EE16B - Spring'20 - Lecture 11B Notes[1]

Murat Arcak

2 April 2020

Applications of SVD

Last time we wrote the SVD for the $m \times n$ matrix A as

$$A = U_1 S V_1^T \tag{1}$$

where $U_1 = \begin{bmatrix}\vec{u}_1 \cdots \vec{u}_r\end{bmatrix}$ is $m \times r$, $V_1 = \begin{bmatrix}\vec{v}_1 \cdots \vec{v}_r\end{bmatrix}$ is $n \times r$, and S is the $r \times r$ diagonal matrix with entries $\sigma_1, \ldots, \sigma_r$:

$$S = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}.$$

We also formed the square and orthonormal matrices $U = [U_1\ U_2]$, $V = [V_1\ V_2]$, and rewrote (1) as

$$A = U \Sigma V^T \tag{2}$$

where Σ is $m \times n$ and subsumes S in its $r \times r$ upper left block:

$$\Sigma = \begin{bmatrix} S & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix}.$$

Before discussing some applications of SVD we note that the columns of U_1 in (1) form an orthonormal basis for the column space of A. Similarly the columns of V_2 span the null space of A. This is because they are orthogonal to the columns of V_1, so $V_1^T \vec{x} = 0$ for any $\vec{x}$ that is in the column space of V_2 and, thus, (1) implies $A\vec{x} = 0$.

Least Squares with SVD

Recall that in Least Squares we consider the equation

$$\vec{y} = A\vec{x} + \vec{e}$$

where $\vec{y} \in \mathbb{R}^m$ represents measurements, $\vec{x} \in \mathbb{R}^n$ unknowns, $\vec{e}$ errors.

Typically there are more measurements than unknowns ($m > n$) so $A \in \mathbb{R}^{m \times n}$ is a tall matrix. We also assume A has linearly independent columns, so the rank is $r = n$. Then the SVD of A has the form

$$A = U \underbrace{\begin{bmatrix} S \\ 0_{(m-n) \times n} \end{bmatrix}}_{= \Sigma} V^T. \tag{3}$$

[1] Licensed under a

The goal in Least Squares is to find $\vec{x}$ such that $\vec{e} = \vec{y} - A\vec{x}$ has the least possible length. Substituting the SVD (3) for A, note

$$\|\vec{e}\| = \|\vec{y} - A\vec{x}\| = \left\| \vec{y} - U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T \vec{x} \right\|.$$

Since $UU^T = I$ we can replace $\vec{y}$ in this expression with $UU^T\vec{y}$, so that we can factor out U:

$$\|\vec{e}\| = \left\| UU^T\vec{y} - U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T \vec{x} \right\| = \left\| U \left(U^T\vec{y} - \begin{bmatrix} S \\ 0 \end{bmatrix} V^T \vec{x} \right) \right\|. \tag{4}$$

Remembering that multiplication with the orthonormal matrix U does not change the length of a vector, we conclude

$$\|\vec{e}\| = \left\| U^T\vec{y} - \begin{bmatrix} S \\ 0 \end{bmatrix} V^T \vec{x} \right\|.$$

Next note

$$U^T\vec{y} - \begin{bmatrix} S \\ 0 \end{bmatrix} V^T \vec{x} = \begin{bmatrix} U_1^T\vec{y} \\ U_2^T\vec{y} \end{bmatrix} - \begin{bmatrix} SV^T\vec{x} \\ 0 \end{bmatrix} = \begin{bmatrix} U_1^T\vec{y} - SV^T\vec{x} \\ U_2^T\vec{y} \end{bmatrix} \tag{5}$$

and our goal is to minimize the length of this vector by choosing $\vec{x}$. The solution is apparent: since S and V^T have inverses S^{-1} and V, we can zero out the top component by choosing:

$$\boxed{\vec{x} = VS^{-1}U_1^T\vec{y}.} \tag{6}$$

There is nothing we can do about the bottom component $U_2^T\vec{y}$, since $\vec{x}$ does not appear there. Therefore, (6) will minimize the norm of (5).

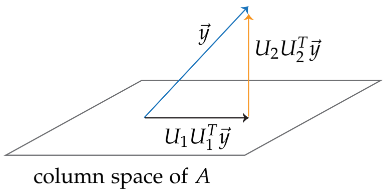
The solution (6) is identical to the familiar Least Squares formula

$$\vec{x} = (A^TA)^{-1}A^T\vec{y}. \tag{7}$$

You can verify this equivalence by substituting (3) in (7), which should give (6) after a little algebra.

The advantage of (6) is the transparency with which we obtained it and the geometric insight it gives. When we substitute $\vec{y} = UU^T\vec{y}$ in (4) we implicitly split $\vec{y}$ into two components:

$$\vec{y} = UU^T\vec{y} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \vec{y} = U_1U_1^T\vec{y} + U_2U_2^T\vec{y}.$$



The first component, $U_1U_1^T\vec{y}$, is the projection of $\vec{y}$ onto the column space of A. This is because the columns of $U_1 = [\vec{u}_1 \;\cdots\; \vec{u}_r]$ form an orthonormal basis for the column space of A. The second component, $U_2U_2^T\vec{y}$, is the remaining part of $\vec{y}$ that is orthogonal to the column space. The Least Squares solution (6) simply matches $A\vec{x}$ to the first component, $U_1U_1^T\vec{y}$, which lies within the column space of A.

Minimum Norm Solution

Above we studied an "overdetermined" problem with more equations than unknowns. We now consider the "underdetermined" equation

$$\vec{y} = A\vec{x} \tag{8}$$

where $\vec{y} \in \mathbb{R}^m$ has smaller dimension than $\vec{x} \in \mathbb{R}^n$; that is $A \in \mathbb{R}^{m\times n}$ is a wide matrix. We assume it has linearly independent rows ($r = m < n$), which means there are infinitely many solutions for $\vec{x}$.

With so many choices for $\vec{x}$ we may want to pick the one with the smallest length. To do so we substitute the SVD

$$A = U \underbrace{\begin{bmatrix} S & 0_{m\times(n-m)} \end{bmatrix}}_{=\Sigma} V^T \tag{9}$$

and write

$$\vec{y} = A\vec{x} = U\begin{bmatrix} S & 0 \end{bmatrix} V^T\vec{x} = U\begin{bmatrix} S & 0 \end{bmatrix}\begin{bmatrix} V_1^T\vec{x} \\ V_2^T\vec{x} \end{bmatrix} = USV_1^T\vec{x}.$$

Since S and U have inverses S^{-1} and U^T, it follows that

$$V_1^T\vec{x} = S^{-1}U^T\vec{y}. \tag{10}$$

Any $\vec{x}$ satisfying (10) solves (8), but which solution has the least norm? Recall that multiplication with V^T does not change the norm, so

$$\|\vec{x}\| = \|V^T\vec{x}\| = \left\| \begin{bmatrix} V_1^T\vec{x} \\ V_2^T\vec{x} \end{bmatrix} \right\|$$

The first component, $V_1^T\vec{x}$, is fixed by (10). The second component, $V_2^T\vec{x}$, is free and we set it to zero so the norm above is minimized. Thus, the minimum norm solution for $\vec{x}$ is given by

$$V^T\vec{x} = \begin{bmatrix} V_1^T\vec{x} \\ V_2^T\vec{x} \end{bmatrix} = \begin{bmatrix} S^{-1}U^T\vec{y} \\ 0 \end{bmatrix}. \tag{11}$$

Since the inverse of V^T is $V = [V_1\ V_2]$, (11) implies

$$\vec{x} = [V_1\ V_2]\begin{bmatrix} S^{-1}U^T\vec{y} \\ 0 \end{bmatrix} \tag{12}$$

or, equivalently,

$$\boxed{\vec{x} = V_1S^{-1}U^T\vec{y}.} \tag{13}$$

Note from the zero entry in (12) that the minimum-norm solution leaves no component in the column space of V_2, which is the null space of A as discussed on page 1. Indeed a nonzero component in the null space would not change $A\vec{x}$ but increase the norm of $\vec{x}$.

As an exercise you can show[2] that (13) is equivalent to the formula:

$$\vec{x} = A^T(AA^T)^{-1}\vec{y}. \tag{14}$$

Principal Component Analysis (PCA)

PCA is an application of SVD in statistics that aims to find the most informative directions in a data set.

Suppose the $m \times n$ matrix A contains n measurements from m samples, for example n test scores for m students. If we subtract from each measurement the average over all samples, then each column of A is an m-vector with zero mean, and the $n \times n$ matrix

$$\frac{1}{m-1}A^TA$$

constitutes what is called the "covariance matrix" in statistics. Recall that the eigenvalues of this matrix are the singular values of A except for the scaling factor $m-1$, and its orthonormal eigenvectors correspond to $\vec{v}_1, \ldots, \vec{v}_n$ in the SVD of A.

The vectors $\vec{v}_1, \vec{v}_2, \ldots$ corresponding to large singular values are called principal components and identify dominant directions in the data set along which the samples are clustered. The most significant direction is $\vec{v}_1$ corresponding to σ_1.

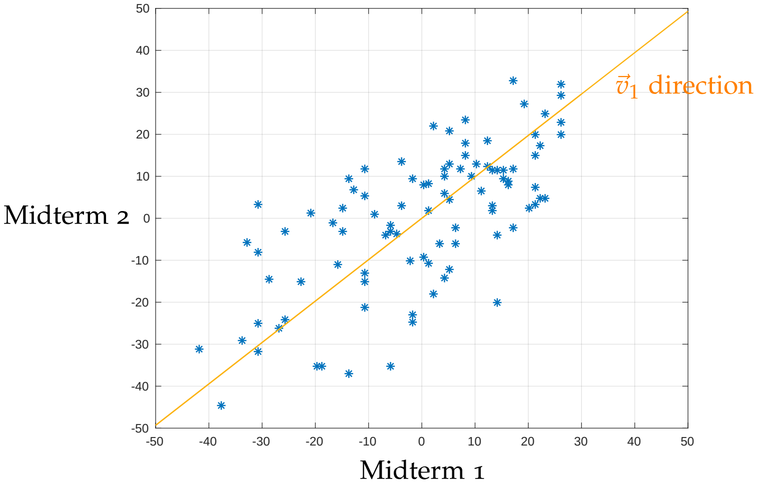
As an illustration, the scatter plot below shows $n = 2$ midterm scores in a class of $m = 94$ students that I taught in the past. The data points are centered around zero because the class average is subtracted from the test scores. Each data point corresponds to a student and those in the first quadrant (both midterms ≥ 0) are those students who scored above average in each midterm. You can see that there were students who scored below average in the first and above average in the second, and vice versa.

For this data set the covariance matrix is:

$$\frac{1}{93}A^TA = \begin{bmatrix} 297.69 & 202.53 \\ 202.53 & 292.07 \end{bmatrix}$$

where the diagonal entries correspond to the squares of the standard deviations 17.25 and 17.09 for Midterms 1 and 2, respectively. The positive sign of the $(1,2)$ entry implies a positive correlation between the two midterm scores as one would expect.

[2] Substitute (9) in (14) and simplify to get (13).



The eigenvalues of A^TA, that is the singular values of A are $\sigma_1 = 215.08$, $\sigma_2 = 92.66$, and the corresponding eigenvectors of A^TA are:

$$\vec{v}_1 = \begin{bmatrix} 0.7120 \\ 0.7022 \end{bmatrix} \quad \vec{v}_2 = \begin{bmatrix} -0.7022 \\ 0.7120 \end{bmatrix}.$$

The principal component $\vec{v}_1$ is superimposed on the scatter plot and we see that the data is indeed clustered around this line. Note that it makes an angle of $\tan^{-1}(0.7022/0.7120) \approx 44.6°$ which is skewed slightly towards the Midterm 1 axis because the standard deviation in Midterm 1 was slightly higher than in Midterm 2. We may interpret the points above this line as students who performed better in Midterm 2 than in Midterm 1, as measured by their scores relative to the class average that are then compared against the factor $\tan(44.6°)$ to account for the difference in standard deviations.

The $\vec{v}_2$ direction, which is perpendicular to $\vec{v}_1$, exhibits less variation than the $\vec{v}_1$ direction ($\sigma_2 = 92.66$ vs. $\sigma_1 = 215.08$).