Lecture 21 - Notes

Rahul Arya

April 2019

1 Overview

In this lecture, we will complete our discussion of image compression from when we first introduced the SVD. In particular, we will quantify the accuracy of our compressed images using a metric known as the *Frobenius Norm*. We will also begin to explore an important application of the SVD to statistics[1], known as *principal component analysis*.

2 Image Compression and the Frobenius Norm

Recall that, given a rectangular image represented as an $m \times n$ matrix A, we need $\Theta(mn)$ space to store it in the naive fashion. However, if we use the SVD to express it as a sum of outer products (assuming WLOG that $m \geq n$)

$$A = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{V}_2^T + \ldots + \sigma_n \vec{u}_n \vec{v}_n^T,$$

we can approximate it as

$$\hat{A} = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{V}_2^T + \ldots + \sigma_k \vec{u}_k \vec{v}_k^T,$$

where k is some constant much smaller than n and m. By representing $\hat{A}$ as a sum of outer products, we can store it in $\Theta(k(m+n))$ space, a significant improvement over the naive approach.

But does this approach really work? That is to say, is the compressed $\hat{A}$ anything at all like A? Last time, we saw that this worked for matrices A that were initially of low rank, where $\hat{A}$ could perfectly represent A. But most matrices, including those representing images, are of full rank, so $A \neq \hat{A}$, so there will be some error term $A - \hat{A}$. We would like to quantify the magnitude of this error, and will do so using the *Frobenius norm*.

Consider some error term of the form

$$\Delta = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

[1] Or "data science" if you wanna be trendy.

We define the Frobenius norm of Δ to be

$$\|\Delta\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

In essence, we sum up the squares of all the terms in the matrix, and then take the square root. This norm has all the properties of norms that we expect: $\|\Delta\|_F = 0 \iff \Delta = 0$, $\|k\Delta\|_F = k\|\Delta\|_F$, and $\|\Delta_1 + \Delta_2\|_F \leq \|\Delta_1\|_F + \|\Delta_2\|_F$[2].

We can interpret this quantity in a number of ways. One would be to stack all the columns or rows of Δ into a single vector, and take its Euclidean norm. Alternatively, it could be thought of as summing the squared Euclidean norms of each of the rows or columns of Δ, and taking the square root of the sum. Regardless of interpretation, what's important is to see intuitively that $\|\Delta\|_F$ gets larger as Δ gets "bigger".

We will now look at how this norm relates to the SVD. To do so, we will need to establish some properties. In particular, we claim that

$$\|\Delta\|_F = \sqrt{\sum_{i=1}^{n}(\Delta^T\Delta)_{ii}} = \sqrt{Tr(\Delta^T\Delta)},$$

where $Tr(X)$ is the sum of the diagonal entries of any matrix X, known as the *trace*. This fact can be proven in a straightforward manner using the definition of matrix multiplication, by observing that

$$(\Delta^T\Delta)_{ii} = \sum_{j=1}^{m}(\Delta^T)_{ij}\Delta_{ji} = \sum_{j=1}^{m}\Delta_{ji}^2.$$

Therefore,

$$Tr(\Delta^T\Delta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\Delta_{ji}^2 = \|\Delta\|_F^2,$$

from which our desired result immediately follows by taking the square roots of both sides.

Now, we will aim to use this property to relate the Frobenius norm of a matrix to its SVD. Let the SVD of Δ be $U\Sigma V^T$, where U and V are both square orthogonal matrices, and Σ is a diagonal matrix. By a direct substitution and

[2]These properties aren't super important for our purposes, which is why we won't prove them rigorously.

application of the above property, we see that

$$\begin{aligned}\|\Delta\|_F &= \sqrt{Tr(\Delta^T\Delta)} \\ &= \sqrt{Tr((U\Sigma V^T)^T(U\Sigma V^T)} \\ &= \sqrt{Tr(V\Sigma^T U^T U\Sigma V^T)} \\ &= \sqrt{Tr(V\Sigma^T\Sigma V^T)} \\ &= \left\|\Sigma V^T\right\|_F\end{aligned}$$

as $U^TU = I$. In other words, we see that pre-multiplication by a orthogonal matrix U does not affect the Frobenius norm.

Now, observe from the definition that taking the transpose of a matrix clearly does not affect its Frobenius norm, since its components are merely being rearranged. Therefore, continuing the above calculation,

$$\begin{aligned}\left\|\Sigma V^T\right\|_F &= \left\|V\Sigma^T\right\|_F \\ &= \sqrt{Tr((V\Sigma^T)^T(V\Sigma^T)} \\ &= \sqrt{Tr(\Sigma V^T V\Sigma^T)} \\ &= \sqrt{Tr(\Sigma\Sigma^T)} \\ &= \sqrt{\sum_i \sigma_i^2},\end{aligned}$$

this time using the fact that $V^TV = I$, where σ_i are the singular values of Δ. Therefore, we find that the Frobenius norm of a matrix can be viewed as the square root of the sum of the squares of its singular values.

Recall that our goal was to produce a good approximation $\hat{A}$ of A using the SVD, in order to minimize the Frobenius norm of the error term $\Delta = A - \hat{A}$. Using our definitions of A and $\hat{A}$ from above, we see that

$$\Delta = A - \hat{A} = \sigma_{k+1}\vec{u}_{k+1}\vec{v}_{k+1}^T + \sigma_2\vec{u}_{k+2}\vec{V}_{k+2}^T + \ldots + \sigma_n\vec{u}_n\vec{v}_n^T,$$

so it clearly has the nonzero singular values σ_{k+1} through σ_n. Thus, its Frobenius norm is

$$\|\Delta\|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + \ldots + \sigma_n^2}.$$

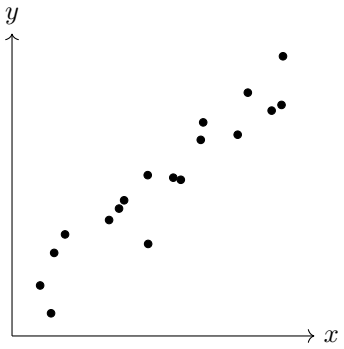
So, is this quantity small? As it turns out, in practice for real images, the singular values drop off rapidly, so $\sigma_1 \gg \sigma_2 \gg \ldots$. Thus, for even small constants k (say, on the order of dozens), σ_j for $j \geq k$ is several orders of magnitude smaller than σ_1, so $\|\Delta\|_F$ is much smaller than $\|A\|_F$, meaning that this approach allows us to compress real-world images quite well.

In fact, there exists a result, known as the *Eckart–Young–Mirsky theorem*, which states that the $\hat{A}$ we produce is the *optimal* rank-k approximation of our input A (in that it minimizes $\left\|A - \hat{A}\right\|_F$), even if we are allowed to use techniques other than the SVD. The proof will not be presented here (though the techniques used to derive it are in scope) since this theorem may be part of a future homework problem.

3 Principal Component Analysis

We will now begin looking at one of the main applications of the SVD in this class, known as *principal component analysis*. Broadly speaking, this technique allows us to consider (potentially noisy) data, and rewrite it in terms of uncorrelated aspects, which can then be considered separately. We will first describe the mechanical computation of the PCA of a dataset, and then discuss its meaning and importance to statistics.

Consider a dataset made up of n data points, each consisting of m scalar observations represented as real numbers. For instance, each observation could be of a student at Berkeley, with the observations being their grades on 61A, 61B, 70, 16A, and 16B (for $m = 5$). We represent this dataset in an $n \times m$ matrix A, where each row of A corresponds to a different data point, and each column contains one particular scalar observation across all data points. We may plot a set of data points as points in m-dimensional space - for instance, for $m = 2$, we could start with the following dataset:



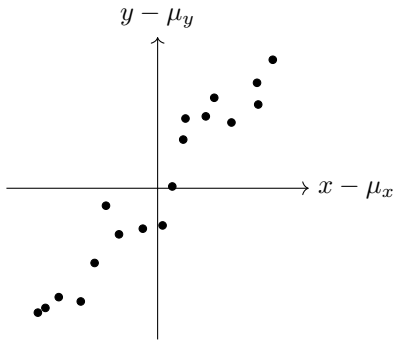
The goal of PCA will be to select a new basis for our observations such that each basis vector represents an uncorrelated characteristic of a data point. Visually, the goal of PCA is to choose orthogonal axes that our data points are "aligned about" - for instance, in the above example, the data appears to be aligned about the line $y = x$, so we would expect a vector in that direction to be one of our basis vectors.

First, though, we will translate our data such that it is centered about the origin, so that a constant offset of our entire dataset can be neglected in favor of the

variation within the dataset. More precisely, letting μ_i be the mean of the ith scalar observation across all data points, we obtain the translated

$$\widetilde{A} = \begin{bmatrix} A_{11} - \mu_1 & A_{12} - \mu_2 & \cdots & A_{1m} - \mu_m \\ A_{21} - \mu_1 & A_{22} - \mu_2 & \cdots & A_{2m} - \mu_m \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} - \mu_1 & A_{n2} - \mu_2 & \cdots & A_{nm} - \mu_m \end{bmatrix}.$$

From a visual perspective, our example plot would become



Now that we have centered our data, we will compute what is known as the *covariance matrix* S, of our dataset, defined as

$$S = \frac{1}{n}\widetilde{A}^T\widetilde{A}.$$

What does this matrix look like, and what does it represent? Well, letting $\vec{\widetilde{A}}_i$ represent the column vector of $\widetilde{A}$ corresponding to the ith measurement across all data points, observe that each entry

$$S_{ij} = \frac{1}{n}(\vec{\widetilde{A}}_i \cdot \vec{\widetilde{A}}_j) = \frac{1}{n}\left\|\vec{\widetilde{A}}_i\right\|\left\|\vec{\widetilde{A}}_j\right\|\cos\theta,$$

applying the geometric definition of the dot product, where θ is the angle between the ith and jth observation vectors.

Thus, S_{ij} will be positive and large if the ith and jth observation vectors are mostly aligned (indicating that they behave very similarly), zero if they are mostly orthogonal (indicating that their values appear to be unrelated) and negative and large if they tend to behave in opposite ways (so if the ith observation is large, the jth will be small, and vice-versa). This quantity is known as the *covariance* between the ith and jth measurements for our dataset. Observe also that $S_{ij} = S_{ji}$, so S is symmetric. This makes sense, since we'd expect the covariance between two measurements to be calculated by treating the two measurements in a symmetric fashion.

Notice that the $\frac{1}{n}$ term means that S_{ij} will not increase with more observations, since it exactly offsets the growth in the magnitudes of the observation

vectors. However, S_{ij} also has a dependence on the magnitudes of the scalar observations, which may not be desirable. For instance, if we switch the units of our measurements from meters to millimeters (as an example), then the magnitude of S_{ij} will increase by a factor of 10^6, even though the data hasn't really changed! Instead, we'd like it to depend only on the relation between the two observations.

To do so, we will simply divide out the unwanted factors. More precisely, let the scalar values

$$S_i = \sqrt{S_{ii}} = \frac{1}{\sqrt{n}}\left\|\tilde{\vec{A}}_i\right\| = \sqrt{\frac{\sum_{j=1}^{n} A_{ij}^2}{n}}.$$

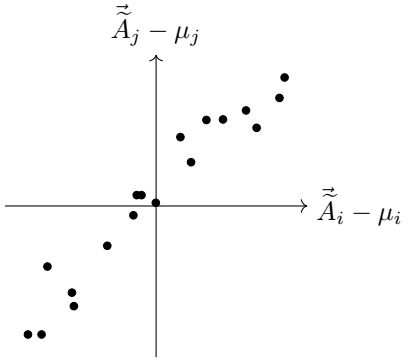
Notice that S_i is in fact the standard deviation of the ith measurement over the entire dataset, though this fact is technically out of scope for the course.

Now, we can divide out to obtain the *correlation* matrix R, defined such that

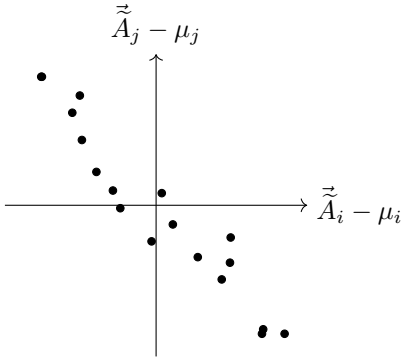
$$R_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{\frac{1}{n}\left\|\tilde{\vec{A}}_i\right\|\left\|\tilde{\vec{A}}_j\right\|\cos\theta}{\left(\frac{1}{\sqrt{n}}\left\|\tilde{\vec{A}}_i\right\|\right)\left(\frac{1}{\sqrt{n}}\left\|\tilde{\vec{A}}_j\right\|\right)} = \cos\theta,$$

where θ is as defined previously. This matrix sounds great! Observe that R is also symmetric, with all its entries lying between -1 and 1, with its diagonal entries all normalized to 1.

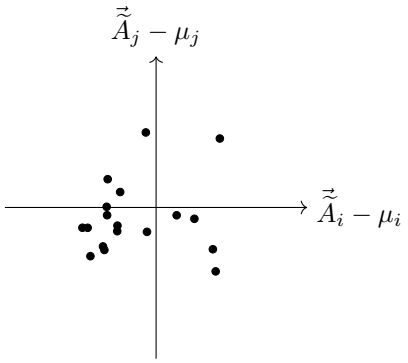
Let's look visually at what the R_{ij} represent. If $R_{ij} \approx 1$, then we expect the ith and jth measurement vectors to be aligned in some fashion, as shown:



In contrast, if $R_{ij} \approx -1$, then we expect the behavior of the jth measurement to be the opposite of the ith measurement, with the two measurement vectors pointing on opposite directions. Plotting this behavior, we'd expect to see something like:



And of course, if $R_{ij} \approx 0$, we'd expect to see no correlation at all, as shown:



Observe, however, that there are a few slight edge cases in our definition of correlation. In particular, notice that it is defined as the ratio

$$R_{ij} = \frac{S_{ij}}{S_i S_j},$$

but what if $S_i = 0$? This would occur, for instance, if all the ith measurements across the dataset were constant. Then we could not say anything about its correlation to the jth measurements, since we could not know how the jth measurements vary as the ith measurements change, since in our dataset the ith measurement is always a constant.

Thus, despite the very intuitive nature of the correlation matrix, we will have to perform PCA with respect to the covariance matrix S (which is always defined, since there's never a risk of dividing by zero), in order to obtain a fully general result. Recall that S is a symmetric matrix, and so can be diagonalized into

$$S = P \Lambda P^T,$$

where P is an orthogonal matrix and Λ consists of nonnegative numbers. As it turns out, the columns of P are our *principal components*, and if we write our

dataset in the basis of P, we will see that these columns provide us with uncorrelated, orthogonal, measurement directions. While intuitively this may "feel" true, a rigorous justification of this result will be presented next lecture.