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

Jason Yi, Daniel Wang, Nicholas Do, Anish Parepalli, Krish Patel

Overview & Problem Statement

- Traditional methods – Principal Component Analysis (PCA) assume linear relationships between stocks
- Real markets often behave nonlinearly – especially during crises and volatility

=> Can deep learning uncover hidden structures in financial time series that PCA misses?

GOALS:

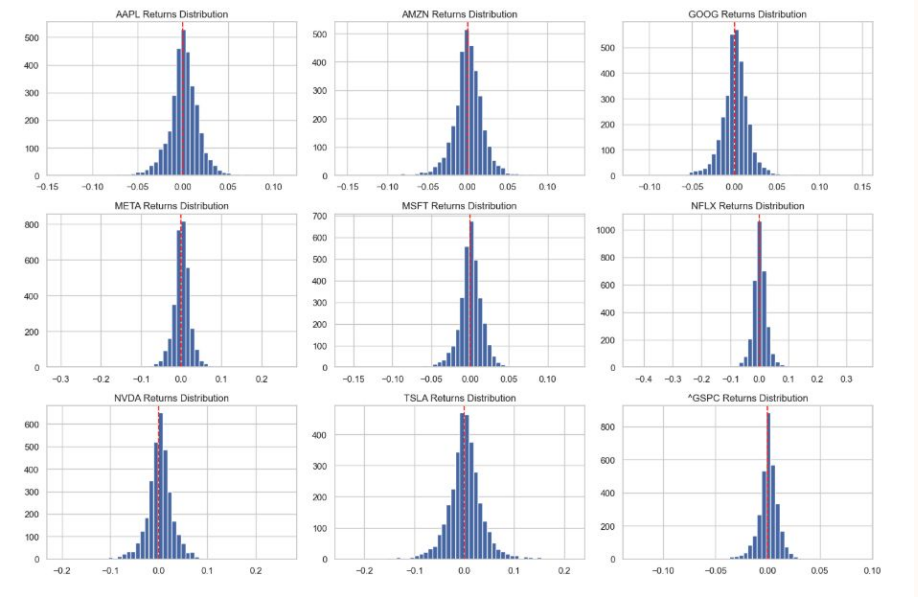
- Compare PCA and autoencoders for extracting latent financial factors
- Evaluate reconstruction performance using R^2 and Mean Squared Error
- Analyze latent factor structure through K-means clustering

Data & Methods

- 13 years (2012–2025) of daily returns
- S&P 500 index + 8 major tech stocks (AAPL, NVDA, TSLA, META, etc.)
- Real-world market conditions including COVID crash, tech boom, etc.

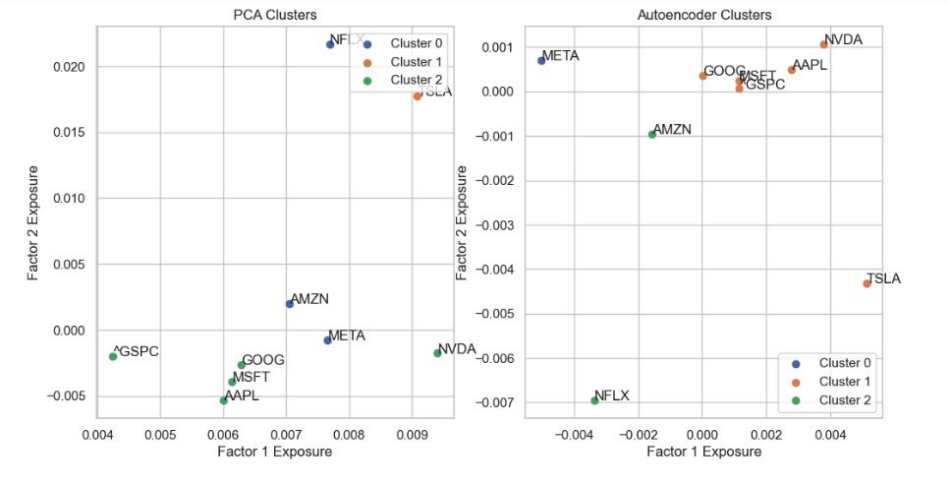
Methods

- PCA (Linear Baseline): Linear dimensionality reduction—finds directions of maximum variance
- Autoencoder (Nonlinear): Encoder–decoder neural network with 128 hidden neurons & a low-dimensional bottleneck
- Clustering: K-means with (k = 3) applied separately to PCA and autoencoder latent factors
- Evaluation: R^2, MSE



Results & Impacts

- PCA is a strong baseline, especially for more linear or index-like series
- Autoencoders achieves competitive R^2 and MSE, matched/exceeded PCA for NVDA, AAPL
- K-means on autoencoder factors reveals different groupings, suggesting the nonlinear embedding captures relationships that PCA does not:
  - TSLA with NFLX (both high-volatility disruptors)
  - META with NVDA (platform + infrastructure)



=> **Applications:** Portfolio Construction, Risk Management, Algorithmic Trading

THANK YOU!