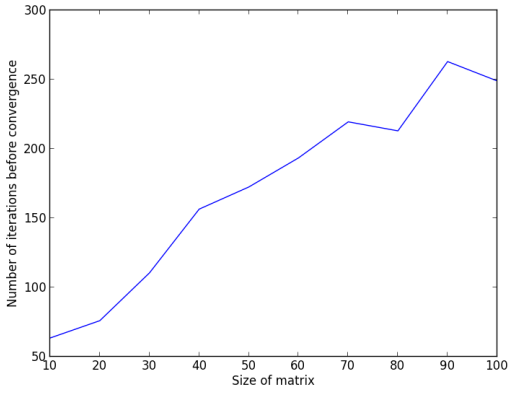
Principal Component Analysis / Singular Value Decomposition / Matrix completion

Andrew Beinstein

June 25, 2013

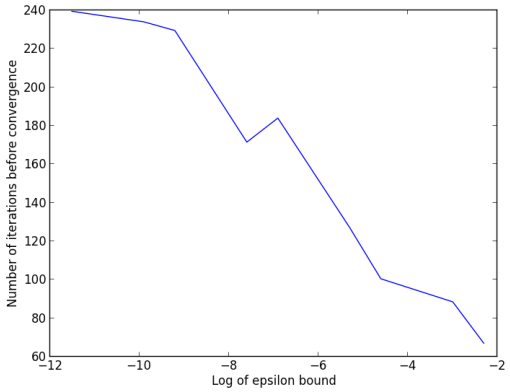
1 eigP

First, we will look at the number of iterations required for convergence changes with respect to the size of the matrix, and the choice of the epsilon bound. The first plot below shows the relationship between the size of the matrix, and the number of iterations required for convergence.



To calculate this, I counted the number of iterations for randomly generated symmetric matrices of size $10 + 10i$ for $0 < i \leq 10$. For each size, I did the procedure 100 times, and took the median of the result (so outliers would not affect the results). I set the epsilon value to .001.

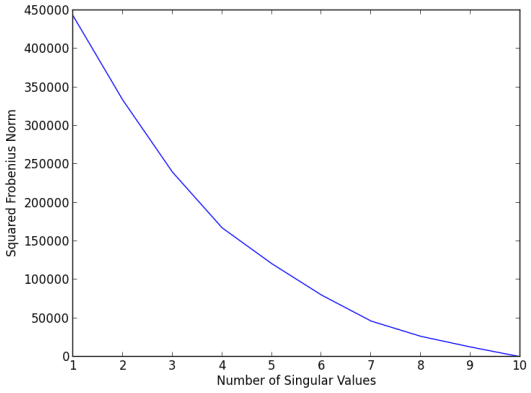
Next, I looked at the relationship between the number of iterations required for convergence and the epsilon bound.



I counted the required number of iterations for convergence for 9 different epsilon values, between .1 and .00001. I used a randomly generated 50 x 50 symmetric matrix. For visual clarity, I plotted the log of the epsilon value over the number of iterations, and we see that the number of iterations increases with a decrease in epsilon value.

2 svdP

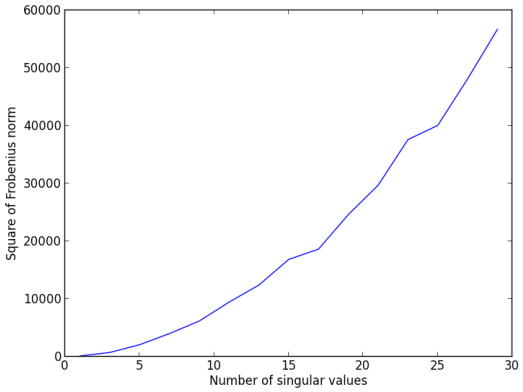
Now, we investigate the Singular Value Decomposition. In particular, the plot below investigates the relationship between the number of singular values, and the squared Frobenius norm of the difference between the reconstructed matrix and the original matrix.



We see that the squared norm decreases exponentially as a function of the number of singular values. Upon further investigation, I noticed that the norm decreases by a factor of between 1.3 and 2.5 every time, with these factors increasing for each singular value.

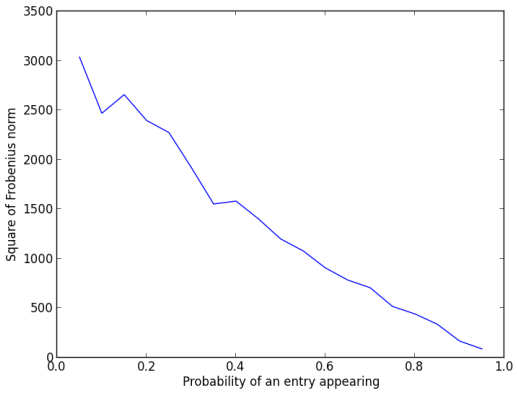
3 genA

Now, we will focus on matrix completion. I generated a 30 x 30 real matrix, with the left and right singular vectors randomly chosen as described in the assignment. Then, I randomly zeroed out some of the entries with probability $1 - p$, and tried to reconstruct the original matrix using the Singular Value Decomposition. The following plot shows the relationship between k, the number of singular values, and the square of the Frobenius norm.



In this plot, I used a probability value of 0.3. For each k value, I took the median of 50 calculations. As you can see, the square of the Frobenius norm increases with the number of singular values used. This means that the SVD gets worse at reconstructing the original matrix with the more singular values used!

Now, here is the relationship between p, the probability that an entry is non-zero, and the Frobenius norm.

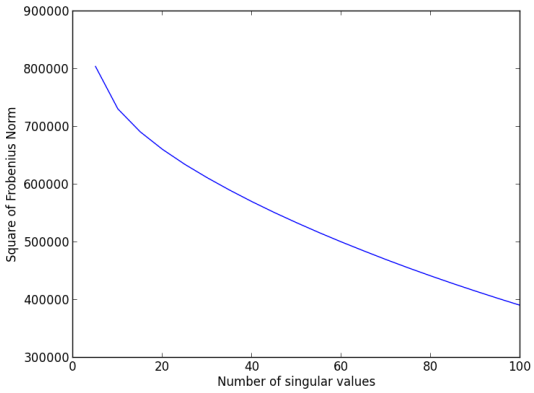


As expected, the SVD does an increasingly better job with reconstructing the matrix with the more entries that are non-zero. I computed this plot similarly to the previous one, with a common k value of 5.

4 Movies!

Finally, we will apply all of this abstract linear algebra stuff to predicting which movies people will enjoy, which is the single most pressing and important problem of our generation.

Below is a plot which showcases the accuracy of the SVD procedure, and reveals a pattern to people's movie-watching habits.



We see that our algorithm gets more accurate with the number of singular values used. To construct this plot, I ran the SVD on the movie matrix, using singular values between 5 and 100. With each singular value, the SVD produced a more accurate reconstruction of the original matrix. The steepest improvement takes place in the beginning, but the movie matrix begins to gradually flatten out, as the noise in the data overpowers the signal.

Then, I analyzed the 3 largest right singular vectors. These right singular vectors will correspond with particular genres, or movies that similar groups of people rate highly. The top 5 movies that contributed the most weight to the first eigenvector are, in order: 1) Star Wars, 2) Raiders of the Lost Ark, 3) Return of the Jedi, 4) Fargo, 5) Silence of the Lambs. This group seems to correspond to the most popular movies of this generation, with a particular bent to action/adventure films. The movies that contributed the most weight to the second eigenvector are, in order: 1) Dr. Strangelove, 2) Casablanca, 3) The Graduate, 4) Citizen Kane, 5) The African Queen. This eigenvector seems to consist of primarily older movies, which may have been popular with parents. However, the genres here vary, as Dr. Strangelove is a comedy, but Casablanca is a romantic drama. All of these movies are infamous and well-known, which

is why they may have a large influence. Finally, the top 5 movies for the 3rd eigenvector are, in order: 1) Jurassic Park, 2) True Lies, 3) Top Gun, 4) Speed, 5) Batman. These movies look like younger action films, that may have been very popular with teenagers of the day.

My preliminary conclusion from analyzing the movie data set is that age has much to do with rating similarity. This makes some sense, as people tend to watch the movies that correspond with their age group.