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

by

Eric A Carlen
Professor of Mathematics
Georgia Tech

Table of Contents

Overview

The *symmetric eigenvalue problem* is the problem of computing the eigenvalues and eigenvectors of a symmetric $n \times n$ matrix A. Knowing how to deal with this problem is fundamentally important in the calculus of several variables since Hessian matrices are always symmetric, and as we have seen, applications of the calculus of several variables often require computing eigenvalues and eigenvectors of Hessian matrices. There are many other reasons why this topic is important in the calculus of several variables, as we shall see. Accepting this fact, consider the problem itself.

How one goes about computing eigenvalues of an $n \times n$ matrix A depends very much on n. If $n = 2$ or 3, it may be quite reasonable to compute the characteristic polynomial, and then factor it to find the eigenvalues. One can then easily solve for the eigenvectors. However, even for $n = 3$, the factorization gets cumbersome in general. The analog of the quadratic formula for cubic polynomials is quite complicated! Moreover, for $n = 5$, there is no such formula. There are 5th degree polynomials whose roots cannot be written in terms of the coefficients using any finite sequences of algebraic operations.

Therefore, when n is larger than 2, it is better to abandon the characteristic polynomial, and do something else. There are several alternative approaches to the problem. The one we choose to explain here is due to Jacobi. It provides an excellent gateway into the realm of *iterative methods* for finding eigenvalues. An especially nice feature is that it "leverages" the fact that we have simple formulas for diagonalizing 2×2 matrices to *progressively diagonalize* larger matrices. This is explained in the first two sections.

Joacobi's algorithm takes any given $n \times n$ symmetric matrix A, and "makes it more and more diagonal" by *similarity transformations*. If at some finite step, it produces an exactly diagonal matrix, the eigenvalues of A are the diagonal entries of the diagonal matrix. In general though, the off diagonal entries will only get very small – very fast in fact – but never actually reach zero. This leads us to a study of "almost diagonal matrices". We will see in the third section that the eigenvalues of almost diagonal matrices are *very* close to the diagonal entries, much closer than one might expect.

The results of the first three sections give us a very satisfactory solution to the symmetric eigenvalue problem: They provide us with the means to compute all of the eigenvalues and eigenvectors of any symmetric matrix – no matter how large – to any desired degree of accuracy.

The final two sections concern the *singular value decomposition*, which is closely related to the diagonalization of symmetric matrices, but can be done for *any* matrix A, square or not. This has applications in the theory of integration in several variables, so it is a calculus topic. However, its utility extends far beyond calculus *per se*.

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 $t = \dfrac{a+d}{2} \pm \sqrt{b^2 + \left(\dfrac{a-d}{2}\right)^2}$. Since $b^2 + \left(\dfrac{a-d}{2}\right)^2$ 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_+.

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 $\mathbf{r}_1^\perp$ since it is an eigenvector of A with eigenvalue μ_+. This means that $(\mathbf{r}_1^\perp)^\perp$ is an eigenvector of A with eigenvalue μ_-. Now, as you can easily check,

$$(\mathbf{r}_1^\perp)^\perp = -\mathbf{r}_1 \ .$$

We have just found both eigenvectors, at least when the two eigenvalues are distinct.

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 $\mathbb{R}^2$ is an eigenvector of A with eigenvalue a. Hence the same formulas apply in this case as well.

Therefore, if $B = A - \mu_+$, and we write $B = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}$, we have that with

$$\mathbf{v}_+ = \mathbf{r}_1^\perp \qquad \text{and} \qquad \mathbf{v}_- = -\mathbf{r}_1 \ ,$$

$\mathbf{v}_+$ is an eigenvector with eigenvalue μ_+, and $\mathbf{v}_-$ is an eigenvector with eigenvalue μ_-.

Finally, define

$$\mathbf{u}_+ = \frac{1}{|\mathbf{v}_+|}\mathbf{v}_+ \qquad \text{and} \qquad \mathbf{u}_- = \frac{1}{|\mathbf{v}_-|}\mathbf{v}_- \ . \tag{1.3}$$

This is an orthonormal basis of $\mathbb{R}^2$ consisting of eigenvectors of A. We summarize 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

$$\mathbf{v}_+ = \mathbf{r}_1^\perp \qquad \text{and} \qquad \mathbf{v}_- = -\mathbf{r}_1 , \tag{1.5}$$

then with $\mathbf{u}_+$ *and* $\mathbf{u}_-$ *defined by (1.3),* $\{\mathbf{u}_+, \mathbf{u}_-\}$ *is a basis of* $\mathbb{R}^2$ *consisting of eigenvectors of* A*. Indeed,* $A\mathbf{u}_+ = \mu_+\mathbf{u}_+$ *and* $A\mathbf{u}_- = \mu_-\mathbf{u}_-$*.*

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 $\mu_\pm = \dfrac{9}{2} \pm \dfrac{5}{2}$; 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

$$\mathbf{v}_+ = \begin{bmatrix} -2 \\ -4 \end{bmatrix} \qquad \text{and} \qquad \mathbf{v}_- = \begin{bmatrix} 4 \\ -2 \end{bmatrix} . \tag{1.6}$$

For any vector $\mathbf{v}$, $|\mathbf{v}^\perp| = |\mathbf{v}|$, so $|\mathbf{v}_+| = |\mathbf{v}_-|$. Computing, we find $|\mathbf{v}_+| = 2\sqrt{5}$. Hence, the orthonormal basis of eigenvectors is

$$\mathbf{u}_+ - \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ -2 \end{bmatrix} \qquad \text{and} \qquad \mathbf{u}_- = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix} . \tag{1.7}$$

You can (and should) check that $A\mathbf{u}_+ = \mu_+\mathbf{u}_+$ and $A\mathbf{u}_- = \mu_-\mathbf{u}_-$.

Once we have the eigenvalues of A, we can *diagonalize* it: Let

$$U = [\mathbf{u}_+, \mathbf{u}_-] . \tag{1.8}$$

Then by a fundamental formula for matrix multiplication,

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

If we define

$$D = \begin{bmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{bmatrix} , \tag{1.9}$$

then we have

$$AU = UD \ . \tag{1.10}$$

Now, since the columns of U are orthonormal, U is invertible, and in fact $U^{-1} = U^t$. To see this, recall that for any i and j.

$$\begin{aligned}(U^tU)_{i,j} &= (\text{row } i \text{ of } U^t) \cdot (\text{column } j \text{ of } U) \\ &= (\text{column } i \text{ of } U) \cdot (\text{column } j \text{ of } U) \\ &= I_{i,j} \ .\end{aligned}$$

This shows that $U^tU = I$, which is what we are after, and all that we have used in this computation is that the columns of U are orthonormal.

Multiplying both sides of (1.10) through on the left by U^t and using $U^tU = I$, we have

$$U^tAU = D \ . \tag{1.11}$$

Since the right hand side is diagonal, this factorization is referred to as a *diagonalization of A*.

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 $\mathbf{u}_+ \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ -2 \end{bmatrix}$ and $\mathbf{u}_- = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Hence from (1.8) and (1.9), 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, $U^tAU = D$.

Theorem 1 provides *one*, together with (1.8) and (1.9), way to diagonalize a 2×2 symmetric matrix with an orthogonal matrix U. Unless the matrix is a multiple of the identity, there are exactly 8 choices for U. For U to diagonalize A, as we have seen, both of its columns must be eigenvectors of A. For U to be orthogonal, the columns must also be unit vectors. (The will automatically be orthogonal to one another since they correspond to distinct eigenvalues). So the columns must be $\pm\mathbf{u}_+$ and $\pm\mathbf{u}_-$. They can be placed in U in either order, and so each of the eight matrices $[\pm\mathbf{u}_+, \pm\mathbf{u}_-]$ and $[\pm\mathbf{u}_-, \pm\mathbf{u}_+]$ will diagonal A. Since the eigenvalues are arranged in D according to the arrangement of the eigenvectors in U, any one of the first four choices leads to $D = \begin{bmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{bmatrix}$, and any one of the second fours choices leads to $D = \begin{bmatrix} \mu_- & 0 \\ 0 & \mu_+ \end{bmatrix}$.

The choice we have made in (1.8) guarantees that D comes out with the largest value on top, and that U is a 2×2 rotation matrix. To see this last point, recall that any unit

vector can be written in the form $\begin{bmatrix}\cos(\theta)\\ \sin(\theta)\end{bmatrix}$. In particular, we can define θ in the interval $[0, 2\pi)$ by

$$\mathbf{u}_+ = \begin{bmatrix}\cos(\theta)\\ \sin(\theta)\end{bmatrix} .$$

Then since we have chosen to take $\mathbf{u}_- = \mathbf{u}_+^\perp$, we have

$$\mathbf{u}_+ = \begin{bmatrix}-\sin(\theta)\\ \cos(\theta)\end{bmatrix} .$$

Therefore,

$$U = [\mathbf{u}_+, \mathbf{u}_-] = \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix} .$$

The linear transformation of $I\!R^2$ induced by this matrix is a counter clockwise rotation through the angle θ.

We have therefore seen that every 2×2 symmetric matrix can be diagonalized by a rotation matrix. We will put this to work for diagonalizing larger matrices in the next section.

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?

There cannot be any closed form formula for the eigenvectors and eigenvalues of $n \times n$ matrices for $n \geq 5$.

By a closed form formula for the eigenvalues, we mean one that expresses them in terms of the entries of the matrix using a finite number of multiplications, divisions, additions, subtractions, and root extractions. For example, we have seen that for 2×2 symmetric matrices $A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$, the eigenvalues are given by

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

As explained in the overview, there can be no such formula for matrices that are 5×5 or larger. This applies even to symmetric matrices, and it tells us that we will have to go about finding eigenvectors and eigenvalues of $n \times n$ matrices using an *iterative procedure*. That is, we will employ an infinite sequence of successive approximations.

The original 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

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

the upper left 2×2 block with U. This gives us the 3×3 matrix G_1:

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

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 Analysis of a stopping rule for the Jacobi algorithm

In general, the Jacobi algorithm does not produce an exactly diagonal matrix in any finite number of iterations, so the formulation we gave above would result in an infinite loop. We need to introduce a stopping rule to get us out of the loop. The stopping rule will be based on checking the value of a function that measure far a matrix is from being diagonal.

Definition For any $n \times n$ matrix B, define the following three quantities:

(1) $\mathrm{Off}(B) = \sum_{i \neq j} |B_{i,j}|^2$ *(The sum of the squares of the off–diagonal entries).*

(2) $\mathrm{On}(B) = \sum_{i} |B_{i,i}|^2$ *(The sum of the squares of the on–diagonal entries).*

(3) $F(B) = \sum_{i,j} |B_{i,j}|^2$. *(The sum of the squares of all entries).*

Notice 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

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

Therefore, for any $\epsilon > 0$,

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

We now modify the algorithm to include a stopping rule. 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 $\mathrm{Off}(B) \leq 2\epsilon^2$, stop. Otherwise, go to *(1)* and repeat.

Now, how do we know that this is a valid stopping rule? Could it be that we might still always have $\mathrm{Off}(B) > 2\epsilon^2$ so that we still get an infinite loop?

The answer is no. In fact, we can calculate an explicit upper bound on how many iterations will be needed before stopping rule "kicks in". This is what the functions $\mathrm{On}(B)$ and $F(B)$ are good for. Here is the main point:

Let B be an $n \times n$ symmetric matrix. Let G be the givens rotation matrix produced for this B in step (3). Then:

$$F(G^t B G) = F(B) \tag{2.2}$$

and

$$\mathrm{On}(G^t B G) = \mathrm{On}(B) + 2 \max_{i \neq j} |B_{i,j}|^2 \ . \tag{2.3}$$

The validity of these equations is not obvious; we will derive them soon. But first, let's accept their validity, and see what they tell us about how many iterations will be required before the stopping rule "kicks in".

First observe that

$$F(B) = \mathrm{On}(G^t B G) + \mathrm{Off}(G^t B G) \ .$$

Therefore, since $F(B)$ does not change and $\mathrm{On}(B)$ increases, it must be that $\mathrm{Off}(B)$ decreases:

$$\mathrm{Off}(G^t B G) = \mathrm{Off}(B) - 2 \max_{i \neq j} |B_{i,j}|^2 \ . \tag{2.4}$$

We now wish to eliminate $\max_{i \neq j} |B_{i,j}|^2$, and express the right hand side in terms of $\mathrm{Off}(B)$ alone. To do this, note that $\mathrm{Off}(B)$ is a sum over the $n^2 - n$ sqaures 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 (2.4), we have

$$\begin{aligned} \mathrm{Off}(G^t B G) &= \mathrm{Off}(B) - \frac{2}{n^2 - n} \mathrm{Off}(B) \\ &= \left(1 - \frac{2}{n^2 - n}\right) \mathrm{Off}(B) \ . \end{aligned} \tag{2.5}$$

For instance, with $n = 3$, this becomes

$$\mathrm{Off}(G^t B G) \leq \frac{2}{3}\mathrm{Off}(B) \ .$$

This means that each iteration decreases $\mathrm{Off}(B)$ by *at least* a factor of 2/3. Let $B^{(k)}$ denote the matrix B produced after k iterations. We see that

$$\mathrm{Off}(B^{(k)}) \leq \left(\frac{2}{3}\right)^k \mathrm{Off}(A) \ . \tag{2.6}$$

This means that $\mathrm{Off}(B^{(k)})$ decreases *exponentially fast*. Indeed, if we define a sequence of numbers $\{a_k\}$ by

$$a_k = \ln\left(\mathrm{Off}(B^{(k)})\right) \ ,$$

Then (2.6) say that

$$a_k \leq \ln\left(\frac{2}{3}\right) k + \ln\left(\mathrm{Off}(A)\right) \ .$$

Since $\ln(2/3)$ is negative, the sequence decreases in steady increments and will eventually become as negative as you like. Our stppong rule kicks in as soon as $a_k \leq \ln(2\epsilon^2) = \ln(2) + 2\ln(\epsilon)$, We see that

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

and this is satisfied as soon as

$$k \geq \frac{2|\ln(\epsilon)| + |\ln\left(\mathrm{Off}(A)\right)|}{\ln(3) - \ln(2)} \ . \tag{2.7}$$

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, and the stopping rule will kick in exponentially fast at that.

It is easy to adapt the preceding discussion from the 3×3 case to the $n \times n$ case; see the problems.

2.3 Some technical points

In this final subsection, we prove (2.2) and (2.3). 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 $F(B)$ in terms of the trace as follows: Using the symmetry of B,

$$\mathrm{tr}(B^2) = \sum_{i,j=1}^{n} B_{i,j}B_{j,i} = \sum_{i,j=1}^{n} |B_{i,j}|^2 = F(B) \ . \tag{2.8}$$

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 (2.8), (2.2) is true.

Now consider (2.3). For any k and ℓ,

$$\begin{aligned}(G^tBG)_{k,k} &= \sum_{r,s=1}^{n} (G^t)_{k,r}B_{r,s}G_{s,k} \\ &= \sum_{r,s=1}^{n} (G)_{r,k}G_{s,k}B_{r,s}cr\end{aligned} \tag{2.9}$$

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 (2.9) is the term with $r = s = k$. That is,

$$(G^tBG)_{k,k} = B_{k,k} \qquad \text{for} \quad k \neq i, j \ . \tag{2.10}$$

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,

$$|(G^tBG)_{i,i}|^2 + |(G^tBG)_{j,j}|^2 = |B_{i,i}|^2 + |B_{j,j}|^2 + 2|B_{i,j}|^2 \tag{2.11}$$

Since, by hypothesis,

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

we have from this, (2.10) and (2.11) that

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

and hence (2.3) is true.

2.4 convergence of the Jacobi algorithm

If we use the stopping rule we discussed at the end of the last subsection, the algorithm terminates in an approximate diagonalization of A. But the inequality (2.5) implies that $\lim_{n\to\infty} f(B_n) = 0$. On this basis, we might expect that

$$\lim_{n\to\infty} B_n = D \tag{2.12}$$

where D is diagonal. We might also expect that

$$\lim_{n\to\infty} V_n = V \tag{2.13}$$

where V is orthogonal, and V_n is the cumulative rotation produced by the algorithm up to the nth stage.

The our expectations about (2.12) and (2.13) will be born out, though to explain this, we need to say what we mean by the limit of a sequence of matrices.

Given a sequence $\{C_n\}$ of $m \times n$ matrices, and another $m \times n$ matrix C, we say that

$$\lim_{n\to\infty} C_n = C$$

in case

$$\lim_{n\to\infty} \|C - C_n\| = 0 \ . \tag{2.14}$$

Convergent sequences of matrices are nicely behaved. For example, the sequence of their norms is a bounded sequence of numbers. This follows from the fact that

$$C_n = C + (C_n - C)$$

and so

$$\|C_n\| \leq \|C\| + \|C - C_n\| \ . \tag{2.15}$$

Every convergent sequence of numbers is bounded, and so from (2.14), $\{\|C - C_n\|\}$ is bounded. Now (2.15) implies that $\{\|C_n\|\}$ is bounded.

In a similar way, you can show that if $\lim_{n\to\infty} C_n = C$ and $\lim_{n\to\infty} M_n = M$, then $\lim_{n\to\infty} C_n M_n = CM$, assuming only that the sizes of C and M are such that they can be multiplied.

Applying this to

$$B_n = V_n^t A V_n$$

we would have

$$D = V^t A V$$

It is clear from (2.5) that D would have to be diagonal. Also, as soon as we have (2.13), it follows from the fact that each V_n is orthogonal that

$$V^t V = \lim_{n\to\infty} V_n^t V_n = \lim_{n\to\infty} I \ .$$

Hence, once we have (2.12) and (2.13), we have that

$$A = V^t D V$$

where D is diagonal and V is orthogonal. This of course means that

$$AV = DV$$

so that if $V = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ and D is the diagonal matrix whose jth diagonal entry is μ_j, we have

$$A\mathbf{u}_j = \mu_j \mathbf{u}_j$$

for $j = 1, 2, \ldots, n$. Since V is orthogonal, its columns are orthonormal, and we have produced an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A. Also, the entries of D are clearly real, so that all of the eigenvalues of A are real. So once we have proved (2.12) and (2.13), we have proved, in a constructive manner, the following theorem:

Theorem 1 (Diagonalization of symmetric matrices) *Let A be any $n \times n$ symmetric matrix. Then there is an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of A, and all of the eigenvalues of A are real.*

Proof: In fact, one just needs to prove something a bit less than (2.12) and (2.13). One just needs to find convergent *subsequences* of the sequences of matrices produced by the Jacobi algorithm. This is always possible since one can always select a convergent subsequence

from any sequence of matrices whose norms are uniformly bounded. Here is how: We recall a fundamental result concerning bounded sequences of numbers; namely that *given any uniformly bonded infinite sequence of numbers, one can select from it a convergent subsequence.*

Now for any matrix A, every entry $A_{i,j}$ satisfies

$$-\|A\| \leq A_{i,j} \leq \|A\| \ .$$

Hence if $\{A^{(k)}\}$ is any sequence of $n \times n$ whose norms are uniformly bounded, then $\{(A^{(k)})_{1,1}\}$ is a bounded sequence of numbers. We may therefore select a subsequence $\{(A^{(k_\ell)})_{1,1}\}$ for which

$$\lim_{\ell \to \infty} \left(A^{(k_\ell)}\right)_{1,1}$$

exists. Throw away the unused terms in the original sequences, but still denote the resulting subsequence by $\{A^{(k)}\}$. This sequence has the property that $\lim_{k\to\infty} A^{(k)}_{1,1}$ exists.

Next, consider the numerical sequence $\{(A^{(k)})_{1,2}\}$. Again, this is bounded, and so we can select a subsequence $\{(A^{(k_\ell)})_{1,2}\}$ for which

$$\lim_{\ell \to \infty} \left(A^{(k_\ell)}\right)_{1,2}$$

Now every subsequence of a convergent sequence converges, and since we selected this subsequence from a sequence of matrices whose $1,1$ entries converge, we now have a subsequence for which both the $1,1$ and the $1,2$ entries converge.

Continuing in this way, we obtain a subsequence with all of the entries convergent. ■

It is nice to know that the Jacobi algorithm converges. If nothing else, this gives us a proof of the fact that every symmetric matrix can be diagonalized. But to really make use of the Jacobi algorithm, we have to come to grips with the fact that in actual applications, we are going to have to stop after some finite number of steps. In general, we will be left not with a diagonal matrix, but an *almost diagonal* matrix.

If we are trying to calculate the eigenvalues of an $n \times n$ symmetric matrix A, and we have nearly, but not quite, diagonalized it, what can we say about the eigenvalues? That is, suppose

$$G^t A G = B$$

where $G^t = G^{-1}$ so that A and B are similar matrices. If B were exactly diagonal, its eigenvalues would be its diagonal entries, and we could just "read off" the eigenvalues of B and hence A. For instance, if $n = 3$ and $B = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, then the eigenvalues of A are 3, 2 and 1.

But what if B is only nearly diagonal? Suppose for example that

$$B = \begin{bmatrix} 3 & 10^{-5} & 10^{-4} \\ 10^{-5} & 2 & 10^{-4} \\ 10^{-4} & 10^{-4} & 1 \end{bmatrix} .$$

Is it still true that B, and hence A, has eigenvalues μ_1, μ_2 and μ_3 with

$$\mu_1 \approx 3 \qquad \mu_2 \approx 2 \qquad \text{and} \qquad \mu_3 \approx 1 \ ,$$

or does the small departure from the diagonal condition have disastrous consequences for the eigenvalues?

In fact, things are much better than one might hope. It turns out that

$$\mu_1 = 3 \pm 8 \times 10^{-8} \qquad \mu_2 = 2 \pm 8 \times 10^{-8} \qquad \text{and} \qquad \mu_3 = 1 \pm 8 \times 10^{-8} \ .$$

Since the largest off–diagonal entry is 10^{-4}, which is *much* larger than 8×10^{-8}, the eigenvalues are given by the diagonal entries of B to an astonishingly good accuracy. In the next section we explain how to do such accuracy calculations. It is very nice mathematics: simple, somewhat surprising, and very useful in many contexts!

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 (2.7) 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*?

It turns out that this is the case, and that we can even say *how close*. Before explaining this, we make some definitions: For each i with $1 \leq i \leq n$, 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

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

and

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

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

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

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. 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

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.

Let B be any $n \times n$ symmetric matrix. Recall that in our analysis of the Jacobi algorithm, we have defined

$$\mathrm{Off}(B) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |B_{i,j}|^2 \ .$$

That is, $\mathrm{Off}(B)$ is the sum of the sqaures of the off diagonal entries of B. This is the natural measure of how close B is to being diagonal for the Jacobi algorithm, since this is what decreases at each step.

On the other hand, as far as eigenvalues are concerned, the relevant measure of of how close B is to being diagonal is given by $r(B)$. To put everything together we ask:

• *Is there a realtion between $r(B)$ and $\mathrm{Off}(B)$? In particular, if we know that $\mathrm{Off}(B)$ is small, what can we say about the size of $r(B)$?*

The following lemma gives us the answer.

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

$$(r(B))^2 \leq (n-1)\mathrm{Off}(B) \ . \tag{3.15}$$

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

$$\begin{aligned}(r(B))^2 &= \max\{\ (r_i(B))^2\ :\ 1 \le i \le n\} \\ &\le \sum_{i=1}^{n} (r_i(B))^2 \\ &\le \sum_{i=1}^{n} (n-1) \sum_{\substack{j=1 \\ j \ne i}}^{n} |B_{i,j}|^2 \\ &\le (n-1)\mathrm{Off}(B)\ .\end{aligned} \tag{3.16}$$

■

Suppose 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. That is, the ith eigenvalue μ_i is given by

$$\mu_i = B_{i,i} \pm \epsilon\ .$$

We can derive conditions for this by applying Theorem 3 to B. Using the inequality(3.16), both conditions of Theorem 3, namely

$$(r(B))^2 < \left(\frac{\delta(B)}{2}\right)^2 \qquad \text{and} \qquad \frac{2}{\delta(B)}(r(B))^2 < \epsilon\ ,$$

are satisfied if

$$\mathrm{Off}(B) < \frac{1}{n-1}\left(\frac{\delta(B)}{2}\right)^2 \qquad \text{and} \qquad \frac{2(n-1)}{\delta(B)}\mathrm{Off}(B) < \epsilon\ . \tag{3.17}$$

Now, we know that the diagonal entries eventually converge to the eigenvalues of A. Let μ_i, $1 \le i \le n$ be the eigenvalues of A, and let $\tilde{\delta}(A)$ be defined by

$$\tilde{\delta}(A) = \min\{\ |\mu_i - \mu_j|\ :\ 1 \le i < j \le n\ \}\ . \tag{3.18}$$

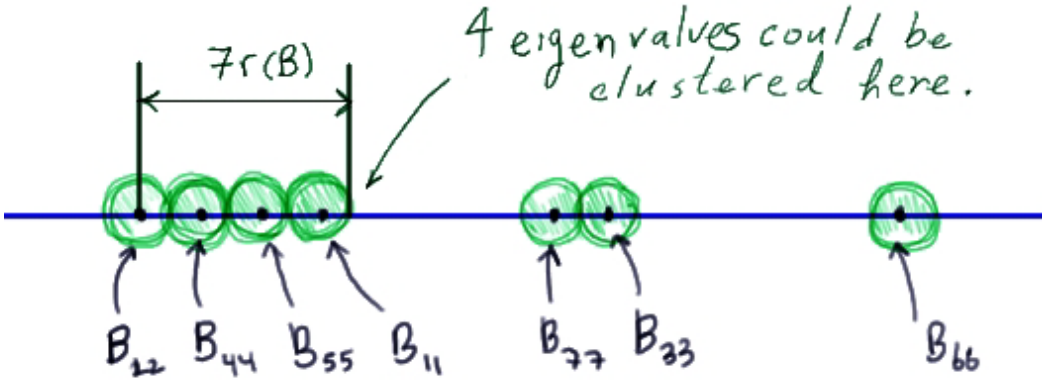
As long as all of the eigenvalues of A are distinct, $\tilde{\delta}(A) > 0$. After B is approximately diagonal, $\delta(B) \approx \tilde{\delta}(A) > 0$. That is, as we keep running the algorithm, $\delta(B)$ tends to a fixed non zero value, but $\mathrm{Off}(B)$ continues to decrease. So eventually, as long as all of the eigenvalues of A are distinct, (3.17) will eventually be satisfied, no matter how small ϵ is taken to be.

This gives us a certain stopping rule for the Jacobi algorithm: Run the algorithm until $\mathrm{Off}(B) < \epsilon$. Run it a few more iterations, and then check (3.17), If it is satisfied, stop.

Otherwise, run it a few more and check again. As long as all of the eigenvalues are distinct, which is the "generic" case, the algorithm will quickly terminate.

In the unlucky case in which there are repeated eigenvalues, $\delta(B)$ will keep shrinking along with $\mathrm{Off}(B)$, and (3.17) will not be satisfied. There will always be some overlap among the Gershgorin 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?



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 $\epsilon > 0$,

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

It therefore follows that when

$$\text{Off}(B) < \frac{1}{n-1}\left(\frac{\epsilon}{2n-1}\right)^2 \ , \tag{3.19}$$

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

So now we know how to calculate the eigenvalues to an accuracy of $\pm\epsilon$ no matter what: We run Jacobi until Off(B) $< \epsilon$. We then run another sweep, and check check (3.17), If it is satisfied, stop. If not, check (3.19). If it is satisifed, stop. Repeat the stepping and checking until the stopping rule licks in.

Since Off(B) goes to zero exponentially fast as we run the algorithm, we can be certain that eventually (3.19) will be satisfied, and the program will terminate – that is, the stopping rule is guaranteed to kick in, and there cannot be an infinite loop. When the stopping rule does kick in, each diagonal entry of B is an eigenvalue of A to an accuracy of $\pm\epsilon$.

3.3 Perturbation theory for eigenvalues of symmetric matrices

Let A and B be two symmetric $n \times n$ matrices. Introduce a parameter t, and define the t dependent matrix $A(t)$ by

$$A(t) = A + tB \ .$$

Notice that $A(0) = A$. The question we want to answer is this:

• *How do the eigenvalues of $A(t)$ depend on t?*

We will generally be concerned with small values of t, so that we can think of tB as a small "perturbation" of A. we then want to see how this small perturbation affects the eigenvalues of A.

Since A is symmetric, it can be diagonalized. Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A. Let μ_i be the eigenvalue corresponding to $\mathbf{u}_i$; that is

$$A\mathbf{u}_i = \mu_i \mathbf{u}_i \ .$$

Let $Q = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ and let D be the diagonal entry whose ith diagonal entry is μ_i. Then $Q^t A Q = D$. Therefore,

$$Q^t A(t) Q = Q^t (A + tB) Q = Q^t A Q + t Q^t B Q = D + tC$$

where

$$C = Q^t B Q \ .$$

Now if t is small, the off diagonal elements of $D + tC$ will be small, and we can apply what we know about Gershgorin disks to study the eigenvalues of $D + tC$. Since $D + tC$ is similar to $A(t)$, we thereby gain information on the eigenvalues of $A(t)$.

For any $1 \leq i, j \leq n$, since $\mathbf{u}_i = Q\mathbf{e}_i$ and $\mathbf{u}_j = Q\mathbf{e}_j$,

$$\begin{aligned} C_{i,j} &= \mathbf{e}_i \cdot C\mathbf{e}_j \\ &= \mathbf{e}_i \cdot Q^t BQ\mathbf{e}_j \\ &= Q\mathbf{e}_i \cdot BQ\mathbf{e}_j \\ &= \mathbf{u}_i \cdot B\mathbf{u}_j \end{aligned}$$

The diagonal entries of $D + tC$ are then given by

$$\begin{aligned} D_{i,i} + tC_{i,i} &= \mu_i + tC_{i,i} \\ \mu_i + t\mathbf{u}_i \cdot B\mathbf{u}_i \end{aligned} \tag{3.20}$$

Since for $i \neq j$, $D_{i,j} = 0$, the off diagonal entries of $D + tC$ are given by

$$\mathbf{u}_i \cdot B\mathbf{u}_j \ .$$

Now by the Schwarz inequality,

$$|\mathbf{u}_i \cdot B\mathbf{u}_j| \leq |\mathbf{u}_i||B\mathbf{u}_j| \leq \|B\| \ . \tag{3.21}$$

Hence for each i, $r_i(D + tC) \leq t(n-1)\|B\|$, which means that

$$r(D + tC) \leq t(n-1)\|B\| \ . \tag{3.22}$$

There are now two cases to consider: The nicest case is when all of the eigenvalues of A are distinct so that $\tilde{\delta}(A) > 0$, where $\tilde{\delta}(A)$ is still given by (3.18), as in the last subsection. Then from (3.22) and Theorem 2, as long as

$$t(n-1)\|B\| \leq \frac{\tilde{\delta}(A)}{2} \ , \tag{3.23}$$

for each i, there is exactly one eigenvalue of $D + tC$ in the Gershgorin disk about the ith diagonal entry of $D + tC$, namely $\mu_i + t\mathbf{u}_i \cdot B\mathbf{u}_i$. Call this eigenvalue $\mu_i(t)$. Since $A(t)$ is similar to $D + tC$, we see that for each i there is an eigenvalue $\mu_i(t)$ satisfying

$$\mu_i(t) \approx \mu_i + t\mathbf{u}_i \cdot B\mathbf{u}_i \ .$$

We can now apply Theorem 3 to say how close this approximation is. We first work out $\delta(D + tC)$. For any $i \neq j$,

$$(D + tC)_{i,i} - (D + tC)_{j,j} = \mu_i - \mu_j - t(C_{j,j} - C_{i,i}) \ ,$$

and hence, using (3.20) and (3.21) once more,

$$|(D+tC)_{i,i} - (D+tC)_{j,j}| \geq |\mu_i - \mu_j| - 2t\|B\| \ .$$

Hence

$$t < \frac{\tilde{\delta}(A)}{4\|B\|} \qquad \Rightarrow \qquad \delta(D+tC) > \frac{\tilde{\delta}(A)}{2} \ . \tag{3.24}$$

As long as $n \geq 3$, the condition on t in (3.24) is satisfied whenever the condition on t in (3.23) is satisfied. Since computing eigenvalues in the 2×2 case is straight forward, we henceforth assume that $n \geq 3$. Then by Theorem 3, for each i, as long as (3.23) is satisfied,

$$\begin{aligned} |\mu_i(t) - (\mu_i + t\mathbf{u}_i \cdot B\mathbf{u}_i)| &\leq \frac{\delta(D+tC)}{2}(r(D+tC))^2 \\ &\leq t^2 \frac{\tilde{\delta}(A)}{4}(n-1)^2\|B\|^2 \ . \end{aligned}$$

In particular, since $\mu_i(0) = \mu_i$,

$$\mu_i(t) - \mu_i(0) = t\mathbf{u}_i \cdot B\mathbf{u}_i + \mathcal{O}(t^2)$$

so that

$$\lim_{t\to 0} \frac{\mu_i(t) - \mu_i(0)}{t} = \mathbf{u}_i \cdot B\mathbf{u}_i \ . \tag{3.25}$$

That is, $\mu_i(t)$ is differentiable at $t = 0$, and we have a formula for the derivative!

In fact it is easy to see that $\mu_i(t)$ is differentiable at any value t_0 of t for which $A(t)$ has distinct eigenvalues – just replace A by $A(t_0)$ in the analysis above. By (3.24), each $\mu_i(t)$ is differentiable in *at least* the interval

$$-\frac{\tilde{\delta}(A)}{4\|B\|} \ < \ t \ < \ \frac{\tilde{\delta}(A)}{4\|B\|} \ .$$

When A has repearted eigenvalues, so that $\tilde{\delta}(A) = 0$, things are more complicated. The formula (3.25) is still valid, provided you make a special choice of the eigenvectors $\mathbf{u}_i$. We shall not go into this case here.

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

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

Once one knows that the eigenvalues are differentiable, one can go on and show that it is possible to choose the eigenvectors so that they too are differentiable. Knowing this, one can go on to show that the eigenvalues are in fact twice differentiable, and can compute a formula for the second derivative. One can keep going in this way, and get formulas for derivatives of every order. The key fact that get all of this going is the fact that the eigenvalues are differentiable, and this, as we have explained, follows from the Gershgorin Disk Theorem.

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 replaces diagonalization for non-square matrices?

We have seen that we can *diagonalize* any square symmetric matrix A. Indeed, whenever A is a square symmetric matrix, There is a diagonal matrix D and an orthogonal matrix U so that

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

The columns of U are the eigenvectors of A, and the corresponding entries of D are the corresponding eigenvalues. It is easy to compute with orthogonal matrices and diagonal matrices, so this decomposition of A is very useful in practice, as we've seen.

Now, what if A is not symmetric, or if A is not even square? There is still something almost as simple, and also very useful: a decomposition

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

where U and V are isometries, and D is diagonal with strictly positive entries. These positive entries of D are called the *singular values* of A, and the decomposition (1.2) is called the *singular value decomposition*.

We want to explain two things in this section: First, we want to explain what the decomposition (1.2) is good for, and second, how to compute V, U and D when given A.

Although the singular value decomposition is related to an eigenvalue decomposition, it not used in *exactly* the same way as (1.1). While one application we made of (1.1) was to compute matrix exponentials, we can't compute powers of A, let alone an exponential of A, when A is not square. For our first example of how we can use (1.2), we ask:

- *What is the minimal length least squares solution of* $A\mathbf{x} = \mathbf{b}$*?*

The singular value decomposition provides a very good way to solve this problem.

4.2 Using a singular value decomposition

Let A be an $m \times n$ matrix, and let $\mathbf{b}$ be a vector in $\mathbb{R}^m$, and suppose we are looking for a least squares solution of

$$A\mathbf{x} = \mathbf{b} \ . \tag{4.3}$$

We know that $\mathbf{x}$ is a least squares solution of (1.3) if and only if it satisfies the normal equations

$$A^tA\mathbf{x} = A^t\mathbf{b} \ . \tag{4.4}$$

The singular value decomposition can be used to find a particularly interesting solution of the normal equations, and moreover, can be used to find it in a numerically stable manner.

Theorem 1 (Singular Values and Least Squares) *Let A be an $m \times n$ matrix, and suppose that $A = VDU^t$ where V and U are isometries and D is a diagonal matrix with strictly positive diagonal entries. Then the columns of V are an orthonormal basis for the image of A, and the columns of U are an orthonormal basis for the image of A^t.*

Moreover, if we define the $n \times m$ matrix A^+ by

$$A^+ = UD^{-1}V^t \ , \tag{4.5}$$

then for any $\mathbf{b}$ in $\mathbb{R}^m$, $A^+\mathbf{b}$ is a least squares solution to $A\mathbf{x} = \mathbf{b}$, its length is less than that of any other least squares solution.

Proof: First of all, suppose that D is $r \times r$. Then V is $m \times r$ and U is $n \times r$. The rank of an isometry is the the number of its columns – the columns are orthonormal, so certainly they are linearly independent. Hence $\text{rank}(U) = \text{rank}(V) = r$.

Since $\text{rank}(U^t) = \text{rank}(U)$ it now follows that $\text{rank}(U^t) = r$. Hence by the dimension formulae $\dim(\text{Ker}(U^t)) = n - r$. Since $\text{Ker}(V) = 0$ and $\text{Ker}(D) = 0$,

$$\text{Ker}(A) = \text{Ker}(VDU^t) = \text{Ker}(U^t) \ .$$

Hence $\dim(\text{Ker}(A)) = n - r$, so that by the dimension formula once again, $\text{rank}(A) = r$. Finally, since D is an invertible $r \times r$ matrix, $\text{rank}(D) = r$. Let's summarize:

$$\text{rank}(A) = \text{rank}(U) = \text{rank}(V) = \text{rank}(D) = r \ . \tag{4.6}$$

Now, from the fact that $\text{Img}(VC) \subset \text{Img}(V)$ for any matrix C it follows from $A = VDU^t$ that $\text{Img}(A) \subset \text{Img}(V)$. Then, since the dimensions of these spaces are the same, it follows that $\text{Img}(A) = \text{Img}(V)$. Since V is an isometry, its columns are an orthonormal basis for its image, and hence for the image of A.

Since $A^t = UDV^t$, the exact same argument applied to A^t shows that the columns of U are an orthonormal basis for the image of A^t. This proves the first paragraph.

The second paragraph is a direct consequence of the first one. Notice that

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

since $U^tU = I$. By the first paragraph, VV^t is the orthogonal projection onto $\text{Img}(A)$, and hence for any $\mathbf{b}$ in $\mathbb{R}^m$, $A(A^+\mathbf{b})$ is the orthogonal projection of $\mathbf{b}$ onto $\text{Img}(A)$; i.e., the vector in $\text{Img}(A)$ 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 $\text{Img}(A^t) = (\text{Ker}(A))^\perp$. That is, $A^+\mathbf{b}$ is orthogonal to very vector in $\text{Ker}(A)$. 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 $\mathbf{w} = 0$. ■

Definition (generalized inverse) The Matrix A^+ defined in (1.5) is called the *generalized inverse* of A.

Since Theorem 1 says that $A^+\mathbf{b}$ is the least length, least squares solution of $A\mathbf{x} = \mathbf{b}$, which is a uniquely determined vector, a matrix cannot have more than one generalized inverse. Since we shall soon see that every matrix has at least one singular value decomposition, every matrix –square or not – has a genenralized inverse!

Let's look at an example.

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

$$V = \frac{1}{3\sqrt{5}} \begin{bmatrix} 2 & -6 \\ 4 & 3 \\ 5 & 0 \end{bmatrix} \quad , \qquad S = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad U = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} . \tag{4.7}$$

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 (1.5) gives us *the* least squares solution to $A\mathbf{x} = \mathbf{b}$.

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

$$\begin{aligned} A^+\mathbf{b} = UD^{-1}V^t\mathbf{b} &= \frac{1}{15} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 2 & 4 & 5 \\ -6 & 3 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{15} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 11 \\ -3 \end{bmatrix} \\ &= \frac{1}{15} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 11/3 \\ -3/2 \end{bmatrix} \\ &= \frac{1}{18} \begin{bmatrix} 8 \\ 7 \end{bmatrix} . \end{aligned} \tag{4.8}$$

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

$$AA^+\mathbf{b} = \frac{1}{18} \begin{bmatrix} 2 & 0 \\ 0 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 8 \\ 7 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 8 \\ 7 \\ 11 \end{bmatrix} = \begin{bmatrix} 0.88\bar{8} \\ 0.77\bar{7} \\ 1.22\bar{2} \end{bmatrix} . \tag{4.9}$$

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 $\left[\begin{array}{cc|c} 2 & 0 & x \\ 0 & 2 & y \\ 1 & 2 & z \end{array}\right]$ leads to $\left[\begin{array}{cc|c} 2 & 0 & x \\ 0 & 2 & y \\ 0 & 0 & z - x/2 - y \end{array}\right]$ and so $\mathrm{Img}(A)$ is the plane in $\mathbb{R}^3$ given by the equation

$$x + 2y - 2z = 0 . \tag{4.10}$$

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,

$$\mathbf{c} = \mathbf{b} - \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|^2}\mathbf{a} = \begin{bmatrix}1\\1\\1\end{bmatrix} - \frac{1}{9}\begin{bmatrix}1\\2\\-2\end{bmatrix} = \frac{1}{9}\begin{bmatrix}8\\7\\11\end{bmatrix} \ . \tag{4.11}$$

Comparing (1.9) and (1.11), 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 2 (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 (1.10). Let

$$V = \frac{1}{3\sqrt{5}}\begin{bmatrix}6 & 2\\-3 & 4\\0 & 5\end{bmatrix} \ , \ D = \begin{bmatrix}2\sqrt{6} & 0\\0 & 3\end{bmatrix} \text{ and } U = \frac{1}{\sqrt{30}}\begin{bmatrix}2 & \sqrt{6}\\-1 & 2\sqrt{6}\\5 & 0\end{bmatrix} \ . \tag{4.12}$$

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 1 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 (1.11): 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

$$D = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix} \ , \tag{4.13}$$

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 2 (A Singular Value Decomposition Always Exists) *Let A be any $m \times n$ matrix, and let $r = \mathrm{rank}(A)$. 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 $\mathbb{R}^n$ 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:

$$\begin{aligned} A^tA &= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] \begin{bmatrix} \mu_1 & 0 & \ldots & 0 \\ 0 & \mu_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_n \end{bmatrix} [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]^t \\ &= \mu_1 \mathbf{u}_1 \mathbf{u}_1^t + \mu_2 \mathbf{u}_2 \mathbf{u}_2^t + \cdots + \mu_n \mathbf{u}_n \mathbf{u}_n^t \ . \end{aligned} \tag{4.14}$$

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 (1.15), 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 (1.15) to

$$A^tA = \mu_1 \mathbf{u}_1 \mathbf{u}_1^t + \mu_2 \mathbf{u}_2 \mathbf{u}_2^t + \cdots + \mu_r \mathbf{u}_r \mathbf{u}_r^t \ . \tag{4.15}$$

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

$$\sigma_j = \sqrt{\mu_j} > 0 \tag{4.16}$$

and then can define D by (1.13). Finally, define the $n \times r$ matrix U by

$$U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] \ . \tag{4.17}$$

Then (1.15) can be written as

$$A^tA = UD^2U^t \ . \tag{4.18}$$

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

$$V = AUD^{-1} \ . \tag{4.19}$$

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 (1.18) and the fact that $U^tU = I$. This shows that if we *define* V by (1.19), V is an isometry. But if we define V by (1.19), 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

$$D = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix} \ . \tag{4.20}$$

(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

$$U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \ . \tag{4.21}$$

(3) Finally, compute

$$V = AUD^{-1} \ . \tag{4.22}$$

Example 3 (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,

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

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

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

so

$$\mu_1 = 24 \quad , \ \mu_2 = 9 \quad \text{and} \quad \mu_3 = 0 \ . \tag{4.23}$$

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 (1.23), 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 $U = [\mathbf{u}_1, \mathbf{u}_2] = \frac{1}{\sqrt{30}} \begin{bmatrix} 2 & \sqrt{6} \\ -1 & 2\sqrt{6} \\ 5 & 0 \end{bmatrix}$. Now that we have U and D, we find V through $V = AUD^{-1} = \frac{1}{3\sqrt{5}} \begin{bmatrix} 6 & 2 \\ -3 & 4 \\ 0 & 5 \end{bmatrix}$. These are exactly the factors we listed in (1.12).

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 $A^t A$, 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 4 (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

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

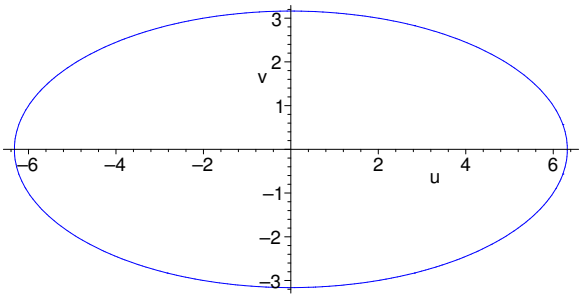
With $\begin{bmatrix} x \\ y \end{bmatrix}$ and $\begin{bmatrix} u \\ v \end{bmatrix}$ realted in this way, $\begin{bmatrix} u \\ v \end{bmatrix}$ is in the image of the unit circle if and only if $x^2 + y^2 = 1$, which in terms of u and v is

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

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

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

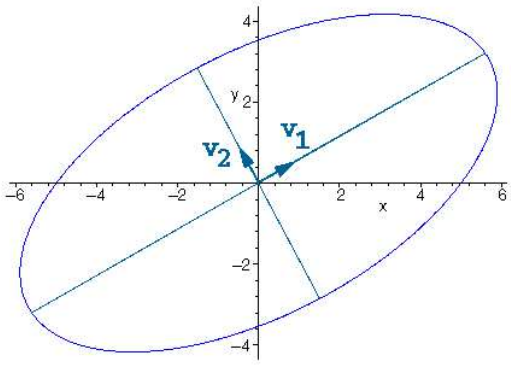
Hence $\begin{bmatrix} x \\ y \end{bmatrix}$ is in the image of the ellipse given by (2.1) if and only if

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

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 (2.4) in terms of x and y using (2.5) 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 (2.5) 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

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

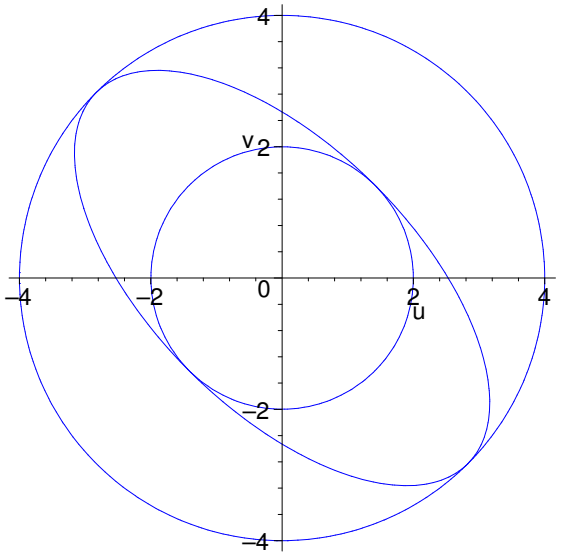
because D is what is responsible for stretching the unit circle to produce these major and minor axes. Thus we can "see"

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

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 (2.6). 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 $\mathbb{R}^n$ under an $n \times n$ matrix A.

We may assume that A is invertible, since otherwise the image of all of $\mathbb{R}^n$, 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 (2.8) 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 $\mathbb{R}^n$ 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

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

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 $\mathbb{R}^n$. 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*

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

---

* This is Cavallieri's principle.

where the unit vectors $\mathbf{x}$ *and* $\mathbf{y}$ *in the right belong to* $\mathbb{R}^n$ *and* $\mathbb{R}^m$ *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 $\mathbb{R}^n$. 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 (2.9) with $B = A^tA$, we see that $|A\mathbf{x}| = \sqrt{\mathbf{x} \cdot A^tA\mathbf{x}} \leq \sigma_1$, 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

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

By definition,

$$A - A_{(s)} = \sigma_{s+1}\mathbf{v}_{s+1}\mathbf{u}_{s+1}^t + \cdots + \sigma_r\mathbf{v}_1\mathbf{u}_r^t \ . \tag{5.13}$$

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 (2.13) 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 the theorem above, this means that $\|A - A_{(s)}\| = \sigma_{s+1}$.

Now, 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 coloring of each of an array of pixels. Notice that A has $60,000$ entries. Now suppose that the first 10 singular values of A are by far the largest. That is, suppose that σ_{11} is "negligibly small" compared to σ_1. Then

$$\frac{\|A - A_{(10)}\|}{\|A\|} \approx 0 \ ,$$

and essentially all of the 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 $\mathbb{R}^{200}$, $\mathbf{v}_1$ through $\mathbf{v}_{10}$, and the 10 unit vectors in $\mathbb{R}^{300}$, $\mathbf{u}_1$ through $\mathbf{u}_{10}$. Then we can reconstruct $A_{(10)}$ using (2.12). 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 $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.