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

Jason Yi
730604615
jasonyi@unc.edu

Anika Sharma
730612993
anika17@unc.edu

Akksharvan Senthilkumar
730674459
akksharvan@unc.edu

Abstract

This project explores the use of autoencoders as a nonlinear alternative to Principal Component Analysis (PCA) for dimensionality reduction and anomaly detection in financial return data. Using standardized asset returns from Yahoo Finance API and Alpha Vantage API, we extract latent factors using both PCA and a neural autoencoder architecture. We evaluate model performance based on R^2 and Mean Squared Error (MSE), and apply clustering via K-Means and anomaly detection using Isolation Forests. Preliminary results show that while PCA remains strong in linear settings, autoencoders demonstrate competitive performance as model capacity increases, particularly in capturing nonlinear structure. These findings support the use of deep learning models for uncovering hidden relationships in financial time series data.

GitHub Repository: https://github.com/yi-json/auto-returns

1 Introduction

Understanding the latent patterns in asset returns is a fundamental problem in financial modeling, with direct applications to risk management, portfolio optimization, and anomaly detection. Financial markets are influenced by a wide variety of economic, political, and behavioral factors, and identifying structure in asset returns can provide crucial insights for both theoretical and applied finance.

In this project, we study a multivariate dataset of U.S. equity returns, aiming to uncover hidden regularities and detect unusual patterns that may suggest structural shifts or market anomalies. The high dimensionality and inherent noise in financial data make unsupervised learning methods particularly appealing, as they can extract patterns without requiring labeled data.

To work towards this goal, we compare the effectiveness of various unsupervised learning techniques in revealing the underlying structure in U.S. equities. Specifically, we analyze the performance of Principal Component Analysis (PCA), non-linear PCA, K-means clustering, and autoencoders. These methods are used to reduce dimensionality and group assets based on latent factors that may influence return co-movement.

Additionally, we investigate anomaly detection by applying the Isolation Forest algorithm to identify outliers in the return distributions. We compare its performance against traditional univariate techniques, such as z-score filtering and the interquartile range (IQR) method, to assess whether the nuanced machine learning approach offers a more insightful understanding of financial anomalies.

Through this analysis, we aim to evaluate whether modern unsupervised learning methods can supplement or even outperform traditional statistical tools in understanding complex financial systems.

Preprint. Work in progress.

2 Related Work

Dimensionality reduction techniques have long been used in finance to extract latent factors that explain asset return variation. The most classical approach is Principal Component Analysis (PCA), which linearly decomposes return matrices into orthogonal components that explain the maximum variance. PCA has been widely adopted in applications such as risk modeling, portfolio construction, and factor investing (Connor & Korajczyk, 1986).

However, PCA is inherently linear and may fail to capture complex, nonlinear relationships present in modern financial data. To address this limitation, recent literature has proposed using neural network-based autoencoders as nonlinear generalizations of PCA (Hinton & Salakhutdinov, 2006). Autoencoders are trained to compress and reconstruct data through a bottleneck architecture, learning latent representations that may encode nonlinear structure in returns. These have shown promise in fields like quantitative trading, market regime detection, and latent factor extraction.

In addition to factor extraction, unsupervised learning methods such as K-Means clustering have been applied to group assets based on latent features. Some recent works (e.g., Sirignano & Cont, 2019) apply deep learning architectures to high-dimensional market data, while others propose using variational autoencoders (VAEs) or denoising autoencoders to improve factor robustness and interpretability (Heaton, Polson & Witte, 2017). Additionally, comparisons between PCA and nonlinear methods often reveal trade-offs between interpretability, training complexity, and reconstruction accuracy.

Nanda, Mahanty, and Tiwari used K-means clustering to minimize risk in a stock portfolio through diversification based on stock returns for stocks from the Bombay Stock Market 2007-2008. They found that the utilization of clustering algorithms allowed for better diversification of stocks in a portfolio by enhancing diversification criteria. Their study shows that k-means "formed well compact clusters compared to Fuzzy C-means and SOM neural network" (Nanda, 2010), which were the other clustering methods used in the study. However, k-means clustering struggles to capture efficient clusters if the underlying market structure contains customer or market segments that have complex patterns rather than simple ones.

In a similar study, Baser and Saini compared the performance of K-means, K-medoids, and fast K-means based on intra-class inertia using data from the Nifty 50 stock market index for the purpose of building diversified stock portfolios. The results of the study show that the compactness of K-means clusters is much more stable than the compactness of K-medoids or fast K-means as k changes.

Rather than drawing insights from k-means, Khoo, Pathmanathan, and Dabo-Niang focused on using PCA and FASPCA to find spatial autocorrelation of stock exchange returns during the 2015-16 global market sell-off. Compared to classical PCA, FASPCA combines PCA with functional data analysis (FDA) to reduce the dimensionality of data with a spatial component. This allows the spatial relationship in the data to be analyzed in the dimensionality reduction. The results of the study found that FASPCA performed better than classical PCA at explaining variability in spatial stock data.

Choi, Gwak, Song, and Chang used bi-dimensional histograms and autoencoders to analyze daily stock trading prices and volumes in the S&P500 from 2006-2019. The stock prices and volumes are plotted using the bi-directional histogram and the autoencoder reduces the dimensionality of the data and extracts meaningful features by obtaining a latent vector that serves as a compressed representation of the histogram. The results of the study show that the latent vector produced by the autoencoder contained less noise/missing values than the histogram. The histogram network generated from decoding the latent vector was found to be more assortative than a traditional price network and the portfolio derived from the network outperformed other benchmarks.

In this work, we build on these ideas by comparing PCA and autoencoder-derived latent factors, evaluating their effectiveness not only in explaining return variance, but also in facilitating downstream tasks such as clustering via K-Means and outlier detection through reconstruction loss. Our findings suggest that while PCA maintains advantages in linear regimes, autoencoders begin to offer competitive—and at times superior—performance as model complexity increases, particularly in capturing nuanced structure across assets.

3 Methodologies

3.1 Data Collection and Preprocessing

We gathered historical financial data using the Yahoo Finance `yfinance` Python library, focusing on the S&P 500 index (`^GSPC`) and individual tech stocks over 12–13 years from 2012 to 2025. The data consisted of adjusted daily close prices, which we transformed into daily returns. These returns were standardized using z-score normalization via the `StandardScaler` module from `scikit-learn` to ensure all input features were on the same scale prior to model training and analysis.

To expand coverage beyond the S&P 500, we accessed daily OHLCV (Open, High, Low, Close, Volume) stock data via the Alpha Vantage `TIME_SERIES_DAILY` API. This API provides raw (unadjusted) daily time series data for global equities and supports full historical coverage of 20+ years. We queried individual symbols (e.g., META, AAPL) and retrieved full-length time series in JSON format, which we then converted to daily returns. While the API returns raw prices, we focused primarily on the daily closing price for return calculation. Returns were standardized before feeding them into dimensionality reduction pipelines.

3.2 Anomaly Detection Techniques

3.2.1 Traditional Methods

We first implemented classical statistical methods for anomaly detection, including Z-score and Interquartile Range (IQR).

The **Z-score** measures how many standard deviations a data point x is from the mean μ:

$$Z = \frac{x - \mu}{\sigma}$$

Observations with $|Z| > 3$ are typically flagged as outliers under the assumption of normality.

The **Interquartile Range (IQR)** method flags values outside the range:

$$[\text{Q1} - 1.5 \cdot \text{IQR}, \quad \text{Q3} + 1.5 \cdot \text{IQR}]$$

where $\text{IQR} = \text{Q3} - \text{Q1}$. This method is robust to skewed distributions but still fails to capture multivariate structure.

3.2.2 Machine Learning Methods

To address the limitations of univariate approaches, we implemented **Isolation Forests**, a tree-based ensemble method that identifies anomalies by recursively partitioning data and measuring the path length required to isolate a point. Isolation Forests do not rely on assumptions of normality and are effective in high-dimensional settings. They also implicitly account for correlations between features, enabling more reliable detection of multivariate anomalies.

3.3 Dimensionality Reduction for Unsupervised Learning

3.3.1 Limitations of PCA

Traditional **Principal Component Analysis (PCA)** identifies directions of maximum variance in the data via eigenvalue decomposition of the covariance matrix. However, it is inherently linear and sensitive to noise and outliers, making it less suitable for financial data with nonlinear dependencies.

We evaluate PCA and its alternatives using two metrics: the coefficient of determination R^2 and the Mean Squared Error (MSE).

The **coefficient of determination** is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the actual return, $\hat{y}_i$ is the predicted return from the model, and $\bar{y}$ is the mean of the actual returns.

The **Mean Squared Error (MSE)** is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

These metrics allow for a direct comparison between PCA and autoencoder-based reconstructions.

3.3.2 Nonlinear Alternatives: Autoencoders

To overcome PCA's limitations, we explored **autoencoders**, which are neural networks trained to reconstruct input data via a low-dimensional bottleneck. This approach is data-driven and can capture complex nonlinear relationships. Our architecture consisted of an encoder and decoder with a single hidden layer of 128 neurons. The latent factors learned from the bottleneck were used as nonlinear analogs to principal components and evaluated against PCA using the R^2 and MSE metrics described above.

4 Data Exploration

4.1 Daily Return Distributions

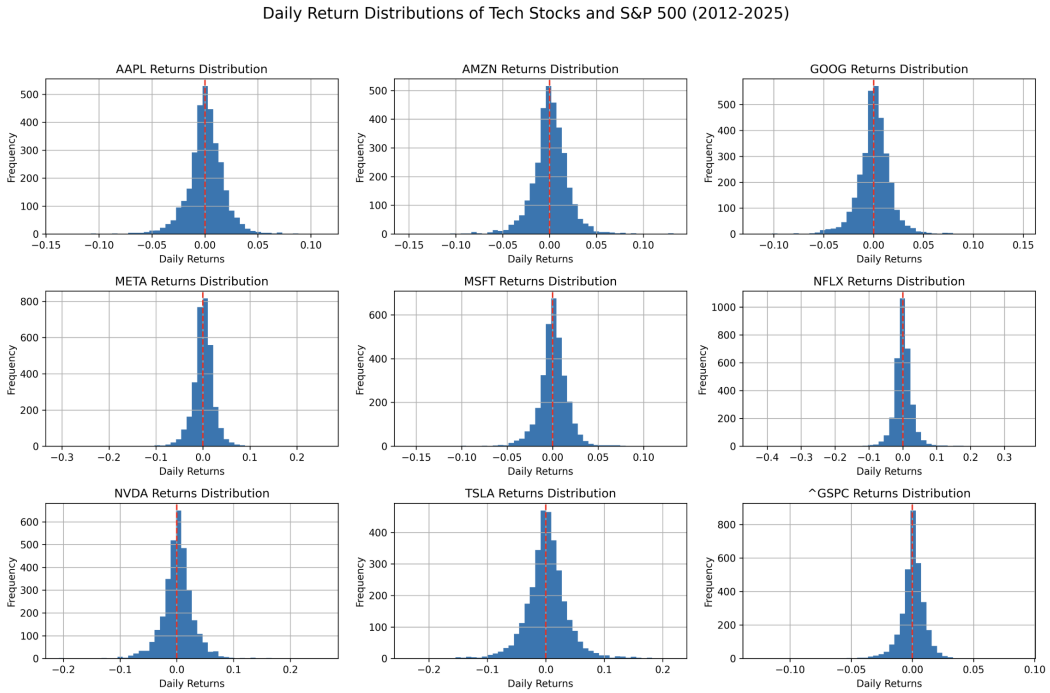


Figure 1: Daily Return Distributions of Tech Stocks and S&P 500

Our dataset comprises 13 years (2012-2025) of daily returns for the S&P 500 index and eight major tech stocks (AAPL, AMZN, GOOG, META, MSFT, NFLX, NVDA, TSLA), with 3,240 observations per security and no missing values. Initial analysis reveals that all stocks exhibit positive mean daily returns, with NVDA showing the highest and the S&P 500 the lowest, demonstrating how tech significantly outperformed the broader market. TSLA displays the highest volatility, over three times that of the S&P 500. The return distributions, as shown in the histograms (Figure 1), are characteristically leptokurtic with more concentration around the mean (taller peaks) and fatter tails than normal distributions would have, a pattern known as excess kurtosis in financial markets that

suggests a higher probability of extreme price movements. All distributions are centered close to zero (as indicated by the red dashed line in each histogram), aligning with the efficient market hypothesis that predicts stock returns should center around zero in the short term. This non-normality, combined with varying volatility profiles across securities, supports our approach of applying dimensionality reduction techniques to uncover latent structures that linear methods might miss, particularly during market stress periods when correlations often shift dramatically.

4.2 Correlation Analysis

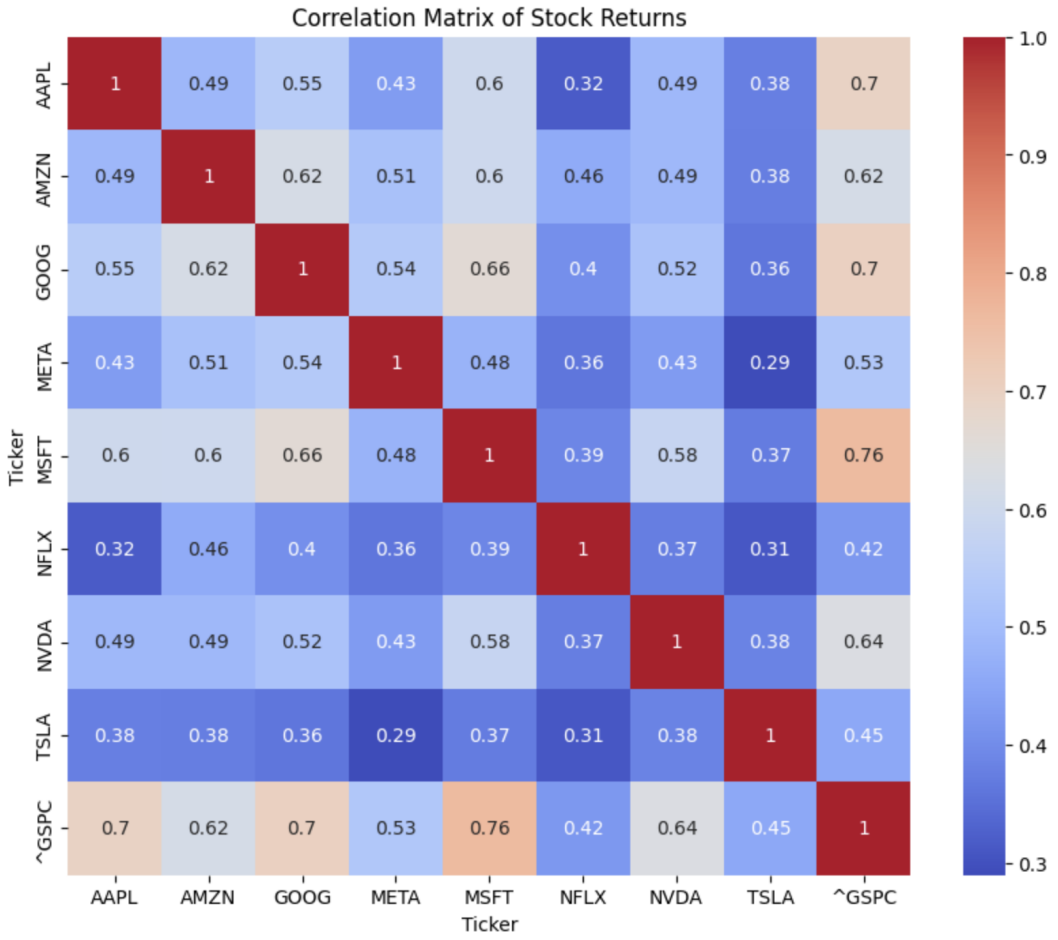


Figure 2: Correlation matrix of daily log returns across tech stocks and the S&P 500 index

Figure 2 presents the correlation matrix of daily log returns across our sample of tech stocks and the S&P 500 index. The matrix reveals predominantly positive and moderate-to-strong correlations (0.3-0.7 range), indicating these securities generally move together. Microsoft demonstrates the strongest relationship with the broader market (0.76 correlation with S&P 500), followed closely by Apple and Google (both at 0.70). In contrast, Tesla and Netflix display notably lower correlations with other stocks, suggesting their returns are driven more by company-specific factors than market-wide movements. This varying correlation structure creates natural groupings: high market sensitivity stocks (MSFT, AAPL, GOOG), moderate sensitivity (NVDA, AMZN), and lower sensitivity (META, NFLX, TSLA) - a pattern that supports our approach of using dimensionality reduction to uncover latent market factors.

4.3 Cumulative Performance Analysis

Figure 3 illustrates the cumulative returns of our selected stocks from 2012 to 2025, revealing dramatic performance disparities. NVIDIA demonstrates extraordinary growth, particularly after

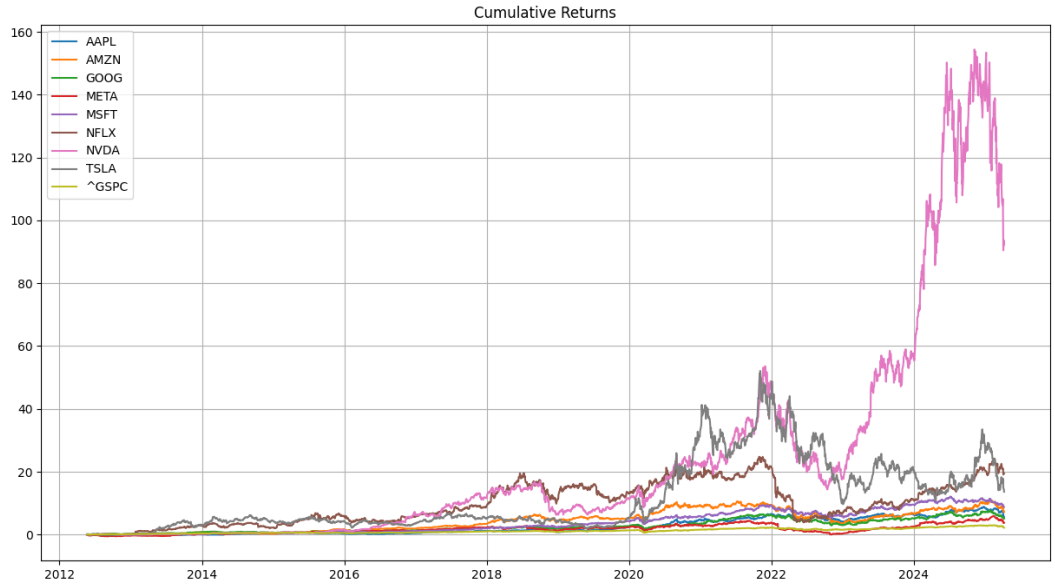


Figure 3: Cumulative returns of selected tech stocks and S&P 500 index

2020, with returns that dramatically outpace all other securities in our sample. Tesla follows as a distant second, though with extreme volatility including a dramatic surge during 2020-2021 followed by significant declines. The remaining stocks form distinct performance tiers: solid performers (AMZN, NFLX, AAPL, MSFT), moderate performers (GOOG), and relative laggards (META) that barely outperform the S&P 500 benchmark. These substantial differences in long-term performance, despite the moderate correlations observed in daily returns (Figure 2), suggest that while these stocks often move together day-to-day, their long-term trajectories are shaped by company-specific factors that accumulate significantly over time.

4.4 Basic outlier analysis

We also conducted basic outlier analysis by calculating z-scores for each daily return and identifying extreme movements exceeding 3 standard deviations. The results reveal Netflix and META as responsible for the most extreme single-day price shocks, with Netflix's 43.26% drop in April 2022 representing the largest outlier (z-score: 14.70) due to its first subscriber loss in a decade (1). META's 30.64% plunge in February 2022 followed, while the S&P 500's 12.77% drop in March 2020 occurred during the COVID-19 market crash. Notably, Tesla experienced the highest frequency of outlier days (55), while NVIDIA had the fewest (39), suggesting Tesla's returns are characterized by frequent extreme movements while NVIDIA's exceptional performance came through more consistent growth.

4.5 Autocorrelation Analysis

Figure 4 displays the autocorrelation function (ACF) of S&P 500 daily returns, which provides insight into temporal dependencies in the market. The plot reveals a statistically significant negative autocorrelation at lag 1, indicating a mild tendency for returns to reverse direction from one day to the next - when the market rises one day, it has a slight tendency to fall the following day, and vice versa. This pattern suggests short-term mean reversion in the S&P 500. Beyond this one-day effect, returns largely behave as a random walk, with autocorrelations at longer lags falling within the confidence bands (blue shaded area), consistent with the efficient market hypothesis that past returns have limited predictive power for future returns.

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

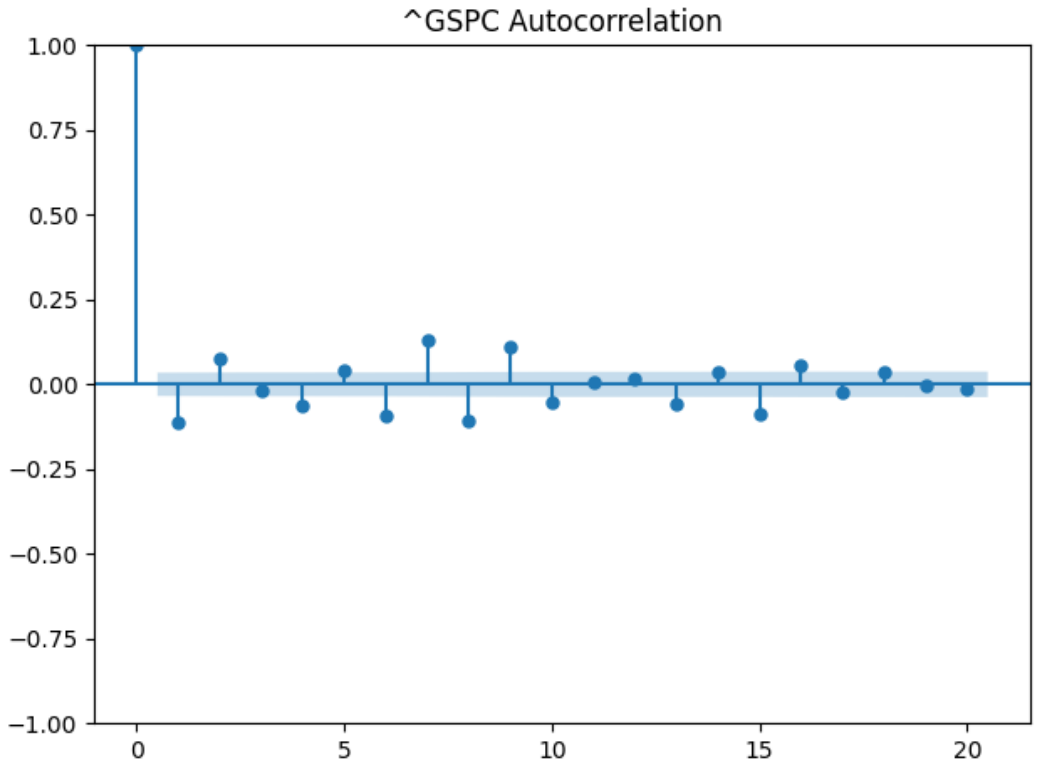
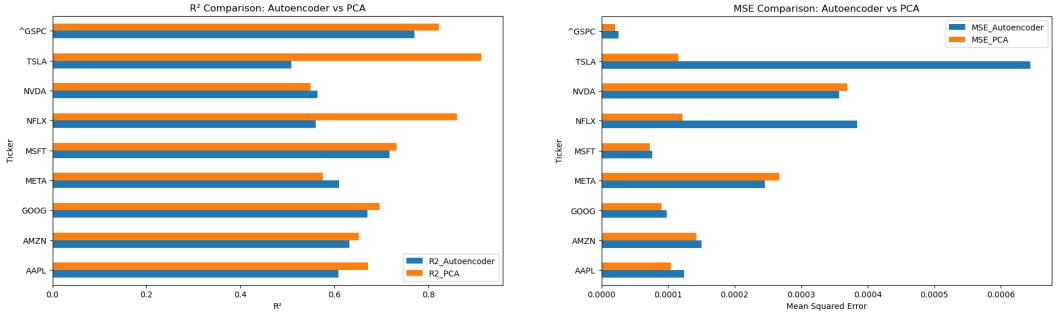


Figure 4: Autocorrelation function (ACF) of S&P 500 daily returns

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).



(a) R^2 scores across assets.

(b) Mean Squared Error (MSE) across assets.

Figure 5: Performance comparison between PCA and Autoencoder based on R^2 and MSE metrics.

The autoencoder demonstrated a substantial improvement in R^2 relative to its earlier 64-neuron version, suggesting enhanced explanatory power for reconstructing asset returns. 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 successfully captured much of the same underlying return structure.

Despite these gains, PCA continued to outperform the autoencoder for certain assets—most prominently GSPC and TSLA—highlighting PCA's continued advantage in modeling highly linear or benchmark-like return profiles.

Evaluation through the MSE metric yielded consistent results. While PCA generally achieved lower reconstruction error across most assets, the autoencoder exhibited a clear reduction in MSE relative to its smaller-capacity variant. This improvement reflects the model's enhanced ability to capture nonlinear relationships in the data. Although PCA retained an edge in terms of reconstruction efficiency, the narrowing gap in MSE between the two methods suggests that the autoencoder is approaching PCA's performance, particularly as its model capacity increases.

Collectively, these findings suggest that while PCA remains a strong and efficient baseline—especially in datasets dominated by linear correlations—autoencoders present a competitive and increasingly viable alternative. As model capacity is scaled up, autoencoders become better suited to capturing complex, nonlinear relationships within financial return data. These preliminary results validate the feasibility of neural factor models and motivate further exploration of deeper, regularized architectures capable of surpassing PCA in more structurally complex environments.

6 Clustering on Latent Features

6.1 Clustering on Latent Features

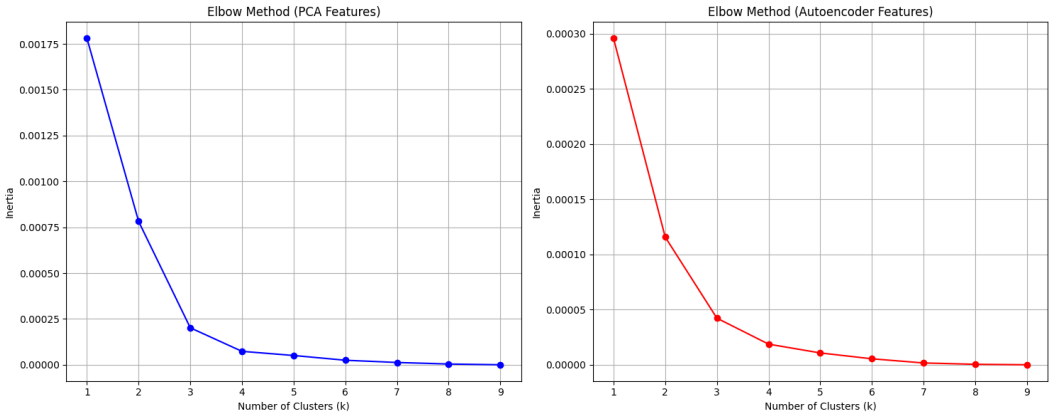


Figure 6: Results for latent feature clustering using the elbow method

Next, we performed clustering on the latent features extracted from both PCA and autoencoder methods to identify natural groupings in our stock market data. Using the elbow method, we plotted inertia against different numbers of clusters to find the optimal number of clusters where adding more clusters yields diminishing returns. This step was crucial for discovering which stocks behave similarly based on their latent factor exposures rather than relying on predefined sector classifications. The analysis revealed that three clusters optimally describe both our PCA and autoencoder embeddings, suggesting that despite their different mathematical approaches, both methods capture similar underlying market structures among these tech stocks.

Following our clustering analysis using the elbow method, we applied K-means clustering (with k=3) to both our PCA and autoencoder factor exposures to identify natural groupings among the stocks. This step goes beyond simple dimensionality reduction by organizing stocks into groups based on similar behavior in the latent space. While factor analysis tells us how stocks relate to underlying market drivers, clustering reveals which stocks move together based on these relationships, potentially uncovering market segments that traditional sector classifications might miss. The results reveal interesting differences between the linear and nonlinear approaches: In the PCA results, most major tech stocks (AAPL, AMZN, GOOG, META, MSFT, NVDA) cluster with the market index (GSPC) in Cluster 0, while TSLA and NFLX each form their own singleton clusters (Clusters 1 and 2 respectively). This suggests that from a linear perspective, most tech stocks behave similarly, while TSLA and NFLX have distinctive return patterns. In contrast, the autoencoder clustering identifies different relationships: META and NVDA form their own cluster (Cluster 0), most large-cap tech

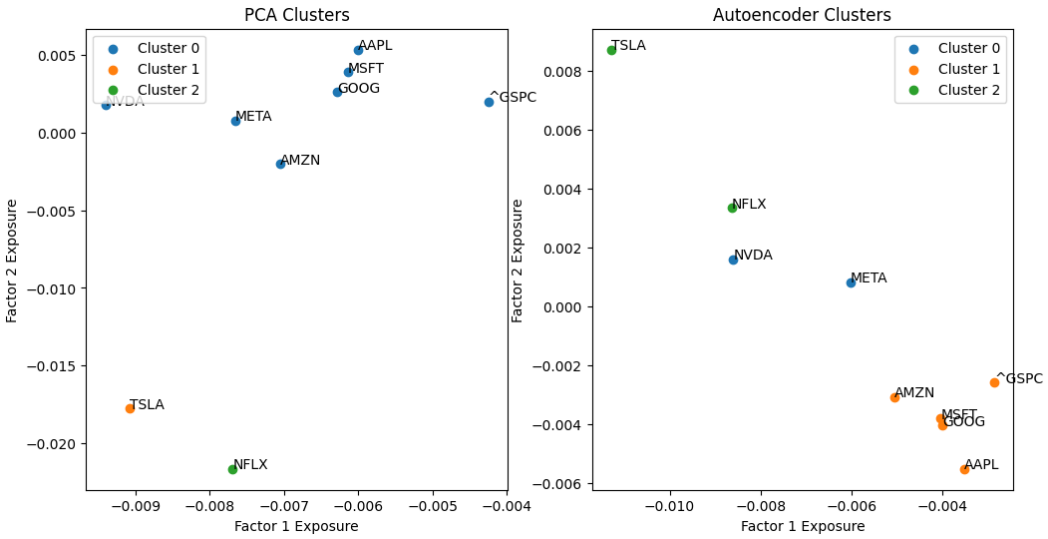


Figure 7: Comparison of K-means clustering results on PCA and autoencoder factor exposures

stocks (AAPL, AMZN, GOOG, MSFT) plus the market index form Cluster 1, and NFLX and TSLA are grouped together in Cluster 2. This suggests the nonlinear autoencoder is capturing subtler relationships between stocks that the linear PCA approach misses, particularly similarities between NFLX and TSLA that weren't apparent in the linear analysis. 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 investigated the effectiveness of nonlinear autoencoders compared to classical Principal Component Analysis (PCA) in modeling latent factors within financial return data. Through a series of experiments involving dimensionality reduction, clustering, and anomaly detection, we evaluated both methods on their ability to reconstruct returns, capture latent structure, and group assets meaningfully.

Our key findings are as follows:

- Autoencoders demonstrated competitive performance with PCA in terms of R^2 and Mean Squared Error (MSE), particularly for assets exhibiting nonlinear patterns.
- Clustering on autoencoder-derived latent features revealed nuanced groupings that PCA did not, suggesting the presence of relationships in return behavior that linear methods may overlook.
- Traditional outlier detection techniques (e.g., Z-scores, IQR) were effective in capturing univariate extremes, while Isolation Forests—augmented by reconstruction error—provided additional insights into multivariate anomalies.

Based on these results, we do not reject our hypothesis. Autoencoders provide a viable and sometimes superior alternative to PCA for uncovering complex structures in financial time series data. While PCA remains efficient and interpretable in linear regimes, autoencoders are better suited for capturing nonlinear co-movements and hidden dynamics as model capacity increases. These results motivate further exploration into deeper and more regularized architectures for financial factor modeling and anomaly detection.

References

Below are all resources that were used.

[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] Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, 19(9), 1449–1459. https://arxiv.org/abs/1803.06917

[4] 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

[5] Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12), 8793–8798. https://doi.org/10.1016/j.eswa.2010.06.026

[6] 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

[7] Khoo, Tzung Hsuen, et al. "Spatial Autocorrelation of Global Stock Exchanges Using Functional Areal Spatial Principal Component Analysis." *MDPI*, Jan. 2023, https://doi.org/10.3390/math11030674.

[8] Choi, Sungyoon, et al. "Stock market network based on bi-dimensional histogram and autoencoder." *Sage Journals*, May 2022, https://doi.org/10.3233/IDA-215819.