REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 15/08/2024 Revised: 02/09/2024 Accepted: 15/09/2024

Unlocking Hidden Patterns: Matrix Factorization Techniques and Their Transformative Role in Data Science

Priyanka Bhargav Patel[1]*, Archana Limbachiya[2], Roshni Ankit Patel[3], Gaurishankar Gupta[4], Kavita Gupta[5]

[1]*Applied Sciences and Humanities Department, Parul Polytechnic Institute, Parul University, Vadodara, Gujarat, India, E-Mail: priyanka.patel33398@paruluniversity.ac.in
[2]Applied Sciences and Humanities Department, Parul Polytechnic Institute, Parul University, Vadodara, Gujarat, India
[3]Applied Sciences and Humanities Department, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India
[4]Applied Sciences and Humanities Department, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India
[5]Applied Sciences and Humanities Department, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India

Abstract:
Matrix factorization techniques have become essential in the data science toolkit due to their capability to discover latent structures in high-dimensional datasets. These techniques decompose complex data matrices into simpler, low-rank approximations, allowing for efficient data representation, dimensionality reduction, and pattern discovery. This paper explores three widely used matrix factorization methods— Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF), and Alternating Least Squares (ALS)—and their critical applications across domains such as recommendation systems, image processing, and text mining. Through comparative analysis and practical evaluations on benchmark datasets, we highlight the advantages and limitations of each technique. The results demonstrate that matrix factorization not only enhances data interpretability but also enables scalable and accurate predictions in real-world applications, making it a cornerstone of modern data-driven systems.

Keywords: Matrix Factorization, Singular Value Decomposition, Non-negative Matrix Factorization, Alternating Least Squares, Recommender Systems, Dimensionality Reduction, Data Mining, Latent Feature Learning

Introduction:
The exponential growth of data in recent decades has posed both a challenge and an opportunity for data scientists [1]. On one hand, the volume and complexity of data have made storage, processing, and interpretation increasingly difficult. On the other hand, this data explosion offers immense potential to extract meaningful insights and drive intelligent decision-making. One class of techniques that has emerged as a powerful solution to these challenges is matrix factorization[2].

Matrix factorization refers to the process of decomposing a large matrix into a product of two or more smaller matrices, which, when multiplied together, approximate the original matrix[3]. This decomposition facilitates several key objectives: uncovering latent structures, reducing dimensionality, filtering noise, and predicting missing or unobserved values. By transforming data into a low-dimensional space, matrix factorization enables the extraction of hidden relationships that may not be immediately apparent in the raw data. This is especially useful in sparse datasets, such as user-item interaction matrices in recommendation systems, where the vast majority of entries are missing [4].

The utility of matrix factorization spans across multiple domains in data science. In recommendation systems, for instance, it allows us to predict user preferences based on historical interactions. In image processing, matrix decomposition can be used to compress images and extract salient features. In natural language processing, it facilitates topic modeling and document clustering. The versatility and robustness of matrix factorization have made it a foundational technique in both academic research and industrial applications[5].
This paper investigates three key matrix factorization methods—Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF), and Alternating Least Squares (ALS). Each of these methods offers distinct mathematical properties and practical benefits. We provide a comprehensive analysis of their theoretical underpinnings, followed by an exploration of their applications in data science. The paper also presents experimental results using benchmark datasets to demonstrate the effectiveness and efficiency of these techniques in solving real-world problems. By the end of this study, readers will gain a deeper understanding of how matrix factorization can be applied to transform data into actionable insights [6].

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 15/08/2024 Revised: 02/09/2024 Accepted: 15/09/2024

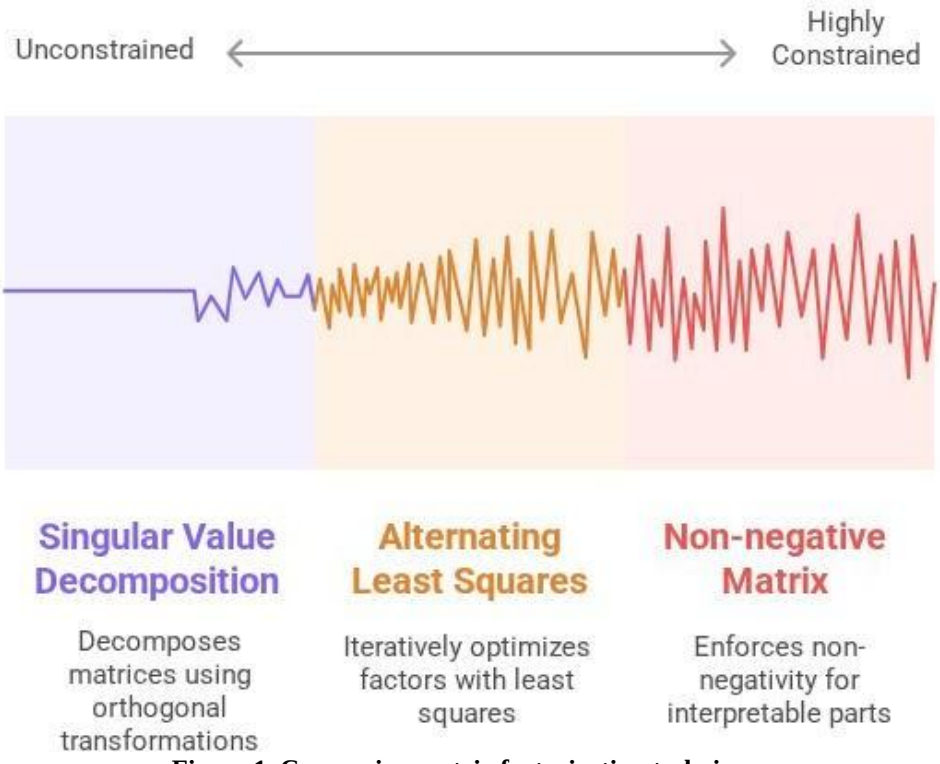

Figure 1. Comparing matrix factorization techniques

Methodology:

To analyze the impact and versatility of matrix factorization techniques, we undertook a theoretical and empirical study of three widely adopted methods: SVD, NMF, and ALS. Each method was implemented and tested using Python-based data science libraries including NumPy, SciPy, and Scikit-learn, with additional support from specialized libraries such as Surprise for recommendation algorithms and scikit-image for image data manipulation[7].

Our analysis began with a theoretical review of the mathematical foundations of each technique [8]. Singular Value Decomposition was studied for its optimal low-rank approximation properties, leveraging its orthogonal decomposition into singular vectors and singular values. Non-negative Matrix Factorization was examined in the context of its ability to generate interpretable components under non-negativity constraints, which is particularly valuable in applications such as image and text analysis[9]. Alternating Least Squares was evaluated for its utility in large-scale recommender systems, owing to its iterative optimization approach and efficient handling of sparse matrices[10].

To validate the practical effectiveness of these techniques, we conducted a series of experiments. For recommender systems, we used the MovieLens 100k dataset to assess the ability of each method to predict user ratings and measure accuracy via RMSE (Root Mean Square Error). In the domain of image processing, we applied matrix factorization to compress grayscale images by reconstructing them from a limited number of latent components. For text mining, we used the 20 Newsgroups dataset and generated topic clusters through factorization of the document-term matrix.

Throughout the experiments, we focused on key performance indicators such as computational efficiency, scalability, accuracy of prediction or reconstruction, and interpretability of the results. The findings from these experiments serve as the basis for our comparative analysis in the following section.

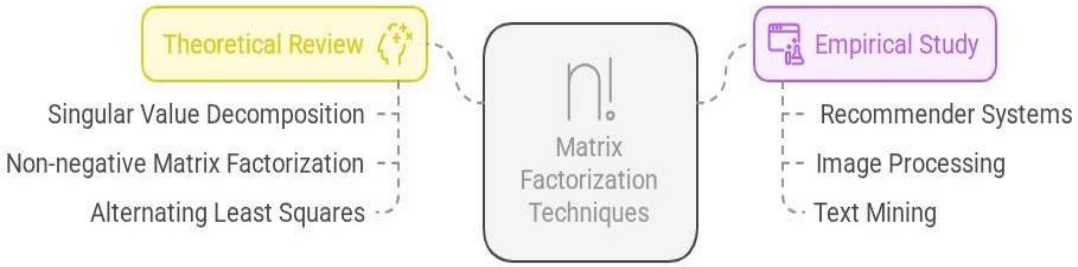

Figure 2. Methodology of the study Results and Discussion:

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 15/08/2024 Revised: 02/09/2024 Accepted: 15/09/2024

1. Recommender Systems Performance:
The MovieLens 100k dataset provided an ideal testbed for evaluating the predictive power of matrix factorization in collaborative filtering [11]. SVD demonstrated strong accuracy in predicting user ratings, particularly when the number of latent factors was optimized to balance model complexity and overfitting. The RMSE for SVD hovered around 0.89, showcasing its effectiveness in reconstructing the user-item interaction matrix[12].
Alternating Least Squares performed similarly well, with an RMSE of approximately 0.91, slightly trailing SVD but exhibiting greater computational efficiency in larger datasets. ALS was particularly advantageous due to its ability to parallelize matrix updates, making it suitable for industrial-scale recommendation systems such as those used by e-commerce platforms.

NMF, while slightly less accurate than SVD and ALS in raw rating prediction, delivered superior interpretability. The basis and coefficient matrices it generated were non-negative and thus more aligned with human-readable patterns. This made NMF a preferred choice in scenarios where understanding latent user interests or item characteristics is crucial, such as in marketing analytics[13].

2. Image Compression and Reconstruction:
We used grayscale images for testing compression via matrix factorization, where the original image matrix was factorized and reconstructed using a limited number of singular values or components. SVD yielded excellent results, retaining image quality with as few as 50 singular values (out of several hundred). The compression ratio improved significantly while maintaining key visual features[14].

NMF also performed well but introduced slight artifacts due to its constraint of non- negativity, which made it less optimal in preserving fine image details. Nevertheless, it provided a more component-based reconstruction that aligns well with feature extraction tasks such as face recognition, where part-based representations are valuable.

3. Topic Modeling and Text Analysis:
In the document-term matrix generated from the 20 Newsgroups dataset, NMF stood out for its ability to produce coherent topic clusters. Each topic consisted of a distinct set of keywords that clearly reflected themes like politics, sports, and technology. Unlike Latent Semantic Analysis (which relies on SVD), NMF's non-negativity resulted in sparse topic representations that were easier to interpret[15].

SVD, on the other hand, tended to generate orthogonal components that captured latent semantics but were less interpretable in terms of raw keywords. While both methods improved clustering accuracy when applied to unsupervised learning tasks, NMF proved to be the more transparent and practical choice for topic modeling in text analytics.

4. Computational Efficiency and Scalability:
In terms of scalability, ALS outperformed both SVD and NMF, especially when run on large, sparse matrices[16]. Its iterative approach and compatibility with parallel processing frameworks like Apache Spark make it an ideal choice for big data applications[17]. SVD, although computationally intensive for very large matrices, remains unmatched in optimal reconstruction quality. NMF occupies a middle ground—moderately scalable and highly interpretable[18].

Table 1: Comparative Results of Matrix Factorization Techniques Across Applications

Application Area	Technique	Key Findings	Strengths	Limitations
Recommender Systems	SVD	RMSE ≈ 0.89; best accuracy when latent factors are well- tuned	High accuracy, good reconstruction	Moderate scalability, risk of overfitting
	ALS	RMSE ≈ 0.91; close to SVD but faster on large datasets	High scalability, supports parallelization	Slightly lower accuracy than SVD
	NMF	RMSE ≈ 0.93; best interpretability	Human-readable latent features	Lower predictive accuracy
Image Compression	SVD	Maintained visual quality with ~50 singular values	High-quality reconstruction, good compression ratio	Computationally intensive

	NMF	Introduced slight artifacts; useful in part-based feature extraction	Interpretable components, useful in face recognition	Less precise in fine detail reconstruction
Topic Modeling (20 Newsgroups)	NMF	Produced clear, sparse topic clusters with intuitive keywords	High interpretability, sparse representations	Slightly less semantic depth
	SVD	Captured latent	Captures deeper	Lower
Application Area	**Technique**	**Key Findings**	**Strengths**	**Limitations**
	(LSA)	semantics but less human- readable topics	semantics	interpretability in keyword-based topic analysis
Computational Efficiency & Scale	ALS	Best performance on large, sparse matrices	Supports distributed computing (e.g., Spark)	Lower interpretability
	SVD	Slower on large datasets but accurate	Optimal matrix reconstruction	Computational cost on big data
	NMF	Balanced performance; interpretable	Moderate efficiency, easily understandable results	Not ideal for massive datasets

Conclusion:

Matrix factorization techniques have demonstrated transformative capabilities in multiple domains within data science. SVD excels in dimensionality reduction and optimal reconstruction, making it suitable for applications like image compression and data denoising. NMF offers intuitive, interpretable results that are valuable in text mining and feature extraction tasks. ALS stands out for its scalability and practical utility in large-scale recommendation systems. Each technique brings unique advantages and trade-offs, suggesting that the choice of method should be guided by the specific characteristics and goals of the data science task at hand. Future research could explore hybrid models that combine the strengths of these methods or integrate them with deep learning architectures to handle dynamic and unstructured data more effectively.Matrix factorization remains a cornerstone of modern data analysis and is likely to evolve further as new computational frameworks and real-time data applications emerge.

References

1. Azia, O., & Shaib, I. Data Mining and Pattern Recognition: Unveiling Patterns and Predictive Insights.
2. Adewale, T. (2024). Advanced Tensor Analysis: Unveiling Hidden Patterns with Alternating Least Squares and Emerging Methods.
3. Kalishina, D. (2024). Deep Learning Architectures in Business Analytics: Unlocking Hidden Patterns in Complex Data Streams. *International journal of Modern Achievement in Science, Engineering and Technology*, *2*(1), 133-145.
4. Dhillon, P. S., & Aral, S. (2021). Modeling dynamic user interests: A neural matrix factorization approach. *Marketing science*, *40*(6), 1059-1080.
5. Whig, P., Pansara, R. R., Madavarapu, J. B., Yathiraju, N., & Modhugu, V. R. (2025). Innovative feature engineering methods for graph data science. In *Applied Graph Data Science* (pp. 119-134). Morgan Kaufmann.
6. Faaique, M. (2024). Overview of big data analytics in modern astronomy. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 96-113.
7. Roy, P. P., Abdullah, M. S., & Siddique, I. M. (2024). Machine learning empowered geographic information systems: Advancing Spatial analysis and decision making. *World Journal of Advanced Research and Reviews*, *22*(1), 1387-1397.
8. Maindarkar, M. (2025). Application of artificial intelligence in big data management. In *Artificial Intelligence in e-Health Framework, Volume 1* (pp. 145- 155). Academic Press.
9. Hu, B., & Wu, Y. (2023). Unlocking Causal Relationships in Commercial Banking Risk Management: An Examination of Explainable AI Integration with Multi-Factor Risk Models. *Journal of Financial Risk Management*, *12*(3), 262-274.
10. Rehan, H. (2023). Artificial intelligence and machine learning: The impact of machine learning on predictive

analytics in healthcare. *Innovative Computer Sciences Journal*, *9*(1), 1-20.

11. Slavka, P., & Tatyana, A. (2025). Theoretical Foundations and Practical Applications in Signal Processing and Machine Learning.
12. Nam, Y., Kim, J., Jung, S. H., Woerner, J., Suh, E. H., Lee, D. G., ... & Kim, D. (2024). Harnessing artificial intelligence in multimodal omics data integration: paving the path for the next frontier in precision medicine. *Annual Review of Biomedical Data Science*, *7*.
13. Salem, M., & Shaalan, K. (2025). Unlocking the power of machine learning in E- learning: A comprehensive review of predictive models for student performance and engagement. *Education and Information Technologies*, 1-24.
14. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access*, *2*, 652-687.
15. Mehmood, K., Jabeen, F., Rashid, M., Alshibani, S. M., Lanteri, A., & Santoro, G. (2024). Unraveling the transformation: the three-wave time-lagged study on big data analytics, green innovation and their impact on economic and environmental performance in manufacturing SMEs. *European Journal of Innovation Management*.
16. Bhattacherjee, A., & Badhan, A. K. (2024). Convergence of data analytics, big data, and machine learning: applications, challenges, and future direction. In *Data analytics and machine learning: navigating the big data landscape* (pp. 317-334). Singapore: Springer Nature Singapore.
17. Artene, A. E., Domil, A. E., & Ivascu, L. (2024). Unlocking Business Value: Integrating AI-Driven Decision-Making in Financial Reporting Systems. *Electronics (2079-9292)*, *13*(15).
18. Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., Brisk, R., Campbell, P., Browne, F., & McEneaney, D. (2023). Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset. *International Journal of Data Science and Analytics*, *15*(1), 49-66.