Notes on Eigenvalues, Singular Values and QR

Michael Overton, Numerical Computing, Spring 2017

March 30, 2017

1 Eigenvalues

Everyone who has studied linear algebra knows the definition: 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

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

of a polynomial of degree 5 or more,[2] so whatever we do we will have to use some kind of approximation algorithm.

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 premultiply 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

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

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

[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`.

Basic information on eigenvalues is also described in Chapter 4 of Ascher and Greif (p. 69–73, p. 77 and p. 79).

---

[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 $Q^{-1} = Q^*$.

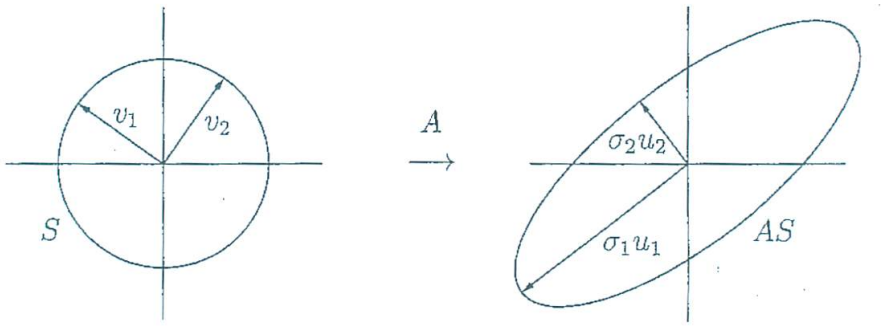
[6] The proof of this is more complicated but can be found in many books, such as the one by Trefethen and Bau.

[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 famous QR algorithm, described in many books including Ascher and Greif.

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

$$A\,[v_1, \ldots, v_n] = [u_1, \ldots, u_n]\,\hat{\Sigma}, \qquad \text{where} \qquad \hat{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_n).$$

[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.

Let $V = [v_1, \ldots, v_n]$ and let $\hat{U} = [u_1, \ldots, u_n]$ (note the "hats" on U and on Σ, but not on V). Then we have

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

Since the n columns of V have length n and form an orthonormal set, we have that the $n \times n$ matrix V is an "orthogonal" matrix, so $V^{-1} = V^T$ and we can write 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 $m > n$. But we can introduce a set of additional $m - n$ 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, using $V^{-1} = V^T$, 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 $A = U\Sigma V^T$, we have $Az = U\Sigma V^T z$. The only way this can be zero is if z is a linear combination of $v_{r+1}, \ldots, v_n$, since anything else will give $\Sigma V^T z \neq 0$ and therefore $Az \neq 0$. So, $v_{r+1}, \ldots, v_n$ 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 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

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

[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.

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

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$$

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.

Pseudo-inverse of a full-rank rectangular matrix. If $A = U\Sigma V^T$ has full rank, its pseudo-inverse is[12]

$$A^\dagger = (A^T A)^{-1} A^T.$$

Note that $A^\dagger b$ is the solution of the least-squares problem

$$\min_x \|Ax - b\|_2$$

(via the normal equations). If A is square and nonsingular,

$$A^\dagger = A^{-1}(A^T)^{-1}A^T = A^{-1},$$

the ordinary inverse, but this does not make sense if A is not square. If $A = U\Sigma V^T$, we have

$$A^T A = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T U^T U\Sigma V^T = V \begin{bmatrix} \hat{\Sigma} & 0 \end{bmatrix} \begin{bmatrix} \hat{\Sigma} \\ 0 \end{bmatrix} V^T = V\hat{\Sigma}^2 V^T$$

[12] This is the "Moore-Penrose" pseudo-inverse. There are many other variants of pseudo-inverse, but this is the one that is most commonly used.

so

$$A^\dagger = (V\hat{\Sigma}^2 V^T)^{-1} V\Sigma^T U^T = V\hat{\Sigma}^{-2} V^T V \begin{bmatrix}\hat{\Sigma} & 0\end{bmatrix} \begin{bmatrix}\hat{U}^T \\ * \end{bmatrix} = V\hat{\Sigma}^{-1}\hat{U}^T = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T,$$

where * denotes the transpose of the matrix of columns $u_{n+1}, \ldots, u_m$ that we appended to $\hat{U}$ to get U.

Condition number of the pseudo-inverse when A has full rank. Just as with the ordinary inverse, we can define

$$\kappa_2(A) = \|A\|_2 \|A^\dagger\|_2 = \frac{\sigma_1}{\sigma_n}.$$

Pseudo-inverse in the non-full-rank case. If A does *not* have full rank, with rank $r < n$, then we can instead define

$$A^\dagger = [v_1, \ldots, v_r] \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}) [u_1, \ldots, u_r]^T = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T.$$

i.e., we invert the positive singular values, but not the zero singular values.

3 QR Decomposition

The QR decomposition (or factorization) offers some of the same features as the SVD (see the exercises below), but it is much simpler to compute. The reduced or economy-size QR decomposition of an $m \times n$ matrix A, with $m \geq n$, is

$$A = \hat{Q}\hat{R}$$

where $\hat{Q}$ is an $m \times n$ matrix with n orthonormal columns of length m, so $\hat{Q}^T\hat{Q} = I_n$, and $\hat{R}$ is an upper triangular $n \times n$ square matrix. As with the SVD, let us append an additional $m-n$ orthonormal columns to the columns of $\hat{Q}$, and an additional $m - n$ rows of zeros below $\hat{R}$, giving the full version

$$QR = \begin{bmatrix}\hat{Q}, q_{n+1}, \ldots, q_m\end{bmatrix} \begin{bmatrix}\hat{R} \\ 0\end{bmatrix} = \hat{Q}\hat{R} = A$$

where now Q is an orthogonal $m \times m$ square matrix, so $Q^T = Q^{-1}$, and R has the same size as A. This notation is different from the notation in A&G, where Q is used for both versions of the factorization, which could be confusing.

The most well known way to compute the reduced QR factorization of a full rank matrix A is Gram-Schmidt. Given the n columns of A, which are vectors of length m, it first normalizes a_1 by scaling it, giving q_1. Then it "orthogonalizes" the pair q_1, a_2 by subtracting the appropriate multiple of q_1 from a_2, and normalizes the result, giving q_2. The details are in many books including A&G. The "classical" version is potentially unstable but there is an improved version called "modified Gram-Schmidt" which is stable. This process is sometimes called "triangular orthogonalization" because the operations generating $\hat{Q}$ from A amount to post-multiplying A by the triangular matrix $\hat{R}^{-1}$, so $\hat{Q} = A\hat{R}^{-1}$ and hence $A = \hat{Q}\hat{R}$.

Householder reflectors provide an elegant and stable way to compute the full QR decomposition. This may be viewed as "orthogonal triangularization" because it amounts to pre-multiplying A by an orthogonal matrix Q^T, reducing it R, a rectangular matrix with its first n rows in triangular form and the other $m-n$ rows equal to zero. Thus $R = Q^T A$, and hence $A = QR$. Again, details may be found in many books including A&G.

In MATLAB, the full QR decomposition is computed by `[Q,R]=qr(A)` and the economy-size QR decomposition is computed by `[Qhat,Rhat]=qr(A,0)`.

QR exercises. (Not to be submitted for now, maybe later.)

1. Show that if A has full rank, meaning that $Ax = 0$ implies $x = 0$, then $\hat{R}$ is a nonsingular matrix, and that the diagonal entries of R are all nonzero.

2. Show that if A has full rank, then $q_1, \ldots, q_n$ form an orthonormal basis for the range space of A.

3. Show that if A has full rank, then $q_{n+1}, \ldots, q_m$ form an orthonormal basis for the null space of A^T (**not** of A: the null space of A is $\{0\}$ when A has full rank). Hence, the range space of A and the null space of A^T are orthogonal to each other.

4. Show that if A has full rank, then the pseudo-inverse of A defined above has the simple form
$$A^\dagger = \hat{R}^{-1}\hat{Q}^T.$$
If A is square and nonsingular, this formula gives A^{-1}.

5. Show that, given u with $\|u\|_2 = 1$, the Householder reflector $P = I - 2uu^T$ is an orthogonal matrix. Hint: multiply out $P^T P$ and observe that the result is the identity matrix.

Note that, unlike the SVD, the QR decomposition cannot be used directly in the rank deficient case. For example, although $\hat{R}$ having a zero on the diagonal tells us that $\hat{R}$ and hence A is rank deficient, knowing that $\hat{R}$ has two zeros on the diagonal 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 is easy to compute, via modified Gram-Schmidt or Householder's method, in a finite number of operations 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 pp. 1-2 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.