MA614 Mini Project
Matrix Completion via SVD in Recommendation Systems

Ranjeet Singh - 2021EEB1203

May 14, 2025

Abstract

This project investigates the use of Singular Value Decomposition (SVD) for matrix completion in recommendation systems, specifically for the MovieLens 100k dataset. We implement truncated SVD from scratch, use it for matrix completion, and evaluate its performance by comparing the Root Mean Squared Error (RMSE) for various values of the number of singular values k. Heatmaps are generated to visually compare the original and completed matrices for different values of k.

1 Introduction

Recommendation systems are a crucial component of many online platforms such as Netflix, helping users discover products, movies, and services. Matrix factorization techniques, including Singular Value Decomposition (SVD), are commonly used for collaborative filtering in recommendation systems. This project explores the application of truncated SVD to perform matrix completion on the MovieLens 100k dataset.

2 Problem Definition

The problem of matrix completion involves predicting missing values in a matrix, given partial observations. In the context of recommendation systems,

the matrix represents user-item interactions (e.g., movie ratings by users) where the rows represent the users, columns represent movie and the values represent the rating given by the user to the particular movie. Since a user will have a ratings for only a small number of movies, the goal of this project is to use SVD to complete the matrix by predicting the missing ratings to help us determine if that movie should be recommended or not.

3 Methodology

3.1 Singular Value Decomposition (SVD)

SVD is a matrix factorization technique that decomposes a matrix A into three matrices:

$$A = U\Sigma V^T$$

where:

- U is an $m \times k$ matrix of left singular vectors,
- Σ is a $k \times k$ diagonal matrix of singular values,
- V^T is a $k \times n$ matrix of right singular vectors.

In the context of matrix completion, we use truncated SVD, which approximates the original matrix A using only the top k singular values and vectors. This allows us to reconstruct a low-rank approximation of A and predict missing values.

3.2 Truncated SVD Implementation

We implemented truncated SVD from scratch. The process includes:

The matrix is then reconstructed by multiplying U, Σ, and V^T to predict the missing entries and adding the mean evaluated above.

Algorithm 1 Matrix Completion using Truncated SVD (Handling Missing Values)

Require: Incomplete user-item matrix $R \in \mathbb{R}^{m \times n}$ with missing entries as NaN, Target rank k

Ensure: Completed matrix $\hat{R}$

Let M be a boolean mask of observed entries: $M_{ij} = 1$ if R_{ij} is observed, else 0

Normalize the User wise data and obtain R_c, keeping NaNs

Replace NaNs in R_c with 0: $R_c[\text{isnan}] \leftarrow 0$

Compute $A = R_c^T R_c$

Perform eigen decomposition of A: $A = V \Lambda V^T$

Sort top-k eigenvalues and corresponding eigenvectors

Compute singular values: $\Sigma = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$

Form matrix V_k of top-k eigenvectors

Compute $U_k = R_c V_k \Sigma^{-1}$

Reconstruct centered matrix: $\hat{R}_c = U_k \Sigma V_k^T$

Denormalize the data

3.3 RMSE Evaluation

To evaluate the performance of the matrix completion, we used the Root Mean Squared Error (RMSE) metric:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R_{ij} - \hat{R}_{ij})^2}$$

where R_{ij} are the observed ratings and $\hat{R}_{ij}$ are the predicted ratings.

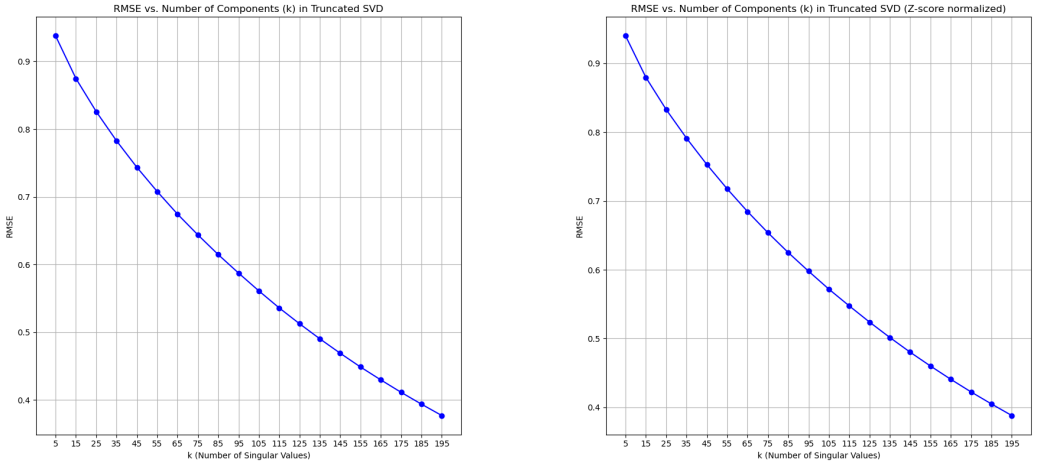
3.4 Dataset

The MovieLens 100k dataset contains 100,000 ratings of 1,682 movies by 943 users. The dataset is sparse, with many missing values.

4 Results

4.1 RMSE Comparison for Different Values of k

The RMSE scores for different values of k were computed to evaluate the accuracy of matrix completion. As shown in the following plot, the RMSE generally decreases as k increases, but at some point, the improvement becomes marginal. Therefore, there is a tradeoff between the accuracy of the predictions and the complexity of the model. In Fig. 4, we can observe the RMSE vs k plots for both mean and z-score normalized data. We observe that the normalization technique does not affect the convergence of RMSE score.

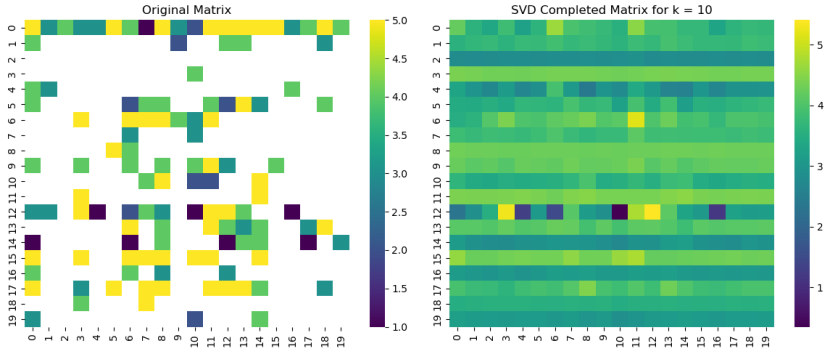


(a) Mean Normalized (b) Z-score Normalized

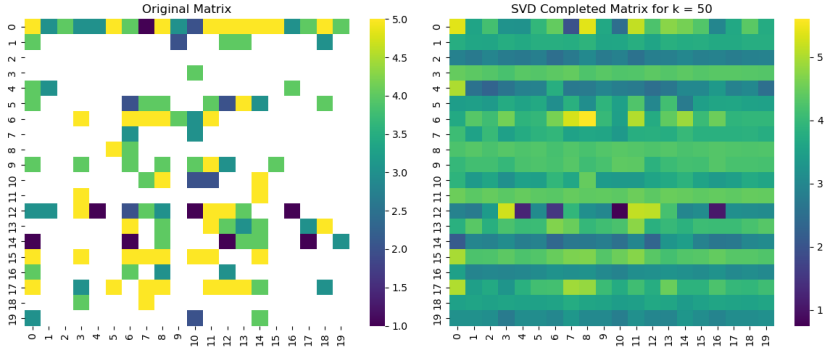
Figure 1: RMSE scores for different values of k .

4.2 Matrix Completion and Heatmap Comparison

We performed matrix completion using truncated SVD for different values of k. The heatmaps below illustrate the subset of original user-item matrix and the corresponding completed matrix for different values of k. As k increases, the completed matrix becomes more accurate in approximating the original matrix. In this case, we have mean normalized the user rating matrix.



(a) Original vs. Completed Matrix for $k = 10$ RMSE : 0.902



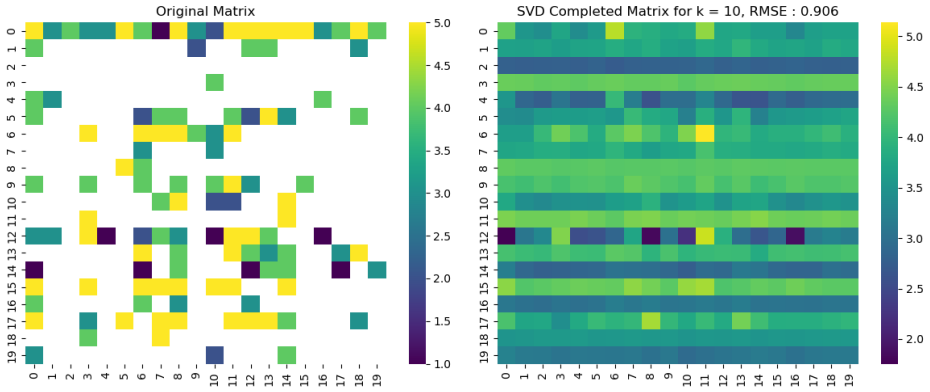
(b) Original vs. Completed Matrix for $k = 50$ RMSE : 0.725

Figure 2: Heatmaps of the original and completed matrices for different values of k using Mean normalization.

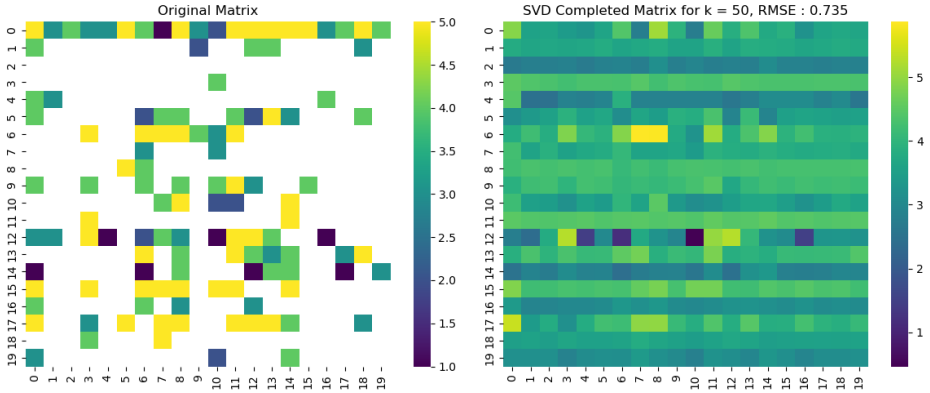
We can observe from the above two heatmaps that increasing the value of k, increases the granularity of the heatmap as it is using more singular values to predict the ratings.

Now we will use Z-score normalization and obtain the same heatmaps for $k = 10$ and $k = 50$.

We can also compare the heatmaps of both these normalization techniques side by side for $k = 50$ in Figure 4. We observe that in case of Z-score normalized heatmap, the predicted values have become less extreme though the change is barely visible for majority of the heatmap. Z-score normalization aims to reduce the effect of outliers to some extent which could be the extreme ratings given by some users. However, this won't account for the noisy ratings given by the users.



(a) Original vs. Completed Matrix for $k = 10$



(b) Original vs. Completed Matrix for $k = 50$

Figure 3: Heatmaps of the original and completed matrices for different values of k using Z score Normalization.

5 Limitations and Improvements

- Truncated SVD considers the user rating matrix as low rank and data relationship as linear. However, real-user preferences can be non-linear and may not conform to low rank model leading to inaccurate predictions.

- This implementation is not robust against outliers and is not able to segregate between actual and non-serious ratings from users. This may result in wrong recommendations to the users which were not intended to them. We may use Robust PCA to solve this problem.

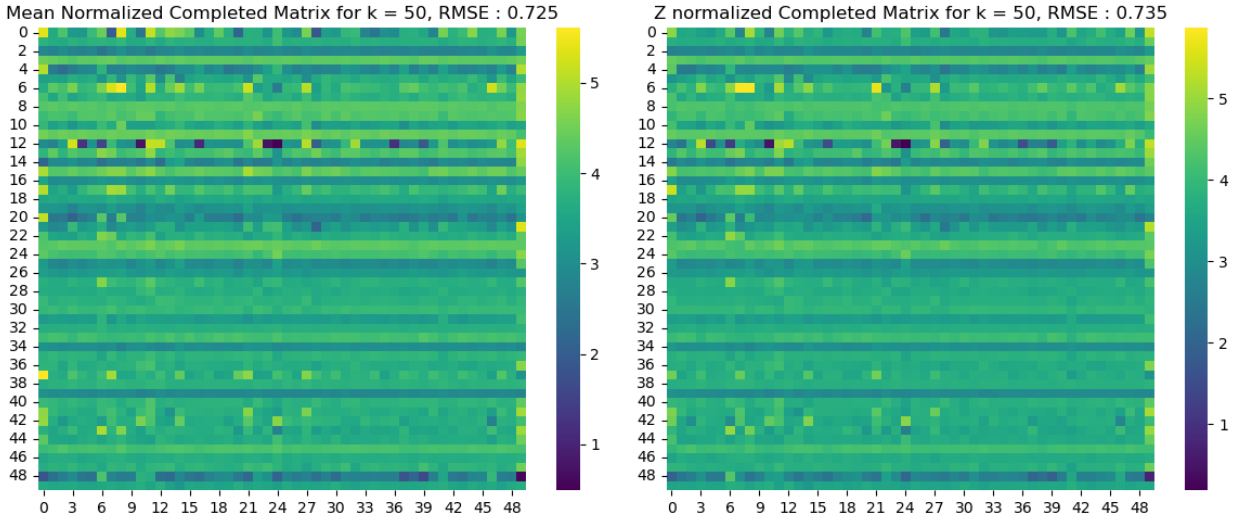


Figure 4: Heatmaps of Completed Matrices for Mean normalized and Z-score normalized data

- Since we are working only on MovieLens 100k dataset, this method may not work against larger datasets as it is computationally and memory intensive.

6 Conclusion

In this project, we applied truncated SVD for matrix completion on the MovieLens 100k dataset. We implemented the algorithm from scratch, evaluated its performance using RMSE, and analyzed the effect of different values of k and normalization techniques. The results show that truncated SVD is effective for predicting missing entries for small user-item matrix, and a suitable k value balances between accuracy and computational cost.