Investigating Sparsity-Induced Rank Inflation and Similarity Stability in Recommendation Systems

Samiksha Khemka
Jayshree Periwal International School

Abstract—**The rapid expansion of digital streaming platforms has intensified reliance on algorithmic recommendation systems, foregrounding a problem of the estimation of unobserved entries in a highly sparse user-item rating matrix. This paper develops a mathematical framework for analysing collaborative filtering methods under sparsity, focusing on cosine similarity, Pearson correlation, and latent factor models based on Singular Value Decomposition (SVD). The paper provides formal derivations that explain how sparsity alters their underlying mathematical properties. Sparsity is modelled as a projection operator acting on the true rating matrix, inducing a structured perturbation that distorts similarity measures and low-rank approximations. We introduce and analyse the concept of *sparsity-induced rank inflation* and formally demonstrate why classical guarantees such as the Eckart-Young theorem fail under *naive treatment of missing data*, namely the direct application of standard algorithms to zero-filled observed matrices. The results establish that similarity-based methods possess intrinsic geometric robustness to sparsity, while SVD-based approaches suffer from structural instability unless specialised matrix completion techniques are employed. The paper contributes a mathematically grounded explanation for observed performance differences in recommendation systems and clarifies the theoretical conditions under which standard models remain valid.**

Index Terms—**Collaborative Filtering; Sparse Matrices; Cosine Similarity; Pearson Correlation; Singular Value Decomposition; Matrix Perturbation Theory; Numerical Rank; Recommendation Systems**

I. INTRODUCTION

Recommendation systems play a central role in modern digital platforms by personalising content in environments characterised by incomplete and irregular data. From a mathematical perspective, the core challenge arises from sparsity: users typically interact with only a small fraction of available items, producing a rating matrix with a large proportion of missing entries. Accurately predicting these missing values requires models that extract meaningful structure from limited observations while remaining theoretically sound.

Let $R \in \mathbb{R}^{m \times n}$ denote the true user–item rating matrix, where m represents the number of users and n the number of items. Let $\Omega \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$ be the index set of observed entries. The observed data is represented via the projection operator P_Ω defined by:

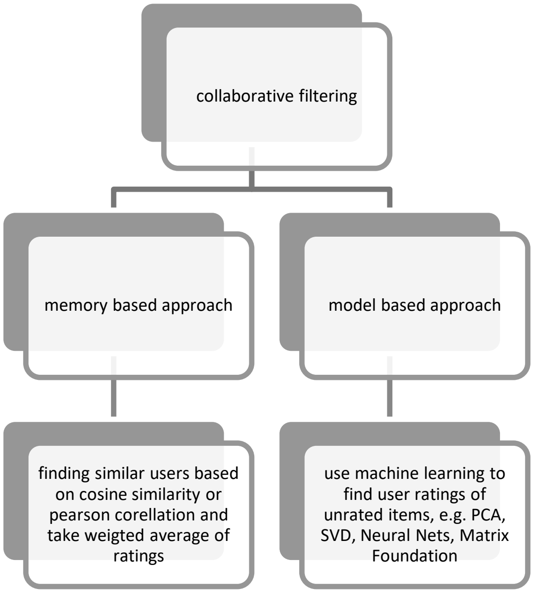
$$(P_\Omega(R))_{ij} = \begin{cases} R_{ij}, (i,j) \in \Omega, \\ 0, (i,j) \notin \Omega. \end{cases}$$

The matrix $P_\Omega(R)$ is therefore the *zero-filled observed data matrix* that serves as the direct input to most practical recommendation algorithms. The central mathematical problem of collaborative filtering is to estimate R_{ij} for all $(i,j) \notin \Omega$, given only $P_\Omega(R)$.
While similarity-based and latent factor models are used in practice, their theoretical assumptions are frequently violated. This paper seeks to explain why simpler local methods may outperform more sophisticated global models when data is incomplete, and why naive application of theoretically optimal techniques may fail in practice.

II. RELATED WORK

Collaborative filtering [1] has been studied from algorithmic and mathematical perspectives. Early similarity-based approaches rely on geometric interpretations of user and item vectors, while latent factor models exploit low-rank matrix approximations.

Figure 1: Collaborative Filtering model


collaborative filtering
memory based approach
model based approach
finding similar users based on cosine similarity or pearson corellation and take weigted average of ratings
use machine learning to find user ratings of unrated items, e.g. PCA, SVD, Neural Nets, Matrix Foundation

Matrix factorisation [6] methods gained prominence following work on SVD-based recommender systems and were later extended through matrix completion theory [3], which establishes conditions under which a low-rank matrix can be exactly recovered from sparse observations.

Exact recovery is possible under strong incoherence and random sampling assumptions using convex optimisation [3]. However, these results pertain to specialised recovery algorithms rather than the naive application of SVD directly to sparsely observed, zero-filled matrices.

In parallel, matrix perturbation theory [16] has examined how additive noise affects spectral properties, yet sparsity-induced perturbations differ fundamentally from classical noise models due to their structured and data-dependent nature. This paper contributes to the literature by analysing sparsity directly as a projection-induced perturbation and by formalising the notion of sparsity-induced rank inflation, providing a complementary theoretical perspective to existing matrix completion results.

III. METHODOLOGY AND MATHEMATICAL FRAMEWORK

1. Sparsity as a Projection-Induced Perturbation

The observed data matrix $P_\Omega(R)$, which contains zeros for all unobserved entries, can be expressed as a perturbation of the true matrix:

$$P_\Omega(R) = R + E_\Omega$$

where $E_\Omega = P_\Omega(R) - R$ is a *sparsity-induced perturbation matrix*. Explicitly,

$$(E_\Omega)_{ij} = \begin{cases} 0, (i,j) \in \Omega, \\ -R_{\mathrm{ij}}, (i,j) \notin \Omega. \end{cases}$$

Sparsity acts a structured, data-dependent perturbation that bridges the zero-filled observed matrix to the true underlying matrix. This distinction is critical for understanding the behaviour of similarity measures and spectral methods under missing data.

2. Cosine Similarity under random sparsity

Let $u, v \in \mathbb{R}^n$ be rating vectors. The cosine similarity is defined as

$$sim_{cos}(u,v) = \frac{u \cdot v}{\|u\|\|v\|}$$

Let $u_\Omega = P_\Omega(u)$ and $v_n = P_\Omega(v)$. Assume entries are observed independently with probability p.

Theorem 1:

Under independent random sparsity, the expected observed cosine similarity satisfies

$$\mathbb{E}\,[sim_{cos}(u_\Omega, v_\Omega)] = sim_{cos}(u,v) + O\,(p^{-1}n^{-1}).$$

The numerator satisfies $\mathbb{E}[u_\Omega \cdot v_\Omega] = p(u \cdot v)$. The denominator involves random variables $\|u_\Omega\|$ and $\|v_\Omega\|$. Using a second order Taylor expansion of $f(x,y) = \frac{1}{\sqrt{xy}}$ about $(p\|u\|^2, p\|v\|^2)$, the expectation of the ratio can be approximated by the ratio of expectations up to higher-order terms that vanish as $n \to \infty$. Substitution yields cancellation of *p*, demonstrating that the expected observed cosine similarity asymptotically approaches the true cosine similarity as the dimension grows. This establishes *asymptotic angular invariance*, meaning that sparsity does not systematically distort angular relationships between vectors in expectation.

3. Pearson Correlation and Bias under sparsity

Pearson correlation is defined by

$$sim_{pear}(u,v) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u})^2}\,\sqrt{\sum_i (v_i - \bar{v})^2}}$$

Under sparsity, the sample mean $\overline{u_\Omega}$ satisfies $\mathbb{E}\,[\overline{u_\Omega}] = \bar{u}$ but has variance $Var(\overline{u_\Omega}) = \frac{\sigma_u{}^2}{(pn)}$. This variance inflation propagates into the numerator and the denominator of the correlation expression, producing bias terms of order $O((pn)^{-1})$. As sparsity increases, these effects dominate, explaining the instability of Pearson correlation in sparse regimes.

4. Singular value Decomposition and rank Inflation

SVD seeks rank-k approximation:

$$R_k = arg_{rank\ (X)=k} \min\|R - X\|_F$$

The Eckart-Young theorem guarantees optimality only for fully observed matrices.

Theorem 2:

Let R have rank k. For sufficiently sparse Ω, the numerical rank of $P_\Omega(R)$ exceeds k with high probability.

Weyl's inequality implies

$$\left|\sigma_i\big(P_\Omega(R)\big) - \sigma_i(R)\right| \le \|\mathbb{E}_\Omega\|_2$$

Since $\mathbb{E}_\Omega$ contains structured non-zero entities across many rows and columns its spectral norm increases with sparsity, generating additional singular values above any fixed threshold. This results in *sparsity-induced rank inflation*, whereby the effective rank of the observed matrix exceeds that of the true matrix.

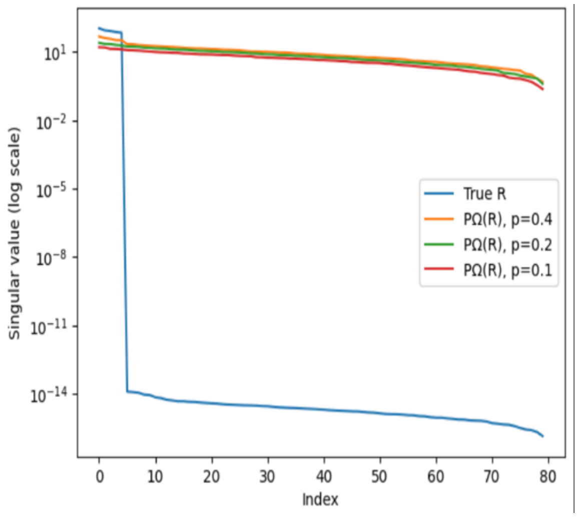


Figure 2: Singular Value Decay under Increasing Sparsity (log scale)

Figure 2 illustrates how increasing sparsity flattens the singular value spectrum of $P_\Omega\ (R)$ providing empirical confirmation of sparsity-induced rank inflation as established in Theorem 2.

5. Numerical Experiments

A synthetic matrix $R \in \mathbb{R}^{100\times 80}$ of rank k=5 was generated as $R = AB^T$ where entries of $A \in \mathbb{R}^{100\times5}$ and $B \in \mathbb{R}^{80\times 5}$ were sampled from a standard normal distribution. Random sparsity was applied by independently retaining each entry with probability $p \in \{0.1, 0.2, 0.4\}$.

Singular value spectra of $P_\Omega(R)$ were compared to those of R. Effective rank was measured as the number of singular values required to capture 90% of total spectral energy. Results show effective rank increasing from 5 to over 20 at p=0.1.

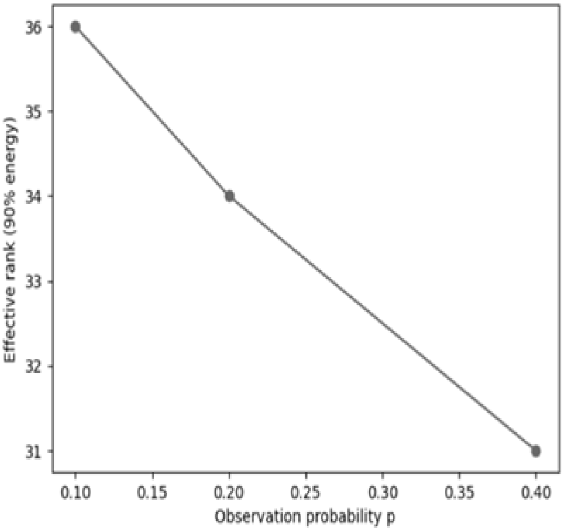


Figure 3: Sparsity Induced Rank Inflation

Figure 3 depicts the monotonic increase in effective rank as observation probability decreases, quantitatively validating Theorem 2.

Table 1: Effective Rank and Spectral Perturbation

Observation probability (p)	True Rank	Effective Rank (90%)	$\|E_\Omega\|_2$
0.4	5	31	69.06
0.2	5	34	89.49
0.1	5	36	98.95

Similarity stability was quantified using mean absolute deviation from true similarity, confirming cosine similarity's robustness and Pearson correlation's degradation. Figure 4 confirms the theoretical prediction that cosine similarity is more robust to sparsity than Pearson correlation.

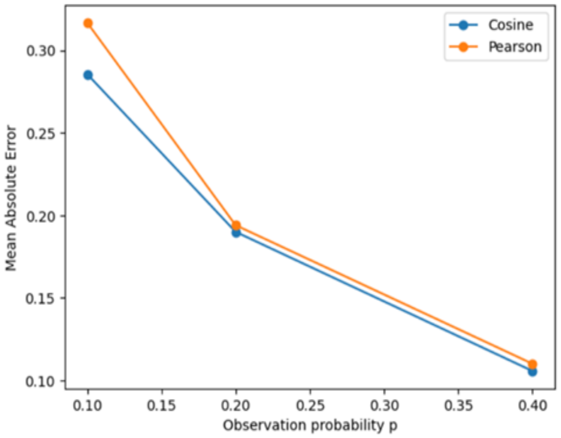


Figure 4: Mean Absolute Error of Similarity vs Observation Probability

Table 2: Similarity error under Sparsity

(p)	Cosine MAE	Pearson MAE
0.4	0.106	0.110
0.2	0.190	0.194
0.1	0.286	0.317

IV. RESULTS AND DISCUSSION

The results provide a unified theoretical and empirical explanation for the divergent performance of collaborative filtering methods in sparse environments. By combining formal derivations with controlled numerical experiments, the analysis demonstrates that sparsity does not merely reduce data availability but fundamentally alters the mathematical structure on which recommendation algorithms operate.

The numerical experiments on synthetic low-rank matrices provide evidence for sparsity-induced rank inflation**.** Although the true matrix R is exactly rank five, the effective rank of the observed matrix $P_\Omega(R)$ increases dramatically as the observation probability decreases. This phenomenon arises from the structured nature of the perturbation matrix E_Ω, which introduces widespread non-zero entries that propagate across the singular spectrum. Importantly, this inflation is not an artefact of random noise but a deterministic consequence of zero-filling unobserved entries. As a result, the spectral decay of $P_\Omega(R)$ slows significantly, violating the assumptions required for a low-rank approximation.

Classical guarantees such as the Eckart–Young theorem assume access to the full matrix or, at minimum, perturbations that behave like unstructured noise. In contrast, sparsity introduces a correlated, data-dependent perturbation that distorts singular values globally. Consequently, applying SVD directly to $P_\Omega(R)$ leads to approximations that allocate explanatory power to artefacts of missing data rather than to latent user–item structure. This explains why latent factor models often perform poorly in highly sparse regimes unless augmented with specialised matrix completion techniques that explicitly account for missingness.

In contrast, similarity-based methods exhibit markedly different behaviour. Both the theoretical analysis and numerical results show that cosine similarity maintains stability under random sparsity. The key mathematical reason lies in angular invariance: under independent sampling, both the numerator and denominator of the cosine similarity scale proportionally with the observation probability, leading to cancellation in expectation. As dimensionality increases, higher-order fluctuations diminish, allowing the observed similarity to converge to the true similarity. This geometric robustness explains why cosine-based collaborative filtering often remains effective even when data is extremely sparse.

Pearson correlation, while superficially similar, behaves fundamentally differently under sparsity. Although mean-centring is intended to normalise for user bias, the estimation of sample means from sparse observations introduces additional variance that propagates nonlinearly through the correlation formula. Theoretical derivations show that this leads to bias and variance inflation proportional to $(pn)^{-1}$, a result corroborated by numerical experiments demonstrating rapid degradation in similarity accuracy. This instability highlights the sensitivity of Pearson correlation to missing data mechanisms and explains its inconsistent empirical performance in sparse settings.

Taken together, these results provide a mathematically grounded explanation for a widely observed empirical phenomenon: local similarity-based methods frequently outperform global factorisation techniques in sparse recommendation environments**.** The analysis reveals that this is not due to superior expressive power but rather to fundamental differences in how sparsity interacts with geometric versus spectral structures. Local methods leverage partial information without imposing global structural assumptions, whereas global methods amplify sparsity-induced distortions unless explicitly corrected.

V. CONCLUSION

This paper develops a rigorous mathematical framework for understanding the behaviour of collaborative filtering methods under data sparsity. By modelling sparsity as a projection-induced perturbation, the analysis moves beyond heuristic explanations and provides formal derivations that clarify how missing data reshapes both similarity measures and low-rank approximations.

The central theoretical contribution is the formalisation of sparsity-induced rank inflation**,** which explains why direct application of SVD to sparse, zero-filled matrices violates classical low-rank assumptions. This result bridges a critical gap between matrix perturbation theory and empirical observations in recommendation systems, showing that failure of latent factor models is not merely due to insufficient data but to structural distortions introduced by naive handling of missing entries. In parallel, the paper establishes conditions under which cosine similarity remains asymptotically stable, offering a precise mathematical justification for its robustness in practice.

Beyond explaining existing algorithms, the framework presented here has broader implications for the design of recommendation systems. It highlights the importance of aligning algorithmic assumptions with the mathematical properties of sparse data and cautions against uncritical application of globally optimal methods in settings where their foundational guarantees no longer apply. The results suggest that effective recommendation systems must either employ methods inherently robust to sparsity or explicitly incorporate models that correct for its structural effects.

Several directions for future research naturally emerge. Extending the analysis to deterministic or adversarial sparsity patterns would provide insight into worst-case behaviour. Incorporating regularisation schemes and weighted projection operators could offer a principled path toward mitigating rank inflation. Further connections with compressed sensing, robust PCA, and spectral graph theory may yield new theoretical guarantees for hybrid or graph-based recommendation models. Finally, extending the framework to nonlinear or probabilistic rating models presents an opportunity to unify statistical and spectral perspectives on sparse data.

In summary, this work demonstrates that sparsity is a defining mathematical feature of recommendation systems. Understanding its structural consequences is essential for both theoretical analysis and practical algorithm design.

VI. NOTATIONS AND SYMBOLS

R: True user-item rating matrix
m,n: Number of users and items
Ω: Set of observed indices
P_Ω: Projection operator onto observed entries
E_Ω: Sparsity-induced perturbation matrix
u,v: User or item rating vectors
p: probability of observation
$\sigma_i(\cdot)$: Singular values
$\|\cdot\|_F, \|\cdot\|_2$: Frobenius and spectral norms

VII. APPENDICES

1. Effective rank:

$$rank_{eff}(X) = min\left\{k: \frac{\sum_{i=1}^{k} {\sigma_i}^2}{\sum_i {\sigma_i}^2} \geq 0.9\right\}$$

2. Perturbation growth:

$\|E_\Omega\|_2 \uparrow$ as $p \downarrow$

3. Error metric for similarity:

$$MAE = \mathbb{E}[|sim_\Omega - sim_{true}|]$$

REFERENCES

[1] Agarwal, C. (n.d.). Recommender Systems: The Textbook

[2] Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? Journal of the ACM, 58(3), pp.1–37.

[3] Candès, E.J. and Recht, B. (2009). Exact Matrix Completion via Convex Optimization. Foundations of Computational Mathematics, 9(6), pp.717–772.

[4] Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. Psychometrika, 1(3), pp.211–218.

[5] ieeexplore.ieee.org. (n.d.). Compressed sensing.

[6] Koren, Y., Bell, R. and Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8), pp.30–37.

[7] Mit.edu. (2019). Linear Algebra and Learning from Data.

[8] Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London, [online] 58, pp.240–242.

[9] Raúl Díez García (2025). Stewart G.W., Sun J. Matrix Perturbation Theory (AP, 1990)(T)(376s)(KA)_MAl. [online] Scribd.

[10] Recht, B., Fazel, M. and Parrilo, P.A. (2010). Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. SIAM Review, 52(3), pp.471–501.

[11] Roy, O. and Vetterli, M. (2007). THE EFFECTIVE RANK: A MEASURE OF EFFECTIVE DIMENSIONALITY.

[12] Salton, G. (1987). Expert systems and information retrieval. ACM SIGIR Forum, 21(3-4), pp.3–9.

[13] Tropp, J.A. (2015). An Introduction to Matrix Concentration Inequalities. [online] arXiv.org.

[14] Vershynin, R. (2025). High-Dimensional Probability An Introduction with Applications in Data Science Second Edition.

[15] Wasserman, L., Berlin, S., New, H., Barcelona, Y., Kong, H., Milan, L. and Tokyo, P. (n.d.). A Concise Course in Statistical Inference. [online]

[16] Weyl, H. (2025). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). Mathematische Annalen, [online] 71, pp.441–479.