CS 6140: Data Mining Assignment 7

Dimensionality Reduction

Yash Gangrade (u1143811), MS in Computing

3$^{\text{rd}}$ April 2019

Contents

1 Singular Value Decomposition

1.1 L2 norm of the difference between A and Ak

Ans: The L2 norm is computed for the difference between A and Ak values for all the values of k in range of [1, 10] using the matlab function, *norm(A - Ak, 2)*. The values are summarized in form of a table below.

k	L2 norm
1	1862.6
2	1525.7
3	1171.9
4	925.1212
5	827.8121
6	815.2254
7	639.6120
8	526.8578
9	327.0397
10	227.2520

1.2 Smallest value of k

The 10% value of L2 norm of A is 636.6475. So the smallest value of k for which the L2 norm of A - Ak is less than 636.6475 is **k = 8** at which the L2 norm of difference is 526.8578.

1.3 Plotting

Here, we are considering the matrix as 5000 points in 40 dimensions. We have to plot the points in 2 dimensions such that it minimizes the sum of the residuals squared. So, we know that $A = USV^T$ and here the dimensions is 40. We are also centering them using PCA where we calculate SVD of $C_n * A$ instead of A where Cn is defined as $I_n - \frac{1}{n} * \mathbf{1}\mathbf{1}^T$ If we want to reduce the number of dimensions to 2, we just have to select the appropriate rows and columns from the U and S matrix. Here we are doing the dot product of U2 and S2 matrix. Thus our question boils down to plotting U2*S2 where $U2 = U(:, 1 : 2)$ and $S2 = S(1 : 2, 1 : 2)$. Essentially, to create the X vector, we do $U(:, 1) * S(1, 1)$ and to create the Y vector we do $U(:, 2) * S(2, 2)$. The resultant matrices are 5000 x 2 double matrix. Then we simply use the **scatter(x, y)** function in MATLAB to plot the X and Y values. Please find the plot below.

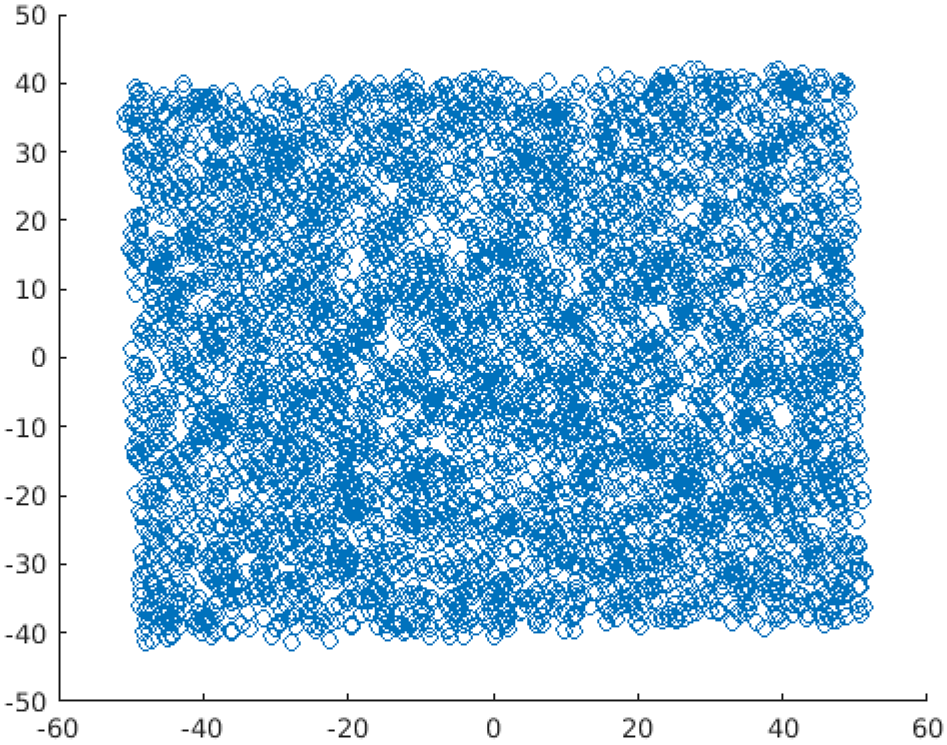

Figure 1: Scatter plot

2 Frequent Directions and Random Projections

2.1 Measuring the error

We are measuring the error using $norm(A' * A - B' * B, 2)$. The error values and the theoretical bound (k = 0) for L = 1 to 10 are summarized in the table below.

l	**Error**	**Theoretical Bound (k = 0)**
1	40504747.6652	50915773.2179
2	6499977.0604	25457886.6089
3	3746744.9112	16971924.4059
4	2317369.6650	12728943.3044
5	1492120.3940	10183154.6435
6	1049785.4826	8485962.2029
7	797950.7974	7273681.8882
8	517808.7327	6364471.6522
9	316702.6556	5657308.1353
10	133994.1245	5091577.3217

2.1.1 How large does L need to be for the above error to be at most $||A||_F^2/10$?

Through the program, we get $||A||_F^2/10 = 5091577.3217$. From the above table and the information about $||A||_F^2/10$, we can conclude that L needs to be set up at 3 for the error to be at most 5091577.3217.

2.1.2 How does this compare to the theoretical bound (e.g. for k = 0)?

From the table above, we can see that at L = 8, the error goes below the theoretical bound. So L should be set up to 8 if we want our error to be at most theoretical bound.

2.1.3 How large does L need to be for the above error to be at most $||A - A_k||_F^2/10$ (for k= 2)?

Through our analysis, we found that $\frac{||A-A_2||_F^2}{10} = 691449.0699$. So from this information and the table, we conclude that L need to be 8 for the error to be at most 691449.0699.

2.2 Random Projections

Here, I am running 500 trials for each value of L (upto 300) and then averaging the error and reporting that average value of error for that L. For this dataset, the error increases after decreasing till around L = 70 and it starts increasing. Please refer to the plots below. We get very close to $||A||_F^2/10 = 5091577$ but never go lower than this. I started with 50 trials for each L but the results were too spiky so I moved on to 150 trials per L and it produced fine results. Finally I increased the number of trials to 500 and it produces good results as the fluctuation has decreased. Please find the attached plots which demonstrates the trend of the graph. From the information above and the graphs, I would say L should be on the order of 70 to make sure that error is at most $||A||_F^2/10$. At L = 70, we get an error of 6049014.8043.

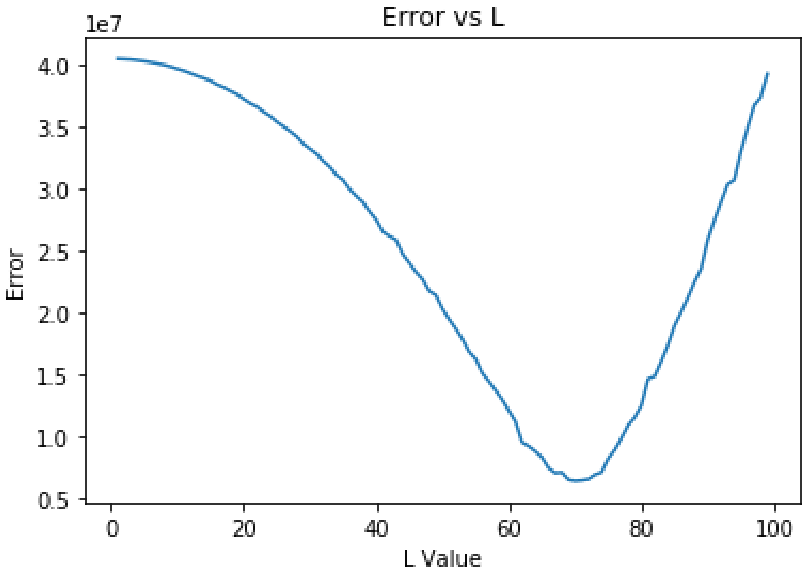

Figure 2: Error vs L (L upto 100)

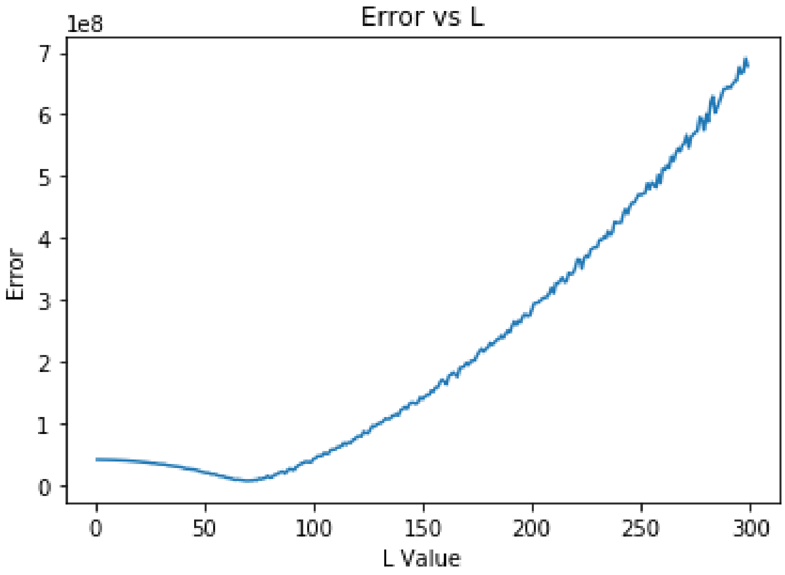

Figure 3: Error vs L (L upto 300)