1 Methodology and Process

1.1 Ledoit–Wolf Correlation-Based Clustering

The first clustering approach relies on the linear dependence structure of asset returns. Let $R \in \mathbb{R}^{T \times N}$ denote the matrix of daily returns, where T is the number of time periods and N is the number of assets. To obtain a stable estimate of the covariance matrix in the presence of limited observations and high dimensionality, we use the Ledoit–Wolf shrinkage estimator as implemented in the `scikit-learn` library. This estimator automatically determines an optimal shrinkage intensity and returns a regularized covariance matrix $\hat{\Sigma}_{LW}$, which reduces estimation noise while retaining essential correlation structure.

The corresponding correlation matrix is computed as

$$\rho_{ij} = \frac{\hat{\Sigma}_{LW,ij}}{\sqrt{\hat{\Sigma}_{LW,ii}\hat{\Sigma}_{LW,jj}}},$$

where ρ_{ij} measures the degree of linear co-movement between assets i and j. The correlation matrix is then transformed into a distance matrix

$$D_{ij} = \sqrt{2(1 - \rho_{ij})},$$

so that more highly correlated assets are closer in distance space. Hierarchical agglomerative clustering with Ward linkage is then applied to D to form groups of assets that exhibit similar correlation structures. This method captures relationships driven primarily by linear dependencies in return dynamics and serves as a baseline for comparison with nonlinear approaches.

1.2 Signature-Based Clustering

The second clustering framework leverages the rough path *signature* representation to capture nonlinear and temporal characteristics of asset return trajectories. Unlike correlation-based methods, which summarize linear co-movements, this approach embeds the entire path evolution of each asset into a high-dimensional feature space derived from iterated integrals.

Each asset's cumulative return path is first constructed as

$$P_t^i = \prod_{\tau \leq t}(1 + r_\tau^i),$$

and normalized to start at one to ensure comparability across assets. To preserve the temporal ordering of price movements, we apply a *lead–lag transformation*, which maps each one-dimensional time series into a two-dimensional path that encodes both the current and lagged values. This transformation enhances the representation of sequential structure, making it suitable for computing path signatures.

We then compute the truncated path signature of each transformed trajectory using the `iisignature` library. For a continuous path $X : [0, T] \to \mathbb{R}^d$, the signature of order m is defined as the sequence of all iterated integrals:

$$S^{(m)}(X) = \left(1, \int dX_t, \int dX_{t_1} \otimes dX_{t_2}, \ldots, \int_{0<t_1<\cdots<t_m<T} dX_{t_1} \otimes \cdots \otimes dX_{t_m}\right).$$

In our implementation, we truncate the expansion at depth $m = 4$, balancing representational richness and computational tractability. Each signature is then flattened into a feature vector, forming a feature matrix

$$\mathbf{S} = [S^{(4)}(X^1), S^{(4)}(X^2), \ldots, S^{(4)}(X^N)]^\top,$$

where each row corresponds to one asset.

We perform agglomerative hierarchical clustering with Ward linkage directly on the Euclidean distances in this signature feature space. This groups assets whose return trajectories exhibit similar temporal shapes and nonlinear behaviors, even if their linear correlations differ. Visualization through t-SNE projection of the signature features provides geometric insight into the structure of these clusters, while cumulative return analysis reveals how each cluster evolves over time.

2 Experiments

2.1 First Experiment: Optimal Cluster Selection based on Silhouette Score

We conducted an initial experiment to compare the Ledoit–Wolf correlation-based clustering and the signature-based clustering methods using ten years of daily return data from 2015 to 2025. For both approaches, the optimal number of clusters was determined using the silhouette coefficient, which evaluates clustering performance by balancing intra-cluster cohesion and inter-cluster separation.

The silhouette analysis indicated that both methods achieved their highest scores when the number of clusters was set to two. This suggests that, under both linear correlation and path-signature representations, the asset universe naturally partitions into two dominant regimes. However, the structural characteristics of the clusters produced by each method differ substantially.

The Ledoit–Wolf correlation-based clustering formed two stable and economically interpretable groups, reflecting similarity in linear co-movement and persistent correlation patterns among assets. By contrast, the signature-based clustering—though designed to capture nonlinear and temporal dependencies—did not yield an ideal partition in this initial experiment. The clusters derived from the signature features appeared less distinct and exhibited weaker separation in terms of return trajectories and sectoral alignment.

This result implies that, while rough path signatures provide a richer representation of time series dynamics, their direct use with Euclidean distance

in hierarchical clustering may not fully exploit the expressive capacity of the signature space.

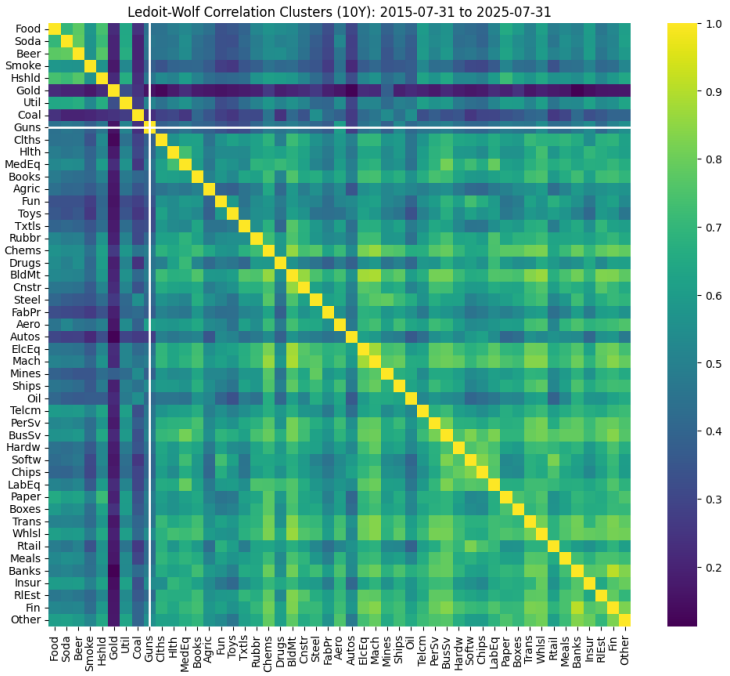
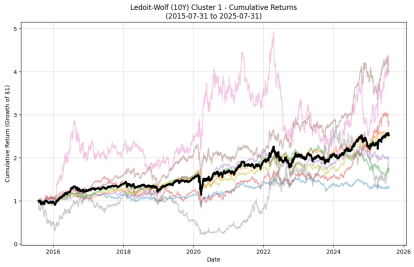
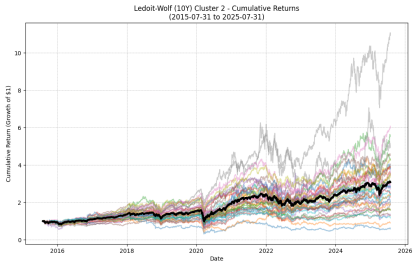


Figure 1: Ledoit–Wolf correlation matrix between industry portfolios.



(a)Cumulative returns of assets within the first cluster and their average.



(b)Cumulative returns of assets within the second cluster and their average.

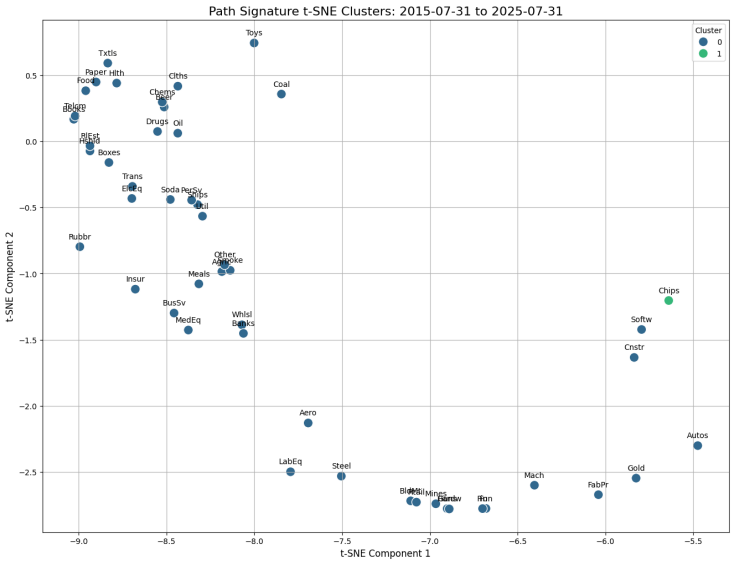
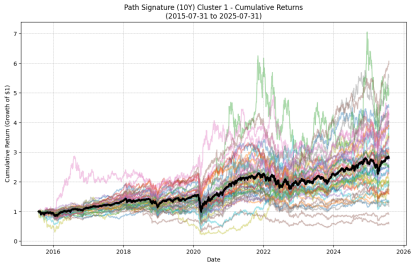
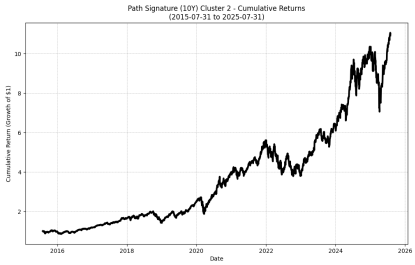


Figure 2: Signature-based industry portfolio clusters.



(a)Cumulative returns of assets within the first cluster and their average.



(b)Cumulative returns of assets within the second cluster and their average.