Chapter 3 of Calculus^{++}: The Symmetric Eigenvalue Problem

by

Eric A Carlen
Professor of Mathematics
Georgia Tech

Table of Contents

Section 1: Diagonalizing 2×2 symmetric matrices

Section 2: Jacobi's Algorithm

Section 3: The eigenvalues of almost diagonal matrices

Section 4: The singular value decomposition

Section 5: Geometry and the singular value decomposition

Section 1: Diagonalizing 2×2 Symmetric Matrices

1.1: An explicit formula

Symmetric matrices are special. For instance, they always have real eigenvalues. There are several ways to see this, but for 2×2 symmetric matrices, direct computation is simple enough: Let A be any symmetric 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix} .$$

Then $A - tI = \begin{bmatrix} a-t & b \\ b & d-t \end{bmatrix}$ so that

$$\det(A - tI) = (a-t)(d-t) - b^2 = t^2 - (a+d)t + ad - b^2 .$$

Hence the eigenvalues of A are the roots of

$$t^2 - (a+d)t + ad - b^2 = 0 . \tag{1.1}$$

Completing the square, we obtain

$$\begin{aligned} \left(t - \frac{a+d}{2}\right)^2 &= b^2 - ad + \left(\frac{a+d}{2}\right)^2 \\ &= b^2 - ad + \left(\frac{a^2+d^2+2ad}{4}\right) \\ &= b^2 + \frac{a^2+d^2-2ad}{4} \\ &= b^2 + \left(\frac{a-d}{2}\right)^2 \end{aligned}$$

Hence, (1.1) becomes $t = \frac{a+d}{2} \pm \sqrt{b^2 + \left(\frac{a-d}{2}\right)^2}$. Since $b^2 + \left(\frac{a-d}{2}\right)^2$ is the sum of two squares, it is positive, and so the square root is real. Therefore, the two eigenvalues are

$$\mu_+ = \frac{a+d}{2} + \sqrt{b^2 + \left(\frac{a-d}{2}\right)^2} \qquad \text{and} \qquad \mu_- = \frac{a+d}{2} - \sqrt{b^2 + \left(\frac{a-d}{2}\right)^2} . \tag{1.2}$$

We have just written down an explicit formula for the eigenvalues of the 2×2 symmetric matrix $A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$. As you can see from the formula, the eigenvalues are both real.

There is even more that is special about $n \times n$ symmetric matrices: They can always be diagonalized, and by an orthogonal matrix at that. Again, in the 2×2 case, direct computation leads to an explicit formula.

Let $B = A - \mu_+ I$. Then an non zero vector $\mathbf{v}$ is an eigenvector of A with eigenvalue μ_+ if and only if $B\mathbf{v} = 0$. Now write B in row vector form: $B = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}$. Now, by a basic formula for matrix multiplication, $B\mathbf{v} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \mathbf{v} = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{v} \\ \mathbf{r}_2 \cdot \mathbf{v} \end{bmatrix}$. So if $\mathbf{v}$ is an eigenvector with eigenvalue μ_+, then

$$\mathbf{r}_1 \cdot \mathbf{v} = 0 \qquad \text{and} \qquad \mathbf{r}_2 \cdot \mathbf{v} = 0 \ .$$

Now $\mathbf{r}_1 \cdot \mathbf{v} = 0$ if and only if $\mathbf{v}$ is a multiple of $\mathbf{r}_1^\perp$. This means that a vector $\mathbf{v}$ is an eigenvector of A with eigenvalue μ_+ if and only if $\mathbf{v}$ is a multiple of $\mathbf{r}_1^\perp$. In particular, $\mathbf{r}_1^\perp$ is an eigenvector of A with eigenvalue μ_+. Normalizing this, we define

$$\mathbf{u}_1 = \frac{1}{|\mathbf{r}_1|}\mathbf{r}_1^\perp \ .$$

This is a unit vector, and an eigenvector of A with eigenvalue μ_+.

Next, we use another basic fact about symmetric matrices: *Eigenvectors corresponding to distinct eigenvalues are orthogonal.* So as long as $\mu_- \neq \mu_+$, the eigenvectors of A with eigenvalue μ_- must be orthogonal to $\mathbf{u}_1$. This means that $\mathbf{u}_1^\perp$ is an eigenvector of A with eigenvalue μ_-. It is also a unit vector, and orthogonal to $\mathbf{u}_1$, so if we define $\mathbf{u}_2$ by

$$\mathbf{u}_2 = \mathbf{u}_1^\perp \ ,$$

then

$$\{\mathbf{u}_1, \mathbf{u}_2\}$$

is an orthonormal basis of $I\!R^2$ consisting of eigenvectors of A.

What if the eigenvalues are the same? You see from (1.2) that the two eigenvalues are the same if and only if $b^2 = 0$ and $(a-d)^2 = 0$, which means that $A = aI$, in which case A is already diagonal, and *every* vector in $I\!R^2$ is an eigenvector of A with eigenvalue a. Hence the same formulas apply in this case as well.

Now form the matrix U defined by

$$U = [\mathbf{u}_1, \mathbf{u}_2] \ .$$

Then

$$AU = A[\mathbf{u}_1, \mathbf{u}_2] = [A\mathbf{u}_1, A\mathbf{u}_2] = [\mu_+\mathbf{u}_1, \mu_-\mathbf{u}_2] = [\mathbf{u}_1, \mathbf{u}_2] \begin{bmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{bmatrix} \ .$$

If we define D to be the diagonal matrix

$$D = \begin{bmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{bmatrix} \ ,$$

then we can rewrite this as

$$AU = UD \ . \tag{1.3}$$

Now since U has orthonormal columns, it is an *orthognal matrix*, and hence U^t is the inverse of U. Therefore, (1.3) can be rewritten as

$$D = U^t AU \ .$$

We summarize all of this in the following theorem:

Theorem 1 (Eigenvectors and eigenvalues for 2×2 symmetric matrices) *Let $A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$ be any 2×2 symmetric matrix. Then the eigenvalues of A are*

$$\mu_+ = \frac{a+d}{2} + \sqrt{b^2 + \left(\frac{a-d}{2}\right)^2} \qquad \text{and} \qquad \mu_- = \frac{a+d}{2} - \sqrt{b^2 + \left(\frac{a-d}{2}\right)^2} \ . \tag{1.4}$$

Moreover, if we define $\mathbf{r}_1$ and $\mathbf{r}_2$ by

$$A - \mu_+ I = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \ ,$$

and put

$$\mathbf{u}_1 = \frac{1}{|\mathbf{r}_1|}\mathbf{r}_1^\perp \qquad \text{and} \qquad \mathbf{u}_2 = \mathbf{u}_1^\perp \ , \tag{1.5}$$

then $\{\mathbf{u}_1, \mathbf{u}_2\}$ is an orthonormal basis of $I\!R^2$ consisting of eigenvectors of A, and with

$$U = [\mathbf{u}_1, \mathbf{u}_2] \qquad \text{and} \qquad D = \begin{bmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{bmatrix} \ , \tag{1.6}$$

$$U^t AU = D \ . \tag{1.7}$$

Example 1 (Finding the eigenvectors and eigenvalues of a 2×2 symmetric matrix) Let $A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$. With $A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$, we have

$$a = 3 \qquad b = 2 \qquad d = 6 \ .$$

Using (1.4), we find that $\mu_\pm = \frac{9}{2} \pm \frac{5}{2}$; i.e.,

$$\mu_+ = 7 \qquad \text{and} \qquad \mu_- = 2 \ .$$

Now,

$$A - \mu_+ I = \begin{bmatrix} 3-7 & 2 \\ 2 & 6-7 \end{bmatrix} = \begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \ .$$

The first row of this matrix – written as a column vector – is $\mathbf{r}_1 = \begin{bmatrix} -4 \\ 2 \end{bmatrix}$. Hence we have

$$\mathbf{u}_1 - \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ -2 \end{bmatrix} \qquad \text{and} \qquad \mathbf{u}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \ . \tag{1.8}$$

Example 2 (Diagonalizing a 2×2 symmetric matrix) Let A be the 2×2 matrix $A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ that we considered in Example 1. There we found that the eigenvalues are 7 and 2, and we found corresponding unit eigenvectors $\mathbf{u}_1 \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ -2 \end{bmatrix}$ and $\mathbf{u}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Hence from (1.6), we have

$$U = \frac{1}{\sqrt{5}} \begin{bmatrix} -1 & 2 \\ -2 & -1 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix} \ .$$

As you can check, $U^t A U = D$, in agreemant with (1.7).

Theorem 1 provides *one* way to diagonalize a 2×2 symmetric matrix with an orthogonal matrix U. However, there is something special about it: The matrix U is not only an orthogonal matrix; it is a rotation matrix, and in D, the eigenvalues are listed in decreasing order along the diagonal.

This turns out to be useful, and to explain it better, we recall a few facts about 2×2 orthogonal matrices.

1.2: 2×2 orthogonal matrices: rotations and reflections

Let $U = [\mathbf{u}_1, \mathbf{u}_2]$ be any orthogonal matrix. Then $\mathbf{u}_1$ is a unit vector, so

$$\mathbf{u}_1 = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}$$

for some θ with $0 \le \theta < 2\pi$.

Next, $\mathbf{u}_2$ is orthogonal to $\mathbf{u}_1$, but there are exactly two unit vectors that are orthogonal to $\mathbf{u}_1$, namely $\pm\mathbf{u}_1^\perp$. Therefore,

$$\text{either} \quad \mathbf{u}_2 = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \qquad \text{or else} \qquad \mathbf{u}_2 = \begin{bmatrix} \sin(\theta) \\ -\cos(\theta) \end{bmatrix} \ .$$

In the first case,

$$U = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \ , \tag{1.9}$$

while in the second case,

$$U = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix} \ . \tag{1.10}$$

The matrix U in (1.9) describes a counterclockwise rotation through the angle θ. Since this is the sort of U we get using Theorem 1, we see that this theorem provides us a diagonalization in terms of a rotation matrix.

what we have said so far is all that is really important in what follows, but you may be wondering what sort of transformation might be encoded in (1.10). There is a simple answer: The matrix U in (1.10) describes a reflection.

To see this, define $\phi = \theta/2$. Then,

$$\begin{aligned}\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ \sin(\theta) & -\cos(\theta)\end{bmatrix} &= \begin{bmatrix}\cos^2(\phi)-\sin^2(\phi) & 2\sin(\phi)\cos(\phi)\\ 2\sin(\phi)\cos(\phi) & \sin^2(\phi)-\cos^2(\phi)\end{bmatrix}\\ &= \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - 2\begin{bmatrix}\sin^2(\phi) & -\sin(\phi)\cos(\phi)\\ -\sin(\phi)\cos(\phi) & \cos^2(\phi)\end{bmatrix} \ .\end{aligned}$$

From here, one easily sees that if $\mathbf{u}_\phi = \begin{bmatrix}\cos(\phi)\\ \sin(\phi)\end{bmatrix}$,

$$\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ \sin(\theta) & -\cos(\theta)\end{bmatrix} = I - 2(\mathbf{u}_\phi^\perp)(\mathbf{u}_\phi^\perp)^t \ .$$

From here it follows that with U given by (1.10),

$$U\mathbf{u}_\phi = \mathbf{u}_\phi \qquad \text{and} \qquad U\mathbf{u}_\phi^\perp = \mathbf{u}_\phi^\perp \ .$$

This shows that the matrix U in (1.10) is the reflection about the line through the origin and $\mathbf{u}_\phi$.

Problems

1 Let $A = \begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}$. Use Theorem 1 to find the eigenvectors and eigenvalues of A, and find an orthogonal matrix U that diagonalizes A.

2 Let $A = \begin{bmatrix}4 & 2\\ 2 & 4\end{bmatrix}$. Use Theorem 1 to find the eigenvectors and eigenvalues of A, and find an orthogonal matrix U that diagonalizes A.

3 Let $A = \begin{bmatrix}1 & 2\\ 2 & 5\end{bmatrix}$. Use Theorem 1 to find the eigenvectors and eigenvalues of A, and find an orthogonal matrix U that diagonalizes A.

4 Let $A = \begin{bmatrix}-1 & 2\\ 2 & 4\end{bmatrix}$. Use Theorem 1 to find the eigenvectors and eigenvalues of A, and find an orthogonal matrix U that diagonalizes A.

Section 2: Jacobi's algorithm

2.1 Why iterate?

We have seen that for 2×2 symmetric matrices $A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$, the eigenvalues are given by

$$\frac{a+d}{2} \pm \frac{\sqrt{(a-d)^2 + 4b^2}}{2} \ .$$

How about $n \times n$ matrices? Is there such an explicit formula for the eigenvalues of $n \times n$ matrices for larger values of n?

In fact, there cannot be any such formula for the eigenvectors and eigenvalues of $n \times n$ matrices for $n \geq 5$. This is because for polynoials of degree 5 and higher there is no formula for computing the roots in terms of the coefficients in a finite number of steps. The eigenvalues of A are, of course, the roots of the characteristic polynomial of A.

Nonetheless, there are very effective *iterative algorithms* for computing the eigenvalues of a symmetric matrix.

The original iterative algorithm for this purpose was devised by Jacobi*. We first explain it in the 3×3 case.

Consider the matrix

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \ ,$$

and focus on the 2×2 block $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ in the upper left corner. We know how to diagonalize every 2×2 matrix, so we can certainly diagonalize this one. Jacobi's idea is to use the similarity transform that diagonalizes this 2×2 matrix to *partially diagonalize* the 3×3 matrix A. We then pick another 2×2 block, and do a further partial diagonalization, and so on. Here is how this goes.

Applying Theorem 1 of the previous section, we find that with

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \ ,$$

$$U^t \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} U = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \ .$$

The idea is now to "promote" U to a 3×3 rotation matrix, which we will denote G_1, and then to work out $G_1^t A G_1$. Specifically, take the 3×3 identity matrix, and overwrite the upper left 2×2 block with U. This gives us the 3×3 matrix G_1:

$$G_1 = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \ .$$

* This is the Jacobi who is the namesake of Jacobian matrices, among other things.

Multiplying out $G_1^t A G_1$ we find

$$G_1^t A G_1 = \begin{bmatrix} 3 & 0 & \sqrt{2} \\ 0 & 1 & 0 \\ \sqrt{2} & 0 & 2 \end{bmatrix} .$$

Now four out of six off–diagonal entries are zero. This is our partial diagonalization, and this is progress towards fully diagonalizing A!

Let's continue. The only non zero off diagonal entries are in the $1,3$ and $3,1$ positions. We therefore focus on the 2×2 block that contains them: The 2×2 block we get by deleting the second column and row is $\begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix}$.

Our general formulas for diagonalizing 2×2 matrices give us

$$U^t \begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix} U = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

where

$$U = \begin{bmatrix} \sqrt{2/3} & -\sqrt{1/3} \\ \sqrt{1/3} & \sqrt{2/3} \end{bmatrix} .$$

Define the 3×3 rotation matrix G_2 by overwriting the 3×3 identity matrix with the entries of U, putting them in the $1,1$, $1,3$, $3,1$ and $3,3$ places, since it was from these places that we took our 2×2 block. We obtain:

$$G_2 = \begin{bmatrix} \sqrt{2/3} & 0 & -\sqrt{1/3} \\ 0 & 1 & 0 \\ \sqrt{1/3} & 0 & \sqrt{2/3} \end{bmatrix} .$$

We then compute

$$G_2^t (G_1^t A G_1) G_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

Defining $V = G_1 G_2$, and $D = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ we can write this as

$$A = V D V^t .$$

The diagonalization is complete! In particular we can now read of the eigenvalues of A; they are 4 and 1 with multiplicities 1 and 2 respectively. We have diagonalized the 3×3 matrix A by repeated use of our 2×2 diagonalization formulas!

The matrix A is not so bad to deal with by analysis of its characteristic polynomial. Indeed,

$$\det(A - tI) = -t^3 + 6t^2 - 9t + 4 \ .$$

You might notice that $t = 1$ is a root, and from there you could factor

$$-t^3 + 6t^2 - 9t + 4 = (4 - t)(1 - t)(1 - t) \ .$$

However, factoring cubic polynomials is not so easy. Using Jacobi's idea, we diagonalized A without ever needing to factor a cubic polynomial. This is even more advantageous for larger matrices.

Moving from this specific example to the general 3×3 symmetric matrix, let's define the three kinds of rotation matrices that we will use to diagonalize 2×2 submatrices. There will be three kinds because there are three ways to choose a pair of rows (and columns) We index these matrices by the pair of row indices that we "keep":

For $1 \leq i < j \leq 3$, define the *Givens rotation matrix* $G(\theta, i, j)$ by

$$G(\theta, 1, 2) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$G(\theta, 1, 3) = \begin{bmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{bmatrix}$$

and

$$G(\theta, 2, 3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix} \ .$$

These are just rotations through the angle θ in the x, y plane, the x, z plane and the y, z plane respectively. In particular, they are orthogonal: their columns are orthonormal, and so their inverses are just their transposes.

The $n \times n$ version is similar:

$$G(\theta, i, j)_{i,i} = \cos(\theta) \qquad G(\theta, i, j)_{i,j} = -\sin(\theta)$$

$$G(\theta, i, j)_{j,i} = \sin(\theta) \qquad G(\theta, i, j)_{j,j} = \ \ \sin(\theta)$$

For all other entires,

$$G(\theta, i, j)_{k,\ell} = I_{k,\ell} \ .$$

With these preparations made, here is the algorithm. We give it in pseudo code, as the sketch of a program. We assume we are given a symmetric matrix A to diagonalize.

Jacobi's Algorithm

Declare two $n \times n$ symmetric matrix variables, B and V. Initialize them as

$$B \leftarrow A \qquad \text{and} \qquad V \leftarrow I \ .$$

Then:

(1) Find the off–diagonal element of B with the largest absolute value. That is, find values of i and j that maximize

$$|B_{i,j}| \qquad \text{with} \qquad i < j \ .$$

(2) For the values of i and j determined in *(1)*, let U be a rotation matrix so that

$$U^t \begin{bmatrix} A_{i,i} & A_{i,j} \\ A_{j,i} & A_{j,j} \end{bmatrix} U$$

is diagonal.

(3) Let θ be the angle of rotation of the matrix U found in *(2)*. For the values of i and j found in *(1)*, assign

$$B \leftarrow G(\theta, i, j)^t B G(\theta, i, j) \qquad \text{and} \qquad V \leftarrow V G(\theta, i, j) \ .$$

(Notice that B is still symmetric after being updated)

(4) If B is diagonal, stop. Otherwise, go to *(1)* and repeat.

In the example with which we began this section, the procedure terminated in two iterations. However, this was beginners luck: The first matrix we looked at was particularly nice. Even with 3×3 symmetric matrices, the program sketched above would go into an infinite loop.

If it does terminate, then we have

$$V^t A V = D$$

where D is diagonal and V is a rotation matrix. Then the diagonal entries of B are the eigenvalues of A and the columns of V are eigenvectors of A.

The good news is that even if it doesn't terminate exactly in any finite number of steps, the off diagonal terms will tend to zero in the limit as the number of iterations goes to infinity. In fact, we can guarantee a limit on the number of iterations it will take before the off diagonal entries all round off to zero if we are working with any fixed number of digits.

Let's take another example. This time, we will simply report the results in decimal form.

Example 1 (Three steps of Jacobi's algorithm) Let $A = \begin{bmatrix} 2 & -4 & 1 \\ -4 & 5 & 1 \\ 1 & -1 & 2 \end{bmatrix}$. Note that $|B_{i,j}|$ is largest for $i = 1$ and $j = 2$. We compute the corresponding angle θ, and after the first step we have

$$B = \begin{bmatrix} 7.77200 & 0 & 1.39252 \\ 0 & -0.77200 & 0.25233 \\ 1.39252 & 0.25233 & 2 \end{bmatrix} .$$

Now, $|B_{i,j}|$ is largest for $i = 1$ and $j = 3$. We compute the corresponding angle θ, and after the second step we have

$$B = \begin{bmatrix} 8.08996 & 0.56207 & 0 \\ 0.56207 & -0.77200 & 0.24599 \\ 0 & 0.24599 & 1.68205 \end{bmatrix} .$$

Now, $|B_{i,j}|$ is largest for $i = 1$ and $j = 2$. We compute the corresponding angle θ, and after the third step we have

$$B = \begin{bmatrix} 8.08996 & 0.00555 & -0.05593 \\ 0.00555 & 1.70646 & 0 \\ -0.05593 & 0 & -0.79642 \end{bmatrix} .$$

The matrix is now almost diagonal. A few more iterations, and the off diagonal entries would all be zero in the decimal places kept here.

2.2 Will the Jacobi algorithm diagonalize any symmetric matrix – or can it "get stuck"?

In general, the Jacobi algorithm does not produce an exactly diagonal matrix in any finite number of iterations, so the formulation of the algorithm that we gave above would result in an infinite loop. However, it will very quickly "almost diagonalize" any matrix. The main goal in this section is to carry out a "worst case analysis" of how fast the Jocobi algorithm eliminates the off diagonal entries.

We now define a quantity that we will use to measure "how nearly diagonal" a matrix is:

Definition For any $n \times n$ matrix B, define the number $\mathrm{Off}(B)$ by

$$\mathrm{Off}(B) = \sum_{i \neq j} |B_{i,j}|^2 \ . \tag{2.1}$$

Notice that this is the sum of the squares of the off–diagonal entries.

The point of the definition is that

$$B \text{ is diagonal} \qquad \Longleftrightarrow \qquad \mathrm{Off}(B) = 0 \ ,$$

and

$$B \text{ is almost diagonal} \qquad \Longleftrightarrow \qquad \mathrm{Off}(B) \approx 0 \ .$$

In fact, since B is symmetric, the largest (in absolute value) off diagonal entry occurs twice, so

$$2 \max_{i \neq j} |B_{i,j}|^2 \leq \mathrm{Off}(B) \ . \tag{2.2}$$

Therefore, for any $\epsilon > 0$,

$$\mathrm{Off}(B) \le 2\epsilon^2 \quad \Rightarrow \quad \max_{i \neq j} |B_{i,j}| \le \epsilon \ .$$

We can now state the main result of this section:

Theorem 1 (Rate of diagonalization for the Jacobi Algorithm) *Let A be any $n \times n$ symmetric matrix. Let $A^{(n)}$ denote the matrix produced by running n steps of the Jacobi algorithm. Then*

$$\mathrm{Off}(A^{(n)}) \le \left(1 - \frac{2}{n^2 - n}\right)^m \mathrm{Off}(A) \ .$$

Since $\left(1 - \frac{2}{n^2 - n}\right) < 1$,

$$\lim_{m \to \infty} \left(1 - \frac{2}{n^2 - n}\right)^m = 0 \ ,$$

and so the theorem says that

$$\lim_{m \to \infty} \mathrm{Off}(A^{(m)}) = 0 \ .$$

This answers the question raised in the title of this subsection: No, the Jacobi algorithm cannot ever get stuck.

In fact, the theorem says that the off diagonal entries are "wiped out" at an exponential rate. For instance, if $n = 3$, $\left(1 - \frac{2}{n^2 - n}\right) = \frac{2}{3}$, so

$$\mathrm{Off}(A^{(m)}) \le \left(\frac{2}{3}\right)^m \mathrm{Off}(A) \ . \tag{2.3}$$

If we define a sequence of numbers $\{a_m\}$ by

$$a_m = \ln\left(\mathrm{Off}(A^{(m)})\right) \ ,$$

Then (2.3) say that

$$a_m \le \ln\left(\frac{2}{3}\right) m + \ln\left(\mathrm{Off}(A)\right) \ . \tag{2.4}$$

Since $\ln(2/3)$ is negative, the sequence decreases in steady increments and will eventually become as negative as you like.

Suppose we want to run the algortihm until $\text{Off}(A^{(m)}) < \epsilon$ for some given value of ϵ, say $\epsilon = 10^{-10}$. How many steps might this take? We can stop as soon as

$$a_m \leq \ln(\epsilon) \ ,$$

and from (2.4), we see that this is guaranteed by

$$\ln\left(\frac{2}{3}\right) m + \ln\left(\text{Off}(A)\right) \leq \ln(\epsilon) \ ,$$

which means

$$m \geq \frac{\ln\left(\text{Off}(A)\right) - \ln(\epsilon)}{\ln(3) - \ln(2)} \ . \tag{2.5}$$

Let m be the smallest integer satisfying (2.5). Then we are guaranteed that $\text{Off}(A^{(m)} < \epsilon$. It will take no more than this many steps for the stopping rule to kick in; we have an *a priori* upper bound on the run time for our algorithm.

In fact, it works much, much better that this in practice for a typical symmetric matrix A. The estimate is actually a much too pessimistic "worst case" analysis. But it still shows that the stopping rule will *always* kick in, so we *never* have an infinite loop.

Accordingly, we now modify the Jacobi algorithm by including a stopping rule in the fourth step. The new version includes a parameter $\epsilon > 0$ to be specified along with A. The only modification is to the fourth step, which now becomes:

---

(4) If $\text{Off}(B) \leq \epsilon$, stop. Otherwise, go to *(1)* and repeat.

There is still an important issue to be dealt with here. When the stopping rule kicks in, the algorithm returns an *almost diagonal* matrix, but in general it will not be exactly diagonal. Hence, the diagonal entries of the matrix we get back will not *exactly equal* the eigenvalues of A. So we have to ask: If a matrix A is almost diagonal, can we be sure that its eigenvalues are very close to its diagonal entries? In the next section, we shall see that the answer is yes, and hence our modified Jacobi algorithm actually returns useful information when it terminates.

In the rest of this section, we prove the theorem. We shall need another definition that is companions to the definition of $\text{Off}(A)$.

Definition For any $n \times n$ matrix B, define the quantity $\text{On}(B)$ by

$$\text{On}(B) = \sum_i |B_{i,i}|^2 \ ,$$

which is the sum of the squares of the on–diagonal entries.

Recall that for any matrix B, the Hilbert–Schmidt norm of B, $\|B\|_{\mathrm{HS}}$ is defined by

$$\|B\|_{\mathrm{HS}} = \sqrt{\sum_{i,j} |B_{i,j}|^2} \ .$$

Notice that

$$\mathrm{On}(B) + \mathrm{Off}(B) = \sum_{i,j} |B_{i,j}|^2 = \|B\|_{\mathrm{HS}}^2 \ . \tag{2.6}$$

The key to proving Theorem 1 lies with the following lemma:

Lemma *Let B be an $n \times n$ symmetric matrix. Let G be the givens rotation matrix produced for this B in the first step of the Jocobi algoithm. Then:*

$$\|G^t BG\|_{\mathrm{HS}} = \|B\|_{\mathrm{HS}} \tag{2.7}$$

and

$$\mathrm{On}(G^t BG) = \mathrm{On}(B) + 2 \max_{i \neq j} |B_{i,j}|^2 \ . \tag{2.8}$$

What this lemma says is that the Hilbert Schmidt norm of a matrix is unchanged as we run the Jacobi algorithm, and the value of $\mathrm{On}(B)$ is increased. But by (2.6), this means that $\mathrm{Off}(B)$ must decrease – by the same amount. That is,

$$\mathrm{Off}(G^t BG) = \mathrm{Off}(B) - 2 \max_{i \neq j} |B_{i,j}|^2 \ . \tag{2.9}$$

As we now explain, this leads directly to the proof of the theorem. We shall then come back, and prove the Lemma.

Proof of Theorem 1: The first step is to eliminate $\max_{i \neq j} |B_{i,j}|^2$ from the right hand side of (2.9), and express it in terms of $\mathrm{Off}(B)$ alone.

To do this, note that $\mathrm{Off}(B)$ is a sum over the $n^2 - n$ squares of the off diagonal entries of B, and each term in the sum is clearly no larger than $\max_{i \neq j} |B_{i,j}|^2$. Therefore,

$$\mathrm{Off}(B) \leq (n^2 - n) \max_{i \neq j} |B_{i,j}|^2 \ .$$

In other words,

$$2 \max_{i \neq j} |B_{i,j}|^2 \geq \frac{2}{n^2 - n} \mathrm{Off}(B) \ .$$

Combining this and (2.9), we have

$$\begin{aligned} \mathrm{Off}(G^t BG) &\leq \mathrm{Off}(B) - \frac{2}{n^2 - n} \mathrm{Off}(B) \\ &= \left(1 - \frac{2}{n^2 - n}\right) \mathrm{Off}(B) \ . \end{aligned} \tag{2.10}$$

This shows that in each step of the algorithm, Off(B) is decreased by the stated factor. ■

2.3 Proof of the Key Lemma

In this final subsection, we prove (2.7) and (2.8). First reacall that for any $n \times n$ matrix B, the *trace* of B, denoted by $\mathrm{tr}(B)$, is defined by

$$\mathrm{tr}(B) = \sum_{i=1}^{n} B_{i,i} \ .$$

We can express $\|B\|_{\mathrm{HS}}$ in terms of the trace as follows: Using the symmetry of B,

$$\mathrm{tr}(B^2) = \sum_{i,j=1}^{n} B_{i,j}B_{j,i} = \sum_{i,j=1}^{n} |B_{i,j}|^2 = \|B\|_{\mathrm{HS}}^2 \ . \tag{2.11}$$

The key fact about the trace that makes this useful is that *similar matrices have the same trace.* Let G be a Givens rotation matrix. Then

$$(G^tBG)^2 = G^tBGF^tBG = G^tB^2G \ .$$

Since G^t is the inverse of G, this says that $(G^tBG)^2$ and B^2 are similar. Hence they have the same trace, and so by (2.11), (2.7) is true.

Now consider (2.8). For any k and ℓ,

$$\begin{aligned}(G^tBG)_{k,k} &= \sum_{r,s=1}^{n} (G^t)_{k,r}B_{r,s}G_{s,k} \\ &= \sum_{r,s=1}^{n} (G)_{r,k}G_{s,k}B_{r,s}cr\end{aligned} \tag{2.12}$$

We suppose that $G = G(\theta, i, j)$; that is the largest off diagonal element in B occurs at the i, j entry. If $k \neq i$ and $k \neq j$, then the kth column of G is the same as the kth column of the identity matrix. Therefore, the only non zero term in the sum (2.12) is the term with $r = s = k$. That is,

$$(G^tBG)_{k,k} = B_{k,k} \qquad \text{for} \quad k \neq i, j \ . \tag{2.13}$$

What happens for $k =$ or $k = j$? To see this, consider the 2×2 matrices

$$\begin{bmatrix} (G^tBG)_{i,i} & (G^tBG)_{i,j} \\ (G^tBG)_{j,i} & (G^tBG)_{j,j} \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix} \ .$$

Let U be the 2×2 rotation matrix out of which G is constructed. Then

$$\begin{bmatrix} (G^tBG)_{i,i} & (G^tBG)_{i,j} \\ (G^tBG)_{j,i} & (G^tBG)_{j,j} \end{bmatrix} = U^t \begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix} U \ .$$

Since U diagonalizes $\begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix}$, we have that $(G^tBG)_{i,j} = (G^tBG)_{j,i} = 0$, and so

$$\begin{bmatrix} (G^tBG)_{i,i} & 0 \\ 0 & (G^tBG)_{j,j} \end{bmatrix} = U^t \begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix} U \ .$$

Therefore, since U is a rotation matrix,

$$\begin{aligned} |(G^tBG)_{i,i}|^2 + |(G^tBG)_{j,j}|^2 &= F\left(\begin{bmatrix} (G^tBG)_{i,i} & 0 \\ 0 & (G^tBG)_{j,j} \end{bmatrix} \right) \\ &= F\left(U^t \begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix} U \right) \\ &= F\left(\begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix} \right) \\ &= |B_{i,i}|^2 + |B_{j,j}|^2 + 2|B_{i,j}|^2 \ . \end{aligned}$$

In short,

$$|(G^tBG)_{i,i}|^2 + |(G^tBG)_{j,j}|^2 = |B_{i,i}|^2 + |B_{j,j}|^2 + 2|B_{i,j}|^2 \qquad (2.14)$$

Since, by hypothesis,

$$|B_{i,j}|^2 = \max k, \ell |B_{k,\ell}|^2 \ ,$$

we have from this, (2.13) and (2.14) that

$$\mathrm{On}(G^tBG) = \mathrm{On}(B) + 2 \max k, \ell |B_{k,\ell}|^2 \ ,$$

and hence (2.8) is true.

Problems

1 Write down the 5×5 Givens rotation matrix $G(\pi/4, 2, 3)$.

2 Write down the 5×5 Givens rotation matrix $G(\pi/3, 1, 4)$.

3 Find the analog of ♣dsl91 that is valid for the $n \times n$ case.

4 (a) Work out by hand one iteration of the Jacobi algorithm for $A = \begin{bmatrix} 2 & 2 & 3 \\ 2 & 1 & 1 \\ 3 & 1 & 2 \end{bmatrix}$.

(b) Using a computer and some software package for linear algebra, let $\epsilon = 10^{-3}$, and run the Jacobi algorithm for A until the stopping rule kicks in. Give the results of this approximate calculation of the eigenvalues of A.

5 (a) Work out by hand one iteration of the Jacobi algorithm for $A = \begin{bmatrix} 4 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{bmatrix}$.

(b) Using a computer and some software package for linear algebra, let $\epsilon = 10^{-3}$, and run the Jacobi algorithm for A until the stopping rule kicks in. Give the results of this approximate calculation of the eigenvalues of A. Are all of the eigenvalues of A positive.

6 Consider the function $f(x, y, z)$ given by

$$f(x, y, z) = x^3yz^2 + 4xy - 3yz \ .$$

Determine whether all of the eigenvalues of the Hessian at $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ are positive or if the are all negative, or neither. (Use the Jacobi algorithm to compute the eigenvalues accurately enough to decide this).

7 Consider the function $f(x, y, z)$ given by

$$f(x, y, z) = xyz^2 + xy^2z + x^2yz \ .$$

Determine whether all of the eigenvalues of the Hessian at $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ are positive or if the are all negative, or neither. (Use the Jacobi algorithm to compute the eigenvalues accurately enough to decide this).

Section 3: The eigenvalues of almost diagonal matrices

3.1 The Gershgorin Disk Theorem

Let A be an $n \times n$ matrix. If A is diagonal, we know what the eigenvalues are – they are the diagonal entries of A. But suppose that the off diagonal entries are very small in absolute value, but not actually zero. Is it then true that each eigenvalue of A is close to one of the diagonal entries of A and *vice versa*?

Let us look at a simple example: consider $A = \begin{bmatrix} 3 & \epsilon \\ \epsilon & 2 \end{bmatrix}$ where ϵ is some small number. Are the eigenvalues of A close to 3 and 2? By Theorem 1 of Section 1, the eigenvalues of A are

$$\mu_\pm = \frac{5}{2} \pm \frac{1}{2}\left(4\epsilon^2 + 1\right)^{1/2} \ .$$

Now, by Taylor's Theorem with remainder

$$\left(1 + 4\epsilon^2\right)^{1/2} = 1 + 2\epsilon^2 + \mathcal{O}(\epsilon^4) \ .$$

Therefore, the eigenvalues are

$$\mu_+ = 3 + \epsilon^2 + \mathcal{O}(\epsilon^4) \qquad \text{and} \qquad \mu_- = 2 - \epsilon^2 + \mathcal{O}(\epsilon^4) \ .$$

This is very nice: For small values of ϵ, ϵ^2 is much, much smaller than ϵ, and so the difference between the eigenvalues and the diagonal entries is much, much smaller that the largest off diagonal entry! Thus, using the diagonal entries to estimate the eigenvalues works very well in this case. Can we prove something like this for general use?

It turns out that this we can. We will prove a theorem saying *how close* the diagonal entries are to the eigenvalues. This is very important in practice. If the matrix we are dealing with is a Hessian of a function f at some critical point $\mathbf{x}_0$, and we want to know whether or not $\mathbf{x}_0$ is a local minimum, we need to know whether or not all of the eigenvalues of the Hessian are positive. If you are using a Jacobi iteration, and at some point all of the diagonal entries are positive, does that mean that all of the eigenvalues are positive? The answer depends on the sizes of the off diagonal terms, but in a rather nice way, as we shall see.

Before explaining this, we make some definitions: For each i with $1 \le i \le n$, define

$$r_i(A) = \sum_{\substack{j = 1 \\ j \neq i}} |A_{i,j}| \ . \tag{3.1}$$

That is, $r_i(A)$ is the sum of the absolute values of all of the *off–diagonal* elements in the ith row of A.

The ith *Gershgorin disk* of A is then defined to be the set of all complex numbers z that are within a distance $r_i(A)$ of $A_{i,i}$, the ith diagonal element of A. The Gershgorin Disk Theorem says that every eigenvalue of A lies within one of the Gershgorin disks of A.

Theorem 1 (Gershgorin Disk Theorem) *Let A be any $n \times n$ matrix, and let μ be any eigenvalue of A. Then for some i with $1 \leq i \leq n$,*

$$|\mu - A_{i,i}| \leq r_i(A)$$

where $r_i(A)$ is given by (3.1).

Proof: Let $\mathbf{v}$ be an eigenvalue of A with eigenvalue μ. Then for each i,

$$\mu v_i = \sum_{j=1}^{n} A_{i,j} v_j \ . \tag{3.2}$$

Let ℓ be chosen so that

$$|v_\ell| = \max\{\ |v_j|\ :\ 1 \leq j \leq n\}\ . \tag{3.3}$$

Taking $i\ = \ell$ in (3.2), we have

$$(\mu - A_{\ell,\ell})\, v_\ell = \sum_{\substack{j\,=\,1 \\ j\,\neq\,\ell}}^{n} A_{\ell,j} v_j \ . \tag{3.4}$$

Since eigenvectors are non zero by definition, $v_\ell \neq 0$, and so, dividing through by v_ℓ, we get,

$$(\mu - A_{\ell,\ell}) = \sum_{\substack{j\,=\,1 \\ j\,\neq\,\ell}}^{n} A_{\ell,j} \frac{v_j}{v_\ell} \ . \tag{3.5}$$

By (3.3), $|v_j/v_\ell| \leq 1$ for all j, so taking absolute values in (3.5),

$$|\mu - A_{\ell,\ell}| \leq \sum_{\substack{j\,=\,1 \\ j\,\neq\,\ell}}^{n} |A_{\ell,j}| = r_\ell \ .$$

This says that μ belongs to the Gershgorin disk about $A_{\ell,\ell}$. ■

Example 1 (Gershgorin disks) Consider the matrix $A = \begin{bmatrix} 3 & 0.1 & -0.1 \\ 0.1 & 0 & 0.1 \\ -0.1 & 0.1 & 2 \end{bmatrix}$. This matrix happens to be symmetric, so all of its eigenvalues are real numbers. We compute the radii of the Gershgorin disks finding

$$r_1(A) = 0.2 \qquad r_2(A) = 0.2 \qquad \text{and} \qquad r_3(A) = 0.2\ .$$

The Gershgorin disks are therefore the disks of radius 0.2 centered on -1, 0 and 2 respectively. These disks contain the eigenvalues. Since in this case we know that the eigenvalues are real numbers, we know that they lie in the intervals

$$[2.8, 3.2] \qquad [-0.2, 0.2] \qquad \text{and} \qquad [1.8, 2.2] \ .$$

This is all well and good, but for all we know at this point, *all three of the eigenvalues might lie in just one of the intervals*. Instead, we might well expect that there is one eigenvalue in each interval. That is, we might well expect that there is one eigenvalue close to 3, one close to 0, and one close to 2.

This is in fact the case. It can be shown that whenever the Gershgorin disks do not overlap, there is one eigenvalue in each of them. Proofs of this in the general case seem to rely more on complex analysis than linear algebra *per se*, and we won't give such a proof here. But many of our applications will be to the case in which A is symmetric, as in the example, and then there is a fairly simple proof. To explain, we first make some more definitions:

Define the numbers $\delta(A)$ and $r(A)$ by

$$\delta(A) = \min\{\ |A_{i,i} - A_{j,j}| \ :\ i \le i < j \le n\ \}\ , \tag{3.6}$$

and

$$r(A) = \max\{\ r_i(A) \quad 1 \le i \le n\ \}\ . \tag{3.7}$$

That is, $\delta(A)$ is the minimum distance between distinct diagonal elements of A, and $r(A)$ the the maximum of the radii of the Gershgorin disks. Clearly, as long as

$$\delta(A) > 2r(A)\ ,$$

the disks do not overlap.

Theorem 2 (One eigenvalue per disk) *Let A be any symmetric $n \times n$ matrix, and suppose that*

$$r(A) < \frac{\delta(A)}{2}\ . \tag{3.8}$$

Then there is exactly one eigenvalue in each Gershgorin disk of A.

Proof: Supose that there is no eigenvalue in the ith Gershgorin disk. Then all of the eigenvalues of A lie in the other Gershgorin disks, and so if μ is any eigenvalue of A,

$$|\mu - A_{i,i}| > \delta(A) - r(A)\ . \tag{3.9}$$

In particular, $A_{i,i}$ is not an eigenvalue of A, so $(A - A_{i,i}I)^{-1}$ is invertible. The eigenvalues of $(A - A_{i,i}I)^{-1}$ are exactly the numbers $(\mu - A_{i,i})^{-1}$ where μ is an eigenvalue of A. By (3.9), none of these eigenvalues is larger than $(\delta(A) - \rho(A))^{-1}$. But since $(A - A_{i,i}I)^{-1}$ is symmetric, its norm is the maximum absolute value of its eigenvalues. Hence

$$\|(A - A_{i,i}I)^{-1}\| \le (\delta(A) - r(A))^{-1}\ .$$

Now $(A - A_{i,i}I)\mathbf{e}_i$ is just the ith column of $A - A_{i,i}I$, which by the symmetry of A is just the ith row of $A - A_{i,i}I$. Hence

$$|(A - A_{i,i}I)\mathbf{e}_i| = \sqrt{\sum_{\substack{j=1 \\ j \neq i}} |A_{i,j}|^2} \ .$$

Clearly, for $j \neq i$, $|A_{i,j}| \leq r_i(A) \leq r(A)$, and so we have

$$\sum_{\substack{j=1 \\ j \neq i}} |A_{i,j}|^2 \leq r(A) \left(\sum_{\substack{j=1 \\ j \neq i}} |A_{i,j}| \right) \leq (r(A))^2 \ .$$

Hence,

$$|(A - A_{i,i}I)\mathbf{e}_i| \leq r(A) \ .$$

Next, since

$$\mathbf{e}_i = (A - A_{i,i}I)^{-1}(A - A_{i,i}I)\mathbf{e}_i \ ,$$

$$\begin{aligned} 1 = \|\mathbf{e}_i\| &\leq \|(A - A_{i,i}I)^{-1}\||(A - A_{i,i}I)\mathbf{e}_i| \\ &\leq \frac{1}{\delta(A) - r(A)} r(A) \ . \end{aligned}$$

This implies that $\delta(A) \leq 2r(A)$. Hence, under the condition (3.8), it is impossible that the ith Gershgorin disk does not contain an eigenvalue. Since i is arbitrary, each Gershgorin disk contains an eigenvalues. Since there can be no more than n eigenvalues, each contains exactly one. ■

Example 2 (Checking for one eigenvalue per disk) Let A be the matrix $\begin{bmatrix} 3 & 0.1 & -0.1 \\ 0.1 & 0 & 0.1 \\ -0.1 & 0.1 & 2 \end{bmatrix}$ form Example 1. From the computations of the $r_i(A)$ done there, we see that $r(A) = 0.2$. It is also clear that $\delta(A) = 1$. Hence

$$r(A) = 0.2 < \frac{1}{2} = \frac{\delta(A)}{2} \ ,$$

and so (3.8) is satisfied in this case, and we now know that there is exactly one eigenvalue in each of the intervals

$$[2.8, 3.2] \qquad [-0.2, 0.2] \qquad \text{and} \qquad [1.8, 2.2] \ . \tag{3.10}$$

The results we have obtained so far are very useful as they stand. But if we actually calculate the eigenvalues of the matrix A considered in Examples 1 and 2, we find that they are:

$$\mu_1 = 3.0125807... \qquad \mu_2 = -0.0086474... \qquad \text{and} \qquad \mu_3 = 1.9960667... \tag{3.11}$$

where all digits shown are exact.

As you see, they are *very* close to 3, 0 and 2, respectively; much closer than is ensured by (3.10). Was this just luck, or is there a chance to say something more incisive about the location of the eigenvalues?

It was not just luck, as our 2×2 example indicates at the beginning of this section indicates. In fact, it turns out that we can be considerably more incisive. The key fact enabling us to squeezing more out of the Gershgorin disk theorem is the fact that *similar matrices have the same eigenvalues.*

Fix any i with $1 \leq i \leq n$, and any $\alpha > 0$, Let S be the diagonal matrix whose ith diagonal entry is α, and whose other diagonal entries are all 1. For instance, if $n = 4$ and $i = 2$, we have

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

Then SAS^{-1} is obtained from A by multiplying the ith row through by α, and the ith column through by $1/\alpha$. In particular, the two factors cancel on the diagonal, and so the ith diagonal entry is unchanged, as are all of the other diagonal entries.

Since every off–diagonal entry in the ith row of A gets multiplied by α we have

$$r_i(SAS^{-1}) = \alpha r_i(A) . \tag{3.12}$$

If $\alpha < 1$, this change shrinks the ith Gershgorin disk. Unfortunately, it expands the others: For $k \neq i$, $|A_{i,k}| \leq r_k(A)$, and so

$$r_k(SAS^{-1}) = r_k(A) + (1/\alpha - 1)|A_{k,i}| \leq (1/\alpha)r_k(A) .$$

That is,

$$r(SAS^{-1}) \leq \frac{1}{\alpha} r(A) .$$

Since the diagonal entries of SAS^{-1} are the same as the diagonal entries of A,

$$\delta(SAS^{-1}) = \delta(A) .$$

Therefore, as long as

$$\frac{1}{\alpha} r(A) < \frac{\delta(A)}{2} , \tag{3.13}$$

Theroem 2 says that each Gershgorin disk of SAS^{-1} contains one eigenvalue of SAS^{-1}, and hence of A. We now choose α as small as possible while still keeping (3.8) satisfied so there will be one eigenvalue in each disk. This shrinks the radius of the ith Gershgorin disk as much as possible while sitll making sure it includes an eigenvalue of A.

By (3.13), the smallest admissible value for α is

$$\alpha_{\min} = \frac{2r(A)}{\delta(A)} \ .$$

Using this value of α in (3.12), the radius of the ith Gershgorin of SAS^{-1} becomes

$$r_i(SAS^{-1}) = \frac{2r(A)}{\delta(A)} r(A) \leq \frac{2}{\delta(A)} (r(A))^2 \ .$$

The disk of this radius about $A_{i,i}$ contains one eigenvalue of SAS^{-1}, and hence of A. Since i is arbitrary, this proves the following result:

Theorem 3 (Small Gershgorin Disks) *Let A be any symmetric $n \times n$ matrix. Suppose that $r(A) < \delta(A)/2$. Then for all i with $1 \leq i \leq n$, the disk of radius*

$$\frac{2}{\delta(A)} (r(A))^2$$

about $A_{i,i}$ contains exactly one eigenvalue of A.

Example 3 (Checking for one eigenvalue per small disk) Let A be the matrix $\begin{bmatrix} 3 & 0.1 & -0.1 \\ 0.1 & 0 & 0.1 \\ -0.1 & 0.1 & 2 \end{bmatrix}$ from Examples 1 and 2. From the computations of the $r_i(A)$ done there, we see that $r(A) = 0.2$. It is also clear that $\delta(A) = 1$. Hence

$$\frac{2}{\delta(A)} (r(A))^2 = 0.08 \ ,$$

and so (3.8) is satisfied in this case, and we now know that there is exactly one eigenvalue in each of the intervals

$$[2.92, 3.08] \qquad [-0.08, 0.08] \qquad \text{and} \qquad [1.92, 2.08] \ . \tag{3.14}$$

These intervals are much narrower than before – the radius is 0.08 instead of 0.2. They still comfortably contain the eigenvalues (3.11), as they must.

3.2 Application to Jacobi iteration

Suppose we are given an $n \times n$ symmetric matrix A, and are asked to compute the eigenvalues of A to 10 digits of accuracy. That is, if μ_i denotes the ith eigenvalue of A, we want to compute explicit numbers d_i so that for each i

$$\mu_i = d_i \pm 10^{-10} \ . \tag{3.15}$$

Of course, the numbers d_i will be the diagonal entries of some matrix $A^{(m)}$ that is obtained from A by running m steps of the Jacobi algorithm.

The results we have obtained enable us to be sure that what we are computing with the Jacobi algorithm are actually the eigenvalues of A. We want to stop the computation at the first m such that (3.15) is satisifed.

• *Can we devise a "stopping rule" that we could code into an implementation of the Jacobi algorithm that would guarantee the validity of (3.15) at the stopping point? Moreover, can we be sure that this stopping rule always leads to termination in a finite number of steps for all matrices A?*

The answer is yes, and the stopping rule is rather simple. To explain it in a generally useful form, let us replace the specific accuracy level 10^{-10} by ϵ. Then we have the following theorem:

Theorem 4 (Stopping rule for the Jacobi algorithm) *Let $\epsilon > 0$ be any given positive number. For any $n \times n$ symmetric matric A, let $\mu_1, \mu_2, \ldots, \mu_n$ be the eigenvalues of A, arranged in decreasing order. Let $A^{(m)}$ be the matrix produced at the mth step by the Jacobi algorithm, and let $d_1^{(m)}, d_2^{(m)}, \ldots, d_n^{(m)}$ be the diagonal entries of $A^{(m)}$, arranged in decreasing order. Then if*

$$\text{Off}(A^{(m)}) < \frac{1}{n-1}\left(\frac{\epsilon}{2n-1}\right)^2 ,$$

it follows that

$$|\mu_i - d_i^{(m)}| < \epsilon$$

for each $i = 1, \ldots, n$.

Since we know from the previous section that

$$\text{Off}(A^{(m)}) \leq \left(1 - \frac{2}{n^2-n}\right)^m \text{Off(A)} , \tag{3.16}$$

it follows that the stopping rule "kicks in" for A no later than the mth step where m is the smallest integer such that

$$\left(1 - \frac{2}{n^2-n}\right)^m \text{Off(A)} < \frac{1}{n-1}\left(\frac{\epsilon}{2n-1}\right)^2 . \tag{3.17}$$

In particular, no matter what accuracy level we set for our computation, we can compute the eigenvalues of A to that accuracy in a finite number of steps. We can even set a "worst case" upper bound on the number of steps before we begin the computation. It must be stressed, however, that this upper bound on the number of steps is indeed based on a "worst case" analysis. In fact, (3.16), while always true, is rather pessimistic. In practice, the decrease of $\text{Off}(A^{(m)})$ is much, much faster. Therefore, in pratice the stopping rule "kicks" much, much more quickly. Still, it is important to know that it always does kick in at some finite step. If you implement the Jacobi algorithm in code, your program will never go into an infinite loop.

The rest of this subsection is devoted to proving Theorem 4. Before we begin the proof, we prove a useful lemma relating Off(B) and $r(B)$ for an $n \times n$ symmetric matrix.

Lemma *Let B be any $n \times n$ matrix. Then*

$$(r(B))^2 \leq (n-1)\text{Off}(B) \ . \tag{3.18}$$

Proof: For any i, by the Schwarz inequality,

$$r_i(B) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |B_{i,j}| \leq \sqrt{n-1} \sqrt{\sum_{\substack{j=1 \\ j \neq i}}^{n} |B_{i,j}|^2} \ .$$

Hence

$$\begin{aligned}(r(B))^2 &= \max\{\ (r_i(B))^2 \ :\ 1 \leq i \leq n\} \\ &\leq \sum_{i=1}^{n} (r_i(B))^2 \\ &\leq \sum_{i=1}^{n} (n-1) \sum_{\substack{j=1 \\ j \neq i}}^{n} |B_{i,j}|^2 \\ &\leq (n-1)\text{Off}(B) \ .\end{aligned} \tag{3.19}$$

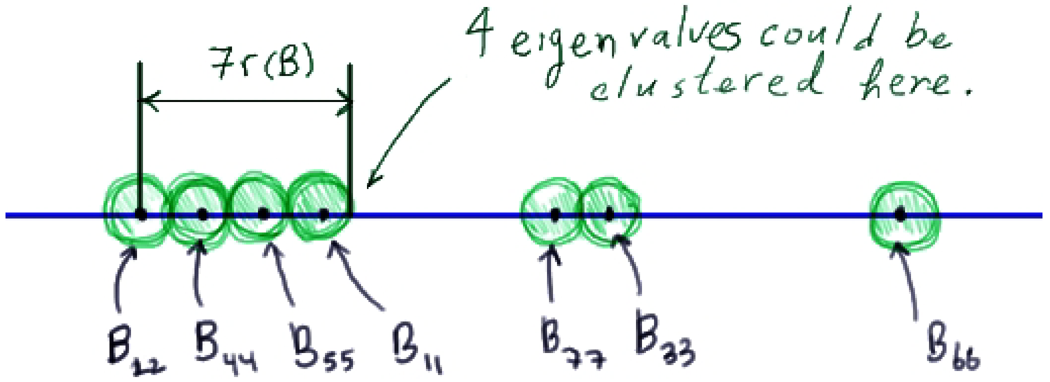
■

Proof of Theorem 4 Our goal is to compute the eigenvalues of A to a given accuracy level $\pm\epsilon$. We run the Jacobi algorithm some number of steps, resulting in an "almost diagonal" matrix B. We would like to be sure that the eigenvalues of B, and hence of A, are given by by the diagonal entries of B up to order ϵ.

It would be nice to apply the small Gershgorin Disk Theorem, except there is no way to tell in advance how small $\delta(B)$ might be – it might even be zero.*

Therefore, we have to deal with overlapping disks. To analyze this case, it is best to draw a picture. The following picture shows 7 disks of radius $r(B)$ with several overlapping clusters. What can we say about the eigenvalues of B in this case?

* If the eigenvlaues of A are all distinct, then after enough steps of the Jacobi algorithm, $\delta(B) \approx \delta(A) > 0$, but we cannot know this in advance: We are trying to compute the eigenvalues.



It can be shown, using the same ideas that were used to prove Theorem 2, that each of the clusters of overlapping disks contains as many eigenvalues as there are disks in the cluster. In the picture, you see three clusters: The cluster on the left has 4 disks, the middle cluster has two, and the cluster on the right has just one.

By what we have explained above, the cluster on the left contains 4 eigenvalues, but they can be anywhere in the cluster – really. So if we are really unlucky, they might all be bunched up at the extreme right end of the cluster, as far away from $B_{2,2}$ as possible. Then the distance from $B_{2,2}$ to the nearest eigenvalue is 3 diameters and one radius; i.e., $7r(B)$, as in the picture. (We have drawn the worst case, where the disks "just barely" overlap. Otherwise, the distance would be less).

In general, you see that in a cluster of k disks covering $B_{i,i}$, the distance from $B_{i,i}$ to the nearest eigenvalue is no more than $k-1$ diameters plus one radius; i.e., $(2k-1)r(B)$. Since $k \leq n$, we have that in any case, no matter how bad the clustering is, there is an eigenvalue μ of B satisfying

$$|B_{i,i} - \mu| \leq (2n-1)r(B) \ .$$

Now by the lemma, for any $\epsilon > 0$,

$$\mathrm{Off}(B) < \frac{1}{n-1}\left(\frac{\epsilon}{2n-1}\right)^2 \qquad \Rightarrow \qquad r(B) < \frac{\epsilon}{2n-1} \ .$$

It therefore follows that when

$$\mathrm{Off}(B) < \frac{1}{n-1}\left(\frac{\epsilon}{2n-1}\right)^2 \ , \tag{3.20}$$

each diagonal entry of B is within ϵ of some eigenvalue of B. ■

3.3 Continuity of Eigenvalues

For any symmetric $n \times n$ matrix A, there is an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ consisting of eigenvectors of A, Let μ_j denote the eigenvalue corresponding to $\mathbf{u}_j$. We may redistribute the labels, if need be, to arrange that

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \ .$$

In our discussion so far, A has been a fixed matrix, and we have learned how to compute the eigenvalues in terms the entries of A. We do this by an iterative procedure, but all the same, given A and j, μ_j is a well determined number. To explicitly indicate that it came from A, let us write $\mu_j(A)$ to denote the jth largest (with repitition) eigenvalue of A.

Let us now shift our points of view, and think of the assignment

$$A \mapsto \mu_j(A)$$

as it a function on the space of $n \times n$ symmetric matrices. We call this function the jth *eigenvalue function.*

Our concern in this subsection is with the *continuity* of these functions. For functions f on sets of matrices, the definition of continuity is very much like the one for functions on $I\!R^n$.

Definition (Continuity of Matrix Functions) Let f be a fucntion defined on a set U of $m \times n$ matrices. If A is in U, then f *is continuous at* A in case for all $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ so that

$$\|A - B\| \leq \delta(\epsilon) \ \Rightarrow \ |f(A) - f(B)| < \epsilon \tag{3.21}$$

whenever B belongs to U, the domain of definition of f.

The importance of the definition is this: In many applications, one is not working with the exact matrix A, but only some decimal approximation to it. This is certainly the case if the entries of A are irrational, and you are doing your computations on a machine. In this case, *even if you made no further round–off errors*, you would be computing the eigenvalues of some other matrix C that is close to A, a "rounded off" version of A, but is not exactly A. Is this good enough?

This is only good enough if, whenever C is close enough to A, then each eigenvalue of C is close to the corresponding eigenvalue of A. In other words, this is good enough if and only if the eigenvalues are continuous functions of the matrix A. Fortunately they are, as we shall see in this subsection.

To examine the continuity of the eigenvalues at A, let C be any other $n \times n$ symmetric matrix. We wish to estimate $|\mu_j(C) - \mu)j(A)|$ in terms of $\|C - A\|$.

If we define $B = C - A$, so that

$$C = A + B \ ,$$

our goal then is to estimate $|\mu_j(A+B) - \mu_j(A)|$ in terms of $\|B\|$.

For this purpose, it is best to use the orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ consisting of eigenvectors of A that we introduced above. Then

$$A\mathbf{u}_j = \mu_j(A)\mathbf{u}_j \ .$$

(Note: the eigenvecotors $\mathbf{u}_j$ also depend on A, but we do not record this in our notation since we are only studying the dependence of the eigenvalues.)

Let $Q = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]$ and let D be the diagonal entry whose jth diagonal entry is $\mu_j(A)$. Then $Q^tAQ = D$. Define G by $G = Q^tBQ$, so that

$$Q^t(A+B)Q = D + G \ .$$

Since Q is orthogonal, so that $Q^t = Q^{-1}$ this says that $A+B$ is similar to $D+G$. Then since similar matrices have the same eigenvalues,

$$\mu_j(A+B) = \mu_j(D+G) \ . \tag{3.22}$$

Furthermore,

$$\|B\| = \|G\| \ . \tag{3.23}$$

To see this, note that $G^tG = Q^tBQQ^tBQ = Q^tB^tBQ$ so that G^tG is similar to B^tB, Hence both B^tB and G^tG have the same largest eigenvalue. Since the norm of a matrix is the square root of the largest eigenvalue, this implies (3.23).

Now our problem is to estimate $|\mu_j(D+G) - \mu_j(D)|$ in terms of $\|G\|$. For this, we use the fact that for each i and j,

$$G_{i,j} \le \|G\| \ .$$

Indeed, by the Schwarz inequality and the definition of the norm,

$$|G_{i,j}| = |\mathbf{e}_i \cdot G\mathbf{e}_j| \le |\mathbf{e}_i||G\mathbf{e}_j| \le \|G\||\mathbf{e}_i||\mathbf{e}_j| = \|G\| \ . \tag{3.24}$$

Therefore, for each i,

$$r_i(D+G) = \sum_{j \ne i} |G_{i,j}| \le (n-1)\|G\| \ . \tag{3.25}$$

Also, by defintion, the ith diagonal entry of G is

$$\mathbf{e}_i \cdot G\mathbf{e}_i = \mathbf{e}_i \cdot Q^tBQ\mathbf{e}_i = (Q\mathbf{e}_i) \cdot B(Q\mathbf{e}_i) = \mathbf{u}_i \cdot B\mathbf{u}_i \ .$$

Therefore,

$$(D+G)_{i,i} = \mu_i(A) + \mathbf{u}_i \cdot B\mathbf{u}_i \ . \tag{3.26}$$

By the first Gershgorin disk Theorem,

$$\mu_i(D+G) - (\mu_i(A) + \mathbf{u}_i \cdot B\mathbf{u}_i)) \,| \le (n-1)\|B\| \ .$$

But by (3.22), and the fact that $|\mathbf{u}_i \cdot B\mathbf{u}_i)| \le \|B\|$, we now have

$$|\mu_i(A+B) - \mu_i(A)| \le n\|B\| \ .$$

We have now proved the following:

Theorem 5 (Continuity of eigenvalues) *Let A and C be any $n \times n$ symmetric matrices. Then for each $i = 1, \ldots, n$,*

$$|\mu_i(A) - \mu_i(C)| \le n\|A - C\| \ . \tag{3.27}$$

In particular, the function $A \mapsto \mu_i(A)$ is continuous on the set of $n \times n$ symmetric matrices.

Notice that (3.27) says that we can use $\delta(\epsilon) = \epsilon/n$ in (3.21) for the ith eigenvalue function on the set on $n \times n$ symmetric matrices.

In the next section we shall discuss the case of non symmetric matrices. The situation there is complicated by the fact that these cannot always be diagonalized. However, the case of symmetric matrices is particularly important, and it is worth observing that we can push our results a bit further in case all of the eigenvalues of A are distinct, so that we may use the Small Gershgorin Disk Theorem.

Suppose that A is not only symmetric, but is also *non degenrate*, meaning that it has not repeated eigenvalues:

$$\mu_1(A) > \mu_2(A) > \cdots > \mu_n(A) \ .$$

Let δ denote the least gap between these eigenvalues: i.e.,

$$\delta(A) = \min_{j=1,\ldots,n-1} (\mu_j(A) - \mu_{j+1}(A)) \ .$$

Then, since by (3.25) and (3.23), $r(A) \le (n-1)\|B\| = (n-1)\|G\|$, whenever

$$2(n-1)\|B\| < \delta(A) \ ,$$

the conditions of the Small Gershgorin Disk Theorem are satisfied, and the distance between $\mu_i(A+B)$ and the ith diagonal entry of $D+G$ is no greater than $2(n-1)^2\|B\|^2/\delta(A)$. By (3.26), the ith diagonal entry of $D+G$ is $\mu_i(A) + \mathbf{u}_i \cdot B\mathbf{u}_i$. Therefore,

$$|\mu_i(A+B) - (\mu_i(A) + \mathbf{u}_i \cdot B\mathbf{u}_i)| \le \frac{2(n-1)^2\|B\|^2}{\delta(A)} \ . \tag{3.28}$$

This formula shows that if $\delta(A) > 0$, then the function $A \mapsto \mu_i(A)$, is not only continuous, it is also it differentiable. In particular, for any no zero value of t, it follows from (3.28) that

$$\left| \frac{\mu_i(A+tB) - \mu_i(A)}{t} - \mathbf{u}_i \cdot B\mathbf{u}_i \right| \le t\frac{2(n-1)^2\|B\|^2}{\delta(A)} \ . \tag{3.29}$$

Therefore, whenever $\delta(A) > 0$,

$$\lim_{t\to 0} \frac{\mu_i(A+tB) - \mu_i(A)}{t} = \mathbf{u}_i \cdot B\mathbf{u}_i \ .$$

This gives a formula for the "directional derivative" of μ_i in the space of $n \times n$ symmetric matrices. In fact, another way to write (3.28) is

$$\mu_i(A+tB) = \mu_i(A) + t\,(\mathbf{u}_i \cdot B\mathbf{u}_i) + \mathcal{O}(t^2) \ . \tag{3.30}$$

We have now proved the following:

Theorem 6 (Differentiability of eigenvalues) *Let A and B be any $n \times n$ symmetric matrices. Suppose that $\delta(A) > 0$, and for each $i = 1, \ldots, n$, let $\mathbf{u}_i$ be a normalized eigenvector corresponding to the ith eigenvalue of A. Then $t \mapsto \mu_i(A+tB)$ is a differentiable function of t at $t = 0$, and*

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mu_i(A+tB) = \mathbf{u}_i \cdot B\mathbf{u}_i \ .$$

In fact, the somewhat stronger staement (3.30) is also true.

Example 4 (Derivatives of eigenvalues) Let A be the matrix $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$, and let B be the matrix $B = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}$. Since A is already diagonal, we can take $\mathbf{u}_1 = \mathbf{e}_1$, $\mathbf{u}_2 = \mathbf{e}_2$ and $\mathbf{u}_3 = \mathbf{e}_3$. Then $\mu_1 = 3$, $\mu_2 = 1$ and $\mu_3 = 2$. Hence the eigenvalues $\mu_i(t)$ of $A + tB$ satisfy

$$\begin{aligned} \mu_1(t) &= 3 + 2t + \mathcal{O}(t^2) \\ \mu_2(t) &= 1 + 3t + \mathcal{O}(t^2) \\ \mu_3(t) &= 2 + 2t + \mathcal{O}(t^2) \end{aligned}$$

and so

$$\mu_1'(0) = 2 \qquad \mu_2'(0) = 3 \qquad \text{and} \qquad \mu_3'(0) = 2 \ .$$

We have seen that $A \mapsto \mu_i(A)$ is differnetiable at A in case $\delta(A) > 0$. what happens if $\delta(A) = 0$? Then, indeed, $\mu_i(A)$ may not be differentiable.

Example 5 (Nondiferentiability of $t \mapsto \mu_i(A+tB)$ where $\delta(A) = 0$) Let A be the matrix $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, and let B be the matrix $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. It is easy to compute the two eigenvalues of $A + tB$, using, for example, the formulas from Section 1. The result is:

$$\mu_\pm = 2 \pm t \ .$$

Now notice that $\mu_1(A+tB)$ is, by definition, the larger of these two eigenvalues. Which one is larger depends on the sign of t, and the result is that

$$\mu_1(A+tB) = 2 + |t| \ .$$

Likewise,

$$\mu_2(A+tB) = 2 - |t| \ .$$

Notice that these are not differentiable at $t = 0$. This example shows that the role of the condition $\delta(A) > 0$ is not artificial: This condition prevent "crossing" of eigenvalues for small enough values of t. At the crossings, nothing dramatic happens to the the eigenvalues, but they "switch tracks" in a non differentiable way.

Problems

1. Let $A = \begin{bmatrix} 1 & 0.01 & -0.01 \\ 0.01 & 5 & 0.01 \\ -0.01 & 0.01 & 3 \end{bmatrix}$.

(a) Compute $r_i(A)$ for $i = 1, 2, 3$.

(b) Compute $r(A)$ and compute $\delta(A)$.

(c) Find three small intervals about 1, 5 and 3 that are guaranteed, by Theorem 3, to contain the eigenvalues of A.

2. Let $A = \begin{bmatrix} -1 & 0.02 & -0.01 \\ 0.02 & 2 & 0.03 \\ -0.01 & 0.03 & 4 \end{bmatrix}$.

(a) Compute $r_i(A)$ for $i = 1, 2, 3$.

(b) Compute $r(A)$ and compute $\delta(A)$.

(c) Find three small intervals about -1, 2 and 4 that are guaranteed, by Theorem 3, to contain the eigenvalues of A.

3. Let $A = \begin{bmatrix} -1 & 0.002 & -0.001 \\ 0.002 & 3 & 0.003 \\ -0.001 & 0.003 & 5 \end{bmatrix}$.

(a) Compute $r_i(A)$ for $i = 1, 2, 3$.

(b) Compute $r(A)$ and compute $\delta(A)$.

(c) Find three small intervals about -1, 3 and 5 that are guaranteed, by Theorem 3, to contain the eigenvalues of A.

4. Let $A = \begin{bmatrix} 31 & 0.001 & -0.001 \\ 0.001 & 8 & 0.001 \\ -0.001 & 0.001 & 9 \end{bmatrix}$.

(a) Compute $r_i(A)$ for $i = 1, 2, 3$.

(b) Compute $r(A)$ and compute $\delta(A)$.

(c) Find three small intervals about 3, 8 and 9 that are guaranteed, by Theorem 3, to contain the eigenvalues of A.

5. Let A be the matrix $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and let B be the matrix $B = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$. Compute expressions for the eigenvalues $\mu_i(t)$ of $A + tB$ that are valid up to corrections of size $\mathcal{O}(t^2)$.

6. Let A be the matrix $A = \begin{bmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and let B be the matrix $B = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$. Compute expressions for the eigenvalues $\mu_i(t)$ of $A + tB$ that are valid up to corrections of size $\mathcal{O}(t^2)$.

7. Let A be the matrix $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 2 \\ 0 & 2 & 4 \end{bmatrix}$, and let B be the matrix $B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \end{bmatrix}$. Compute expressions for the eigenvalues $\mu_i(t)$ of $A + tB$ that are valid up to corrections of size $\mathcal{O}(t^2)$.

8. Let B be the $n \times n$ matrix that has 1 in every entry in the first row, and 0 elsewhere. Compute $\mathrm{Off}(B)$ and $r(B)$, and compare with the inequality $(r(B))^2 \leq (n-1)\mathrm{Off}(B)$ that was derived in this section.

9. Show that for symmetric $n \times n$ matrices B, the inequality $(r(B))^2 \leq (n-1)\mathrm{Off}(B)$ that was derived in this section can be improved to

$$(r(B))^2 \leq \frac{(n-1)}{2}\mathrm{Off}(B) \ ,$$

and find an $n \times n$ symmetric matrix B for which equality holds in this inequality.

Section 4: The singular value decomposition

4.1 What is a singular value decomposition?

We can *diagonalize* any symmetric $n \times n$ matrix A: Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an orthonormal basis of $I\!R^n$ consisting of eigenvectors of A. Let U be the $n \times n$ matrix whose jth column is U. Then $U^t A U = D$ where D is the diagonal matrix whose j diagonal entry is μ_j, the eigenvalue corresponding to $\mathbf{u}_j$. Multiplying on the left by U, and on the right by U^t, we get

$$A = UDU^t \tag{4.1}$$

since U is orthogonal, so that $U^t = U^{-1}$.

In (4.1), we have a factorization of A into simple pieces – we have "taken it apart" into simple pieces that are easy to work with and understand. "Decompositions" of matrices into simple pieces are very useful in a wide variety of problems involving matrices, and they will be very useful to us in the next chapters when, for example, the matrices in questions are the Jacobians of some non linear transformation

In general, Jacobian matrices are not symmetric, and not even square. However, there is a decomposition very much like (4.1) that is valid for *all* matrices – square or not.

Definition (Singular Value Decomposition) Let A be an $m \times n$ matrix. Suppose D is an $r \times r$ diagonal matrix with strictly positive diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_1 > 0$, Suppose that U and V are $n \times r$ and $m \times r$ matrices, respectively, with orthonormal columns. Suppose finally that

$$A = VDU^t \tag{4.2}$$

Then the positive entries of D are called the *singular values* of A, and the decomposition (4.2) is called the *singular value decomposition.*

As we shall soon see, *every* matrix has a singular value decomposition. First, let us try to understand what information about A is encoded into a singular value decomposition of A.

For this purpose, another way of writing a singular value decomposition is very illuminating. To explain, first recall that if $\mathbf{u}$ is any vector in $I\!R^n$, and $\mathbf{v}$ is any vector in $I\!R^m$, then we may regard them both as matrices – an $n \times 1$ matrix, and and $m \times 1$ matrix respectively. Then the matrix product $\mathbf{vu}^t$ makes sense: It is the product of an $m \times 1$ matrix and a $1 \times n$ matrix. Evidently the result is an $m \times n$ matrix. This sort of matrix product of two vectors is sometimes called their *outer product.*

Example 1 (Outer products) Let $\mathbf{v} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ and let $\mathbf{u} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$. Then

$$\mathbf{vu}^t = \begin{bmatrix} 2 \\ 4 \end{bmatrix} [\,1 \quad 3 \quad 5\,] = \begin{bmatrix} 2 & 6 & 10 \\ 4 & 12 & 20 \end{bmatrix} \ .$$

As you can easily see by pondering the example, in any matrix of the form $\mathbf{v}\mathbf{u}^t$, all of the columns are multiples of a single vector, namely $\mathbf{v}$. Hence the rank of any such matrix is one.

Conversely, if B is an $m \times n$ matrix with rank one, then all of the columns must be multiples of a single vector $\mathbf{v}$ – or else the span of the columns would be at least two dimensional. Hence for some numbers $a_1, a_2, \ldots, a_n$,

$$B = [a_1\mathbf{v}, a_2\mathbf{v}, \ldots, a_n\mathbf{v}] = \mathbf{v}\mathbf{u}^t \qquad \text{where} \qquad \mathbf{u} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} .$$

Hence, an $m \times n$ matrix is a rank one matrix if and only if it has the form $\mathbf{v}\mathbf{u}^t$.

The relevance of this to us here is that a singular value decomposition of A provides a decomposition of an $m \times n$ matrix A into a sum of r rank one matrices, where r turns out to be the rank of A. Indeed, Let $A = VDU^t$ be a singular value decomposition of and $m \times n$ matrix A, and let us write

$$V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r] \qquad \text{and} \qquad U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] .$$

Then

$$A = \sigma_1\mathbf{v}_1\mathbf{u}_1^t + \sigma_2\mathbf{v}_2\mathbf{u}_2^t + \cdots + \sigma_r\mathbf{v}_r\mathbf{u}_r^t . \tag{4.3}$$

As we now explain, this *additive* decomposition is just another way of writing the *multiplicative* decomposition (4.2)

To see that (4.3) is just another way of writing (4.2), we need to make one more observation about rank one matrices. We know that matrices represent linear transformations. What linear transformation does $\mathbf{v}\mathbf{u}^t$ represent? To answer that, apply $\mathbf{v}\mathbf{u}^t$ to a general vector $\mathbf{x}$ in $I\!R^n$, and see what happens. By associativity, $(\mathbf{v}\mathbf{u}^t)\mathbf{x} = \mathbf{v}(\mathbf{u}^t\mathbf{x})$, and by the rule for matrix–vector multiplication in row form, $\mathbf{u}^t\mathbf{x} = \mathbf{u} \cdot \mathbf{x}$. Therefore,

$$(\mathbf{v}\mathbf{u}^t)\mathbf{x} = (\mathbf{u} \cdot \mathbf{x})\mathbf{v} . \tag{4.4}$$

Now, let us apply the terms in $A = VDU^t$ to some general vector $\mathbf{x}$ in $I\!R^n$ one at a time, and see what happens:

First, for any $\mathbf{x}$ in $I\!R^n$,

$$U^t\mathbf{x} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_r \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{u}_1 \cdot \mathbf{x} \\ \mathbf{u}_2 \cdot \mathbf{x} \\ \vdots \\ \mathbf{u}_r \cdot \mathbf{x} \end{bmatrix} .$$

Next,

$$DU^t\mathbf{x} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \cdot \mathbf{x} \\ \mathbf{u}_2 \cdot \mathbf{x} \\ \vdots \\ \mathbf{u}_r \cdot \mathbf{x} \end{bmatrix} = \begin{bmatrix} \sigma_1\mathbf{u}_1 \cdot \mathbf{x} \\ \sigma_2\mathbf{u}_2 \cdot \mathbf{x} \\ \vdots \\ \sigma_r\mathbf{u}_r \cdot \mathbf{x} \end{bmatrix}$$

Finally then,

$$\begin{aligned} VDU^t\mathbf{x} &= [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r] \begin{bmatrix} \sigma_1\mathbf{u}_1 \cdot \mathbf{x} \\ \sigma_2\mathbf{u}_2 \cdot \mathbf{x} \\ \vdots \\ \sigma_r\mathbf{u}_r \cdot \mathbf{x} \end{bmatrix} \\ &= (\sigma_1\mathbf{u}_1 \cdot \mathbf{x})\mathbf{v}_1 + (\sigma_2\mathbf{u}_2 \cdot \mathbf{x})\mathbf{v}_2 + \cdots + (\sigma_r\mathbf{u}_r \cdot \mathbf{x})\mathbf{v}_r \ . \end{aligned} \tag{4.5}$$

By (4.4), this is equivalent to (4.3).

Many applications of the singular value decomposition start from the additive form (4.3). Here is a hint of one that we shall explore more thoroughly later on:

Consider an image in grayscale with a size of, say 1000 by 2000 pixels. This corresponds to a 1000 × 2000 matrix, with 2 million entries, each being an integer in the range 0 to 255. To store or transmit the image faithfully, one must store or transmit all 2 million pixel values. But the *information content* in the image can be stored or transmitted much more efficiently. Let

$$A = \sum_{j=1}^{r} \sigma_j \mathbf{v}_j \mathbf{u}_j^t$$

be a singular value decomposition of the matrix. As we shall see, r is the rank of A, and so this is no more than 1000. Now, here is the key fact: If A comes from an image with structure, like a picture of a face, The first few singular values will be much, much larger than the rest. Hence, we can get a good approximation to A by changing σ_j to zero for all j greater than, say, 20. This gives us the approximation

$$A \approx \sum_{j=1}^{r} \sigma_j \mathbf{v}_j \mathbf{u}_j^t \tag{4.6}$$

Now, to store or transmit the right hand side, you just need to store or transmit the 20 singular values $\sigma_1, \dots, \sigma_{20}$, the 20 vectors $\mathbf{u}_1, \dots, \mathbf{u}_{20}$ with 2000 entries each, and the 20 vectors $\mathbf{v}_1, \dots, \mathbf{v}_{20}$ with 1000 entries each. This is a total of 60,020 numbers, which is a lot less than the 2 million entires of A. Given this data, you can reconstruct the approximation A in (4.6), and hence an approximation to the original image. The basis idea here has many applications that will be explored in projects.

Before delving further into applications of the singular value decomposition, or expaling how to compute one, we close this section with a useful theorem that sheds more light on our current focus – what the singular value decomposition is.

Theorem 1 (Singular value decomposition, rank and bases) *Let A be an $m \times n$ matrix, and let $A = VDU^t$ be a singular value decomposition of A where D is an $r \times r$ matrix. Then r is the rank of A, the columns of V are an orthonormal basis for* Img(A)*, and the the columns of U are an orthonormal basis for* Img$(A^t) = (\text{Ker}(A))^\perp$. *In particular, the orthogonal projections*

$$VV^t \qquad \text{and} \qquad UU^t$$

project onto Img(A) *and* $(\text{Ker}(A))^\perp$*, respectively.*

Proof: Every vecotr in Img(A) is, by definition, of the form $A\mathbf{x}$ for some $\mathbf{x}$ in $I\!R^n$. But by (4.5),

$$A\mathbf{x} = (\sigma_1\mathbf{u}_1\cdot\mathbf{x})\mathbf{v}_1 + (\sigma_2\mathbf{u}_2\cdot\mathbf{x})\mathbf{v}_2 + \cdots + (\sigma_r\mathbf{u}_r\cdot\mathbf{x})\mathbf{v}_r\ ,$$

which is a linear combination of the columns of V. Hence, the columns of V span Img(A). Since they are orthonormal, they are also linearly independent, and hence are an orthonormal basis for Img(A). The rank of A is the dimension of its image, and since we have found a basis with r elements, this shows that rank$(A) = r$.

We also know that to compute the orthognal projection $P_{\text{Img}(A)}$ onto Img(A), we can use $P = QQ^t$ for any matrix Q whose columns are an orthonormal basis for Img(A). V is such a matrix, and hence $P_{\text{Img}(A)} = VV^t$.

To deduce the statements about U, take the transpose of $A = VDU^t$, getting $A^t = UDV^t$. Thus, U for A is V for A^t. Hence what we have just seen shows that the columns of U are an orthonormal basis for Img(A^t), and $P_{\text{Img}(A^t)} = VV^t$. Finally, recall that Img$(A^t) = (\text{Ker}(A))^\perp$. ■

4.2 The singular value decomposition and least square solutions

In this subsection we explain an important application of the singular value decomposition to the solution of least squares problems.

Fist notice that since the diagonal entries of D are all strictly positive, D is invertible, and

$$D^{-1} = \begin{bmatrix} 1/\sigma_1 & 0 & \dots & 0 \\ 0 & 1/\sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_r \end{bmatrix}\ .$$

Hence the matrix product $UD^{-1}V^t$ is always well defined whenver $A = VDU^t$ is a singular value decomposition of A. This matrix is called the *generalized inverse* of A:

Definition(Generalized Inverse) Let $A = VDU^t$ be a singular value decompostion of A. Then the *generalized inverse* of A is the matrix

$$A^+ = UD^{-1}V^t\ . \tag{4.7}$$

Theorem 2 (Singular Values and Least Squares) *Let A be an $m\times n$ matrix, and suppose that $A = VDU^t$ is a singular value decomposition of A. Let A^+ be the generalized inverse of A. Then for any* $\mathbf{b}$ *in* $I\!R^m$*,* $A^+\mathbf{b}$ *is a least squares solution to* $A\mathbf{x} = \mathbf{b}$*, its length is less than that of any other least squares solution.*

Proof: Notice that

$$AA^+ = VDU^tUD^{-1}V^t = VD^{-1}DV^t = VV^t$$

since $U^tU = I$. By Theorem 1, VV^t is the orthogonal projection onto $\mathrm{Img}(A)$, and hence for any $\mathbf{b}$ in $\mathbb{R}^m$, $A(A^+\mathbf{b})$ is the orthogonal projection of $\mathbf{b}$ onto $\mathrm{Img}(A)$; i.e., the vector in $\mathrm{Img}(A)$ that is closest to $\mathbf{b}$. Hence $A^+\mathbf{b}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$. Next, from the formula for A^+,

$$A^+\mathbf{b} = U(D^{-1}V^t\mathbf{b})$$

is a linear compbination of the columns of U, and hence belongs to $\mathrm{Img}(A^t) = (\mathrm{Ker}(A))^\perp$. That is, $A^+\mathbf{b}$ is orthogonal to very vector in $\mathrm{Ker}(A)$. Now since $A^+\mathbf{b}$ is a particular least sqaures solution to $A\mathbf{x} = \mathbf{b}$, any other least squares solution $\mathbf{x}$ can be written as

$$\mathbf{x} = A^+\mathbf{b} + \mathbf{w}$$

where $\mathbf{w}$ belongs to $\mathrm{Ker}(A)$. But then by the orthogonality,

$$|\mathbf{x}|^2 = |A^+\mathbf{b}|^2 + |\mathbf{w}|^2$$

which is larger than $|A^+\mathbf{b}|^2$ unless $\mathbf{w} = 0$. ■

Example 2 (*SVD* **and least squares)** Let A be the matrix $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \\ 1 & 2 \end{bmatrix}$. In this case, let

$$V = \frac{1}{3\sqrt{5}}\begin{bmatrix} 2 & -6 \\ 4 & 3 \\ 5 & 0 \end{bmatrix} \quad , \quad S = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad U = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \ . \tag{4.8}$$

Then, as you can check, $A = VDU^t$. You should also check at this point that U and V are isometries.

Let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. It was claimed above that $UD^{-1}V^t\mathbf{b}$ is the least squares solution to $A\mathbf{x} = \mathbf{b}$ with the least length. In this case, you can see that the columns of A are linearly independent so that $\mathrm{Ker}(A) = 0$ and there is exactly one least squares solution. So in this example, we can ignore the least length part. We just want to see that $A^+\mathbf{b}$ gives us *the* least squares solution to $A\mathbf{x} = \mathbf{b}$.

Working $A^+\mathbf{b} = UD^{-1}V^t\mathbf{b}$ out,

$$\begin{aligned} A^+\mathbf{b} = UD^{-1}V^t\mathbf{b} &= \frac{1}{15}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} 2 & 4 & 5 \\ -6 & 3 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{15}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} 11 \\ -3 \end{bmatrix} \\ &= \frac{1}{15}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 11/3 \\ -3/2 \end{bmatrix} \\ &= \frac{1}{18}\begin{bmatrix} 8 \\ 7 \end{bmatrix} \ . \end{aligned} \tag{4.9}$$

Now let's verify that A^+bb is the least squares solution to $A\mathbf{x} = \mathbf{b}$.

$$AA^+\mathbf{b} = \frac{1}{18}\begin{bmatrix} 2 & 0 \\ 0 & 2 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 8 \\ 7 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 8 \\ 7 \\ 11 \end{bmatrix} \ . \tag{4.10}$$

This isn't $\mathbf{b}$, but that must mean the $\mathbf{b}$ doesn't belong to $\mathrm{Img}(A)$. Let's find the equation for $\mathrm{Img}(A)$. Row reduction of $\begin{bmatrix} 2 & 0 & \mid & x \\ 0 & 2 & \mid & y \\ 1 & 2 & \mid & z \end{bmatrix}$ leads to $\begin{bmatrix} 2 & 0 & \mid & x \\ 0 & 2 & \mid & y \\ 0 & 0 & \mid & z - x/2 - y \end{bmatrix}$ and so $\mathrm{Img}(A)$ is the plane in $I\!R^3$ given by the equation

$$x + 2y - 2z = 0 \ . \tag{4.11}$$

Evidently $\mathbf{b}$ does not lie in this plane, so $A\mathbf{x} = \mathbf{b}$ has no solution, as we thought. Now since $\mathrm{Img}(A)$ is a plane, it is easy to find $\mathbf{c}$, the vector in $\mathrm{Img}(A)$ that is closest to $\mathbf{b}$: By Theorem 2.2.1,

$$\mathbf{c} = \mathbf{b} - \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|^2}\mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{9}\begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 8 \\ 7 \\ 11 \end{bmatrix} \ . \tag{4.12}$$

Comparing (4.10) and (4.12), we see that $A^+\mathbf{b}$ satisfies $A\mathbf{x} = \mathbf{c}$, so it is the least squares solution to $A\mathbf{x} = \mathbf{b}$. Since the columns of A are evidently linearly independent, $\mathrm{Ker}(A) = 0$, and there is just one solution. In this case, there certainly is no solution of lesser length.

Next we look at an example in which the columns of A are not independent, so there is more than one least squares solution.

Example 3 (*SVD* and the least length least sqaures solution) Let $A = \begin{bmatrix} 2 & 0 & 4 \\ 0 & 2 & -2 \\ 1 & 2 & 0 \end{bmatrix}$. This is closely related to the matrix of Example 1: The first two columns are the same, and the new third column is twice the first column minus the second. Since the columns are not linearly independent, the kernel is not zero. Moreover, the span of the columns is the same as in Example 1, so it is still the case that the image of A is the plane given by (4.11). Let

$$V = \frac{1}{3\sqrt{5}}\begin{bmatrix} 6 & 2 \\ -3 & 4 \\ 0 & 5 \end{bmatrix} \ , \ D = \begin{bmatrix} 2\sqrt{6} & 0 \\ 0 & 3 \end{bmatrix} \text{ and } U = \frac{1}{\sqrt{30}}\begin{bmatrix} 2 & \sqrt{6} \\ -1 & 2\sqrt{6} \\ 5 & 0 \end{bmatrix} \ . \tag{4.13}$$

As you can check, $A = VDU^t$. You should also check at this point that U and V are isometries.

Now, we know from Example 2 that $\mathbf{b}$ does not lie in $\mathrm{Img}(A)$, so there is no solution. But we claim that $A^+\mathbf{b} = UD^{-1}V^t\mathbf{b}$ is the least squares solution of least length. To see this, lets first compute $A^+\mathbf{b}$:

$$\begin{aligned} A^+\mathbf{b} = UD^{-1}V^t\mathbf{b} &= \frac{1}{15\sqrt{6}}\begin{bmatrix} 2 & \sqrt{6} \\ -1 & 2\sqrt{6} \\ 5 & 0 \end{bmatrix}\begin{bmatrix} 1/(2\sqrt{6}) & 0 \\ 0 & 1/3 \end{bmatrix}\begin{bmatrix} 6 & -3 & 0 \\ 2 & 4 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{36}\begin{bmatrix} 10 \\ 17 \\ 3 \end{bmatrix} \end{aligned}$$

Now let's apply A to this vector and see what we get:

$$A\mathbf{x}_* = \frac{1}{36}\begin{bmatrix} 2 & 0 & 4 \\ 0 & 2 & -2 \\ 1 & 2 & 0 \end{bmatrix}\begin{bmatrix} 10 \\ 17 \\ 3 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 8 \\ 7 \\ 11 \end{bmatrix}$$

which we recognize from (4.12): The right hand side is $\mathbf{c}$, the orthogonal projection of $\mathbf{b}$ onto $\mathrm{Img}(A)$. Hence $\mathbf{x}_*$ is a least squares solution to $A\mathbf{x} = \mathbf{b}$. We can also directly check that $\mathbf{x}_*$ is the minimal length least squares solution.

To do this, you first determine, in the usual way, that the kernel of A is spanned by $\mathbf{w} = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$, and so the set of all least square solutions of $A\mathbf{x} = \mathbf{b}$ is given by $A^+\mathbf{b} + t\mathbf{w}$. Since

$$A^+\mathbf{b} \cdot \mathbf{w} = \frac{1}{36} \begin{bmatrix} 10 \\ 17 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = 0 \ ,$$

it follows that $|A^+\mathbf{b} + t\mathbf{w}|^2 = |A^+\mathbf{b}|^2 + t^2|\mathbf{w}|^2$. Clearly, we get the minimal length solution by taking $t = 0$.

4.3 Finding a singular value decomposition

Now we've seen two examples in which A has a singular value decomposition. The next theorem tells us that every matrix has a singular value decomposition, and moreover, it tells us *one way* to find U, V and D. We will discuss methods for finding U, V and D after the proof of Theorem 2.

Before going into the proof, let's relate finding singular values to something with which we are more familiar: finding eigenvalues.

Let A be an $m \times n$ matrix, and suppose that it has some singular value decomposition $A = VDU^t$. Then $A^tA = UD^2U^t$, or, what is the same,

$$(A^tA)U = U(D^2) \ .$$

Writing U in the form $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]$, and writing

$$D = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix} \ , \tag{4.14}$$

this is the same as

$$A^tA\mathbf{u}_j = \sigma_j^2\mathbf{u}_j \ .$$

In other words, the columns of U must be eigenvectors of A^tA, and the corresponding entries of D^2 must be the corresponding eigenvalues. Having made this observation, it is very easy to prove that every matrix has a singular value decomposition.

Theorem 3 (A Singular Value Decomposition Always Exists) *Let A be any $m \times n$ matrix, and let $r = \text{rank}(A)$. Then there exist an $r \times r$ diagonal matrix D with strictly positive diagonal entries, an $m \times r$ isometry V and an $m \times r$ isometry U so that $A = VDU^t$. The diagonal entries of D are the square roots of the r strictly positive eigenvalues of A^tA, arranged in decreasing order.*

Proof: Since A^tA is *symmetric*, there is an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ of $I\!R^n$ consisting of eigenvectors for A^tA.

Let $\{\mu_1, \mu_2, \ldots, \mu_n\}$ be the eigenvalues of A^tA arranged in decreasing order. That is, we arrange them so that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$. Each one is non negative since if $\mathbf{u}_j$ is a normalized eigenvector corresponding to μ_j, then

$$\mu_j = \mathbf{u}_j \cdot (\mu_j \mathbf{u}_j) = \mathbf{u}_j \cdot \left(A^tA\right)\mathbf{u}_j = (A\mathbf{u}_j) \cdot (A\mathbf{u}_j) = |A\mathbf{u}_j|^2 \geq 0 \ .$$

We have the following diagonalization of A^tA:

$$\begin{aligned} A^tA &= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] \begin{bmatrix} \mu_1 & 0 & \ldots & 0 \\ 0 & \mu_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_n \end{bmatrix} [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]^t \\ &= \mu_1 \mathbf{u}_1 \mathbf{u}_1^t + \mu_2 \mathbf{u}_2 \mathbf{u}_2^t + \cdots + \mu_n \mathbf{u}_n \mathbf{u}_n^t \ . \end{aligned} \tag{4.15}$$

Since $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ is invertible, the rank of A^tA is the rank of the diagonal matrix in (4.16), which is the number of non-zero eigenvalues. Let r denote the rank of A. Then since $\text{rank}(A^tA) = \text{rank}(A)$, there are exactly r non–zero eigenvalues. In particular, $\mu_j = 0$ for $j > r$, and so we can shorten (4.16) to

$$A^tA = \mu_1 \mathbf{u}_1 \mathbf{u}_1^t + \mu_2 \mathbf{u}_2 \mathbf{u}_2^t + \cdots + \mu_r \mathbf{u}_r \mathbf{u}_r^t \ . \tag{4.16}$$

For $j \leq r$, $\mu_j > 0$, and so we can define

$$\sigma_j = \sqrt{\mu_j} > 0 \tag{4.17}$$

and then can define D by (4.14). Finally, define the $n \times r$ matrix U by

$$U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] \ . \tag{4.18}$$

Then (4.16) can be written as

$$A^tA = UD^2U^t \ . \tag{4.19}$$

We now claim that there is a singular value decomposition of A with this D and this U. If so, we would have $A = VDU^t$, and hence

$$V = AUD^{-1} \ . \tag{4.20}$$

Our claim is correct if and only if the tight hand side defines an isometry. To check this out, we compute

$$\begin{aligned} (AUD^{-1})^t(AUD^{-1}) &= (D^{-1}U^tA^t)(AUD^{-1}) \\ &= D^{-1}U^t(A^tA)UD^{-1} \\ &= D^{-1}U^t(UD^2U^t)UD^{-1} \\ &= D^{-1}(U^tU)D^2(U^tU)D^{-1} \\ &= D^{-1}D^2D^{-1} \\ &= I \ . \end{aligned}$$

The first equality is from the properties of the transpose, and the rest from (4.19) and the fact that $U^tU = I$. This shows that if we *define* V by (4.20), V is an isometry. But if we define V by (4.20), we have

$$VDU^t = AUU^t \ .$$

By Theorem 1, UU^t is the orthogonal projection onto $\mathrm{Img}(A^t)$, which is the orthogonal complement of $\mathrm{Ker}(A)$, and so $AUU^t = A$. Hence V, U and D as defined above give us a singular value decomposition of A. ■

Let's make an important observation that justifies the use of the phrase "the singular values of A". First, according to the theorem every matrix has a singular value decomposition. It will have more than one. For example, consider an extreme case: $A = I$, where I is the $n \times n$ identity matrix. Then

$$I = III^t$$

is a singular value decomposition of I. But then so is

$$I = WIW^t$$

where W is *any* $n \times n$ orthogonal matrix. However, the theorem says that the diagonal matrix in the middle is uniquely determined.

• *The diagonal entires of D are the same in all singular value decompositions of A Therefore, it makes sense to talk about "the" singular values of A.*

The discussion that preceded Theorem 2 not only shows us that U, V and D exist – it gives us a way to find them! Let's recapitulate the steps:

(1) Form the $n \times n$ matrix A^tA, and diagonalize it. Let $\{\mu_1, \mu_2, \ldots, \mu_n\}$ be the eigenvalues, arranged in decreasing order. Define $\sigma_j = \sqrt{\mu_j}$ for $j = 1, 2, \ldots, n$. Let r be the number of these that are strictly positive, and define

$$D = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix} \ . \tag{4.21}$$

(2) Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$ be an orthonormal set eigenvectors for A^tA with $A^tA\mathbf{u}_j = \mu_j\mathbf{u}_j$ for $j = 1, 2, \ldots, n$. Ignore the last $n - r$ of these, and define

$$U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \ . \tag{4.22}$$

(3) Finally, compute

$$V = AUD^{-1} \ . \tag{4.23}$$

Example 4 (Computing an *SVD*) Let's again consider the matrix $A = \begin{bmatrix} 2 & 0 & 4 \\ 0 & 2 & -2 \\ 1 & 2 & 0 \end{bmatrix}$ from Example 2. In this case,

$$A^tA = \begin{bmatrix} 5 & 2 & 8 \\ 2 & 8 & -4 \\ 8 & -4 & 20 \end{bmatrix}$$

Computing the characteristic polynomial, we find $\mu^3 - 33\mu^2 + 216\mu$. This factors as

$$\mu(\mu - 9)(\mu - 24) \ ,$$

so

$$\mu_1 = 24 \quad , \ \mu_2 = 9 \quad \text{and} \quad \mu_3 = 0 \ . \tag{4.24}$$

Evidently, $r = 2$, and since $\sqrt{24} = 2\sqrt{6}$ and $\sqrt{9} = 3$,

$$D = \begin{bmatrix} 2\sqrt{6} & 0 \\ 0 & 3 \end{bmatrix} \ .$$

Eigenvectors corresponding to the eigenvalues in (4.24), found in the usual way, and listed in the corresponding order, are

$$\begin{bmatrix} 2 \\ -1 \\ 5 \end{bmatrix} \quad , \quad \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} \ .$$

Since $r = 2$, we are only concerned with the first two eigenvectors. Normalizing them we have

$$\mathbf{u}_1 = \frac{1}{\sqrt{30}} \begin{bmatrix} 2 \\ -1 \\ 5 \end{bmatrix} \quad \text{and} \quad \mathbf{u}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{30}} \begin{bmatrix} 1\sqrt{6} \\ 2\sqrt{6} \\ 0 \end{bmatrix} \ .$$

This gives us $U = [\mathbf{u}_1, \mathbf{u}_2] = \dfrac{1}{\sqrt{30}} \begin{bmatrix} 2 & \sqrt{6} \\ -1 & 2\sqrt{6} \\ 5 & 0 \end{bmatrix}$. Now that we have U and D, we find V through $V = AUD^{-1} = \dfrac{1}{3\sqrt{5}} \begin{bmatrix} 6 & 2 \\ -3 & 4 \\ 0 & 5 \end{bmatrix}$. These are exactly the factors we listed in (4.13).

The method that we have just illustrated works as long as all computations are done exactly. However, it is not a good method to use for larger matrices when computations are being done on a computer. Computer arithmetic involves round-off, and the method we have explained can break down badly when numbers are rounded off during the computations.

Here is the point. Suppose you are working on a computer, and and doing your computations to 16 decimal places, a fairly standard accuracy. If $\sigma_1/\sigma_r \approx 10^{10}$, then when we compute $\sigma_1 + \sigma_r$, we get a result that differs from σ_1 in the last six decimal places. However, $\mu_1/\mu_r = (\sigma_1/\sigma_r)^2 \approx 10^{20}$ Then when we add $\mu_1 + \mu_r$, we just get μ_1. As far as the computer is concerned, μ_r is zero compared to μ_1. It is thrown away in roundoff.

Now, $\mu_1 + \mu_r$ is not exactly something you would compute in diagonalizing A^tA, but you are likely to be adding numbers of similarly disparate sizes. Since the computer would use roundoff rules giving $\mu_1 + \mu_r = \mu_1$, which is not quite right, things can go wrong. They can go far enough wrong that you might not even get the right value for r, the number of non–zero eigenvalues! Then your matrices U, V and D would even have the *wrong sizes*. This is much more serious than a few decimal places of error in the entries.

Definition The *condition number* of an $m \times n$ matrix A with rank r is the ratio σ_1/σ_r of the largest to the smallest singular values of A.

We have hinted at the significance of this number in at the end of our discussion on how to find singular value decompositions. *The bigger the condition number, the more careful you have to be about roundoff error.* The condition number of A^tA is the square of the condition number of A.

Therefore, computer programs for computing singular value decompositions avoid computing A^tA. They proceed more directly to the singular values σ_j. Then, since

$$\frac{\sigma_1}{\sigma_r} \leq \frac{\mu_1}{\mu_r}$$

and is usually much, much less, like 10^{10} instead of 10^{20}. By avoiding the introduction of A^tA into the analysis, one avoids serious problems with round–off.

We won't go into the methods you would use to write an effective program; that is a beautiful problem but it is not the subject of this book. Our main concern is with understanding how to use the singular value decomposition. In practical application a computer program, hopefully well written, will be used to compute U, V and D. However, there a geometric way of understanding singular value decompositions that shed light on this and other questions.

Example 5 (Condition number) Let A be the 3×3 matrix considered in Examples 2 and 3. We found there that $r = 2$, $\sigma_1 = 2\sqrt{6}$ and $\sigma_2 = 3$. Hence, the condition number of A is

$$\frac{\sigma_1}{\sigma_2} = \frac{2\sqrt{6}}{3} \approx 1.632993162\ .$$

This is a *well conditioned matrix* since the condition number is not "too large". What does "too large" mean in this context? A condition number C is too large if roundoff on your computer causes it to evaluate $C + 1$ as C. In fact, even if C is large enough that your computer would evaluate $C + 10^{-6}$ as C, you are probably skating on thin ice. However, whether you are or not depends on the particular problem at hand.

Exercises

4.1 Let $A = \begin{bmatrix} 22 & -4 \\ -13 & 16 \\ 2 & -14 \end{bmatrix}$.

(a) Compute a singular value decomposition of A.

(b) Use the singular value decomposition of A to compute a least squares solution to $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

(c) Compute the condition number of A.

4.2 Let $A = \begin{bmatrix} 22 & 21 \\ -10 & -30 \\ 17 & 6 \end{bmatrix}$.

(a) Compute a singular value decomposition of A.

noindent**(b)** Use the singular value decomposition of A to compute a least squares solution to $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

(c) Compute the condition number of A.

4.3 Let $A = \begin{bmatrix} 16 & -4 & 14 \\ 13 & -22 & 2 \end{bmatrix}$.

(a) Compute a singular value decomposition of A.

(b) Use the singular value decomposition of A to compute a least length solution to $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. (Recall that any true solution is also a least squares solution).

(c) Compute the condition number of A.

4.4 Let $A = \begin{bmatrix} 28 & 20 & -16 \\ 29 & 10 & -38 \end{bmatrix}$.

(a) Compute a singular value decomposition of A.

(b) Use the singular value decomposition of A to compute a least length solution to $A\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. (Recall that any true solution is also a least squares solution).

(c) Compute the condition number of A.

Section 5: Goemetry and the Singular Value Decomposition

5.1 The image of the unit circle under a linear transformation

The image of the unit circle under an invertible 2×2 matrix A is always an ellipse. An easy proof of this can be given using the singular value decomposition, and this fact in turn can help us understand the nature of the singular value decomposition.

If A has rank one, then $\mathrm{Img}(A)$ is a line, and so the image of the unit circle under A will just be a line segment. This is a “degenerate” sort of ellipse. So let’s suppose that A has rank 2. Then if $A = VDU^t$ is a singular vlaue decompostion of A, V, U and D are all invertible 2×2 matrices. In particular, V and U are orthogonal 2×2 matrices. Let’s work out the effect of A on the unit circle by working out the effects of applying, in succession, U^t, D and then V.

Now any orthogonal transformation is just a rotation or a reflection. Both rotations and reflections leave the unit circle unchanged, and so applying U^t to the unit circle has no effect. At the end of this step we still have a unit circle. Next, what does D do to the unit circle? It just stretched or compresses it along the axes, producing an elipse whose axes are allinged with the x, y axes. Finally, applying V to this ellipse just rotates or reflects it, which gives another ellipse. So the result is always an ellipse.

Example 1 Let $A = \begin{bmatrix} 5 & 3 \\ 0 & 4 \end{bmatrix}$. Then $A^tA = \begin{bmatrix} 25 & 15 \\ 15 & 25 \end{bmatrix}$. Eigenvectors of A^tA are found in the usual way: $A^tA \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 40 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $A^tA \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 10 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. This tell us that $D = \begin{bmatrix} 2\sqrt{10} & 0 \\ 0 & \sqrt{10} \end{bmatrix}$ and $U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ Finally, we find $V = AUD^{-1} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$.

Now that we have our singular value decomposition $A = VDU^t$, we can work out the image of the unit circle under A in three steps.

As explained above, the image of the unit circle under U^t is still the unit circle, since U is a rotation, possibly folllowed by a reflection. Indeed,

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

for $\theta = \pi/4$. Thus, U^t rotates the unit circle clockwise though the angle $\pi/4$, and this doesn’t affect its graph.

Next apply D. Hence a vector $\begin{bmatrix} u \\ v \end{bmatrix}$ is the image under D of a vector $\begin{bmatrix} x \\ y \end{bmatrix}$ if and only if

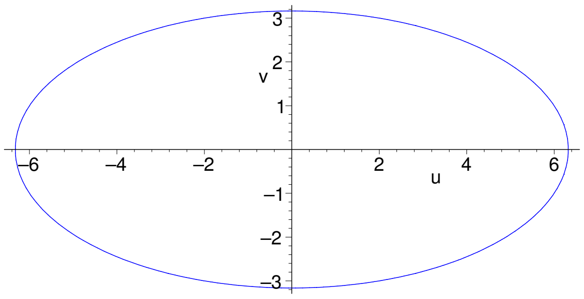
$$\begin{bmatrix} x \\ y \end{bmatrix} = D^{-1} \begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{\sqrt{40}} \begin{bmatrix} u \\ 2v \end{bmatrix} .$$

With $\begin{bmatrix} x \\ y \end{bmatrix}$ and $\begin{bmatrix} u \\ v \end{bmatrix}$ realted in this way, $\begin{bmatrix} u \\ v \end{bmatrix}$ is in the image of the unit circle if and only if $x^2 + y^2 = 1$, which in terms of u and v is

$$u^2 + 4v^2 = 40 . \tag{5.1}$$

That is, the image of the unit circle under D is the ellipse centered at the origin in the u, v plane whose major axis has length $2\sqrt{40} = 4\sqrt{10}$ and runs along the u-axis, and whose whose minor axis has length $2\sqrt{10}$ and runs along the v-axis. Here is a graph*:

* Clearly D stretches the x component on $\begin{bmatrix} x \\ y \end{bmatrix}$ by a factor of $2\sqrt{10}$, and stretches the y component on $\begin{bmatrix} x \\ y \end{bmatrix}$ by a factor of $\sqrt{10}$, so we could draw the ellipse without even working out the equation.



Finally, apply V. This too is an isometry, and

$$V = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

for $\theta = \cos^{-1}(2/\sqrt{5}) \approx .4636476086$ radians. Therefore, V rotates the ellipse we have found though this angle in the counterclockwise direction. The resulting "rotated ellipse" is the image of the unit circle under A. Thinking of V as a mapping from the u, v plane to the x, y plane, we see that $\begin{bmatrix} x \\ y \end{bmatrix} = V \begin{bmatrix} u \\ v \end{bmatrix}$ if and only if

$$\begin{bmatrix} u \\ v \end{bmatrix} = V^t \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 + y \\ 2y - x \end{bmatrix} .$$

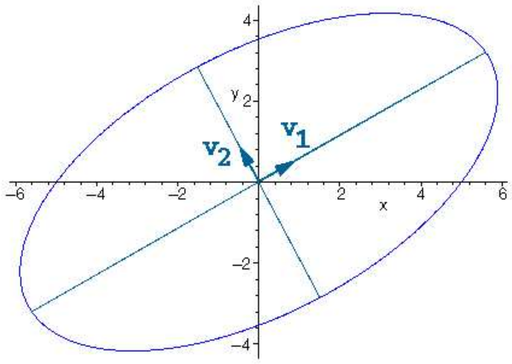
Hence $\begin{bmatrix} x \\ y \end{bmatrix}$ is in the image of the ellipse given by (5.1) if and only if

$$\frac{(2x+y)^2}{5} + 4\frac{(2y-x)^2}{5} = 40 ,$$

which simplifies to

$$2x^2 - 3xy + 4y^2 = 50 . \tag{5.2}$$

Here is a graph of this ellipse with the major and minor axes drawn in, together with two unit vectors pointing along them.



It is not necessary to use the singular value decomposition to find the equation of the ellipse that is the image of the unit circle. Indeed, a point (x, y) belongs to the image of the unit circle under A if and only if

$$\begin{bmatrix} x \\ y \end{bmatrix} = A \begin{bmatrix} u \\ v \end{bmatrix} \tag{5.3}$$

for some u and v with

$$u^2 + v^2 = 1 \ . \tag{5.4}$$

But then

$$\begin{bmatrix} u \\ v \end{bmatrix} = A^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \tag{5.5}$$

and expressing (5.4) in terms of x and y using (5.5) gives us the equation.

Example 2 Consider the matrix $A = \begin{bmatrix} 1 & 3 \\ -3 & -1 \end{bmatrix}$. Then

$$A^{-1} = \frac{1}{8}\begin{bmatrix} -1 & -3 \\ 3 & 1 \end{bmatrix} \ ,$$

and so (5.5) becomes

$$u = -\frac{1}{8}x - \frac{3}{8}y$$
$$v = \frac{3}{8}x + \frac{1}{8}y \ .$$

Substituting this into $u^2 + v^2 = 1$ gives

$$(x + 3y)^2 + (3x + y)^2 = 64$$

or, in simpler terms,

$$5u^2 + 6uv + 5v^2 = 32 \ . \tag{5.6}$$

Not only can you find the ellipse without using the singular value decomposition, once you have the ellipse, you can "see" the singular value decomposition of A by looking at this ellipse. In particular, let L_1 and L_2 be the lengths of the major and minor axes of the ellipse. Then the singular values of A are given by

$$\sigma_1 = \frac{L_1}{2} \qquad \text{and} \qquad \sigma_2 = \frac{L_2}{2} \ , \tag{5.7}$$

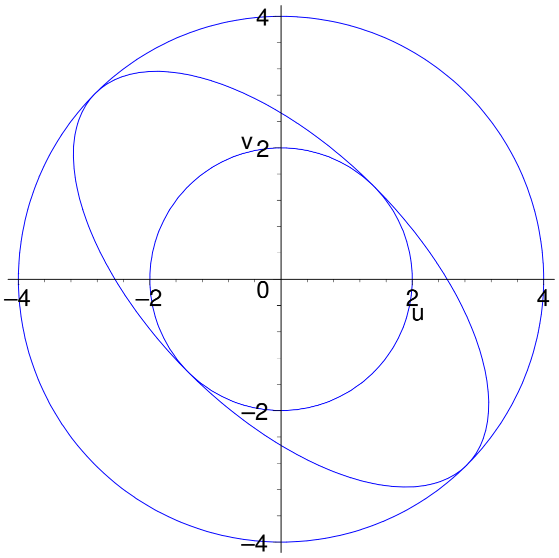
because D is what is responsible for stretching the unit circle to produce these major and minor axes. Thus we can "see"

$$D = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

by looking at the ellipse. We can also see what $\mathbf{v}_1$ and $\mathbf{v}_2$ are: They are the unit vectors pointing in the direction of the major and minor axes. These are only determined up to a sign, but that is fine. We know that we can always change a sign in any of the columns of V if we change the sign in the corresponding column of U. So, making *any* choice for the signs, we have $\mathbf{v}_1$ and $\mathbf{v}_2$, and hence $V = [\mathbf{v}_1, \mathbf{v}_2]$. Now we know that A has a singular value decomposition $A = VDU^t$, and we've determined V and D. Once D and V are known, $A = VDU^t$ gives us $U^t = D^{-1}V^tA$, and hence U is known.

Therefore, from a good graph of the image of the unit circle under A, and careful measurement, you can “read off” the singular value decomposition of A.

Example 3 Consider the matrix $A = \begin{bmatrix} 1 & 3 \\ -3 & -1 \end{bmatrix}$ As we saw in Example 2, the image of the unit circle under this matrix is the ellipse whose equation in the u, v plane is (5.6). Here is a graph, together with circles that inscribe and circumscribe the ellipse:



The diameter of the of the circumscribing circle is the length of the major axis, L_1, while the diameter of the inscribed circle is the diameter of the minor axis, L_2. In this diagram, you see that $L_1 = 8$ and $L_2 = 4$. Hence, we can “see” that for this matrix,

$$D = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} .$$

Next you can see the possible choices for $\mathbf{v}_1$ and $\mathbf{v}_2$. These are where the ellipse touches the outer and inner circle respectively.

$$\mathbf{v}_1 = \pm\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{v}_2 = \pm\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} .$$

Just to concrete, let’s take the plus signs so that we have

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} .$$

Finally, from $U^t = D^{-1}V^tA$,

$$\begin{aligned} U^t &= \frac{1}{4\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ -3 & -1 \end{bmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} . \end{aligned}$$

You can now easily check, that with these definitions of V, D and U, we do indeed have $A = VDU^t$.

As we’ll soon see, the remarkable fact that you can “see” the singular vlaue decomposition of a 2 × 2 matrix extends to higher dimensions, and is very useful.

5.2 The singular value decomposition and volume

We have seen the the image of the unit circle under an invertible 2×2 matrix A is an ellipse whose major axis has length $2\sigma_1$, and whose minor axis has length $2\sigma_2$. The area of such an ellipse is $\pi\sigma_1\sigma_2$. Now if $A = VDU^t$ is a singular value decomposition of A,

$$|\det(A)| = |\det(V)\det(D)||\det(U^t)| = |\det(D)|$$

since V and U^t are orthogonal. But $|\det(D)| = \sigma_1\sigma_2$, and so we see that the area of the ellipse is $|\det(A)|\pi$, or, in other words, $|\det(A)|$ times the area of the unit circle.

The singular value decomposition can be used in the same way to determine the volume of the image of the unit cube in $I\!R^n$ under an $n \times n$ matrix A.

We may assume that A is invertible, since otherwise the image of all of $I\!R^n$, and hence of the unit cube in particular, lies in a subspace of lower dimension, and has zero volume. Then, if $A = VDU^t$ is a singular value decomposition of A, V and U^t are orthogonal.

Now, orthogonal transformations preserve lengths and angles, so the image of the unit cube is just another unit cube, congruent to the original one. In particular, U^t does not affect the volume at all. Next, as we saw above, D is just a scale change – applying D changes the volume by a factor* of

$$\sigma_1\sigma_2\cdots\sigma_n = \det(D) = |\det(A)| \ . \tag{5.8}$$

Finally, applying V produces another congruent region, and this transformation, like U^t, has no effect on the volume. Hence the final volume is given by (5.8) since the unit cube itself, by definition, has unit volume. This gives us a proof of the following Theorem.

Theorem 1 *Let A be an $n \times n$ matrix. The n dimensional volume of the the image of the unit cube in $I\!R^n$ under A is $|\det(A)|$.*

This fact is very important in the theory of integration in several variables.

5.3 Singular Values, norms and low rank approximation

Recall that when B is a symmetric $n \times n$ matrix, the largest eigenvelue μ_1 of B is given by

$$\mu_1 = \max\{\mathbf{x} \cdot B\mathbf{x} \ : \ \mathbf{x} \text{ in } I\!R^n \text{ with } |\mathbf{x}| = 1 \ \} \ . \tag{5.9}$$

That is, μ_1 is the maximum value of the function $\mathbf{v} \to \mathbf{x} \cdot B\mathbf{x}$ on the set of unit vectors in $I\!R^n$. Moreover, if $\mathbf{x}$ is any unit vector with $\mathbf{x} \cdot B\mathbf{x} = \mu_1$, then $A\mathbf{x} = \mu_1\mathbf{x}$.

There is a similar result for singular values. Let A be any $m \times n$ matrix, and let $A = VDU^t$ be a singular value decomposition for it. We know that σ_1^2 is the largest eigenvalue of A^tA, which is the square of the norm of $\|A\|$.

Theorem 2 *Let A be any $m \times n$ matrix, and let σ_1 be the largest singular value of A. Then*

$$\sigma_1 = \max\{ \ |A\mathbf{x}| \ : \ \mathbf{x} \text{ in } I\!R^n \text{ with } |\mathbf{x}| = 1 \ \} = \|A\| \ , \tag{5.10}$$

[*] This is Cavallieri's principle.

where the unit vectors $\mathbf{x}$ *and* $\mathbf{y}$ *in the right belong to* $I\!R^n$ *and* $I\!R^m$ *respectively. Moreover,*

$$|A\mathbf{x}| = \sigma_1$$

for unit a vector $\mathbf{x}$ *if and only if* $A^tA\mathbf{x} = \sigma_1^2\mathbf{x}$.

Proof: Let $\mathbf{x}$ be any unit vector in $I\!R^n$. Then

$$|A\mathbf{x}|^2 = A\mathbf{x}\cdot A\mathbf{x} = \mathbf{x}\cdot A^tA\mathbf{x} = \mathbf{x}\cdot A^tA\mathbf{x}\ , \tag{5.11}$$

Applying (5.9) with $B = A^tA$, we see that $|A\mathbf{x}| = \sqrt{\mathbf{x}\cdot A^tA\mathbf{x}} \le \sigma_1$, and there is eqaulity if and only if $A^tA\mathbf{x} = \sigma^2\mathbf{x}$. ■

This very simple theorem provides an optimal way to approximate an arbitrary matrix A by a matrix of low rank. Suppose that A is an $m\times n$ matrix of rank r, and that $A = VDU^t$ is a singular value decomposition of A. Let

$$U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \qquad \text{and} \qquad V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]\ .$$

Then

$$\begin{aligned} A = VDU^t &= [\sigma_1\mathbf{v}_1, \sigma_2\mathbf{v}_2, \ldots, \sigma_r\mathbf{v}_r]\left([\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]\right)^t \\ &= \sigma_1\mathbf{v}_1\mathbf{u}_1^t + \sigma_2\mathbf{v}_2\mathbf{u}_2^t + \cdots + \sigma_r\mathbf{v}_1\mathbf{u}_r^t\ . \end{aligned}$$

Now pick any $s < r$, and define $A_{(s)}$ by

$$A_{(s)} = \sigma_1\mathbf{v}_1\mathbf{u}_1^t + \sigma_2\mathbf{v}_2\mathbf{u}_2^t + \cdots + \sigma_s\mathbf{v}_s\mathbf{u}_s^t\ . \tag{5.12}$$

Definition (Best rank s approximation) for any $n\times n$ matrix A, let $A = \sum_{j=1}^r \sigma_j\mathbf{v}_j\mathbf{u}_j^t$ be a singular value decompostion of A with the singular values arranged in decreasing order as usual. Then for any $s < r$, define the matrix

$$A_{(s)} = \sum_{j=1}^{s} \sigma_j\mathbf{v}_j\mathbf{u}_j^t\ . \tag{5.13}$$

Then $A_{(s)}$ is the *best rank s approximation of A.*

Note that the matrix $A_{(s)}$ clearly has rank s. Indeed, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$ is an orthonormal basis for its image. As to the question of why we cal it the "best" such approximation, let us first look at how good an approximation it is.

Theorem 3 *Let A be any $m\times n$ matrix, and let $A_{(s)}$ be its best rank s approximation. Then*

$$\|A - A_{(s)}\| = \sigma_{s+1}\ ,$$

where σ_{s+1} is the $(s+1)$st singular value of A. Consequently, for all i and j the absolute difference between the i,jth entries of A and $A_{(s)}$ is no greater than σ_{s+1}:

$$|A_{i,j} - [A_{(s)}]_{i,j}| \leq \sigma_{s+1} \ . \tag{5.14}$$

Proof: By definition,

$$A - A_{(s)} = \sigma_{s+1}\mathbf{v}_{s+1}\mathbf{u}_{s+1}^t + \cdots + \sigma_r\mathbf{v}_1\mathbf{u}_r^t \ . \tag{5.15}$$

If we define $\tilde{U} = [\mathbf{u}_{s+1}, \ldots, \mathbf{u}_r]$, $\tilde{V} = [\mathbf{v}_{s+1}, \ldots, \mathbf{v}_r]$, and define $\tilde{D}$ to be the diagonal matrix with entries $\sigma_{s+1}, \ldots, \sigma_r$, we can rewrite (5.15) as

$$A - A_{(s)} = \tilde{V}\tilde{D}\tilde{U}^t \ .$$

This is a singular value decomposition of $A - A_{(s)}$, and clearly the largest singular value is σ_{s+1}. By Theorem 2, this means that $\|A - A_{(s)}\| = \sigma_{s+1}$.

Now, for any matrix B, $B_{i,j} = \mathbf{e}_i \cdot B\mathbf{e}_j$, and so by the Schwarz inequality and the definition of the norm,

$$|B_{i,j}| = |\mathbf{e}_i \cdot B\mathbf{e}_j| \leq |\mathbf{e}_i||B\mathbf{e}_j| \leq \|B\||\mathbf{e}_i||\mathbf{e}_j| = \|B\| \ .$$

Applying this with $B = A - A_{(s)}$ leads to (5.14). ■

Let us return to the image compression topic that we discussed in the previous section ,and apply this theorem to it.

Suppose that A is a large matrix, say 200×300. Such a matrix might record an image by letting the entries be a numerical designation for the shading of each of an array of pixels. In a standard grayscale image, each entry would be an integer in the range 0 to 255.

Notice that A has $60,000$ entries. Now suppose that the first 10 singular values of A are by far the largest, but that $\sigma_{11} \leq 3$. Then by Theorem 3,

$$\frac{\|A - A_{(10)}\|}{\|A\|} \leq 3 \ ,$$

and for each i,j,

$$\frac{\|A_{i,j} - [A_{(10)}]_{i,j}\|}{\|A\|} \leq 3 \ .$$

Therefore, if we took the matrix $A_{(10)}$, and rounded of the entries to the nearest integer in the range 0 to 255, the result would differ from $A_{i,j}$ by no more than 3. The eye can certainly detect a shift of pixel values of 3 on the 0 to 255 scale, but the image would still be very recognizable as being essentially the same.

that is, essentially all of the visual information in A is in $A_{(10)}$. But $A_{(10)}$ can be expressed very efficiently: We just need to know the 10 numbers σ_1 through σ_{10}, the 10 unit vectors in $I\!R^{200}$, $\mathbf{v}_1$ through $\mathbf{v}_{10}$, and the 10 unit vectors in $I\!R^{300}$, $\mathbf{u}_1$ through $\mathbf{u}_{10}$. Then we can reconstruct $A_{(10)}$ using (5.12). The singular value description of the 200×300 matrix $A_{(10)}$ thus requires only

$$10(1 + 200 + 300) = 5010$$

numbers.

You could use this as a method for image compression, though there are more efficient algorithms. Nonetheless, the idea described here are the basis of important applications of the singular value decomposition in computer vision and data analysis.

Exercises

5.1 Let $A = \begin{bmatrix} 22 & -4 \\ -13 & 16 \\ 2 & -14 \end{bmatrix}$.

(a) Compute the best rank 1 approximation of A, $A_{(1)}$.

(b) Compute $\|A - A_{(1)}\|$.

5.2 Let $A = \begin{bmatrix} 22 & 21 \\ -10 & -30 \\ 17 & 6 \end{bmatrix}$.

(a) Compute the best rank 1 approximation of A, $A_{(1)}$.

(b) Compute $\|A - A_{(1)}\|$.

5.3 Let $A = \begin{bmatrix} 16 & -4 & 14 \\ 13 & -22 & 2 \end{bmatrix}$.

(a) Compute the best rank 1 approximation of A, $A_{(1)}$.

(b) Compute $\|A - A_{(1)}\|$.

5.4 Let $A = \begin{bmatrix} 28 & 20 & -16 \\ 29 & 10 & -38 \end{bmatrix}$.

(a) Compute the best rank 1 approximation of A, $A_{(1)}$.

(b) Compute $\|A - A_{(1)}\|$.

5.5 Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 1+a \\ 1 & 1-a \end{bmatrix}$.

(a) Compute a singular value decomposition of A.

(b) Compute the best rank 1 approximation of A, $A_{(1)}$.

(c) Compute $\|A - A_{(1)}\|$.

(d) Compute the least length, least squares solutions to both $A\mathbf{x} = \mathbf{b}$ and $A_{(1)}\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.