
Nonlinear Factor Models in Financial Markets: A Comparative Study of Autoencoders and PCA

Jason Yi
jasonyi@unc.edu

Daniel Wang
dzw@unc.edu

Anish Parepalli
anishpa@unc.edu

Nicholas Do
khanhdo@unc.edu

Krish Patel
kpatel18@unc.edu

Abstract

This project explores the use of autoencoders as a nonlinear alternative to Principal Component Analysis (PCA) for dimensionality reduction in financial return data. Using daily equity returns, we extract latent factors with both PCA and a neural autoencoder, evaluate reconstruction performance via R^2 and Mean Squared Error (MSE), and apply K-means clustering to the learned latent factors. Preliminary results show that while PCA remains strong in linear settings, autoencoders demonstrate competitive performance, particularly in capturing nonlinear structure, supporting the use of deep learning models for uncovering hidden relationships in financial time series.

GitHub Repository: https://github.com/dzw154/COMP560_F25_Final_Project

1 Introduction

Understanding latent structure in asset returns is central to financial modeling, with implications for risk management and portfolio construction. In this project, we study U.S. equity return data and compare two approaches for uncovering latent factors: Principal Component Analysis (PCA), a standard linear technique, and neural autoencoders, which can model nonlinear relationships.

Both methods compress return series into low-dimensional representations and attempt to reconstruct the original data. By evaluating reconstruction quality and comparing the structure of the learned latent factors, we assess whether nonlinear autoencoders provide advantages over classical PCA in capturing complex co-movements in financial time series.

2 Related Work

Dimensionality reduction has long been used in finance to extract latent factors that explain asset return variation. Principal Component Analysis (PCA) remains the standard linear approach for risk modeling and factor investing (Connor and Korajczyk, 1986), but its linear structure limits its ability to capture nonlinear dependencies in financial data.

Neural autoencoders have been introduced as nonlinear generalizations of PCA (Hinton and Salakhutdinov, 2006), learning compressed latent representations through bottleneck architectures. They have been applied to tasks such as latent factor extraction, regime characterization, and denoising of financial time series, with deeper variants improving robustness and flexibility (Heaton, Polson, and Witte, 2017).

Prior work also explores clustering and extended PCA variants, but the key distinction remains between linear factor models and nonlinear neural embeddings.

3 Methodologies

3.1 Data Collection and Preprocessing

We collected price data for the S&P 500 index and U.S. tech stocks from 2012–2025 using the yfinance and Alpha Vantage APIs.

3.2 Dimensionality Reduction

3.2.1 Principal Component Analysis (PCA)

PCA extracts orthogonal directions of maximum variance from the return matrix, providing a linear low-dimensional representation of asset co-movements. We evaluate PCA reconstructions using two metrics: the coefficient of determination R^2 and the Mean Squared Error (MSE):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad \text{MSE} = \frac{1}{n} \sum_i (y_i - \hat{y}_i)^2.$$

These metrics quantify how well the reduced representation explains variation in the original returns.

3.2.2 Autoencoders

To capture nonlinear dependencies that PCA cannot, we implemented a neural autoencoder with an encoder–decoder architecture and a 128-neuron hidden layer. The network compresses returns into a low-dimensional bottleneck representation and reconstructs them from this latent space. The autoencoder’s latent factors were evaluated using the same R^2 and MSE metrics to allow direct comparison with PCA.

4 Data Exploration

4.1 Daily Return Distributions

Our dataset includes 13 years (2012–2025) of daily returns for the S&P 500 and eight major tech stocks, totaling 3,240 observations per security with no missing data. Figure 1 shows that the stocks exhibit differing volatility levels, with TSLA being the most volatile and NVDA the strongest performer over the sample period. Return distributions are leptokurtic and centered near zero, reflecting the non-normality typical of financial markets. These characteristics motivate the use of dimensionality reduction methods capable of capturing nonlinear structure that may emerge during periods of market stress.

Daily Return Distributions of Tech Stocks and S&P 500 (2012-2025)

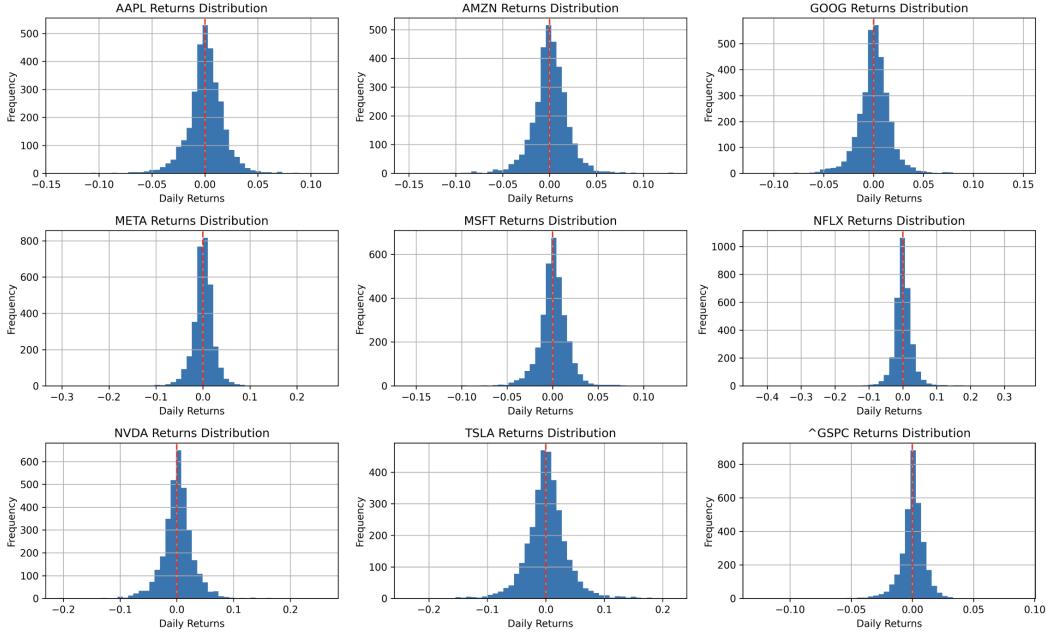


Figure 1: Daily Return Distributions of Tech Stocks and S&P 500

5 Autoencoder vs PCA Results

To assess the viability of autoencoders as nonlinear alternatives to traditional Principal Component Analysis (PCA) in financial factor modeling, we trained an autoencoder architecture with a single hidden layer comprising 128 neurons. Both PCA and the autoencoder models were constrained to extract three latent factors from standardized daily asset returns over a consistent time window. We evaluated performance using two quantitative metrics: the coefficient of determination (R^2) and Mean Squared Error (MSE).

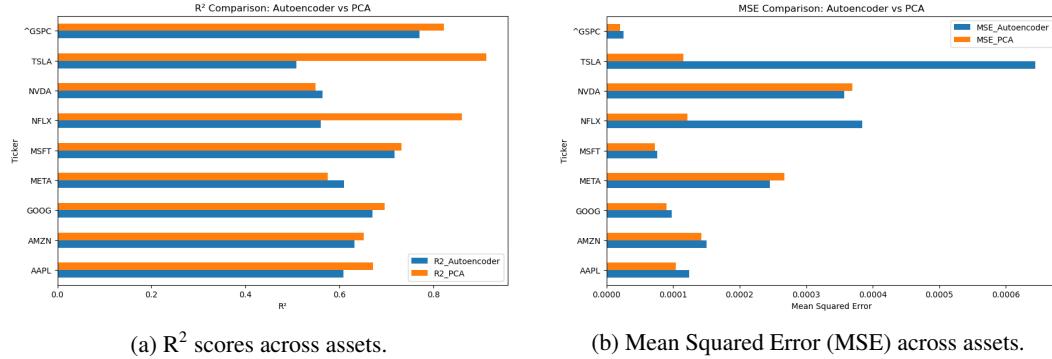


Figure 2: Performance comparison between PCA and Autoencoder based on R^2 and MSE metrics.

Notably, assets such as NVDA and AAPL were reconstructed with R^2 scores that nearly matched or slightly exceeded those achieved by PCA. For GOOG, MSFT, and META, the performance gap between the two methods was minimal, indicating that the autoencoder captured much of the same underlying return structure.

PCA continued to outperform the autoencoder for certain assets. Most prominently GSPC and TSLA, which highlights PCA's advantage in modeling highly linear or benchmark-like return profiles. MSE

results were consistent: PCA generally achieved lower reconstruction error across most assets, but the autoencoder’s errors were close, reflecting its ability to capture nonlinear relationships in the data.

Overall, these results suggest that while PCA remains a strong and efficient baseline, especially in datasets dominated by linear correlations, autoencoders are a competitive nonlinear alternative and warrant further exploration in financial factor modeling.

6 Clustering on Latent Features

We clustered the latent factors from both PCA and the autoencoder to study how stocks group in the reduced spaces. Based on an elbow analysis of K-means inertia (not shown), we selected $k = 3$ clusters for both embeddings.

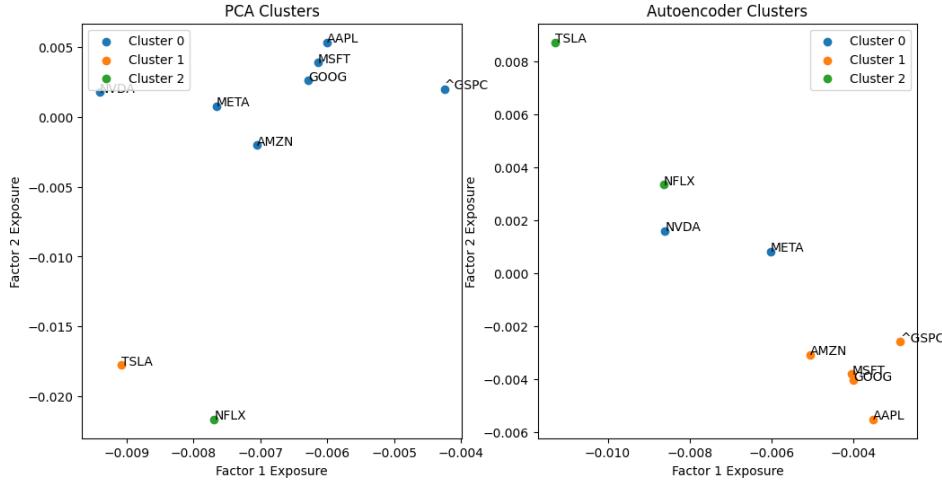


Figure 3: K-means clustering results on PCA and autoencoder latent factors.

With $k = 3$, K-means clustering on PCA factors places most large-cap tech stocks and the market index in a single cluster, with TSLA and NFLX each forming distinct singleton clusters. In contrast, clustering on autoencoder factors groups META and NVDA together, keeps the index with most large-cap tech names, and pairs TSLA with NFLX. This suggests the nonlinear autoencoder uncovers subtler relationships between stocks, particularly between TSLA and NFLX, that the linear PCA representation does not, supporting our hypothesis that nonlinear embeddings can reveal alternative market structures.

These findings support our project hypothesis that nonlinear dimensionality reduction techniques can uncover market structures that classical linear methods might overlook.

7 Conclusion

This project compared nonlinear autoencoders with classical PCA for modeling latent factors in financial return data, using dimensionality reduction and clustering experiments to evaluate how well each method reconstructs returns and captures underlying structure.

Our findings show that:

- Autoencoders achieved competitive R^2 and MSE performance, particularly for assets with nonlinear return patterns.
- Clustering on autoencoder latent features revealed relationships that PCA did not capture, such as grouping TSLA and NFLX together and pairing META with NVDA.

Overall, autoencoders provide a viable and often advantageous alternative to PCA, especially when data exhibit nonlinear co-movements. This motivates further exploration of deeper, regularized autoencoder architectures for financial factor modeling.

References

- [1] Connor, G., & Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15(3), 373–394. <https://www.sciencedirect.com/science/article/pii/0304405X86900279>
- [2] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507. <https://doi.org/10.1126/science.1127647>
- [3] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1), 3–12. <https://doi.org/10.1002/asmb.2209>
- [4] Baser, P., & Saini, J. R. (2015). Agent-based stock clustering for efficient portfolio management. *International Journal of Computer Applications*, 113(3), 6–12. <https://doi.org/10.5120/20317-2381>