Programming Assignment - 2

Manjulata Chivukula

February 25, 2019

1 Variational EM

Implemented variational EM algorithm described in the paper - [1] on provided datasets:

1. dataset-40-20-2.txt
2. dataset-500-500-3.txt
3. ml-100k.txt

Dataset are split into equal parts train and test and following RMSE are taken:

- RMSE on the baseline (training set)
- RMSE on the train set
- RMSE on the test set

1.1 Results - dataset-40-20-2.txt

The algorithm was run for the dataset: data-40-20-2.txt and the RMSE values after 10 iterations are shown in the table: 1. Plot for the results is shown in the figure: 1

RMSE Baseline	1.6045
RMSE train	1.1134
RMSE test	1.5224

Table 1: RMSE values after 10 iterations

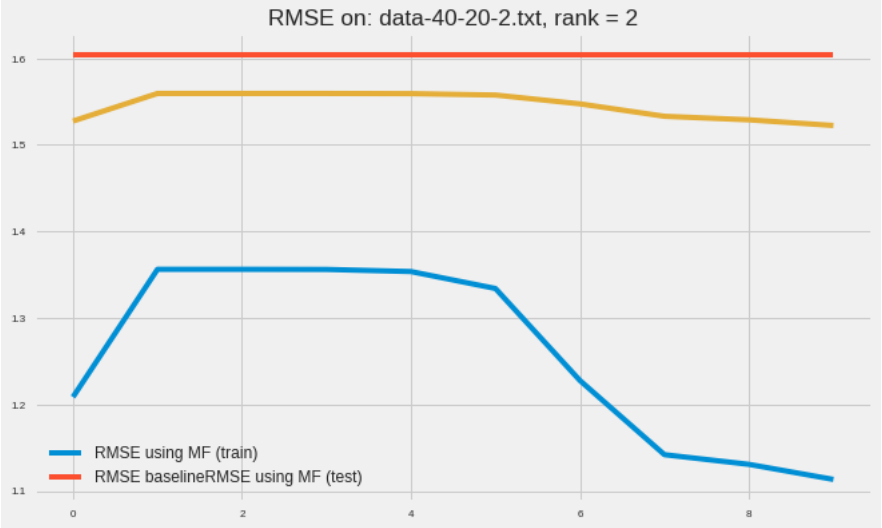

Figure 1: Plot of the results

1.2 Results - dataset-500-500-3.txt

Variational EM algorithm based on probabilistic matrix factorization is run for dataset-500-500-3.txt for following values of rank and results were captured to analyze the performance of the algorithm

1. rank = 3, table: 2 , figure: 2
2. rank = 1, table: 3, figure: 3
3. rank = 2, table: 4, figure: 4
4. rank = 5, table: 5, figure: 5
5. rank = 10, table: 6, figure: 6
6. rank = 20, table: 7, figure: 7

In addition, to above results, also captured comparison of RMSE values on test set for various values of rank over iterations of the algorithm. This is shown in the figure 9

RMSE Baseline	1.8086
RMSE train	1.7948
RMSE test	1.8021

Table 2: RMSE values after 18 iterations - data-500-500-3.txt, rank = 3

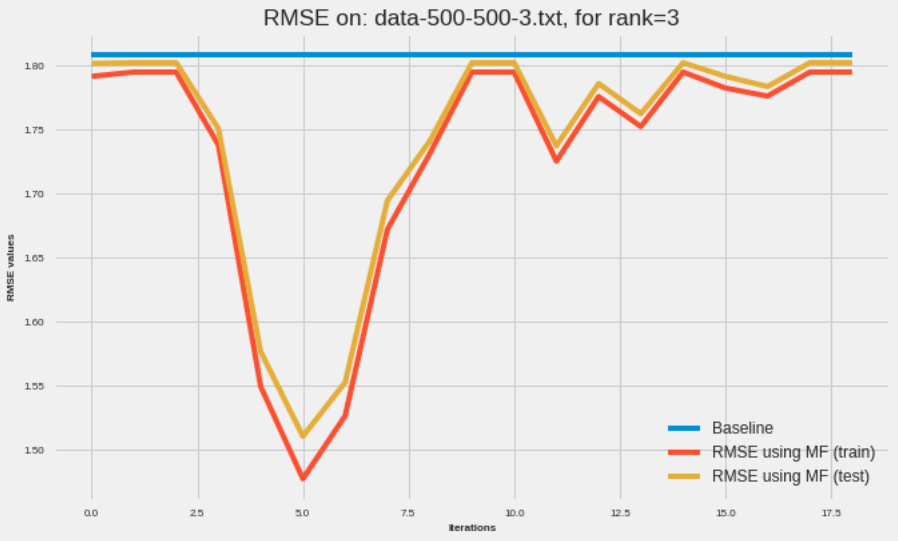

Figure 2: Plot of the results

RMSE Baseline	1.80
RMSE train	1.79
RMSE test	1.79

Table 3: RMSE values after 100 iterations - data-500-500-3.txt, rank = 1

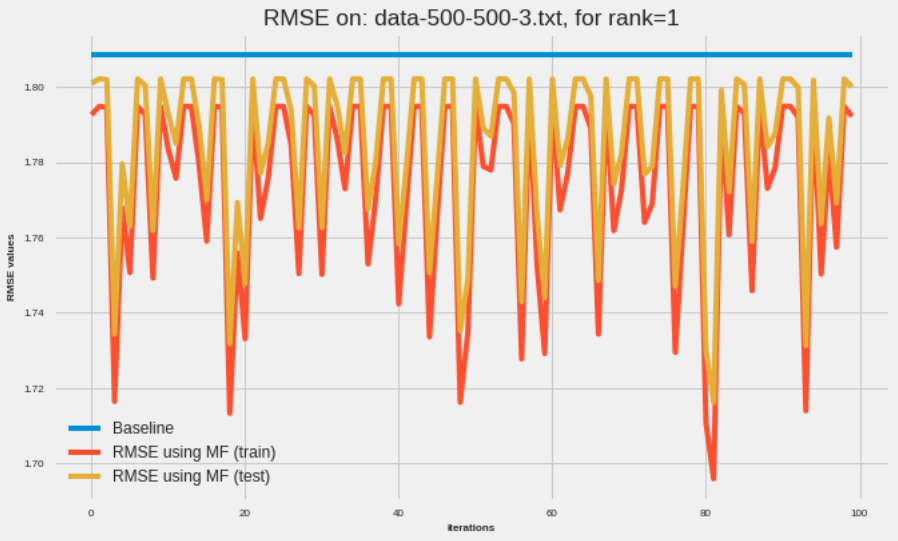

Figure 3: Plot of the results

RMSE Baseline	1.80
RMSE train	1.75
RMSE test	1.76

Table 4: RMSE values after 100 iterations - data-500-500-3.txt, rank = 2

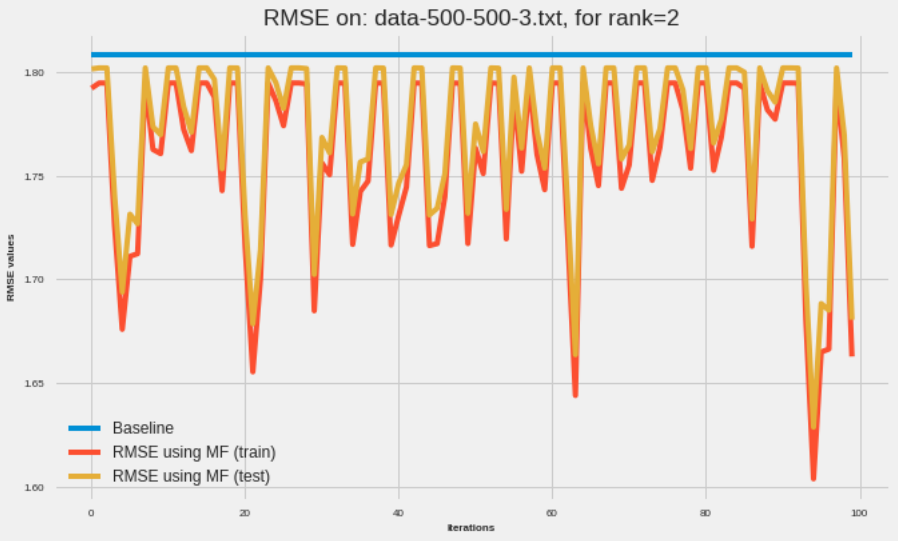

Figure 4: Plot of the results

RMSE Baseline	1.8086
RMSE train	1.7948
RMSE test	1.8021

Table 5: RMSE values after 67 iterations - data-500-500-3.txt, rank = 5

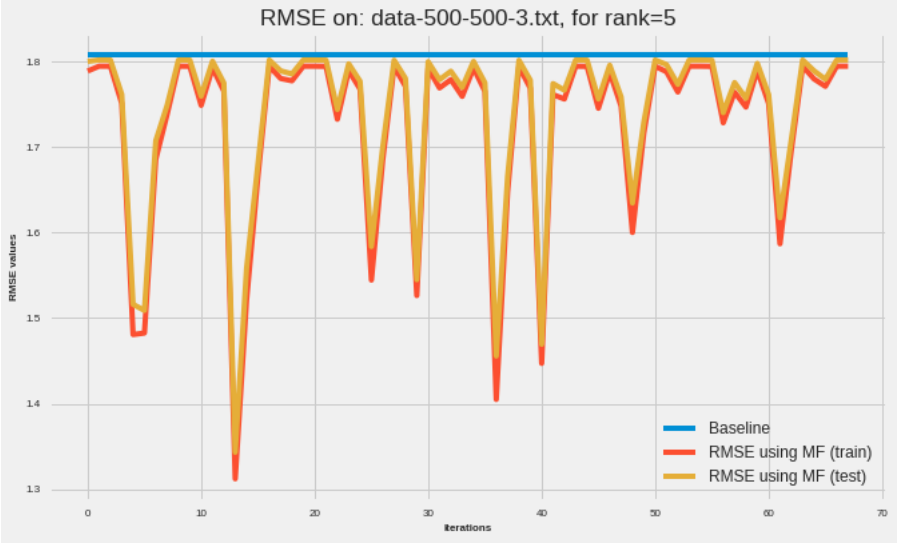

Figure 5: Plot of the results

RMSE Baseline	1.8086
RMSE train	1.7948
RMSE test	1.8021

Table 6: RMSE values after 32 iterations - data-500-500-3.txt, rank = 10

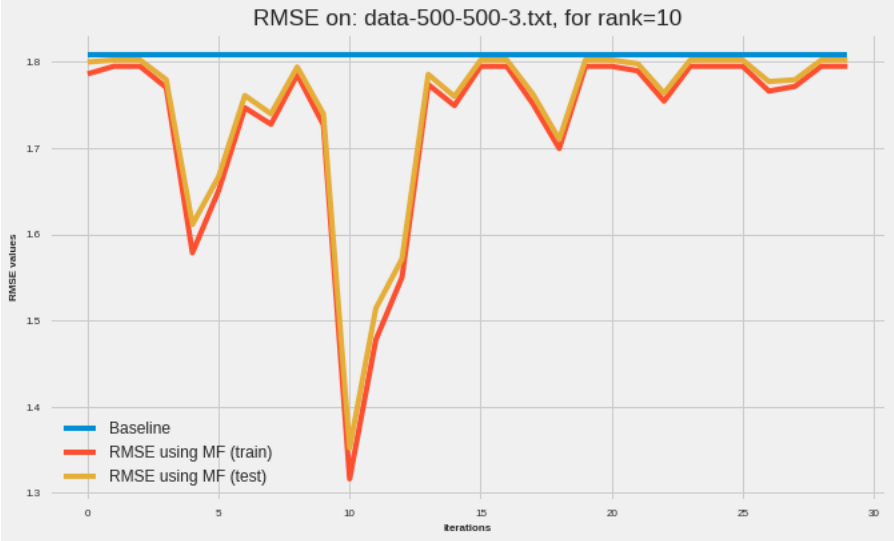

Figure 6: Plot of the results

RMSE Baseline	1.8086
RMSE train	1.7533
RMSE test	1.7673

Table 7: RMSE values after 100 iterations - data-500-500-3.txt, rank = 20

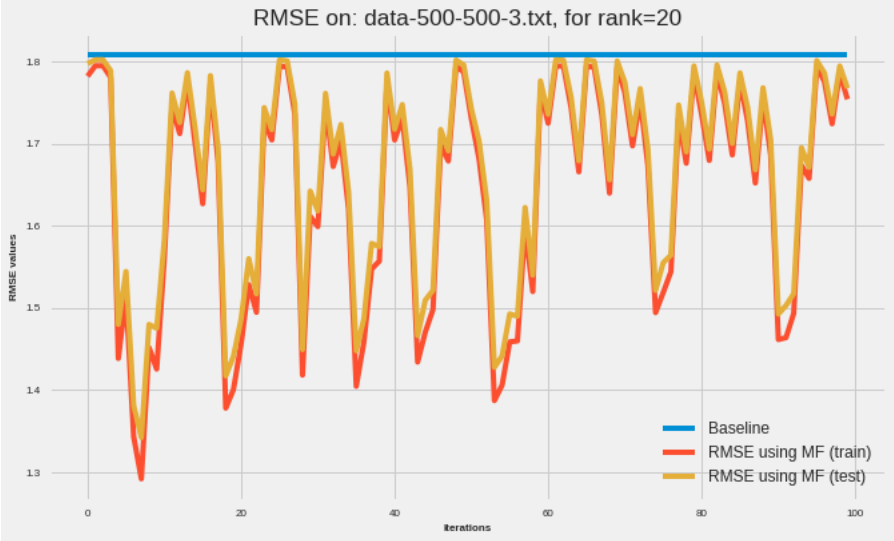

Figure 7: Plot of the results

Rank	Iteration (Convergence)
1	100 iterations (did not converge)
2	100 iterations (did not converged)
3	18 iterations (converged)
5	67 iterations (converged)
10	32 iterations (converged)
20	100 iterations (did not converge)

Table 8: Convergence of the algorithm - various values of the rank on the dataset - data-500-500-3.txt

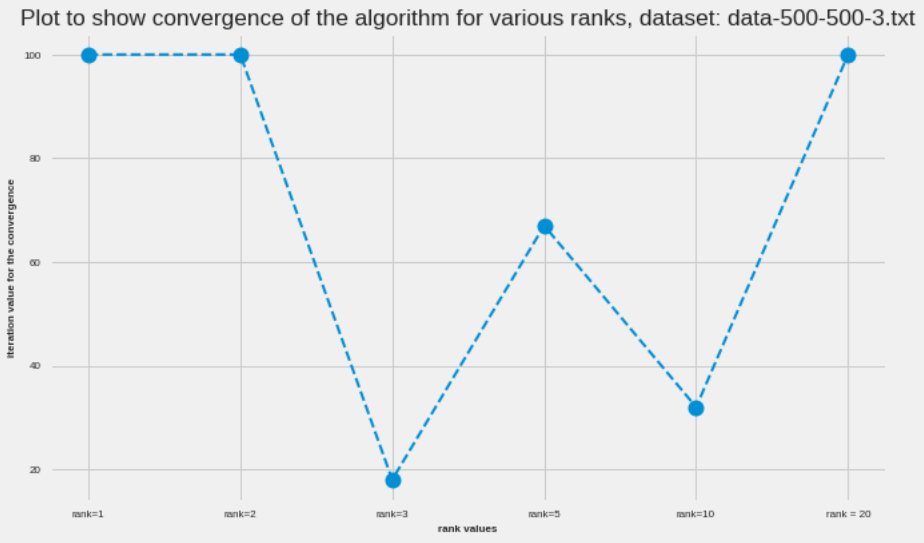

Figure 8: Convergence of the algorithm - various values of the rank on the dataset: data-500-500-3.txt

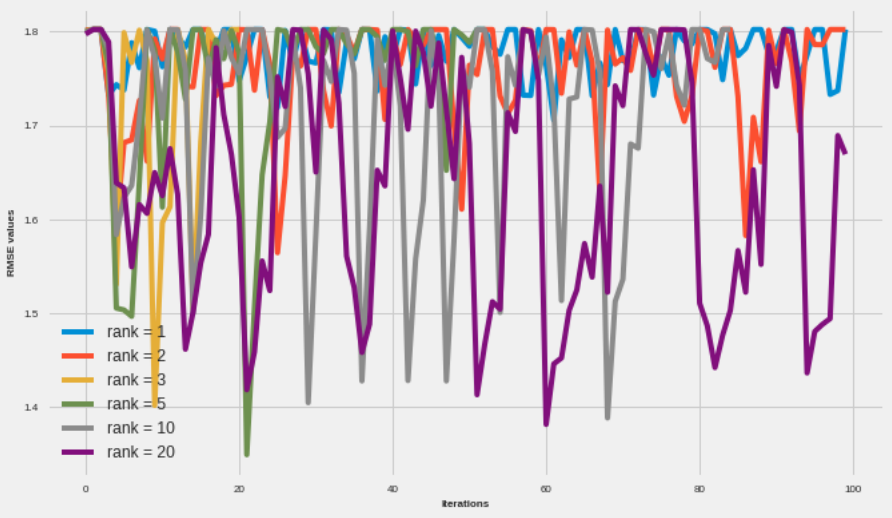

Figure 9: Performance for various ranks - dataset 500-500-3.txt

1.2.1 Convergence of the Algorithm

Algorithm converged for certain values of rank like 3, 5, and 10 and did not converge for other rank values = 1, 2, 20. The table 8 shows the iteration value at which algorithm converged and figure 8 convergence plot for the results

1.2.2 Observations

The point of convergence of the algorithm is different for different values of rank and also noticed slightly different on each run (due to the sampled data). The rmse of the matrix factorization algorithm was very close to the baseline prediction. Run time varied of the algorithm as a function of the rank.

RMSE Baseline	0.9877
RMSE train	3.56
RMSE test	3.561

Table 9: RMSE values after 100 iterations - ml-100k.txt, rank = 5

1.3 Results - ml-100k.txt

Variational EM algorithm based on probabilistic matrix factorization is run for ml-100k.txt dataset for the following values of rank and results were captured to and described below:

1. rank = 5, table: 9 , figure: 10
2. rank = 20, table: 10, figure: 11
3. rank = 100, table: 11, figure: 12

In addition, to above results, also captured comparison of RMSE values on test set for various values of rank over iterations of the algorithm. This is shown in the figure 14

1.3.1 Convergence of the Algorithm

The convergence of the algorithm for different values of the rank is illustrated in the table: 12, and figure: 13

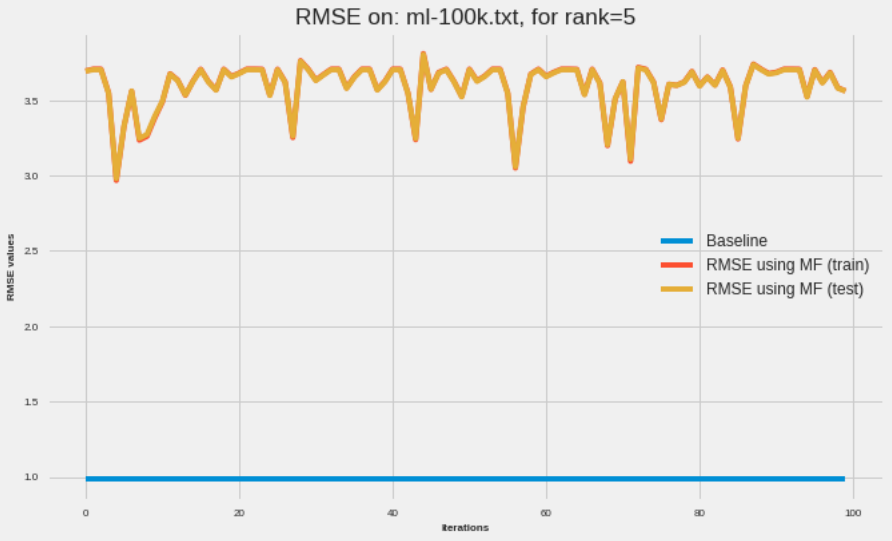

Figure 10: Plot of the results

1.3.2 Observations

The point of convergence of the algorithm is different for different values of rank and also noticed slightly different on each run (due to the sampled data). The rmse of the matrix

RMSE Baseline	0.9877
RMSE train	3.5021
RMSE test	3.50

Table 10: RMSE values after 100 iterations - ml-100k.txt, rank = 20

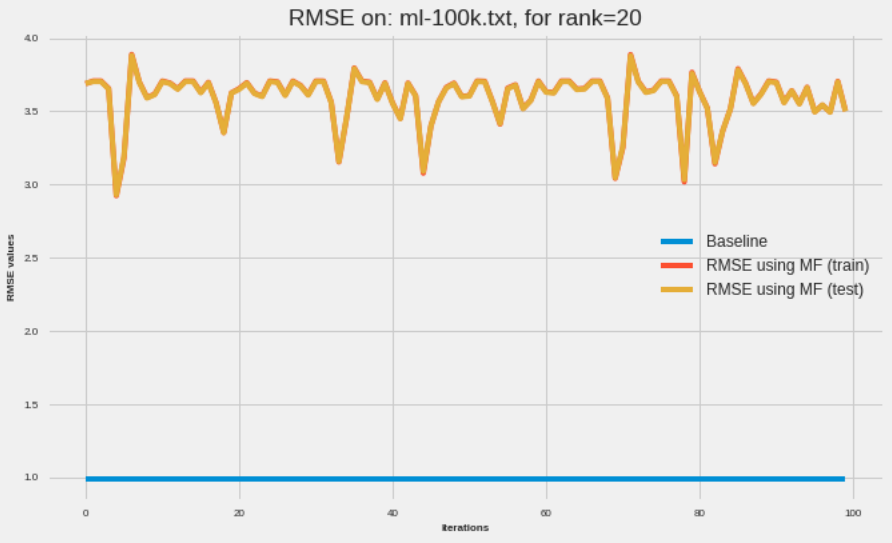

Figure 11: Plot of the results

factorization algorithm was very far from the baseline prediction. The reason for this might be from initialization of u and v matrices. Run time varied of the algorithm as a function of the rank.

RMSE Baseline	0.9877
RMSE train	3.70
RMSE test	3.702

Table 11: RMSE values after 7 iterations - ml-100k.txt, rank = 100

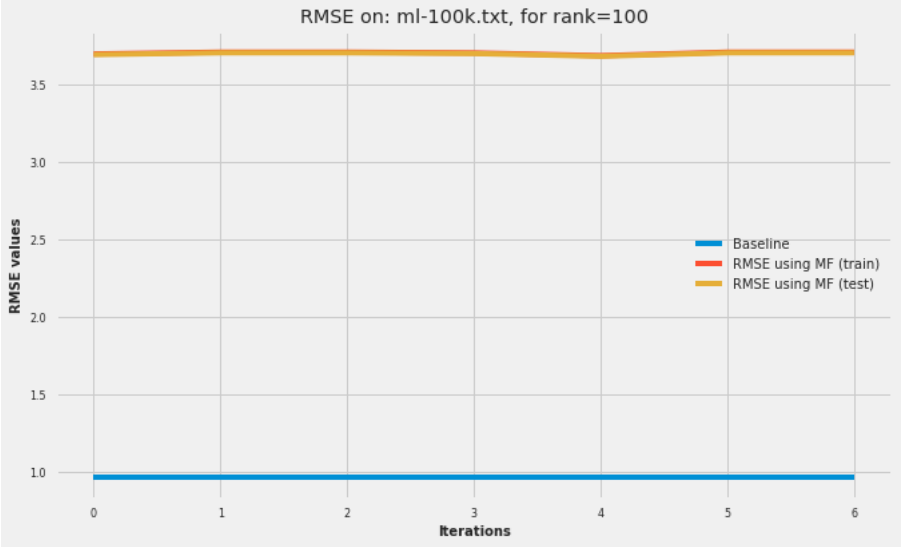

Figure 12: Plot of the results

Rank	Iteration (Convergence)
5	100 iterations (did not converge)
20	100 iterations (did not converge)
100	7 iterations (converged)

Table 12: Convergence for the algorithm - various rank values on the dataset - ml-100k.txt

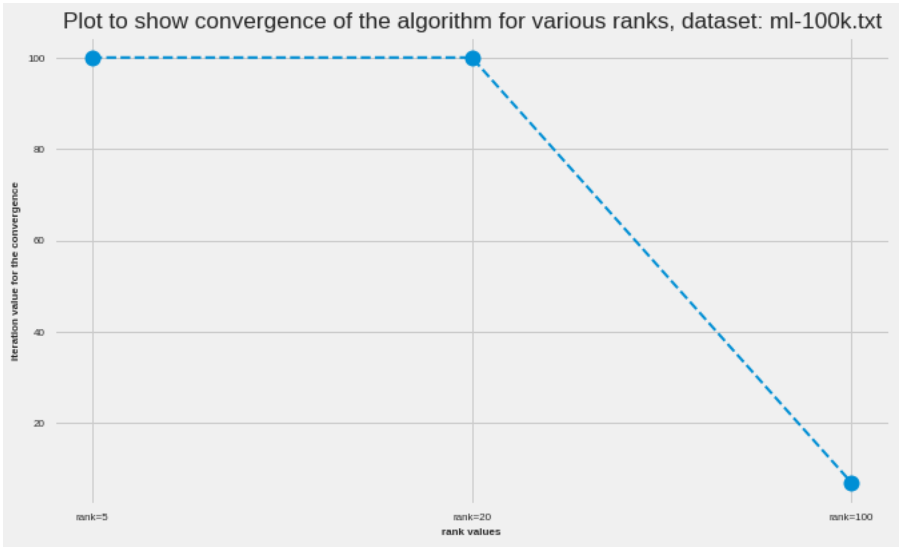

Figure 13: Convergence of the algorithm - various rank values on the dataset: ml-100k.txt

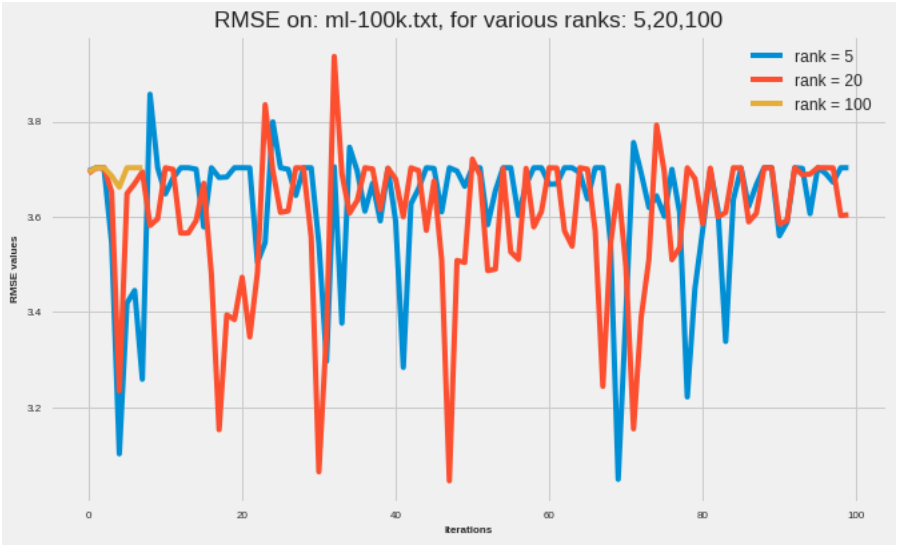

Figure 14: Performance for various ranks - ml-100k dataset

References

[1] Yew Jin Lim. “Variational Bayesian Approach to Movie Rating Prediction”. In: 2007.