Singular Vector Decomposition

January 19, 2015

Abstract

This is my writeup for Strang's Introduction to Linear Algebra, Chapter 6, and the accompanying video lectures. I have never had an intuitive understanding about SVDs... now I sorta do. Let's go over the basics, derivation, and intuition.

1 Eigenvalues and Eigenvectors

1.1 Introduction

Let A be a $n \times n$ matrix. Then $x \in \mathbf{R}^n$ is an eigenvector of A iff:

$$Ax = \lambda x \tag{1}$$

for some number λ. In laymen's terms, x does not change its direction when multiplied by - it merely scales by λ. λ is called the eigenvalue associated with x. Of course, any nonzero multiple of x is going to be a eigenvector associated λ; it is customary we take unit vectors.

1.1.1 Calculation

How do we find eigenvalues and eigenvectors given A? Rewrite (1) as:

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

In other words, x must be in null space of $A - \lambda I$. In order to have nontrivial x, $A - \lambda I$ must be singular. That's how we find eigenvalues - solve:

$$\det |A - \lambda I| = 0$$

This boils down to solving a nth order polynomial. The roots can be anything - they can be repeated, negative, or even complex.

1.1.2 Properties

These nice properties hold:

- If A is singular, 0 is an eigenvalue - of course, because $\det A = \det |A - 0I| = 0$.

- Determinant of A equals the product of all eigenvalues.
- Trace of A equals the sum of all eigenvalues.

1.1.3 Eigenvectors Are In $C(A)$

Something that was not immediately obvious was that each x was in the column space $C(A)$. x comes from null space of $A - \lambda I$, which doesn't seem to be related to A - so it sounded surprising to me first. However, recall the definition: $Ax = \lambda x$. Therefore, $\lambda x \in C(A)$ and so is x.

1.2 Diagonalization

Everything we want to do with eigenvectors work fine when we have n linearly independent eigenvectors. For now, let us assume we can choose such eigenvectors. Then, we can perform something that is very nice called diagonalization. Let $x_1 \cdots x_n$ be n eigenvectors, each associated with eigenvalues $\lambda_1 \cdots \lambda_n$. It is defined:

$$Ax_i = \lambda_i x_i \text{ for } i = 1 \cdots n$$

We can write this in matrix form. Let S be a matrix with x_i as column vectors. This is a $n \times n$ matrix, and the above LHS can be written as:

$$A \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} = AS$$

Let Λ be a diagonal matrix where $\Lambda_{ii} = \lambda_i$, then RHS can be written as:

$$\begin{bmatrix} \lambda_1 x_1 & \lambda_2 x_2 & \lambda_3 x_3 & \cdots & \lambda_n x_n \end{bmatrix} = S\Lambda$$

Joining the two gives

$$AS = S\Lambda$$

S is invertible since its columns are independent. Then we can write:

$$A = S\Lambda S^{-1}$$

which is called the diagonalization of A.

1.2.1 Matrices That Are Not Diagonalizable

When is A not diagonalizable? Say we have a repeated eigenvalue, λ_d which is repeated r times. $A - \lambda_d$ will be singular, but its rank might be greater than $n - r$. In such case, rank of the null space of the matrix is less than r, and we won't be able to choose r vectors that are independent.

1.3 Intuition

A $n \times n$ matrix can be seen as a function (actually it's a linear transformation): it takes a vector, and returns a vector. Eigenvectors can be regarded as "axis" of this function, where those on this axis are not changed (at least in terms of directions). If we use eigenvectors as axis, we can represent any vector in $C(A)$ in this coordinate system, by representing it as a linear combination of eigenvectors. After we do this change of coordinates, the transformation is done by multiplying each coordinate with its eigenvalues *individually*. There are no interactions between each axis anymore.

What happens when $x \notin C(A)$? Of course, even if $x \notin C(A)$, $Ax \in C(A)$. Say $x = a + b$ where $a \in C(A)$ and $b \notin C(A) \iff b \in N(A)$. Then $Ax = Aa + Ab = Aa$. Therefore we can just project x onto $C(A)$ first and proceed as if x was in $C(A)$.

So, now we see x as: projecting x onto $C(A)$ and do a change of coordinates, multiply each coordinate, and convert back to the original coordinate system. Note the diagonalization:

$$A = S\Lambda S^{-1}$$

Multiplying x by A is multiplying 3 matrices - S^{-1} which maps x onto the eigenvector coordinate system, Λ that multiplies each coordinates (remember that Λ is diagonal!), and S that maps the result onto the original coordinate system again.

1.3.1 Application: Powers of Matrix

Using this diagonalization, powers of matrices can be calculated easily. Let what is A^{100}? Since A^2 is

$$A^2 = \left(S\Lambda S^{-1}\right)\left(S\Lambda S^{-1}\right) = S\Lambda^2 S^{-1}$$

$A^{100} = S\Lambda^{100}S^{-1}$. And exponential of a diagonal matrix is really trivial.

2 Symmetric And Positive Definite-ness

2.1 Orthonormal Eigenvectors

When A is symmetric, we have special characteristics that make the diagonalization even better. Those are:

1. All eigenvalues are real.
2. We can always choose n orthonormal eigenvectors; not only they are independent, but they can be chosen to be perpendicular to each other!

The latter point now means that $S^T S = I \iff S = S^{-1}$. We should actually call this matrix Q now, as in QR decomposition. That gives the following diagonalization:

$$A = Q\Lambda Q^T$$

2.1.1 Proof

The first property is proven by taking complex conjugate of λ and discovering that $\lambda = \bar{\lambda}$. The second property is more complex to prove; it comes from Schur's theorem. Schur's theorem states that:

> Every square matrix can be factored into QTQ^{-1} where T is upper triangular and Q is orthonormal.

I think this is very intuitive - it's an analogue of using QR decomposition to find an orthonormal basis for the column space. Now, R is upper triangular - because first column in A is represented as a multiple of the first column of Q. Second column in A is represented as linear combinations of first two columns in R, etc. Now, we can think of decomposing R into

$$R = TQ^{-1}$$

where Q^{-1} maps x onto this coordinate system, and T working in this coordinate system. Then QTQ^{-1} works just like the diagonalization in 1.3.

Now we use induction and prove the first row and column of T are zeros, except for the top left element. If we continue like this, we get a diagonal T which means we will have:

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^T$$

2.1.2 Sum of Rank One Matrices

The symmetric diagonalization can be rewritten as

$$A = Q\Lambda Q^T = \lambda_1 x_1 x_1^T + \lambda_2 x_2 x_2^T + \cdots$$

Yep, it's a weighted sum of rank one matrices! Nice! Even better, $x_1 x_1^T$ are projection matrices! Because $\|x_i\| = 1$ since orthonormal, the projection matrix formula

$$A\left(A^T A\right)^{-1} A^T$$

becomes just $x_i x_i^T$ in this case.

So, we do have a great interpretation of this; multiplication by a symmetric matrix can be done by mapping x onto different, orthogonal 1D spaces. Then, each coordinate will be multiplied by a certain number. Later we can map it back to the original coordinates. This provides another insight regarding what eigenvalues and eigenvectors actually mean. The eigenvector associated with the largest eigenvalue is the axis where multiplying A makes the biggest change!

Since truncated SVD is just a generalization of this diagonalization, it is very intuitive how PCA can be done by SVD.

2.2 Positive Definite Matrices

2.2.1 Definition

A special subset of symmetric matrices are called positive definite. The following facts are all equivalent; if one holds, everything else will.

1. If we do elimination, all pivots will be positive.
2. All eigenvalues are positive.

3. All n upper-left determinants are positive.
4. $x^T Ax$ is positive except at $x = 0$.

2.2.2 Gaussian Elimination is Equivalent to Completing the Square

How on earth is property 4 related to other properties? Why are they equivalent? The course proceeds with an example, and states the following without further generalization.

$x^T Ax$ is effectively a second order polynomial: $\sum_{ij} x_i x_j A_{ij}$. How on earth do we know if this is bounded at 0? We complete the squares! Represent that formula by a sum of squares - and since squares are non-negative we know the polynomial will be positive unless all numbers are 0. The big revealation is that this process is same as Gaussian elimination. Positive coefficients on squares mean positive pivots! Holy crap. Thus, if 1-3 holds, 4 will hold. If 4 holds, the result of the completed squares will show that 1-3 holds.

2.2.3 Relation to Second Order Test in Calculus

Recall the awkward second order test in Multivariate Calculus, which was stated without proof? It actually checks if the Hessian is positive definite. Hahaha!

2.2.4 $A^T A$ is Always PosDef

A special form, $A^T A$, always yields positive definite matrix regarless of the shape of R - if R has independent columns. This is why covariance matrices are always positive!

Actually, the converse always holds too: if B is posdef, there always exists an A such that $A^T A = B$. This can be found by the Cholesky decomposition of B!

Recall: “all covariance matrices are positive definite, and all positive definite matrices are covariance matrices of some distribution”.

3 Singular Value Decomposition

3.1 Definition and Intuition

SVD is a generalization of the symmetric diagonalization. It is applicable to any matrix (doesn’t have to be symmetric, or even square). Therefore it is the most useful. How can this be done? Instead of using a single set of orthonormal basis, we use two sets U and V. Then a matrix A can be factored as:

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

where U and V are orthonormal. U is a basis of column space of A, V is a basis of row space of A. Σ is a (sort-of) diagonal matrix with singular values on its diagonals.

3.1.1 Intuition

The same intuition from symmetric diagonalization applies here. Given x, we can first map it to the row space. (Since V is orthonormal, $V^T = V^{-1}$ - so think of it as the mapping to the row space.) Now, x is represented as a linear combination of an orthnormal basis of the row space of A. We now multiply each coordinate by a number, and convert back to original coordinates by U. Since Ax is always in $C(A)$, columns of U are the basis of $C(A)$.

Hmm, sounds like magic. I know I can map to a space, and recover original coordinate system from that space. But I am mapping x to row space and recovering from column space as if we mapped x to the column space as if we mapped to column space in the first place. WTF? This holds because columns of V are sort-of eigenvectors of A; if we transform v_i, we get a multiple of u_i. Those are called singular vectors, and the relation goes like this:

$$Av_i = \sigma_i u_i$$

where σ_i are called singular values. Also, orthogonality gives that all v_i and u_is are orthogonal to each other! Woah, there's too much magic in here.. :)

3.1.2 Analogue to Rank-one Decomposition

We can also write:

$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots$$

we can easily see how PCA works with SVD.

3.2 Derivation

When $m \neq n$ for $A \in \mathbf{R}^{m\times n}$, the dimensions are kind of match up. But I will gloss over the details. (It's past midnight)

Let U be the eigenvector matrix for AA^T, V be the eigenvector matrix for A^TA and everything works out. We can choose these matrices to be orthnormal, because A^TA and AA^T are symmetric. Now we assert we can find Σ such that:

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

Multiply both sides by A^T on the left:

$$A^TA = \left(U\Sigma V^T\right)^T U\Sigma V^T = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$$

the last equality coming from that we chose Σ to be orthonormal. Yep, RHS is the symmetric decomposition of A^TA - I said it works out. So the diagonal entries of Σ will contain square roots of eigenvalues of A^TA.

Now we can close the loop by proving

$$AV\Sigma^{-1} = U \tag{2}$$

First, we prove that columns of $AV\Sigma^{-1}$ have unit length:

$$\begin{aligned} A^TAV = V\Sigma^2 &\iff V^TA^TAV = V^TV\Sigma^2 \\ &\iff (AV)^T AV = \Sigma^2 \end{aligned}$$

The last equality shows that columns of AV are orthogonal, and the ith singular value gives its length squared.

Next, we prove $AV\Sigma^{-1}$ is indeed an eigenvector of AA^T. Multiply (2) by A on both sides, on the left, and rearrange to get:

$$AA^T AV = \Sigma AV$$

Note that column vectors of AV are actually eigenvectors of matrix AA^T. (Smart parenthesis will let you realize this).

4 That Was It!

Wasn't too hard right?