**AIE425 Intelligent Recommender Systems, Fall Semester 24/25**  
**Assignment #3: Dimensionality Reduction methods**

**Student ID:** A20001136  
**Full Name:** Abdelrahman El-Sayyad

**Table of Contents**

1. [**Introduction**](#introduction)
2. [**Dataset and Preprocessing**](#dataset-and-preprocessing)
3. [**Methodology**](#methodology)  
   3.1. [Part 1 — PCA with Mean-Filling](#part-1--pca-with-mean-filling)  
   3.2. [Part 2 — PCA with Maximum Likelihood Estimation (MLE)](#part-2--pca-with-maximum-likelihood-est)  
   3.3. [Part 3 — Singular Value Decomposition (SVD)](#part-3--singular-value-decomposition-sv)
4. [**Results**](#results)  
   4.1. [Part 1 (PCA + Mean-Filling)](#part-1-pca--mean-filling-results)  
   4.2. [Part 2 (PCA + MLE)](#part-2-pca--mle-results)  
   4.3. [Part 3 (SVD)](#part-3-svd-results)
5. [**Summary and Comparison**](#summary-and-comparison)
6. [**Conclusion**](#conclusion)
7. [**References**](#references-optional)

**1. Introduction**

Matrix factorization and dimensionality reduction techniques are at the core of many recommendation systems, data compression tasks, and pattern recognition problems. In this assignment, we perform three different methods of dimensionality reduction—namely **PCA (Principal Component Analysis) with mean-filling**, **PCA with Maximum Likelihood Estimation (MLE)**, and **Singular Value Decomposition (SVD)**—on a dataset previously scraped in Assignment #1.

The goal of this assignment is to:

1. Identify and handle missing values in a ratings matrix.
2. Determine how each method predicts missing ratings for two specifically chosen items (the lowest-rated items in the dataset).
3. Compare the results from these three methods with a focus on accuracy, computational considerations, and practical implications.

**2. Dataset and Preprocessing**

The dataset, **modified\_dataset.csv**, was generated from Assignment #1 (and refined in Assignment #2). It has the following characteristics:

* **Number of Users (Tnu)**: 50
* **Number of Items (Tni)**: 138
* **Rating Scale**: 1 to 5
* **Missingness**: Approximately 10% of the entries are missing due to intentional random removal in Assignment #2.

A quick analysis revealed:

* The **matrix sparsity** (ratio of missing entries) is about **10.22%**.
* The **average rating** across all items is approximately **2.975**, indicating a mid-range bias (slightly below the midpoint of 3).
* The two **lowest-rated items** (based on average rating) are **“aliteracy”** (I1) and **“understanding”** (I2).

These two items (I1, I2) become the focal points for predicting missing ratings.

**3. Methodology**

**3.1. Part 1 — PCA with Mean-Filling**

1. **Identifying Target Items**  
   We computed the mean rating of each item. The two items with the lowest means were identified as **I1** = aliteracy and **I2** = understanding.
2. **Mean-Filling for Target Items**
   * We specifically replaced missing entries in I1 and I2 with their **respective mean rating**.
   * We then computed the **covariance matrix** for the entire rating matrix (item–item covariance) after mean-filling those two columns.
3. **Top-k Peers and Prediction**
   * For each target item, we identified the **top-5** and **top-10** peer items based on the **highest absolute covariance**.
   * To predict a user’s missing rating for I1 (or I2), we took a **simple average** of the ratings from those top-k peer items, ignoring any that are missing.
4. **Comparison**
   * We compared **top-5** and **top-10** predictions to observe differences.

**3.2. Part 2 — PCA with Maximum Likelihood Estimation (MLE)**

1. **MLE-Based Covariance**  
   Instead of globally mean-filling, we computed the covariance between two items **only over the users** who rated both. If no users overlapped, covariance was set to zero. This mimics a **Maximum Likelihood Estimate** of covariance for incomplete data.
2. **Top-k Peers and Prediction**
   * Similar to Part 1, we derived **top-5** and **top-10** peer items for each target item from the new MLE covariance matrix.
   * Missing ratings for I1 or I2 were predicted by averaging the user’s ratings on those top peers.
3. **Comparison**
   * We compared **top-5 vs. top-10** under the MLE covariance.
   * We also compared these MLE-based predictions to the **mean-fill PCA** predictions from Part 1.

**3.3. Part 3 — Singular Value Decomposition (SVD)**

1. **Global Mean-Filling**  
   For SVD, we filled **all missing ratings** in the matrix by each item’s mean rating.
2. **SVD Decomposition**
   * We performed via np.linalg.svd(...).
   * We chose a **truncation rank** k=10k=10k=10 (or the maximum if the dimension was smaller).
3. **Rating Reconstruction**
   * We formed ​.
   * Any originally missing rating in I1 or I2 was replaced by its corresponding entry in Ȓ .

**4. Results**

**4.1. Part 1 (PCA + Mean-Filling) Results**

* **Top-5** vs. **Top-10**  
  For aliteracy, some predicted values (for the 5 users who missed it) changed notably when we increased from 5 to 10 peers (e.g., from **3.20** to **2.75**).  
  Similarly, understanding predictions shifted for certain users (e.g., from **4.00** down to **3.00** for user 17).
* **Observation**  
  Increasing peer count sometimes **decreases** or **increases** the rating estimate, reflecting the item correlations in the covariance matrix.

**4.2. Part 2 (PCA + MLE) Results**

* **MLE Covariance**  
  The top peers for aliteracy and understanding overlapped somewhat with Part 1, but also revealed differences (e.g., inspirational appeared among aliteracy’s top-5 in MLE but was absent in the Part 1 approach).
* **Top-5 vs. Top-10**  
  We again noted subtle changes in predicted values for some users, but in a few cases, the top-10 predictions matched or nearly matched the top-5 predictions.
* **Comparisons with Mean-Filling**
  + For some users, the MLE-based predictions were **0.0** to **0.4** points higher or lower than the mean-fill PCA approach.
  + In many top-10 scenarios, interestingly, both approaches agreed, because the same set of overlapping users or items influenced the predictions.

**4.3. Part 3 (SVD) Results**

* After **global mean-filling** for **all** items, the **SVD** predicted missing values for aliteracy and understanding quite differently.
* For instance, user 13’s rating for aliteracy was around **1.785** via SVD, whereas the PCA-based methods often gave values above **2.0**.
* The discrepancy arises because SVD captures **latent factors** that do not always align with simple covariance-based item similarities.

**5. Summary and Comparison**

(Per Assignment **Section 3.5**)

1. **Accuracy**
   * **Part 1 (Mean-Filling PCA)** relies on a simplified covariance matrix that partially skews the relationship among items when missing data exist. Its predictions can be reasonable but sometimes overfit or underfit if the mean-filled values distort correlations.
   * **Part 2 (MLE PCA)** refines the covariance estimates by **excluding** non-overlapping users for each item pair. This often yields more “accurate” item correlations, as it does not assume rating values for missing entries. However, in very sparse areas of the matrix, it might produce **0** covariance if no overlap exists.
   * **Part 3 (SVD)** uses a **global factorization** to capture latent structure. This method can yield quite different predictions, especially if the data contains strong hidden factors. SVD often performs well in recommendation contexts but might underperform if the global mean-filling overly distorts the rating pattern.
2. **Pros and Cons**
   * **Mean-Filling PCA**
     + **Pros**: Simple to implement, straightforward to interpret.
     + **Cons**: Potentially **biased** by the mean substitution, might inflate correlations artificially.
   * **MLE PCA**
     + **Pros**: More **statistically principled** for incomplete data, uses actual co-ratings only.
     + **Cons**: If items share few overlapping users, covariance can become **zero**, leading to less stable predictions in extremely sparse segments.
   * **SVD**
     + **Pros**: Captures **latent factors** across the entire matrix, commonly used in advanced recommender systems, may generalize well.
     + **Cons**: Requires global mean-filling (possibly introducing bias) and can be computationally intensive for large matrices.
3. **Overall Observations**
   * For many users, the top-5 and top-10 predictions are close, with differences typically in the **0.2–0.6** range for certain users.
   * MLE-based PCA often shows subtle improvements over mean-filling PCA (if enough overlapping users exist).
   * SVD yields a distinctly different approach, sometimes giving significantly **lower** or **higher** rating predictions for the same missing entries.

**6. Conclusion**

In conclusion, all three dimensionality reduction methods—**PCA with mean-filling**, **PCA with MLE**, and **SVD**—successfully predict missing ratings for the identified target items (aliteracy and understanding). However, they differ in how they handle missing data and in the assumptions they make:

* **PCA with Mean-Filling** is easy to implement but can embed biases via the mean substitution.
* **PCA with MLE** provides a more robust covariance estimation in many cases by restricting calculations to true co-ratings, avoiding artificial data.
* **SVD** (with truncated rank) captures global structure and often leads to more complex patterns in the reconstructed ratings, at the cost of additional computation and a large-scale mean-fill assumption.

Each method has its own advantages, limitations, and best-use scenarios. For extremely **sparse** datasets, MLE-based or factor-model approaches (like SVD) may be superior; for moderate missingness, even a straightforward mean-fill PCA can offer a quick, interpretable solution. Ultimately, method selection should consider **dataset size**, **missingness**, **desired accuracy**, and **computational resources**.