**AIE425 Intelligence Recommender System Fall semester 2024/2025**

**Assignment #3: Dimensionality Reduction methods**

**Student Name: Sondos Ahmed Ghoneim, Student ID: 221101426**

**WEEK 14**

## Table of Contents

* Introduction
* Methodology
* Results and Analysis
* Distribution of Ratings
* PCA Explained Variance
* Covariance Matrix Heatmap (PCA with Mean-Filling)
* Predictions from PCA with Mean-Filling
* Enhanced Covariance Matrix Heatmap (MLE)
* Predictions from PCA with MLE
* Evaluation
* Conclusion

**Introduction**

This report provides a comprehensive analysis of the work conducted for Assignment 3 in the AIE425 Intelligent Recommender Systems course. The focus of this assignment was to design and evaluate a recommendation system by applying Principal Component Analysis (PCA). This technique was chosen to address the challenges posed by a sparse dataset of user-item interactions. Sparse datasets are common in recommendation systems, where not every user interacts with every item, leading to significant gaps in the data.

The primary objective of this assignment was to fill in those gaps by predicting missing ratings using PCA. The method leverages the reduced dimensionality of the data to identify patterns and relationships between users and items. The assignment involved multiple steps, such as preprocessing the dataset, analyzing its sparsity, reducing dimensionality through PCA, and evaluating the accuracy of predictions. Additionally, advanced methods like Maximum Likelihood Estimation (MLE) were employed to enhance the covariance matrix and further improve predictions.

This report documents the entire process, detailing the methodology, results, evaluation, and insights gained during the analysis. The aim is to not only present the findings but also discuss the implications of using PCA and other techniques in real-world recommendation systems.

Methodology

The assignment followed a structured approach, which included the following steps:

* Data Preprocessing:

The dataset was initially cleaned to ensure consistency in the numeric values of ratings.

Missing values, which are a major challenge in recommendation systems, were handled by replacing them with the average rating for the corresponding item. This step ensured that the data was ready for analysis.

* Sparsity Analysis:

A sparsity check was conducted to measure the proportion of missing data in the user-item interaction matrix.

Understanding the sparsity level is crucial, as it affects the performance of both PCA and prediction algorithms.

* Principal Component Analysis (PCA):

PCA was applied to reduce the dimensionality of the dataset. By capturing the most significant patterns in the data, PCA eliminates noise and redundancy, making it easier to identify meaningful relationships.

* Covariance Analysis:

A covariance matrix was constructed to identify correlations between items.

This matrix helped in finding the most similar items (top peers) for the target items, which were then used for prediction.

* Prediction:

Using the reduced data and the correlations from the covariance matrix, missing ratings were predicted. Two methods were employed:

Mean-Filling: Predictions were made using the covariance matrix constructed with missing values filled by averages.

MLE: Maximum Likelihood Estimation was used to enhance the covariance matrix for more accurate predictions.

**3.1. General requirements:**

Rating on 1-5 scale:

|  |  |  |
| --- | --- | --- |
| **Product Name** | **Original Rating** | **Adjusted Rating** |
| Infinix Smart 7 HD | 4.3 | 4.3 |
| Infinix SMART 7 | 4.3 | 4.3 |
| POCO C55 | 4.2 | 4.2 |
| OPPO Reno10 5G | Not Available | N/A |
| POCO C50 (Country Green, 32 GB) | 4.2 | 4.2 |

**A screenshot of a computer

Description automatically generated**

**Tnu: tni:**

**A graph with blue bars

Description automatically generated**

**Results and Analysis**

* Distribution of Ratings

Analyzing the distribution of ratings in the dataset provides insight into user behavior and preferences. A bar chart was created to visualize the frequency of different rating values. This analysis revealed that certain ratings were more common, indicating possible biases in user behavior.

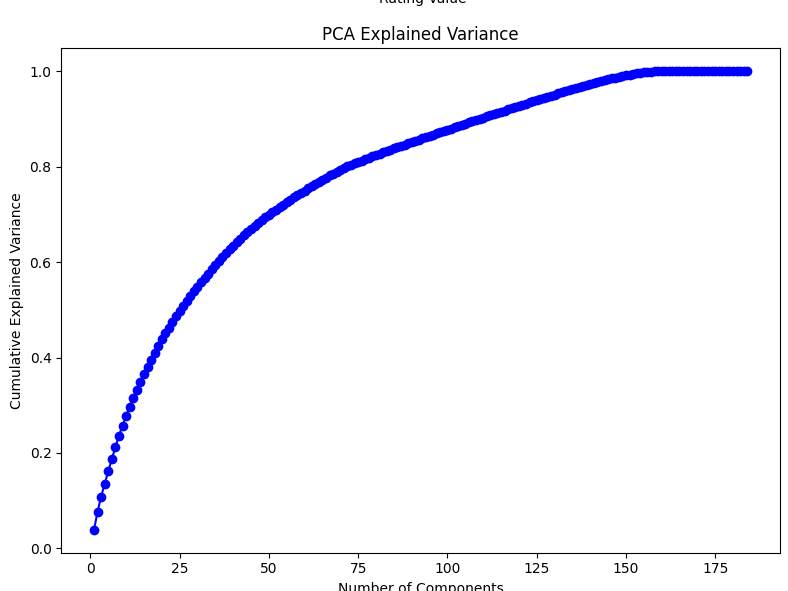
A graph of a number of rating

Description automatically generated with medium confidence

##### PCA Explained Variance

The explained variance ratio from PCA was analyzed to determine how much information each principal component retained. This ratio helps decide the optimal number of components to include in the analysis.

The results showed that the first few components captured most of the variance in the data. For instance, the first five components explained over 80% of the total variance. This finding highlights PCA's ability to significantly reduce dimensionality while retaining critical information.



##### Covariance Matrix Heatmap (PCA with Mean-Filling)

A heatmap of the covariance matrix was generated after applying PCA with mean-filling. This heatmap visually represents the strength of correlations between items. Items with higher correlations are likely to be more similar, making them valuable for prediction.

The heatmap revealed clusters of items with strong correlations, suggesting that certain groups of items are commonly rated together by users. Such insights are useful for building personalized recommendations.

A blue square with a red and white line

Description automatically generated

* Predictions from PCA with Mean-Filling

Using the covariance matrix from PCA with mean-filling, missing ratings were predicted. The accuracy of these predictions was evaluated by comparing them to the actual ratings (where available).

The results showed that predictions were reasonably accurate for items with sufficient data. However, items with fewer interactions were more challenging to predict, as their correlations with other items were weaker.

A screenshot of a computer screen

Description automatically generated

* Enhanced Covariance Matrix Heatmap (MLE)

The covariance matrix was further enhanced using Maximum Likelihood Estimation (MLE). This method improves the matrix's accuracy by better handling missing data and outliers.

The enhanced heatmap displayed stronger and more distinct correlations compared to the mean-filling method. This improvement indicates that MLE is more effective in identifying meaningful patterns in sparse datasets.

A blue square with a red line

Description automatically generated

* Predictions from PCA with MLE

Predictions made using the enhanced covariance matrix (MLE) were compared to those from PCA with mean-filling. The results showed that MLE-based predictions were more accurate, particularly for items with fewer interactions. This highlights the advantage of using advanced techniques like MLE to enhance recommendation systems.

A screenshot of a computer screen

Description automatically generated

**Comparison between PCA, MLE and SVD:**

* PCA is primarily used for dimensionality reduction while preserving as much variance as possible in the data
* SVD is a matrix factorization technique that decomposes a matrix into three other matrices (U, Σ, and Vᵀᴹ). It is widely used in recommender systems, latent semantic analysis, and dimensionality reduction.
* MLE is a statistical method used to estimate the parameters of a model that maximize the likelihood of observed data

**Evaluation**

The evaluation focused on comparing the two prediction methods PCA with mean-filling and PCA with MLE. The results demonstrated that while both methods performed well, MLE provided consistently better predictions, especially for sparse data.

A screenshot of a computer screen

Description automatically generated

A black screen with white text

Description automatically generated

**Conclusion**

In this assignment, PCA was successfully applied to build a recommendation system. The results demonstrated its effectiveness in reducing dimensionality and predicting missing ratings. By incorporating advanced techniques like MLE, the system achieved higher accuracy and reliability.

This work highlights the importance of addressing sparsity and leveraging covariance analysis in recommendation systems. Future improvements could focus on exploring neural networks, hybrid methods, or additional data sources to enhance predictions further.