Project 2 NLA: SVD Applications

Iris Vukovic

December 9, 2024

A brief report with some theoretical comments, details of implementation, and difficulties found. There are some theoretical commentaries and implementation details found within the notebook itself because in some cases I felt that it made more chronological sense to add explanations among the code implementations and results.

1. Least Squares Problem

   Here I will discuss some ideas and methods I found useful to explore when I was working through the least squares problem with QR and SVD decompositions. When I was using QR to decompose a well conditioned, full rank matrix A in the first part of the problem, I found that I could use either **Gram-Schmidt** orthogonalization or **Householder reflections** (which are utilized in the np.linalg.qr() function) to orthogonalize the input matrix and both worked fine. Simply put, Gram-Schmidt produces a new vector that is orthogonal to the previous ones and scales the orthogonalized vector to unit norm, creating an orthonormal family of vectors, the columns of Q. R is constructed from the dot product of the original columns of our input matrix and the columns of Q. However, it is numerically unstable and thus can only be utilized when our input vectors are linearly independent. This is because part of the algorithm consists of subtracting the projected vectors from the previous orthogonal vectors and when the vectors are linearly dependent and very similar, this can result in cancellations that can be catastrophic. On the other hand, Householder reflections are much more numerically stable. Essentially, the method is a sequence of orthogonal transformations, reflections over a hyperplane, that result in an upper triangular matrix, R for our decomposition. Each transformation produces an orthogonal Householder matrix and the product of these matrices gives us Q. With each transformation, we are trying to zero out all the components of the input vector except one of them. Since it calculates the reflection directly and avoids the subtractions that make Gram-Schmidt unstable, the method is more numerically sound.

   When I was doing the polynomial fitting portion of the exercise, I had to construct my own matrix from the given data points. Each column of the matrix progressively increased the degree of the given input variable, i.e. in the case of a 1xn matrix, the matrix would be as such $[1, x_1, x_1^2, ...x_1^{n-1}]$. This concept is formally called a **Vandermonde matrix**.

To calculate the best fit solution to minimize the least squares problem, we had $x = A^+b$. In this equation, A^+ represents the **Moore-Penrose pseudoinverse** which is a generalization of the inverse matrix. Whereas the inverse of a matrix can only be calculated for square, non-singular (full rank) matrices, the psuedoinverse can be computed for all matrices, includnig rectangular, singular ones. In the case of SVD decomposition, $A^+ = V\Sigma^+U^T$, so in fact we end up taking the pseudoinverse of Σ, which takes the inverse of all nonzero values in the matrix and leaves the zero values as zeroes.

2. Graphics Compression

Eckhart-Young Theorem: The partial sum $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U_k \Sigma_k V_k^T$, where $r = rank(A)$ and $1 \leq k \leq r$, is the best rank k approximation of A with respect to the 2-norm and the Frobenius norm. Thus, A_k minimizes both the 2-norm error and the Frobenius norm error.

Proof (2-norm):Let's take matrtix B (also of size mxn same as A) and say that rank(B) = k, then we will try to show that $||A - B||_2 \geq ||A - A_k||_2$ in order to demonstrate that the latter term represents the smallest possible error for a matrix of rank k.

We find, using Singular Value Decomposition, that $||A - A_k||_2 = ||\sum_{i=k+1}^{n} \sigma_i u_i v_i^T||_2 = \sigma_{k+1}$.

Let's define $B = CD^T$ where C is a matrix of size mxk and D is a matrix of size nxk, so by this definition rank(B)$\leq k$.

Now we choose a vector w such that $D^T w = 0$, which by definition implies that w is orthogonal to all columns of D and so A along the vector w is part of the error that can not be captured by B since B is restricted to a k-dimensional subsace.

The vector w can be written as a linear combination of singular vectors $w = \gamma_1 v_1 + ... + \gamma_{k+1} v_{k+1}$ (where $v \in V^T$ from SVD of A and γ is a scalar coefficient). We choose w so that $||w||_2 = 1$, implying that $\sum_i^{k+1} \gamma^2 = 1$.

We take $||(A-B)w||_2^2$ which is equivalent to $||Aw||_2^2$ (since w is what B can't capture). Now, the action of A on w, w which is composed of right singular vectors v_i, produces left singular vectors $\sigma_i u_i \in \Sigma U$, and so we have $Aw = \gamma_1 \sigma_1 u_1 + ... + \gamma_{k+1} \sigma_{k+1} u_{k+1}$.

We know that the values of σ_i are decreasing, i.e. $\sigma_1 \geq ... \geq \sigma_{k+1}$, which implies that $||Aw||_2^2 \geq \sigma_{k+1}^2$.

Thus we have $||A - B||_2^2 \geq ||(A - B)w||_2^2 \geq= ||Aw||_2^2 \geq \sigma_{k+1}^2$. Taking the square root, we finally reach that $||A - B||_2 \geq \sigma_{k+1}$. Remember that $||A - A_k||_2 = \sigma_{k+1}$, so it follows that $||A - B||_2 \geq ||A - A_k||_2$ and our proof is complete.

Proof (Frobenius norm):Once again, we are defining the same B as from the above proof and we are trying to show the same thing: $||A - B||_F \geq ||A - A_k||_F$, except this time with Frobenius norm instead of the 2-norm.

We have that $||A - A_k||_F^2 = ||\sum_{i=k+1}^{n} \sigma_i u_i v_i^T||_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$, the error being the sum of singular value σ_i starting at σ_{k+1}.

Let's decompose A into $A_1 = A - B$ and $A_2 = B$. When we use the notation $\sigma_i(A)$, we are referring to the singular value at index i of the matrix A. So, with $i, j \geq 1$, we consider

$$\sigma_i(A_1) + \sigma_j(A_2) \geq \sigma_1(A_1 - A_{1_{i-1}}) + \sigma_1(A_2 - A_{2_{j-1}})$$

noting that σ_1 represents the largest singular values of the truncated matrices since the singular values go in descending order. But since the matrices have been truncated, their largest singular values must be smaller than those of the non-truncated matrices. By triangle inequality,

$$\sigma_1(A_1 - A_{1_{i-1}}) + \sigma_1(A_2 - A_{2_{j-1}}) \geq \sigma_1(A - A_{1_{i-1}} - A_{2_{j-1}})$$

Since $A - A_{1_{i-1}} - A_{2_{j-1}}$ is the residual matrix matrix after removing the first $i-1$ and $j-1$ singular values respectively, we are left with a matrix of rank at most $i - 1 + j - 1 = i + j - 2$ (since the rank is the number of nonzero singular values),

$$\sigma_1(A - A_{1_{i-1}} - A_{2_{j-1}}) \geq \sigma_1(A - A_{i+j-2}) = \sigma_{i+j-1}(A)$$

because the largest singular value of the truncated matrix $A - A_{i+j-2}$ will be $\sigma_{i+j-1}(A)$. So we have reached that

$$\sigma_i(A_1) + \sigma_j(A_2) \geq \sigma_{i+j-1}(A)$$

Let's set $j = k + 1$ and $i \geq 1$. Since $A_2 = B$ and B is limited to a k-dimensional subspace, $\sigma_j(A_2) = \sigma_{k+1}(B) = 0$, so, plugging $A - B$ back in for A_1, we are left with $\sigma_i(A - B) \geq \sigma_{k+i}(A)$.
From this inequality, it follows that $||A - B||_F^2 = \sum_{i=1}^{n} \sigma_i^2 \geq \sum_{i=k+1}^{n} \sigma_i^2 = ||A - A_k||_F^2$, which is exactly what we set out to prove.

3. Principal Component Analysis: Choosing a sufficient number of principal components to explain the dataset, some methods.

   Kaiser Rule: A principal component with variance equivalent to 1 offers the same amount of information about the dataset as the original variable does. So, the idea is to only retain principal components whose variance is greater than 1 so that we can be sure that we are analyzing components that tell us more about the dataset than the original features of the dataset do. Depending on the dataset though, it's possible that this method selects too many components.

   Scree Plot: A graph that plots the principal component and the associated variance that it captures. We look for the point where the variance attached to a principal component drops off and plateaus, often called the elbow. Then we keep the components that come prior to the elbow/drop-off point. This way we are keeping only the principal components that have high and significantly differing variances. However, some scree plots have multiple elbows, which can make the test too subjective in terms of choosing the best elbow to stop at. In the image below, the chart on the left shows a scree plot example in which we would choose the first three principal components.

3/4 total variance: With this method, we set a threshold that represents the maximum percentage of the total variation, stop at the principal component that accounts for a percentage that is greater than the threshold, and collect all principal components up to that point for analysis. Note that the '3/4' in '3/4 total variance' represents a threshold of 0.75 and this can be changed depending on the dataset. Common threshold values follow the normal distribution. A lower threshold reduces the dimensionality of the space more, because we have less principal components, so choosing a threshold value can depend on how much we want to compress the original dataset and how much information we want to retain. In the image below, in the chart on the right, if we set the threshold to 0.95, we would collect the first four principal components.