Math 312, Spring 2014 Jerry L. Kazdan

Singular Value Decomposition

Positive Definite Matrices

Let C be an $n \times n$ positive definite (symmetric) matrix and consider the quadratic polynomial $\langle \mathbf{x},\, C\mathbf{x}\rangle$. How can we understand what this"looks like"? One useful approach is to view the image of the unit sphere, that is, the points that satisfy $\|\mathbf{x}\| = 1$. For instance, if $\langle \mathbf{x},\, C\mathbf{x}\rangle = 4x_1^2 + 9x_2^2$, then the image of the unit disk, $x_1^2 + x_2^2 = 1$, is an ellipse.

For the more general case of $\langle \mathbf{x},\, C\mathbf{x}\rangle$ it is fundamental to use coordinates that are adapted to the matrix C. Since C is a symmetric matrix, there are orthonormal vectors $\mathbf{v}_1,\ \mathbf{v}_2, \ldots,\ \mathbf{v}_n$ in which C is a diagonal matrix. These vectors $\mathbf{v}_j$ are eigenvectors of C with corresponding eigenvalues λ_j: so $C\mathbf{v}_j = \lambda_j \mathbf{v}_j$. In these coordinates, say

$$\mathbf{x} = y_1\mathbf{v}_1 + y_2\mathbf{v}_2 + \cdots + y_n\mathbf{v}_n. \tag{1}$$

and

$$C\mathbf{x} = \lambda_1 y_1\mathbf{v}_1 + \lambda_2 y_2\mathbf{v}_2 + \cdots + \lambda_n y_n\mathbf{v}_n. \tag{2}$$

Then, using the orthonormality of the $\mathbf{v}_j$,

$$\|\mathbf{x}\|^2 = \langle \mathbf{x},\, \mathbf{x}\rangle = y_1^2 + y_2^2 + \cdots + y_n^2$$

and

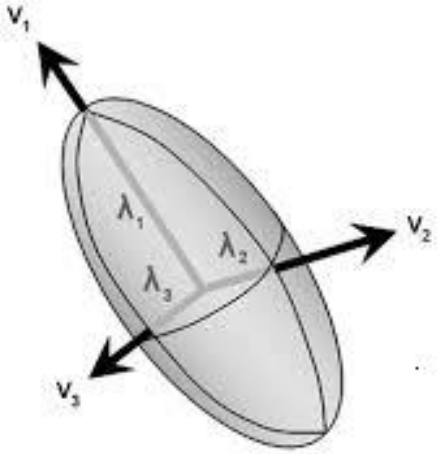
$$\langle \mathbf{x},\, C\mathbf{x}\rangle = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2. \tag{3}$$

For use in the singular value decomposition below, it is conventional to number the eigenvalues in *decreasing* order:

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = \lambda_{\min},$$

so on the unit sphere $\|\mathbf{x}\| = 1$

$$\lambda_{\min} \leq \langle \mathbf{x}, C\mathbf{x} \rangle \leq \lambda_{\max}.$$



As in the figure, the image of the unit sphere is an "ellipsoid" whose axes are in the direction of the eigenvectors and the corresponding eigenvalues give half the length of the axes.

There are two closely related approaches to using that a self-adjoint matrix C can be orthogonally diagonalized.

METHOD 1: Use that there is an orthogonal matrix R (whose columns are the orthonormal eigenvectors of C) and a diagonal matrix Λ (with the corresponding eigenvalues of C on the diagonal) so that

$$C = R\Lambda R^* \tag{4}$$

. Then, letting $\vec{y} = R^*\vec{x}$ we have

$$\begin{aligned} \langle \mathbf{x}, C\mathbf{x} \rangle =& \langle \mathbf{x}, R\Lambda R^*\mathbf{x} \rangle = \langle R^*\mathbf{x}, \Lambda R^*\mathbf{x} \rangle \\ =& \langle \mathbf{y}, \Lambda \mathbf{y} \rangle = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \end{aligned}$$

just as in equation (3).

METHOD 2: Use that $\mathbb{R}^n$ has a basis of orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. This is what we did in equations (1)-(3). For me, Method 2 is usually more intuitive.

It is illuminating to rewrite equation (2) using the formula $y_j = \langle \mathbf{v}_j, \mathbf{x} \rangle$ in equation (1). Thinking of $\mathbf{v}_j$ and $\mathbf{x}$ as column vectors, then $\langle \mathbf{v}_j, \mathbf{x} \rangle$ is the same as the matrix product $\mathbf{v}_j^* \mathbf{x}$ so

$$y_j \mathbf{v}_j = \mathbf{v}_j \langle \mathbf{v}_j, \mathbf{x} \rangle = \mathbf{v}_j (\mathbf{v}_j^* \mathbf{x}) = (\mathbf{v}_j \mathbf{v}_j^*) \mathbf{x} \tag{5}$$

so

$$C\mathbf{x} = \sum \lambda_j (\mathbf{v}_j \mathbf{v}_j^*) \mathbf{x}, \qquad \text{that is,} \qquad C = \sum \lambda_j (\mathbf{v}_j \mathbf{v}_j^*) \tag{6}$$

Notice that the matrices $\mathbf{v}_j \mathbf{v}_j^*$ each have rank one.

Singular Value Decomposition

We will use this to help understand an $m \times n$ matrix $A : \mathbb{R}^n \to \mathbb{R}^m$. We will assume that $m \geq n$. Let $r = \operatorname{rank}(A)$. Note $r \leq n$. This matrix A is not assumed to be symmetric – or even square. We want to measure how the size of $\|A\mathbf{x}\|$ changes as $\mathbf{x}$ moves around the unit sphere, $\|\mathbf{x}\| = 1$ in $\mathbb{R}^n$. For instance,

- What is the largest that $\|A\mathbf{x}\|$ can be? In what direction $\mathbf{x}$ is $\|A\mathbf{x}\|$ largest?
- What is the smallest that $\|A\mathbf{x}\|$ can be? In what direction $\mathbf{x}$ is $\|A\mathbf{x}\|$ smallest?
- How does size of $A\mathbf{x}$ vary with the direction of $\mathbf{x}$?

To answer these we observe that

$$\|A\mathbf{x}\|^2 = \langle A\mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, (A^*A)\mathbf{x} \rangle.$$

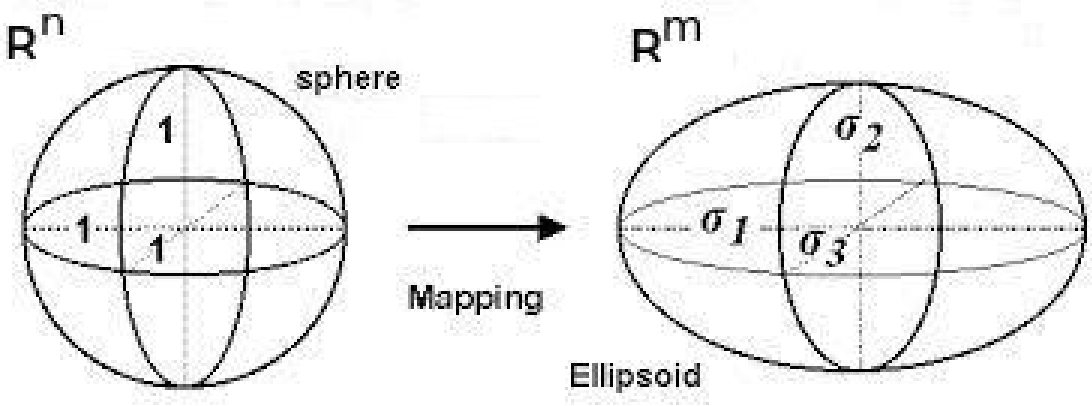
Thus we can answer our questions by investigating the simpler $n \times n$ positive semi-definite symmetric matrix $C = A^*A$ – which is what we just did at the top of this page. As above, let $\lambda_j \geq 0$ and $\mathbf{v}_j$

be the eigenvalues and corresponding orthonormal eigenvectors of $C = A^*A$. Then on the sphere $\|\mathbf{x}\| = 1$

$$\begin{aligned} \max\|A\mathbf{x}\|^2 &= \lambda_1 \qquad \text{is at} \qquad \mathbf{x} = \mathbf{v}_1 \\ \min\|A\mathbf{x}\|^2 &= \lambda_n \qquad \text{is at} \qquad \mathbf{x} = \mathbf{v}_n \end{aligned}$$

Because of the squares on the left side, we let $\sigma_j = \sqrt{\lambda_j}$. These σ_j's are called the *singular values* of A, so we have $\|A\mathbf{v}_j\| = \sigma_j$ and

$$\sigma_n \le \sigma_{n-1} \le \cdots \le \sigma_1. \qquad \text{On } \|\mathbf{x}\| = 1, \quad \sigma_n \le \|A\mathbf{x}\| \le \sigma_1.$$



We define the vectors $\mathbf{u}_j \in \mathbb{R}^m$ as the images of the eigenvectors $\mathbf{v}_j$:

$$A\mathbf{v}_j = \sigma_j \mathbf{u}_j, \quad j = 1, \ldots, r := \text{rank}\,(A). \tag{7}$$

Note that if $\sigma_j = 0$, then the above does not define $\mathbf{u}_j = A\mathbf{v}_j/\sigma_j$ because it would involve division by zero. However, for $j = 1, \ldots, r$, since the v_j are orthonormal, then so are the $\mathbf{u}_j$ because

$$\langle \mathbf{u}_i,\, \mathbf{u}_j \rangle = \frac{\langle A\mathbf{v}_i,\, A\mathbf{v}_j \rangle}{\sigma_i \sigma_j} = \langle \mathbf{v}_i,\, \mathbf{v}_j \rangle.$$

We can now extend the orthonormal vectors $\mathbf{u}_j,\ j = 1, \ldots, r$ to an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_r,\ \mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$ for $\mathbb{R}^m$. Also let V

be the $n \times n$ matrix whose columns are the orthonormal eigenvectors $\mathbf{v}_j,\ j = 1, \ldots, n$. This gives us two orthogonal matrices: V, an $n \times n$ matrix whose columns are the orthonormal eigenvectors $\mathbf{v}_j,\ j = 1, \ldots, n$ and U, an $m \times m$ matrix whose columns are the orthonormal vectors $\mathbf{u}_k,\ k = 1, \ldots, m$. Finally let Σ be the $m \times n$ "diagonal" matrix

$$\Sigma := \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & & \ddots & \\ & & & & \sigma_n \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

whose last $m - n$ rows are all zeroes. Note also that $\sigma_{r+1} = \cdots = \sigma_n = 0$.

Assembling these pieces from equation (7) we find

$$AV = U\Sigma \tag{8}$$

that is,

$$\begin{pmatrix} & \\ & A \\ & \end{pmatrix} \left(\begin{pmatrix} \\ \\ \end{pmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} \begin{pmatrix} \\ \cdots \\ \end{pmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} \right)$$

$$= \left(\begin{pmatrix} \\ \\ \end{pmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} \begin{pmatrix} \\ \cdots \\ \end{pmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} \right) \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Since V is an orthogonal matrix, $V^{-1} = V^*$ so equation (8) can be rewritten as

$$A = U\Sigma V^*. \tag{9}$$

This version of the *Singular Value Decomposition* is the analog of equation (4) for self-adjoint matrices. See the example just below. The following is an equivalent version of SVD that is quite similar to equation (6) for self-adjoint matrices. It uses that $\sigma_k = 0$ for $k > r$.

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^* + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^*. \tag{10}$$

To verify this we begin with equation (2) and use equation (5) to find

$$\mathbf{x} = \sum \mathbf{v}_i \langle \mathbf{v}_i,\, \mathbf{x} \rangle = \sum \mathbf{v}_i (\mathbf{v}_i^* \mathbf{x}).$$

Therefore, using $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$ for $i \leq r$ while $A\mathbf{v}_i = 0$ for $i > r$, we

obtain

$$A\mathbf{x} = \sum_1^n A\mathbf{v}_i(\mathbf{v}_i^*\mathbf{x}) = \sum_1^r \sigma_i \mathbf{u}_i \mathbf{v}_i^* \mathbf{x}.$$

This is exactly equation (10).

The matrices $\mathbf{u}_j\mathbf{v}_j^*$ each have rank 1. It is significant that since $\sigma_k = 0$ for $k > r$, this does not involve $\mathbf{u}_k$ or $\mathbf{v}_k$ for $k > r$. In a sense that can be made precise, *the sum of the first k matrices here is the matrix of rank k that best approximates A*. This is called the *principal component analysis* of A.

EXAMPLE Before going further, it is essential that we compute an explicit example.

a). Find the singular value decomposition (SVD) of $A := \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix}$ in both versions of equations (9) and (10).

b). Find the best rank 1 approximation to A.

SOLUTION: By a routine computation, the matrix $A^*A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ has eigenvalues 3 and 1 with corresponding eigenvectors $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The singular values of A are $\sigma_1 = \sqrt{3}$ and $\sigma_2 = \sqrt{1} = 1$ with corresponding *orthonormal* eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$. Note that it is traditional to label these so that σ_1 is the *largest* singular value. [If you number these differently, then

you must use a consistent convention in part b) since the best rank 1 approximation is always associated with the largest singular value.] Then the orthonormal $\mathbf{u}_j$'s are

$$\mathbf{u}_1 = \frac{A\mathbf{v}_1}{\sqrt{3}} = \frac{1}{\sqrt{3}} \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix},$$

$$\mathbf{u}_2 = \frac{A\mathbf{v}_2}{1} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

The SVD of A is $A = U\Sigma V^T$, where U is an orthogonal 3×3 matrix whose columns are $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ with $\mathbf{u}_3$ a unit vector orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$ (we never need to compute $\mathbf{u}_3$ explicitly), V an orthogonal 2×2 matrix whose columns are $\mathbf{v}_1$ and $\mathbf{v}_2$, and Σ a 3×2 matrix containing the singular values of A. Thus

$$U = \begin{pmatrix} -1/\sqrt{6} & -1/\sqrt{2} & \vdots \\ 2/\sqrt{6} & 0 & \mathbf{u}_3 \\ -1/\sqrt{6} & 1/\sqrt{2} & \vdots \end{pmatrix}, \quad \Sigma := \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad V := \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}.$$

The (equivalent) singular value decomposition of A as in equation (10) is then

$$\begin{aligned} A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T &= \frac{\sqrt{3}}{\sqrt{6}} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \end{pmatrix} + \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} -1 & 1 \\ 2 & -2 \\ -1 & 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{pmatrix} \end{aligned} \tag{11}$$

b). To find the best rank 1 approximation to A, from equation (11), it is $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T = \frac{1}{2}\begin{pmatrix} -1 & 1 \\ 2 & -2 \\ -1 & 1 \end{pmatrix}$

Bird Example [based on notes by J. Jauregui]

This example and data is courtesy of Adam Kapelner, from Wharton Statistics. Adam used Sibley's Bird Database of North American birds to gather data on a simple random sample of 100 bird species. Three factors were measured: length (inches), wingspan (inches), and weight (ounces). Thus, in the data matrix A n = 3 and m = 100, so A is a 100×3 matrix. For each of the three columns we subtracted the mean. Thus the average of each column of A is zero. We imagine plotting the data in A as a cloud of 100 points in a three dimensional space. We are seeking a pattern. Does the data cluster around a line? Does it cluster around a plane – or is it scattered like a random cloud?

We use the singular value decomposition to help. Let $S = A^*A$, which is the 3×3 matrix

$$S = \begin{pmatrix} 91.43 & 171.92 & 297.99 \\ & 373.92 & 545.21 \\ & & 1297.26 \end{pmatrix}$$

As is customary, the entries below the diagonal were omitted, since the matrix is symmetric. We can use MATLAB or octave [1], for instance, to compute the eigenvalues and orthonormal eigenvectors of S. In this case:

$$\lambda_1 = 1626.52, \qquad \lambda_2 = 128.99, \qquad \lambda_3 = 7.10$$

[1] octave is a free, open source alternative to MATLAB.

so the singular values of A are the square roots

$$\sigma_1 = 40.3, \qquad \sigma_2 = 14.4, \qquad \sigma_3 = 2.67.$$

while the eigenvectors of S are

$$\mathbf{v}_1 = \begin{pmatrix} 0.22 \\ 0.41 \\ 0.88 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0.25 \\ 0.85 \\ -0.46 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 0.94 \\ -0.32 \\ -0.08 \end{pmatrix}$$

The first thing to notice is that σ_1 is much larger than σ_2 and σ_3. In fact, the first $\mathbf{v}_1$ principal component accounts for a significant amount of the variation in the data while the second $\mathbf{v}_2$ accounts for less and the remaining principal component, explaining only a small amount of the data. It is negligible compared to the first two.

Now, how to interpret all of this? In studying the sizes (length, wingspan, weight) of North American birds, there are apparently only two factors that are important (corresponding to $\mathbf{v}_1$ and $\mathbf{v}_2$). We might think of $\mathbf{v}_1$ as giving a generalized notion of "size" that incorporates length, wingspan, and weight. Indeed, all three entries of $\mathbf{v}_1$ have the same sign, indicating that birds with larger "size" tend to have larger length, wingspan, and weight.

We could also ask: which of the factors (length, wingspan, weight) is most significant in determining a bird's "size"? In other words, does the first principal component $\mathbf{v}_1$ point the most in the direction of the length axis, the wingspan axis, or the weight axis in $\mathbb{R}^3$? Well, the third entry, weight, of $\mathbf{u}_1$ is the largest, so weight is the most significant. This means a change in one unit of weight tends to affect the size more so than a change in one unit of length or wingspan. The second entry of $\mathbf{v}_1$ is the next largest, which corresponds to wingspan.

Thus, wingspan is the next most important factor in determining a bird's size (followed lastly by length).

Now, what does the second principal component mean? It is mostly influenced by wingspan and weight, as these entries in $\mathbf{v}_2$ have the greatest absolute values. However, they also have opposite signs. This indicates that $\mathbf{v}_2$ describes a feature of birds corresponding to relatively small wingspan and large weight, or vice versa. We might call this quality "stoutness."

For each of these birds, is the "size" large or small? What about the "stoutness"?

In other words, to a very good approximation, this sample of North American birds is described by only two parameters: the "size" (most important) and the "stoutness" (less important). We discovered this by looking at the eigenvalues λ_j (or singular values, σ_j) and the corresponding eigenvectors $\mathbf{v}_j$ of the matrix $S = A^*A$.

Noise in Data

If

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ 3 & 3 & -3 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 1.00001 & 1 & -1 \\ 2 & 2.00001 & -2 \\ 3 & 3 & -3.00001 \end{pmatrix},$$

then A has rank 1 while B has rank 3 We should probably view B as essentially the same as A except for some noise. This is revealed if we compute the singular values of B. We find

$$\sigma_1 = 6.481, \qquad \sigma_2 \approx \sigma_3 \approx 0.000001.$$

This is a very simple example showing how a singular value decomposition can help reveal the essential structure of data.

Last Revised June 30, 2020