IRS

Statistical Analysis, Case Studies 1–3, and Comparative Discussion

Ahmed Alaa Eldin Mostafa
Student ID: 222100123

1 Introduction

This technical report presents a detailed analytical discussion of three user-based collaborative filtering (CF) approaches applied to a sparse ratings dataset. The document follows the required structure: statistical analysis, comments for each case study, and a final comparative reflection across all three approaches.

The focus is on exploring matrix sparsity, rating bias, long-tail effects, similarity distortions, and the impact of different similarity measures: Raw Cosine, Mean-Centered Cosine, and Pearson Correlation Coefficient (PCC).

2 Statistical Analysis: Comments on Points 13 & 14

This section provides insights into the dataset by discussing matrix sparsity, rating bias, and the long-tail problem. These observations help explain results in later case studies.

2.1 Matrix Sparsity

The dataset exhibits extremely high sparsity. Only a tiny percentage of all possible user–item interactions contain ratings. This sparsity leads to:

- Very few overlapping items between most user pairs.
- Unstable similarity scores due to limited co-rated items.
- A high risk of falsely inferring similarity from very little evidence.

As a result, even when cosine similarity is mathematically high, it may not represent real agreement. This observation directly explains many inconsistencies observed in Case Study 1.

2.2 Rating Bias

Users exhibit different rating behaviors:

- Some rate generously (mostly 4s and 5s).
- Some rate harshly (mostly 1s and 2s).
- Some only rate favorite items, leading to inflated averages.

Rating bias strongly affects similarity metrics. Raw cosine similarity does not correct for individual user tendencies, which causes:

- Generous users appearing artificially similar.
- Harsh users appearing dissimilar even if they agree in pattern.
- Users with a narrow rating range dominating similarity calculations.

This motivates the need for mean-centering (Case 2) or Pearson correlation (Case 3).

2.3 Long-Tail Problem

A small portion of items receive most of the ratings (popular items), while the majority receive very few. This leads to:

- Many users with non-overlapping item histories.
- Sparse, unreliable user–user similarity.
- Predictions dominated by frequent items.

The long-tail effect explains why DF (Damped Factor) and DS (Discounted Similarity) become necessary—these mechanisms down-weight unreliable similarity signals based on low overlap.

3 Case Study 1: Comments Section

3.1 Should a High Cosine Similarity Always Mean Strong Agreement?

No. A high cosine similarity does **not** always indicate strong agreement. Key issues:

- Cosine similarity does not consider rating scale differences.

- Users may appear similar due to rating only a few common items.
- Cosine treats zeros as meaningful information, even though zeros indicate missing data.

Therefore, cosine similarity alone can produce misleading signals in sparse datasets.

3.2 Comments for Case Study 1

- Raw cosine similarity overestimates similarity when overlap is small.
- Users who rate only popular items appear overly similar.
- No correction exists for user bias, generous raters distort similarity.
- Without DF/DS, similarity is dominated by noise.

4 Case Study 2: Comments Section

4.1 Unfair Closeness Between Users

When some users rate only their favorite items and others rate many items, they often appear unfairly close or unfairly far depending on the similarity measure.

- Mean-centering partially fixes rating bias, but not rating coverage.
- Users who rate only a few favorite items still produce exaggerated centered values.
- Users with a long rating history get more stable similarity, while short-rating users produce volatile similarity.

4.2 Comments for Case Study 2

- Mean-centering reduces generosity/harshness bias.
- Similarity improves because scale differences are removed.
- However, overlap issues still remain—DF and DS help mitigate them.
- Rating patterns, not absolute values, dominate similarity.

5 Case Study 3: Comments Section

5.1 Cases Where Pearson Detects Opposite Patterns Compared to Cosine

Pearson correlation may be negative even when cosine similarity was positive. This happens when:

- Users rate the same items but with opposite deviations from their own averages.
- Cosine sees only direction in rating space, not deviations from means.
- Pearson emphasizes centered agreement, not absolute rating values.

Pearson is more sensitive to rating tendencies, which can flip similarity signs.

5.2 Should We Trust Pearson When Overlap Is Small?

Generally, no. When fewer than 20% of items are co-rated:

- Pearson becomes unstable.
- The correlation may be artificially high or low.
- The sign of correlation may not reflect actual behavioral similarity.

5.3 Comments for Case Study 3

- Pearson is best for correcting rating bias.
- It is unreliable when overlap is extremely small.
- It can detect patterns missed by cosine, but may overreact to limited data.

6 Final Comparison Across All Case Studies

6.1 Impact of Similarity Metrics and Bias Adjustment

- **Cosine (Case 1)** is simplest but most misleading under sparsity and bias.
- **Mean-Centered Cosine (Case 2)** removes rating-scale bias, producing more stable similarity.
- **Pearson (Case 3)** is strongest for detecting pattern agreement but unreliable under sparse overlap.

6.2 Overall Observations

- Bias adjustment significantly improves meaningful similarity.
- Overlap-based discounting (DF, DS) is essential for reliability.
- Sparsity remains the main challenge—no metric fully solves it.
- Pearson provides the best conceptual measure but is not always practically stable.

6.3 Final Insight

No single similarity metric is universally superior. Performance depends on:

- Rating density,
- Rating scale consistency,
- Overlap size,
- User behavior patterns.

A hybrid approach that combines:

- mean-centering,
- overlap thresholding,
- discounted similarity,
- and Pearson correlation,

is generally the most reliable for sparse real-world datasets.