1-x)^a \geq 1-ax$

$$V(d) \geq \frac{2}{\sqrt{d-1}} A(d-1)(1-\frac{d-2}{2}\frac{1}{d-1}) \geq \frac{V(d-1)}{\sqrt{d-1}}$$

and

$$\text{Prob}(|v_1| \leq \alpha) \leq \frac{2\alpha V(d-1)}{\frac{1}{\sqrt{d-1}} V(d-1)} \leq \frac{\sqrt{d-1}}{10\sqrt{d}} \leq \frac{1}{10}.$$

Thus the probability that $v_1 \geq \frac{1}{20\sqrt{d}}$ is at least 9/10. ■

Theorem 4.11 *Let A be an $n \times d$ matrix and $\mathbf{x}$ a random unit length vector. Let V be the space spanned by the left singular vectors of A corresponding to singular values greater than $(1-\varepsilon)\,\sigma_1$. Let k be $\Omega\left(\frac{\ln(n/\varepsilon)}{\varepsilon}\right)$. Let $\mathbf{w}$ be unit vector after k iterations of the power method, namely,*

$$\mathbf{w} = \frac{\left(AA^T\right)^k \mathbf{x}}{\left|\left(AA^T\right)^k \mathbf{x}\right|}.$$

The probability that $\mathbf{w}$ has a component of at least ϵ perpendicular to V is at most 1/10.

Proof: Let

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

be the SVD of A. If the rank of A is less than n, then complete $\{\mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u}_r\}$ into a basis $\{\mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u}_n\}$ of n-space. Write $\mathbf{x}$ in the basis of the $\mathbf{u_i}'s$ as

$$\mathbf{x} = \sum_{i=1}^{n} c_i \mathbf{u_i}.$$

Since $(AA^T)^k = \sum\limits_{i=1}^{n} \sigma_i^{2k} \mathbf{u_i} \mathbf{u_i}^T$, it follows that $(AA^T)^k \mathbf{x} = \sum\limits_{i=1}^{n} \sigma_i^{2k} c_i \mathbf{u_i}$. For a random unit length vector $\mathbf{x}$ picked independent of A, the $\mathbf{u_i}$ are fixed vectors and picking $\mathbf{x}$ at random is equivalent to picking random c_i. From Lemma 4.10, $|c_1| \geq \frac{1}{20\sqrt{n}}$ with probability at least 9/10.

Suppose that $\sigma_1, \sigma_2, \ldots, \sigma_m$ are the singular values of A that are greater than or equal to $(1-\varepsilon)\,\sigma_1$ and that $\sigma_{m+1}, \ldots, \sigma_n$ are the singular values that are less than $(1-\varepsilon)\,\sigma_1$. Now

$$|(AA^T)^k \mathbf{x}|^2 = \left| \sum_{i=1}^{n} \sigma_i^{2k} c_i \mathbf{u_i} \right|^2 = \sum_{i=1}^{n} \sigma_i^{4k} c_i^2 \geq \sigma_1^{4k} c_1^2 \geq \frac{1}{400n} \sigma_1^{4k},$$

with probability at least 9/10. Here we used the fact that a sum of positive quantities is at least as large as its first element and the first element is greater than or equal to