Notes on Eigenvalues and Singular Values

Michael Overton

Spring 2022

1 Eigenvalues

Given a *square* $n \times n$ matrix A, an eigenvalue is a scalar (real or complex number) λ satisfying

$$Ax = \lambda x$$

for some *nonzero* vector x called an eigenvector.[1] This is equivalent to writing

$$(\lambda I - A)x = 0$$

so, since $x \neq 0$, $A - \lambda I$ must be *singular*, and hence

$$\det(\lambda I - A) = 0.$$

From the (complicated!) definition of determinant, it follows that $\det(\lambda I - A)$ is a *polynomial* in the variable λ with degree n, and this is called the *characteristic polynomial*. By the fundamental theorem of algebra (a nontrivial result), it follows that the characteristic polynomial has n *roots* which we denote $\lambda_1, \ldots, \lambda_n$, but these *may not be distinct* (different from each other). For example, the identity matrix I has characteristic polynomial $(\lambda - 1)^n$ and so all its eigenvalues are equal to one. Note that if A is real, the eigenvalues may not all be real, but those that are not real must occur in complex conjugate pairs $\lambda = \alpha \pm \beta \mathbf{i}$. It does not matter what order we use for numbering the λ_j. Although in principle we could compute eigenvalues by finding the roots of the characteristic polynomial, in practice there are much better algorithms, and in any case there is no general formula for finding the roots of a polynomial of degree 5 or more,[2] so whatever we do we will have to use some kind of approximation algorithm.

[1] If x is an eigenvector, so is αx for any nonzero scalar α.

[2] This was an open question for centuries that was finally resolved in the 19th century.

If all the eigenvalues are distinct, each λ_j corresponds to an eigenvector x_j (the null vector of $\lambda_j I - A$), which is unique except for scalar multiplication, and in this case the eigenvectors $x_j, j = 1, \ldots, n$, are *linearly independent*.[3] So, in this case, the $n \times n$ matrix

$$X = [x_1, x_2, \ldots, x_n]$$

is *nonsingular*. By the eigenvalue-eigenvector definition, we have

$$AX = X\Lambda, \qquad \text{where} \qquad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

the diagonal matrix of eigenvalues, so since X is nonsingular we can premutliply both sides by X^{-1}, or postmultiply both sides by X^{-1}, to obtain

$$X^{-1}AX = \Lambda \qquad \text{and} \qquad A = X\Lambda X^{-1},$$

the *eigenvalue decomposition* or *spectral decomposition* of A. We say that X defines a *similarity* transformation that *diagonalizes* A, displaying its eigenvalues in the diagonal matrix Λ. However, if the eigenvalues are not distinct, this may not be possible. For example, the *Jordan block*

$$J = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

has eigenvalues $\lambda_1 = \lambda_2 = 0$, and there is only one linearly independent eigenvector, namely $x = [1 \quad 0]^T$, or any scalar multiple of this vector. We say that J is *not diagonalizable*.

If A is real symmetric ($A = A^T$) then *all eigenvalues are real* and, regardless of whether they are distinct or not, there is always a set of n eigenvectors, say $q_j, j = 1, \ldots, n$, which are not only linearly independent, but also *orthonormal*, that is, with $q_i^T q_j = 0$ if $i \neq j$ and $q_i^T q_i = 1$.[4] So, the matrix

$$Q = [q_1, \ldots, q_n]$$

[3]Suppose $Ax = \lambda x$, $Ay = \mu y$, with $\lambda \neq \mu$, and $y = \alpha x$, with $\alpha \neq 0$. Then $A(\alpha x) = \mu(\alpha x)$, so $Ax = \mu x$, which is not possible since $Ax = \lambda x$ and $\lambda \neq \mu$. This argument can be extended to show that the set of all n eigenvectors is linearly independent.

[4]The proof that the eigenvalues must be real when $A = A^T$ is not difficult but we do not give it here. However, let's show why eigenvectors for distinct eigenvalues of $A = A^T$ must be orthogonal. Suppose $Ax = \lambda x$, $Ay = \mu y$, with $\lambda \neq \mu$. Then (1) $y^T Ax = y^T(\lambda x) = \lambda y^T x$ and (2) $x^T Ay = x^T(\mu y) = \mu x^T y$. Also, (3) the scalar $y^T Ax = (y^T Ax)^T = x^T A^T y = x^T Ay$. Combining (1), (2) and (3), we have $\lambda y^T x = \mu x^T y = \mu y^T x$, so, since $\lambda \neq \mu$, we must have $y^T x = 0$, i.e., x and y are orthogonal. This argument can be extended to show that there are n mutually orthogonal eigenvectors when $A = A^T$.

is an *orthogonal matrix* with inverse Q^T. We have

$$AQ = Q\Lambda, \qquad \text{where} \qquad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

the diagonal matrix of eigenvalues, and hence

$$Q^T AQ = \Lambda \qquad \text{and} \qquad A = Q\Lambda Q^T.$$

Thus, Q defines an *orthogonal similarity transformation* that diagonalizes A.[5] Orthogonal matrices have very nice properties and lead to numerical algorithms with optimal stability properties.

In general, when A is nonsymmetric, it does not have orthogonal eigenvectors. However, there is a very nice property called the Schur decomposition.[6] Assuming A is real, the Schur decomposition is

$$A = QUQ^T$$

where Q is orthogonal and U is *quasi-upper triangular*, which means upper triangular except that there may be 2×2 blocks along the diagonal, with one subdiagonal entry per block. Each real eigenvalue appears on the diagonal of U, as a 1×1 block, and each complex conjugate pair of eigenvalues of A consists of the eigenvalues of a 2×2 diagonal block. The columns of Q are called *Schur vectors*, but these are generally not eigenvectors.[7] This property is exploited by algorithms for computing eigenvalues of nonsymmetric matrices.[8]

The main MATLAB function for computing eigenvalues is `eig`. See also functions `roots` and `schur`.

[5]The same property holds for *complex Hermitian* matrices ($A = A^*$, where the superscript $*$ denotes complex conjugate transpose), and in fact for all *normal matrices* (satisfying $AA^* = A^*A$): then the q_i are complex and we must write $q_i^* q_j = 0$ if $i \neq j$ and $q_i^* q_i = 1$ and we say that Q is *unitary* instead of orthogonal, with $Q^{-1} = Q^*$.

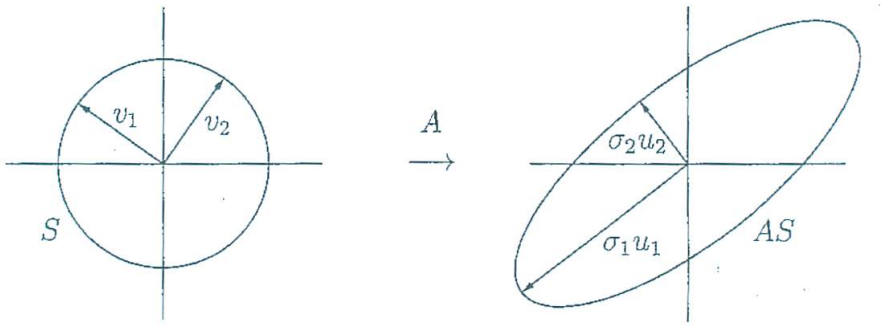
[6]The proof of this is more complicated but can be found in many books, such as *Numerical Linear Algebra* by Trefethen and Bau.

[7]If A is complex, so the eigenvalues do not occur in complex conjugate pairs, then Q is unitary, with $Q^*Q = I$, and U is an upper triangular complex matrix, with no 2×2 blocks.

[8]Including the "QR algorithm," described at a high level by A&G and in more detail by Trefethen and Bau.

2 Singular Values

Let A be an $m \times n$ real[9] matrix, with $m \geq n$. The **key idea** of the singular value decomposition (SVD) is that *multiplication by A* maps the *unit sphere* in $\mathbb{R}^n$ to a "hyper-ellipse" in $\mathbb{R}^m$:



Multiplication by A takes the unit sphere in $\mathbb{R}^n$ to a hyper-ellipse in $\mathbb{R}^m$.
From *Numerical Linear Algebra* by Trefethen and Bau, SIAM.

This means that there is a set $v_j, j = 1, \ldots, n$ of *orthonormal vectors* in $\mathbb{R}^n$ (the "right singular vectors") such that

$$Av_j = \sigma_j u_j, \quad j = 1, \ldots, n$$

where $u_j, j = 1, \ldots, n$ is a set of *orthonormal vectors* in $\mathbb{R}^m$ (the "left singular vectors"), and σ_j are *nonnegative real numbers* ("the singular values").[10]

We assume for convenience that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

So, we can write

$$A\,[v_1, v_2, \ldots, v_n] = [u_1, u_2, \ldots, u_n] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix},$$

[9] Everything applies to the complex case too, just by changing the transpose operations to "complex conjugate transpose" and "orthogonal matrices" to "unitary matrices".

[10] The proof of this fundamental fact is not too difficult but not trivial either. A good reference for this is the book by Trefethen and Bau. The derivation of the SVD now follows from this fact.

that is,

$$AV = \hat{U}\hat{\Sigma}.$$

where $V = [v_1, v_2 \ldots, v_n]$, $\hat{U} = [u_1, u_2 \ldots, u_n]$ and $\hat{\Sigma} = \text{diag}(\sigma_1, \sigma_2 \ldots, \sigma_n)$ (note the "hats" on U and on Σ, but not on V). The $n \times n$ matrix V is square with orthonormal columns, so V is an "orthogonal" matrix, with $V^{-1} = V^T$, and hence we can multiply both sides of this equation on the right by V^T to obtain the"reduced" form of the singular value decomposition:

$$A = \hat{U}\hat{\Sigma}V^T.$$

MATLAB calls this the "economy size" SVD and it can be computed by `[Uhat,Sigmahat,V]=svd(A,0)`. The words "reduced" and "economy size" are used because the matrix $\hat{U}$ has only n columns of length $m \geq n$, so it is not square if $m > n$. Note that its n columns are orthonormal with $\hat{U}^T\hat{U} = I$. We can introduce an additional set of $m-n$ orthonormal vectors, all orthogonal to $u_1, \ldots, u_n$, so now we have a square matrix

$$U = [u_1, \ldots, u_n, u_{n+1}, \ldots, u_m]$$

that *is* an orthogonal matrix, so $U^{-1} = U^T$. Also, define the $m \times n$ matrix

$$\Sigma = \begin{bmatrix} \hat{\Sigma} \\ 0 \end{bmatrix},$$

where we have appended another $m-n$ zero rows to $\hat{\Sigma}$ to obtain Σ, a matrix with the same dimension as A. Then we have

$$U\Sigma = \begin{bmatrix} \hat{U}, u_{n+1}, \ldots, u_m \end{bmatrix} \begin{bmatrix} \hat{\Sigma} \\ 0 \end{bmatrix} = \hat{U}\hat{\Sigma},$$

so using the equation $AV = \hat{U}\hat{\Sigma}$ given above, we get

$$AV = U\Sigma.$$

Finally, multiplying on the right by $V^{-1} = V^T$, we have

$$A = U\Sigma V^T,$$

the "full" SVD, which is computed by `[U,Sigma,V]=svd(A)`. Note that the reduced and full SVD are the same for square matrices (the case $m = n$).

Another useful way to interpret the SVD is that A can be written as the following sum of rank-one matrices:

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

The SVD tells us many things about a matrix. Here are some of them:

Rank. From the equation $A = U\Sigma V^T$, since U and V are $m \times m$ and $n \times n$ orthogonal matrices respectively, it follows that the number of linearly independent rows of A, and the number of linearly independent columns of A, are *the same*, namely, the number of nonzero singular values of A. This number is called the *rank* of A. Let's say the rank of A is r, where $n \geq r \geq 0$. Then we have

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

If $r = n$, there are no singular values equal to zero, and we say A has "full rank", or "full column rank": its columns are linearly independent, so $Ax = 0$ implies $x = 0$. (Of course the rows cannot be linearly independent if $m > n$).

Range. The range of A is the set of all vectors y such that $y = Az$ for some $z \in \mathbb{R}^n$. Since $A = U\Sigma V^T$, we have $Az = U\Sigma V^T z$. Whatever z is, $\Sigma V^T z$ is a linear combination of the first r columns of Σ, since the rest of them are zero, so $U\Sigma V^T z$ is a linear combination of $u_1, \ldots, u_r$. So, $u_1, \ldots u_r$ form an orthonormal basis for the range of A.

Null space. The null space of A is the set of all vectors z such that $Az = 0$. Since $A = U\Sigma V^T$, this means $Az = U\Sigma V^T z = 0$. Let $w = V^T z$, so $z = Vw$, and then $Az = U\Sigma V^T z = (U\Sigma)w$. This vector is a linear combination of the columns $\sigma_1 u_1, \ldots, \sigma_r u_r$, since the rest of the columns of $U\Sigma$ are zero. So, for this linear combination to be zero, we need the first r components of w to be zero. So, z is a linear combination of $v_{r+1}, \ldots, v_n$, which therefore form an orthonormal basis for the null space of A.

Nearest low rank matrix. If $A = U\Sigma V^T$ has full rank, so $\sigma_n > 0$, the nearest[11] rank-deficient matrix (that is, with rank less than n), or nearest singular matrix if $m = n$, is obtained by replacing σ_n in Σ by zero, and the

[11] Using either the 2-norm or the Frobenius norm to define "nearest", i.e., the matrix B minimizing $\|A - B\|_2$ or $\|A - B\|_F$, over all rank-deficient matrices, or over all matrices with rank s.

nearest rank s matrix is obtained by replacing $\sigma_{s+1}, \ldots, \sigma_n$ by zero. Thus, the nearest rank s matrix is

$$\sum_{i=1}^{s} \sigma_i u_i v_i^T.$$

The proof of this can be found in many books including Trefethen and Bau.

Two-norm. The definition of $\|A\|_2$ is

$$\|A\|_2 = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=1} \|U\Sigma V^T x\| = \max_{\|x\|=1} \|\Sigma V^T x\| = \max_{\|y\|=1} \|\Sigma y\| = \sigma_1,$$

where the vector norm is the 2-norm, so the matrix 2-norm of A is its largest singular value. Another way to see the same thing is that pre- and post-multiplication by orthogonal matrices preserves the 2-norm, so

$$\|A\|_2 = \|U\Sigma V^T\|_2 = \|\Sigma\|_2 = \sigma_1.$$

Inverse of a square matrix. When A is square and nonsingular, with $A = U\Sigma V^T$, we have

$$A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T.$$

Condition number in the two-norm. When A is square and nonsingular, its 2-norm condition number is

$$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$$

because the 2-norm of A is σ_1 and, from the formula for the inverse given above, the 2-norm of A^{-1} is $\|\Sigma^{-1}\|_2 = \sigma_n^{-1}$, the *largest* of the reciprocals of the singular values of A. If A is singular, so that $\sigma_n = 0$, we sometimes say that the condition number of A is ∞. If A is not square, the usual definition $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$ does not make sense, but we can still define the two-norm condition number to be $\kappa(A) = \sigma_1/\sigma_n$.

Condition number of the normal equations using the two-norm. We already observed in the QR notes that, for an $m \times n$ matrix A, with $m \geq n$, using $A = \hat{Q}\hat{R}$, we have $A^T A = \hat{R}^T \hat{R}$. We can explore this further using the reduced singular value decomposition $A = \hat{U}\hat{\Sigma}V^T$. We have

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

This means that the singular values (and also the eigenvalues) of A^TA are σ_i^2, $i = 1, \ldots, n$, so its two-norm condition is

$$\kappa(A^TA) = \frac{\sigma_1^2}{\sigma_n^2} = [\kappa(A)]^2.$$

This explains why solving least squares problems via QR is sometimes much more accurate than using the normal equations.

Solving the least squares problem using the SVD. In fact, we can alternatively solve the least squares problem using the SVD. Using the full SVD, $A = U\Sigma V^T$, we have for the two-norm that

$$\|Ax - b\| = \|U^T(Ax - b)\| = \|\Sigma V^T x - U^T b\|.$$

Letting $y = V^T x \in \mathbb{R}^n$ and $z = U^T b \in \mathbb{R}^m$, this becomes

$$\|Ax - b\| = \|\Sigma y - z\| = \left\| \begin{bmatrix} \hat{\Sigma} \\ 0 \end{bmatrix} y - z \right\|.$$

If A has linearly independent columns, i.e., A has rank n, as we usually assume for least squares problems, then $\sigma_n > 0$, and we can solve the least squares problem by setting $y_i = z_i/\sigma_i$, $i = 1, \ldots, n$, and then setting $x = Vy$; the remaining residual is $\|[z_{n+1}, \ldots, z_m]\|$. This is just as accurate a method as using QR, but it is much more work to compute the SVD than to compute the QR factorization. The big advantage of using the SVD is that even if the rank of A is r with $r < n$, we can solve the "rank-deficient" least squares problem by setting $y_i = z_i/\sigma_i$ only for $i = 1, \ldots, r$, and setting the remaining components of y to zero. We can even solve an "approximately" rank-deficient problem by "truncating" or "thresholding" some of the small singular values to zero, as discussed in Ascher and Greif p. 203; this results in a larger residual, but reduces the norm of the solution x.

Relationship Between Eigenvalues and SVD. Several important relationships between eigenvalues and singular values are explored in questions in the homework.

Relationship Between SVD and QR. Unlike the SVD, the QR decomposition cannot be used directly in the rank deficient case. One reason is that, although $\hat{R}$ having a zero on the diagonal tells us that $\hat{R}$ and hence A is rank deficient, knowing how many zeros are on the diagonal of $\hat{R}$ does not tell us the rank of $\hat{R}$ and hence does not tell us the rank of A. There is

a variant of QR called the "full orthogonal decomposition" that does reveal the rank, but it is not used much because the SVD is so convenient.

Note also that the QR factorization can be computed in a finite number of transformations, via Householder's method (or Gram-Schmidt), which would give the exact result if there were no rounding errors. In this sense QR is like LU. We know this is not possible for SVD or eigenvalue computations, for the reason given on p. 1 of these notes.

Another thing that QR has in common with LU is that it can exploit sparsity, but in this case the preferred computational method is to use Givens rotations instead of Householder reflectors, since they can eliminate one nonzero at a time.