**Advanced Methodologies for Clustering Macroeconomic Variables and Identifying Economic Regimes**

**1.0 Introduction**

**1.1 Context and User Objective**

The identification and analysis of distinct economic regimes constitute a critical task in quantitative finance, informing strategic asset allocation, risk management, and economic forecasting. This report addresses the objective of classifying macroeconomic conditions into specific regimes—namely Recession, Expansion, Slowdown, Recovery, and High Inflation—utilizing clustering techniques. The input variables considered for this task include factors derived from the bond yield curve (specifically, level, slope, and curvature, denoted d1, d2, d3), Vanguard's Leading Economic Indicator (LEI), and equity market indicators such as the Cyclically Adjusted Price-to-Earnings (CAPE) ratio and the equity payout ratio. The current methodology employed involves k-means clustering based on Euclidean distance, with all input variables being equally weighted.

**1.2 The Need for Methodological Exploration**

While the k-means algorithm with Euclidean distance provides a straightforward approach to partitioning data, its application to complex macroeconomic variables warrants careful consideration of its inherent limitations. Macroeconomic time series often exhibit non-linear relationships, time-varying volatility, and non-spherical correlation structures that may not align well with the assumptions underpinning standard k-means.1 The algorithm's sensitivity to outliers and the initial placement of cluster centroids can impact the stability and robustness of the resulting regime classification.3 Furthermore, the use of Euclidean distance implicitly assumes that variables are independent and equally scaled, which is often not the case for economic indicators that possess inherent correlations and differing units or volatilities.4 The equal weighting of variables may also obscure the potentially differential importance of certain indicators in defining specific economic states. Exploring alternative clustering algorithms, distance metrics, and variable weighting schemes is therefore a crucial step towards enhancing the accuracy, robustness, and economic interpretability of the identified regimes.

**1.3 Addressing the Vanguard Reference (Zhang & Ahluwalia)**

The work of Yu Zhang and Harshdeep Ahluwalia at Vanguard has been referenced in the context of this analysis.6 It is important to clarify the scope of their research as presented in the available materials. The provided documentation primarily details their significant contributions to the field of portfolio rebalancing, particularly for Target Date Funds (TDFs).8 Their work focuses on optimizing rebalancing policies (e.g., threshold-based versus calendar-based) by balancing transaction costs against deviations from target asset allocations, often employing utility-maximization frameworks and sophisticated modeling of returns and costs.8 Techniques mentioned include regression-based Monte Carlo simulations and modeling volatility clustering.8 While a specific publication by Zhang and Ahluwalia focusing directly on the clustering of macroeconomic variables for *regime identification* was not identified within the provided sources, the rigorous quantitative approach they employ is highly relevant. Methodologies potentially used in broader Vanguard research, such as the Vanguard Capital Markets Model (VCMM) for simulating asset returns and risks 10, often involve sophisticated statistical techniques, including factor models and potentially regime-based analysis, aligning with the spirit of exploring advanced methods discussed in this report.

**1.4 Report Scope and Structure**

This report provides a technical overview and comparison of advanced quantitative methodologies relevant to clustering macroeconomic variables for regime identification. The analysis aims to equip practitioners with the knowledge to refine their existing frameworks by considering alternatives to standard k-means and Euclidean distance.

The report is structured as follows:

* **Section 2.0** explores alternative clustering algorithms beyond k-means, including Hierarchical Clustering, Gaussian Mixture Models (GMM), density-based methods, and model-based approaches like Hidden Markov Models (HMM).
* **Section 3.0** delves into advanced distance and similarity measures suitable for economic and financial time series, examining correlation-based distances, Dynamic Time Warping (DTW), Mahalanobis distance, and others.
* **Section 4.0** addresses strategies for variable weighting and assessing feature importance, covering Principal Component Analysis (PCA), feature-based clustering, and explicit feature weighting algorithms.
* **Section 5.0** synthesizes the findings, providing guidance on method selection and specific recommendations tailored to the user's objective and dataset, including considerations for validation and determining the appropriate number of regimes.
* **Section 6.0** concludes the report, summarizing the key takeaways.

The overall aim is to furnish a comprehensive technical resource that facilitates informed decisions in the development and enhancement of macroeconomic regime identification models.

**2.0 Exploring Alternative Clustering Algorithms for Macroeconomic Regimes**

The choice of clustering algorithm significantly influences the resulting regime classification. While k-means is widely used, its limitations in the context of macroeconomic data motivate the exploration of alternative techniques that offer greater flexibility or incorporate different assumptions about the data structure.

**2.1 Limitations of K-Means in Macroeconomic Context**

The k-means algorithm is a partitioning method that aims to divide N observations into k pre-specified clusters by minimizing the within-cluster sum of squares, often referred to as inertia.13 Each cluster is represented by its centroid, which is the mean of the observations within that cluster.13 This centroid may not correspond to an actual data point.15

However, several characteristics of k-means can be disadvantageous when applied to macroeconomic regime identification:

1. **Sensitivity to Initialization and Outliers:** The final clustering solution can depend on the initial placement of centroids, and the algorithm can be sensitive to outliers, which might unduly influence centroid positions.3
2. **Assumption of Spherical Clusters:** K-means implicitly assumes that clusters are spherical and have roughly equal variance.13 Macroeconomic regimes, when represented in a multi-dimensional feature space, may exhibit more complex, potentially elliptical shapes, violating this assumption.1
3. **Hard Assignments:** K-means assigns each data point to exactly one cluster.13 Economic transitions, however, are often gradual, and a specific point in time might exhibit characteristics of multiple regimes. K-means' "hard" assignments fail to capture this ambiguity or the fuzzy boundaries that might exist between states like 'Slowdown' and 'Recession'.1
4. **Euclidean Distance Default:** The standard implementation relies on Euclidean distance, which treats all dimensions (variables) independently and equally, ignoring potential correlations and differences in scale or volatility among macroeconomic indicators.4 While standardization addresses scale issues, it does not account for correlation structures.17

These limitations suggest that algorithms offering more flexibility in cluster shape, handling uncertainty through probabilistic assignments, or incorporating different distance metrics might yield more meaningful macroeconomic regimes.

**2.2 Hierarchical Clustering**

Hierarchical clustering algorithms build a nested sequence of partitions, represented visually as a dendrogram.15 Unlike k-means, they do not require the number of clusters, k, to be specified beforehand.19 The most common approach is agglomerative (bottom-up), where each observation starts as its own cluster, and pairs of clusters are iteratively merged based on a chosen linkage criterion until only one cluster remains.19

**Linkage Methods:** The core difference between hierarchical methods lies in how the distance between two clusters is defined:

* **Single Linkage:** Uses the minimum distance between any point in the first cluster and any point in the second cluster.21 It is known for its tendency to form elongated clusters (the "chaining" effect) and is closely related to the Minimum Spanning Tree (MST) algorithm.15
* **Complete Linkage:** Uses the maximum distance between any point in the first cluster and any point in the second cluster.21 This method tends to find more compact, roughly spherical clusters.
* **Average Linkage (UPGMA):** Calculates the average distance between all pairs of points, one from each cluster.21 It often provides a balance between the extremes of single and complete linkage.
* **Ward's Method:** Merges the pair of clusters that leads to the minimum increase in the total within-cluster variance (sum of squared errors).23 This method aims to produce clusters that are relatively compact and of similar size, and is often favored for creating well-separated groups suitable for interpretation via dendrograms.23

**Application & Suitability:** Hierarchical clustering is frequently applied to financial time series, particularly for grouping assets based on correlation patterns.19 Its primary advantage is the dendrogram, which allows visualization of cluster relationships at various levels of similarity. This hierarchical structure can be particularly insightful for understanding macroeconomic regimes. For instance, the way the user's defined regimes (Recession, Expansion, Slowdown, Recovery, High Inflation) group together at different levels in the dendrogram can reveal natural affinities or transition pathways. 'Slowdown' might merge with 'Recession' at a higher similarity level than 'Expansion' merges with 'Recovery', providing insights beyond a single flat partition like that produced by k-means.15 The choice of linkage method and the underlying distance metric significantly impacts the results.21 While powerful, traditional hierarchical methods can be computationally intensive, often exhibiting quadratic complexity with respect to the number of observations, which may be a concern for very large datasets.20

**2.3 Gaussian Mixture Models (GMM)**

Gaussian Mixture Models (GMMs) offer a probabilistic approach to clustering.1 The core assumption is that the observed data points are generated from a finite mixture of k Gaussian distributions.2 Each Gaussian component (cluster) is characterized by its mean vector (μk​), covariance matrix (Σk​), and a mixing coefficient (πk​) representing the prior probability or weight of that component in the overall mixture, where ∑k=1K​πk​=1.1 The probability density function of the mixture is given by:

p(x)=k=1∑K​πk​N(x∣μk​,Σk​)

where N(x∣μk​,Σk​) is the multivariate Gaussian density function for component k.

**Soft Clustering and Flexibility:** A key advantage of GMM is its ability to perform "soft" clustering.1 Instead of assigning each data point definitively to a single cluster (as in k-means), GMM calculates the posterior probability (or responsibility) that each data point belongs to each Gaussian component.1 This probabilistic assignment naturally handles uncertainty and is well-suited for situations where clusters overlap or where data points lie near boundaries, reflecting the often ambiguous nature of economic regime transitions.1 Furthermore, by allowing each component to have its own covariance matrix (Σk​), GMMs can model clusters with elliptical shapes and varying sizes and orientations, offering greater flexibility than the spherical assumption of k-means.1 Different constraints can be placed on the covariance matrices (e.g., spherical, diagonal, tied, or full) to control model complexity.29 A GMM with spherical covariance is closely related to k-means.16

**Estimation:** The parameters (μk​,Σk​,πk​) are typically estimated using the Expectation-Maximization (EM) algorithm.1 EM is an iterative procedure that alternates between an E-step (calculating the posterior probabilities or responsibilities for each point belonging to each component, given the current parameters) and an M-step (updating the parameters to maximize the expected log-likelihood, given the responsibilities calculated in the E-step). Initialization of the parameters (often done using k-means results) can influence the final solution.29

**Application & Suitability:** GMMs are widely applied in clustering, density estimation, and pattern recognition across various fields.1 They are explicitly used for identifying market or economic regimes in finance.2 The ability to model non-Gaussian features like fat tails and skewness, common in financial data, by combining multiple Gaussian components is a significant advantage.33 The probabilistic output allows for quantifying regime uncertainty; for instance, a specific time point might be assigned a 60% probability of being in a 'Slowdown' regime and a 40% probability of being in a 'Recession' regime.33 These probabilities can be directly incorporated into downstream tasks like risk management or dynamic asset allocation.36 While the number of components (k) needs to be specified, information criteria like the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) can aid in model selection.29 A potential drawback is the risk of singularities (infinite likelihood) if there are insufficient data points per component, necessitating regularization or constraints on the covariance matrices.29

**2.4 Density-Based Methods (e.g., DBSCAN)**

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) represents a different paradigm for clustering.39 Instead of partitioning the data based on minimizing distances to centroids (like k-means) or merging based on linkage criteria (like hierarchical), DBSCAN identifies clusters as dense regions of points separated by sparser regions.40 Key concepts include:

* **Core Points:** Points with at least a minimum number (min\_samples) of other points within a specified radius (epsilon).
* **Density-Reachability:** A point p is density-reachable from a core point q if p is within epsilon distance of q.
* **Density-Connectivity:** Two points are density-connected if they are mutually density-reachable from a common core point.

A cluster in DBSCAN is a set of density-connected points. Points not reachable from any core point are classified as noise or outliers.

**Application & Suitability:** DBSCAN's main strengths are its ability to discover clusters of arbitrary shapes (not limited to spherical or elliptical) and its robustness to noise, as it explicitly identifies outliers.40 It also does not require the number of clusters to be specified beforehand. While mentioned as a general clustering technique 14, its application specifically to macroeconomic regime identification appears less common in the reviewed literature compared to GMM or HMM. It could be suitable if there is a strong prior belief that economic regimes manifest as non-elliptical structures in the feature space, or if outlier detection (identifying highly unusual economic periods) is a primary goal. However, DBSCAN can be sensitive to the choice of its parameters (epsilon and min\_samples), and performance can degrade in high-dimensional spaces or with varying cluster densities.

**2.5 Model-Based Approaches (Markov Switching Models / HMM)**

Hidden Markov Models (HMMs) provide a powerful framework for modeling time series data that exhibits distinct regimes or states.2 The core idea is that the system transitions between a finite number of unobserved (hidden) states according to a Markov process, meaning the probability of transitioning to the next state depends only on the current state.41 The observed data (e.g., the macroeconomic variables) at any given time are generated from a probability distribution that depends on the hidden state active at that time.30

**Application & Suitability:** HMMs are naturally suited for modeling economic cycles and financial market regimes, as these systems often display periods of distinct behavior (e.g., high vs. low volatility, expansion vs. recession) with persistence and transitions.5 They explicitly model the dynamics of regime switching through the estimation of a transition probability matrix, which quantifies the likelihood of moving from one regime to another in the next time step.30 This contrasts with static clustering methods like k-means or GMM, which typically treat each time point independently based on feature similarity.33 The ability to forecast future regime probabilities conditional on the current state is a key advantage for applications like tactical asset allocation.38

Variations exist, such as Markov-switching Vector Autoregressions (MS-VAR), which model the dynamics within each regime using a VAR structure 39, or Gaussian HMMs, where the emission probabilities within each state are Gaussian.30 Estimation often involves algorithms like the Baum-Welch algorithm (a form of EM) or Markov Chain Monte Carlo (MCMC) methods, which can be computationally intensive.39 HMMs have been used, for example, to identify inflation regimes 45 and model yield curve dynamics at the zero lower bound.47

**2.6 Other Approaches Mentioned**

Several other clustering paradigms warrant mention:

* **Directed Bubble Hierarchical Tree (DBHT):** A novel hierarchical clustering method derived from graph theory, specifically utilizing the topological properties of Planar Maximally Filtered Graphs (PMFGs).21 It has shown promise in outperforming traditional linkage methods in retrieving industrial sector information from stock correlation data.21 Its applicability relies on having a suitable similarity or distance matrix (like correlation) and its performance on macroeconomic indicator clustering requires further investigation.
* **Fuzzy Clustering:** Methods like Fuzzy C-Means allow data points to have partial memberships in multiple clusters, represented by membership degrees.49 This provides a way to handle ambiguity similar to GMM's soft clustering but often based on different optimization criteria.
* **Ensemble Clustering:** This approach combines the results from multiple different clustering algorithms or multiple runs of the same algorithm with different initializations.28 The goal is to produce a more robust and stable consensus clustering, which can be particularly beneficial when dealing with noisy or complex data like financial time series.28 Techniques like feature weighting can be incorporated within ensemble frameworks.53

**Table 1: Comparison of Clustering Algorithms for Macroeconomic Regimes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Key Characteristics** | **Advantages** | **Disadvantages** | **Suitability for Macro Time Series** |
| **K-Means** | Partitioning, Hard Assignment, Assumes Spherical Clusters, Requires k | Simple, computationally efficient 13, scalable.13 | Sensitive to outliers & initialization 3, restrictive shape assumption 1, hard assignments ignore ambiguity.16 | Baseline method, but limitations often make alternatives preferable for complex economic data. |
| **Hierarchical (e.g., Ward)** | Hierarchical, Hard Assignment (at cut level), No Shape Assumption (depends on linkage/distance), Doesn't require k | Provides dendrogram for visualizing relationships 15, flexible linkage criteria 21, Ward finds compact groups.23 | Can be computationally expensive (O(N^2) or O(N^3)) 20, sensitive to distance/linkage choice 21, dendrogram interpretation complex for large N. | Widely used.19 Dendrogram useful for exploring regime relationships. Ward often preferred.23 Requires appropriate distance metric. |
| **Gaussian Mixture (GMM)** | Partitioning (probabilistic), Soft Assignment, Assumes Gaussian (Elliptical) Clusters, Requires k (can use BIC/AIC) | Probabilistic output handles uncertainty 1, models elliptical shapes 29, can model fat tails/skewness 33, principled framework (EM). | Can be sensitive to initialization 29, potential singularity issues 29, assumes Gaussian components. | Highly suitable. Explicitly used for regime detection.30 Probabilistic output valuable for finance. Handles non-normality well. |
| **DBSCAN** | Density-based, Hard Assignment (incl. noise), Finds Arbitrary Shapes, Doesn't require k | Finds non-globular shapes 40, robust to noise/outliers 40, doesn't need k specified. | Sensitive to parameters (epsilon, min\_samples), struggles with varying density clusters and high dimensions. | Less commonly cited for macro regimes. Potentially useful if regimes have complex shapes or outlier detection is key. Parameter tuning may be difficult. |
| **Hidden Markov (HMM)** | Model-based, Probabilistic Assignment, Models Regime Dynamics, Requires k states | Explicitly models regime transitions & persistence 41, allows forecasting regime probabilities 38, flexible state distributions (e.g., Gaussian HMM 30). | Can be complex to estimate (EM/MCMC) 42, assumes Markov property, requires specifying model structure (number of states, emission distributions). | Very suitable for modeling dynamic systems like economies.41 Directly addresses regime switching dynamics. Used for inflation 45 & yield curve 47 regimes. |

**3.0 Advanced Distance and Similarity Measures for Economic Time Series**

The choice of distance or similarity measure is fundamental to the outcome of any clustering algorithm, as it defines what constitutes "similarity" between observations. For macroeconomic variables, which often exhibit complex temporal dependencies, correlations, and varying scales, the standard Euclidean distance may be suboptimal.

**3.1 Limitations of Euclidean Distance**

Euclidean distance calculates the straight-line distance between two points in a multidimensional space: d(x,y)=∑i=1p​(xi​−yi​)2​. Its application in clustering implicitly assumes that:

1. **Independence:** The dimensions (variables) are uncorrelated.
2. **Equal Importance/Scale:** All dimensions contribute equally to the distance calculation.
3. **Point-wise Comparison:** For time series, it compares values at the exact same time points.20

These assumptions are often violated by macroeconomic data. Variables like yield curve level, slope, and curvature are inherently correlated. Indicators such as LEI, CAPE, and payout ratios operate on different scales (indices, ratios, percentages) and exhibit varying degrees of volatility.4 While standardization (scaling variables to have zero mean and unit variance) addresses the issue of differing scales 17, it does not resolve the problem of ignoring correlations or the potential importance of time lags and leads. Consequently, Euclidean distance might overweight volatile variables or fail to recognize similarity between series that co-move but at different absolute levels or with slight time shifts.

**3.2 Correlation-Based Distances**

An alternative approach is to define distance based on the correlation between variables or time series. A common transformation uses the Pearson correlation coefficient (ρij​) to define a distance metric, such as Dij​=2(1−ρij​)​.21

**Application & Suitability:** This approach focuses on the similarity in the *direction* and *pattern* of movement between variables, rather than their absolute levels. It is widely used in finance, particularly for clustering stocks based on the co-movement of their returns, often revealing underlying economic sector structures.19 Given that the user's input variables (yield curve factors, LEI, CAPE, payout ratio) are likely to exhibit significant correlations driven by underlying economic forces, a correlation-based distance could be highly relevant. It would group periods where these indicators show similar patterns of co-movement, potentially defining regimes based on the nature of these interrelationships rather than just their levels. This measure is inherently scale-invariant.

**3.3 Dynamic Time Warping (DTW)**

Dynamic Time Warping (DTW) is a powerful algorithm designed specifically for measuring similarity between time series, particularly when they may be warped or shifted in time relative to each other.54 Instead of comparing points at identical time indices t, DTW finds an optimal non-linear alignment (the "warping path") between two series, X=(x1​,...,xn​) and Y=(y1​,...,ym​), that minimizes the cumulative distance between the aligned points.57 This alignment allows one point in a series to be matched with multiple points in the other series, effectively stretching or compressing segments of the time axis to find the best match in shape.54 The optimal path and the resulting DTW distance are typically found using dynamic programming.57

**Application & Suitability:** DTW is extensively used for time series clustering across diverse domains, including finance and economics.56 Its key advantage lies in its ability to handle phase differences and local distortions in timing.54 This is particularly relevant for macroeconomic variables where lead-lag relationships are common and economically significant. For example, the yield curve is known to invert *before* recessions 60, and the LEI is designed to *lead* economic activity.63 A regime like 'Recovery' might be characterized by a steepening yield curve *followed* by a rising LEI. Euclidean distance might fail to group instances of this regime if the timing between the yield curve steepening and the LEI rise varies slightly. DTW, by finding the optimal alignment, can recognize the similarity in this *phased pattern* even with temporal variations.58 It can also compare series of different lengths.54 While computationally more intensive than Euclidean distance, DTW can be implemented with various clustering algorithms, including k-means (often referred to as k-DBA 66) and hierarchical methods.54

**3.4 Mahalanobis Distance**

The Mahalanobis distance provides a way to measure the distance of a point x from a distribution (or cluster) characterized by a mean vector μ and a covariance matrix Σ.4 It is defined as:

DM​(x)=(x−μ)TΣ−1(x−μ)​

Unlike Euclidean distance, Mahalanobis distance explicitly accounts for the correlations between variables (embedded in Σ−1) and is scale-invariant.4 Geometrically, points with the same Mahalanobis distance from the mean form an ellipsoid whose shape and orientation are determined by the covariance matrix, whereas points with the same Euclidean distance form a sphere.4

**Application & Suitability:** This distance measure is particularly useful in multivariate settings where variables are correlated and may have different scales, making it well-suited for macroeconomic data.4 It is often used for outlier detection, as it identifies points that are unusual given the correlation structure of the data.4 Kritzman and colleagues utilized Mahalanobis distance to measure financial turbulence (how far current returns are from their historical average, considering correlations) 67 and later applied it to measure the similarity of current economic conditions to historical recession/growth "clusters" 4 and to attribute the drivers of inflation regimes.45 By incorporating the covariance structure, Mahalanobis distance recognizes that a deviation in one variable might be typical if accompanied by a correlated move in another, penalizing uncorrelated deviations more heavily. This aligns well with the idea that economic regimes are often defined by characteristic patterns of *interdependence* among indicators, not just their individual levels. Applying it in clustering requires estimating the covariance matrix Σ, potentially for each cluster, which might necessitate an iterative approach (e.g., estimating Σ within clusters found using Euclidean distance, then re-clustering using Mahalanobis distance).

**3.5 Other Measures**

* **Compression-Based Dissimilarity Measure (CDM):** This measure, based on the idea of co-compressibility of time series, was found to be particularly suitable for clustering macroeconomic variables in one study.23 It demonstrated robustness across various data perturbations and seemed effective at identifying similarities based on frequency rather than volatility magnitude.23
* **Wasserstein Distance:** Also known as the Earth Mover's Distance, it measures the distance between two probability distributions.69 It has been applied in "Wasserstein k-means" for market regime clustering, potentially offering advantages over methods based on moments by comparing the entire distributions of data segments.2
* **Maximum Mean Discrepancy (MMD):** Another metric for comparing probability distributions, often used in conjunction with or as an alternative to Wasserstein distance in machine learning contexts.69

**Table 2: Comparison of Distance/Similarity Measures for Macroeconomic Time Series**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure** | **Key Characteristics** | **Advantages** | **Disadvantages** | **Suitability for Macro Time Series** |
| **Euclidean** | Point-wise comparison, Ignores correlation, Sensitive to scale 5 | Simple, computationally fast. | Unrealistic assumptions for macro data (independence, equal scale/importance).4 Ignores time shifts. | Often suboptimal due to correlations and varying scales/volatility. Requires careful standardization. Baseline only. |
| **Correlation-based** | Measures similarity in movement direction 22, Scale-invariant. | Captures co-movement patterns, insensitive to level differences. Widely used in finance.21 | Ignores magnitude information. Sensitive to outliers affecting correlation calculation. | Suitable if regimes are defined by patterns of co-movement among indicators, rather than absolute levels. Relevant for correlated macro variables. |
| **DTW** | Optimal non-linear alignment, Handles time shifts/warping 54, Shape-based. | Robust to phase differences, leads/lags.58 Can compare series of different lengths.54 Captures similarity in shape. | Computationally more expensive than Euclidean. Can sometimes produce counter-intuitive alignments (path constraints help). | Highly suitable when lead-lag relationships (e.g., LEI leading growth, yield curve predicting recession) are key regime characteristics. Captures dynamic patterns. |
| **Mahalanobis** | Accounts for covariance structure 4, Scale-invariant.5 | Considers variable interdependencies. Robust to differing scales. Statistically grounded measure of distance from a distribution. | Requires estimation of covariance matrix (potentially per cluster), sensitive to covariance matrix estimation errors (esp. high dimensions). | Very suitable for correlated macroeconomic variables where interdependencies define regimes. Used in practice for economic regime analysis.4 |
| **CDM** | Compression-based, Focuses on frequency/structure.23 | Found robust for macro data.23 May capture causal links.23 | Less common than others, interpretation might be less direct. Not a true metric. | Promising alternative based on empirical studies 23, warrants consideration if standard metrics prove inadequate. |
| **Wasserstein/MMD** | Compares probability distributions.69 | Captures full distributional information, not just moments. Used in advanced clustering (Wasserstein k-means).69 | Requires representing data segments as distributions. Computationally intensive. | Advanced option. Suitable if regimes are best characterized by the entire distribution of variables within a period, not just their mean/covariance. |

**4.0 Strategies for Variable Weighting and Feature Importance**

Assigning appropriate weights to input variables or selecting the most relevant features is crucial for building robust and interpretable clustering models, especially when dealing with macroeconomic indicators where some variables might be more informative than others for distinguishing between specific economic regimes.

**4.1 Rationale for Differential Weighting**

The default approach in many clustering algorithms, particularly k-means with Euclidean distance on standardized data, implicitly assigns equal weight to all variables. However, this assumption may not be appropriate for macroeconomic regime identification.

* **Economic Significance:** Certain indicators are widely recognized as having strong predictive power for specific economic states. For instance, the slope of the yield curve (specifically, the spread between long-term and short-term rates) is a well-documented predictor of recessions 60, while measures of inflation are definitional for identifying high-inflation regimes.32 Equal weighting might dilute the signal from these critical indicators.
* **Statistical Properties:** Variables differ in their statistical properties, such as variance, persistence, and signal-to-noise ratio. Variables with higher noise or idiosyncratic movements might obscure the underlying regime structure if given equal weight to more stable or informative indicators.
* **Regime Specificity:** The importance of a variable might be regime-dependent. For example, equity valuation metrics like CAPE might be highly relevant for identifying periods of market froth or potential crises ('Precarious' regimes in some studies 32), but less critical in distinguishing between standard Expansion and Recovery phases.

Therefore, methods that allow for differential weighting or selection of variables based on their statistical properties or contribution to distinguishing regimes are highly desirable.

**4.2 Principal Component Analysis (PCA) for Dimensionality Reduction & Implied Weighting**

Principal Component Analysis (PCA) is a widely used linear algebra technique for dimensionality reduction and feature extraction.17 It transforms an original set of potentially correlated variables into a new set of uncorrelated variables, called principal components (PCs), ordered by the amount of variance they explain from the original data.17 The first PC (PC1) is the linear combination of the original variables that captures the maximum possible variance; the second PC (PC2) captures the maximum remaining variance while being orthogonal (uncorrelated) to PC1, and so on.17

Mathematically, PCs are the eigenvectors of the covariance (or correlation) matrix of the data, and the corresponding eigenvalues represent the variance explained by each PC.17 Prior to applying PCA, it is essential to standardize the variables (typically to zero mean and unit variance) if they are measured on different scales, as PCA is sensitive to variance and variables with larger ranges could otherwise dominate the analysis.3

**Application and Interpretation:** PCA is frequently used in finance and economics for reducing the dimensionality of datasets with many correlated variables (like multiple points on the yield curve, or numerous economic indicators), identifying underlying factors driving asset returns or economic activity, visualizing high-dimensional data, and creating decorrelated inputs for subsequent models.72

The contribution of each original variable to a given PC is indicated by its coefficient, known as the **loading**.68 By examining the loadings on the first few PCs (which capture the majority of the total variance), one can infer the importance of the original variables.68 Variables with high absolute loadings on components that explain significant variance are considered more influential in driving the overall structure of the data.68 For example, analyzing the loadings of the user's variables (yield curve level, slope, curvature; LEI; CAPE; payout ratio) on the primary PCs would reveal which variables co-move strongly and contribute most to the dominant patterns of variation in the macroeconomic system represented by these indicators. PC1 might represent a broad 'economic activity' factor driven by LEI and yield level, while PC2 captures 'yield curve shape', and PC3 relates to 'valuation/payouts'. The loadings quantify each variable's role in these underlying economic dimensions.

**Clustering on PCs:** A common strategy is to perform clustering not on the original (potentially numerous and correlated) variables, but on the first few PCs that retain a substantial portion of the original variance (e.g., 80-95%).68 This approach offers several benefits:

* Reduces dimensionality, potentially improving cluster quality and computational efficiency.
* Uses uncorrelated inputs for the clustering algorithm.
* Implicitly weights the original variables based on their contribution to the systematic variance captured by the selected PCs. Variables that contribute heavily to the dominant PCs will have a greater influence on the clustering outcome. This provides a data-driven method for moving beyond equal weighting and focusing on the most significant sources of variation in the macroeconomic data.

**4.3 Feature-Based Clustering**

An alternative strategy is to explicitly engineer features from the raw time series data and then perform clustering on these derived features.20 Instead of using the point-in-time values of, say, the yield curve slope or LEI, one could extract characteristics such as:

* **Trend:** Rolling average, slope of a local linear regression.
* **Volatility:** Rolling standard deviation, GARCH model parameters.
* **Autocorrelation/Momentum:** Lagged correlations (inertia).
* **Shape:** Skewness, kurtosis.
* **Cross-correlations:** Rolling correlations between different input variables.

**Application & Suitability:** This approach allows the analysis to focus on specific dynamic properties or characteristics deemed most relevant for defining economic regimes.20 For instance, a 'Recession' might be characterized not just by a low level of LEI, but by a negative *trend* in LEI and high *volatility* in the yield curve slope. Clustering on these extracted features directly targets these dynamic concepts. This method naturally handles the time-series nature of the data and can accommodate series of different lengths or with missing values, as features are calculated for each series independently.20 The choice of which features to extract is critical and should be guided by economic theory and the specific research question.20 Feature selection or weighting techniques can then be applied to this derived feature set.53 This provides a flexible framework to tailor the clustering process to capture specific economic phenomena beyond simple level comparisons.

**4.4 Explicit Feature Weighting Algorithms**

Several clustering algorithms directly incorporate mechanisms for weighting features, allowing the algorithm itself to determine or utilize differential importance.

* **Weighted K-Means:** Variants of the k-means algorithm exist where the contribution of each feature to the distance calculation is explicitly weighted.69 These weights might be pre-specified based on domain expertise or learned iteratively as part of the clustering process, aiming to assign higher weights to features that better discriminate between clusters. Some methods also allow weighting individual samples.13
* **Regularized Clustering:** Techniques inspired by regularized regression methods like LASSO (Least Absolute Shrinkage and Selection Operator) can be adapted for clustering. For example, sparse k-means or LASSO k-means incorporate an L1 penalty on feature weights within the clustering objective function.90 This encourages sparsity, effectively performing feature selection by driving the weights of irrelevant or redundant features towards zero, while simultaneously performing clustering.39 This offers a principled way to identify the most salient variables for defining cluster structure.
* **Subspace Clustering:** These algorithms recognize that different clusters might exist in different subspaces of the original feature space. They aim to identify clusters along with the specific subset of features (and potentially their weights) that best define each cluster. This is relevant if, for example, yield curve dynamics define recession clusters while valuation metrics define precarious/bubble clusters.
* **Other Approaches:** Methods like Cluster-Induced Ordered Weighted Averaging (CIOWA) have been used for aggregating financial data, assigning weights within identified clusters.51 While perhaps more of an aggregation tool, it reflects the concept of cluster-specific importance. Feature importance derived from supervised methods like Random Forests, if trained on a related predictive task, could potentially inform prior weights for unsupervised clustering.51

**4.5 Considerations for User's Variables**

Applying weighting or feature selection to the user's specific variables (Yield Curve Level, Slope, Curvature; Vanguard LEI; CAPE; Payout Ratio) requires considering their economic roles:

* **Yield Curve Factors:** These three factors (level, slope, curvature) are derived, often via PCA themselves 72, to capture the primary modes of yield curve variation. Their inclusion suggests they are already deemed important. Weighting might focus on the relative importance of *shape* (slope, curvature) versus *level* for different regimes. The slope is particularly noted for its recession forecasting ability.60 PCA applied to these three factors plus the other variables could reveal their joint importance structure.
* **Vanguard LEI:** As a composite forward-looking indicator 63, its weight might be increased, particularly when identifying turning points or forecasting transitions between regimes like Expansion/Slowdown or Recovery/Expansion.
* **CAPE Ratio:** A long-term valuation metric. Its relevance might be higher for identifying potentially unsustainable 'Expansion' or 'Precarious' regimes characterized by high valuations, and perhaps lower during deep 'Recession' phases where immediate activity indicators dominate.
* **Payout Ratio:** Reflects corporate financial policy and confidence. Its importance might be greater in distinguishing mature 'Expansion' phases (stable/high payouts) from early 'Recovery' (low/recovering payouts) or 'Slowdown' (potentially falling payouts).

A data-driven approach like PCA or feature selection via regularized clustering could help quantify these hypothesized differential importances.

**Table 3: Variable Weighting / Feature Importance Approaches**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **Mechanism** | **Advantages** | **Disadvantages** | **Applicability to User's Variables** |
| **Equal Weighting** | All variables contribute equally (often after standardization). Implicit in standard Euclidean k-means. | Simple, no extra parameters. | Unrealistic assumption for macro variables.60 May overweight noisy or less relevant variables. | Current user approach. Likely suboptimal given the nature of the variables. |
| **PCA-based (Cluster on PCs)** | Cluster on first few PCs. Variables weighted implicitly by their contribution (loadings) to major PCs. | Data-driven weighting based on variance explained.68 Reduces dimensionality & noise.79 Decorrelates inputs.79 Reveals co-movement structure.82 | PCs can be hard to interpret directly.74 Assumes linear combinations capture important structure. Variance ≠ Importance for clustering. | Highly applicable. PCA is standard for yield curves.72 Can reveal importance structure of all variables jointly. Clustering on PCs is a robust alternative. |
| **Feature-Based Clustering** | Cluster on features extracted from original variables (e.g., trend, volatility).20 | Focuses on specific dynamics/characteristics relevant to regimes.20 Handles time series nature well. Can bypass missing data issues.20 | Choice of features is crucial and subjective.83 Feature extraction adds a modeling step. | Very relevant. Allows focusing on dynamics (e.g., LEI trend, yield slope volatility) rather than just levels, potentially defining regimes more meaningfully. |
| **Explicit Weighting (Algorithmic)** | Algorithms (e.g., Weighted k-means, Regularized k-means) assign/learn weights during clustering.69 | Directly optimizes for cluster structure considering feature relevance. Can perform simultaneous clustering & feature selection (Regularized).90 | Can be more complex to implement and tune. Learned weights might lack clear economic interpretation. | Applicable. Weighted k-means allows incorporating prior beliefs. Regularized methods offer data-driven feature selection specifically for the clustering task, potentially identifying key drivers for each regime. |

**5.0 Synthesis and Recommendations**

The preceding sections have explored a range of alternative methodologies for clustering macroeconomic variables to identify economic regimes, moving beyond the baseline approach of k-means with Euclidean distance and equal weighting. This section synthesizes these alternatives and provides guidance for selecting and validating appropriate methods for the user's specific objective and dataset.

**5.1 Summary of Alternatives**

The exploration highlighted several key areas for potential improvement:

* **Clustering Algorithms:** Hierarchical clustering offers an interpretable structure (dendrogram) without pre-specifying the number of clusters.15 Gaussian Mixture Models (GMM) provide probabilistic assignments, handle elliptical cluster shapes, and model data density.1 Hidden Markov Models (HMM) explicitly model the dynamics and transitions between regimes.41 Density-based methods like DBSCAN can find arbitrary shapes and handle noise.40
* **Distance/Similarity Metrics:** Correlation-based distances focus on co-movement patterns.22 Dynamic Time Warping (DTW) accommodates time lags and leads inherent in economic indicators.54 Mahalanobis distance accounts for correlations and differing scales among variables.4 Compression-based measures (CDM) have shown promise for macro data.23
* **Variable Weighting/Feature Importance:** Principal Component Analysis (PCA) provides a data-driven way to reduce dimensionality and implicitly weight variables based on variance contribution.17 Feature-based clustering shifts focus from raw levels to dynamic characteristics.20 Explicit weighting algorithms (e.g., weighted k-means, regularized clustering) directly incorporate variable importance into the clustering process.69

**5.2 Guidance on Method Selection**

The optimal choice of methodology depends on the specific characteristics of the data and the analytical objectives. Key considerations include:

* **Data Characteristics:**
* *Correlation:* If variables (yield curve factors, LEI, CAPE, payout ratio) are significantly correlated, methods accounting for this (Mahalanobis distance, PCA, correlation-based distance) are preferable to Euclidean distance.4
* *Time Lags/Leads:* If the timing relationships between indicators are crucial for defining regimes (e.g., LEI leading growth, yield curve leading recessions), DTW is a strong candidate.58
* *Non-Normality/Fat Tails:* If variables exhibit non-Gaussian behavior, GMM (using multiple components) or clustering based on robust features might be more appropriate than methods assuming normality.33
* *Scales/Units:* If variables have different scales, standardization is essential for methods like Euclidean k-means or PCA.17 Scale-invariant metrics like Mahalanobis or correlation-based distances avoid this requirement.
* **Analytical Objectives:**
* *Probabilistic Assignment:* If quantifying the uncertainty of regime classification or using probabilities in downstream models is desired, GMM or HMM are suitable choices.1
* *Modeling Dynamics:* If understanding and forecasting regime transitions is a primary goal, HMM is the most direct approach.41
* *Interpretability:* Hierarchical clustering provides an interpretable dendrogram.15 PCA loadings offer insights into variable contributions.68 Feature-based clustering can be interpretable if features are economically meaningful.83 Complex models like HMM or GMM might be less transparent.
* *Feature Importance:* If identifying the key drivers of different regimes is important, PCA, feature selection methods (e.g., regularized clustering), or feature-based clustering are relevant.20
* **Complexity and Cost:** Balance the sophistication of the model with implementation effort and computational resources. DTW, some hierarchical methods, and complex HMMs can be computationally intensive.20

**5.3 Specific Recommendations for User**

Based on the user's goal of identifying five distinct economic regimes (Recession, Expansion, Slowdown, Recovery, High Inflation) using yield curve factors, LEI, CAPE, and payout ratio, the following steps are recommended:

1. **Algorithm Exploration:** Move beyond k-means.
* **GMM:** Strongly recommended due to its ability to handle elliptical clusters, provide probabilistic assignments (useful for fuzzy regime boundaries), and model potentially non-Gaussian data via mixtures.1 Experiment with different covariance structures (diagonal, full).
* **Hierarchical Clustering:** Use Ward's linkage method 23 to generate a dendrogram. This allows visualizing the relationships between the five target regimes and assessing if the data naturally supports this structure.15
* **HMM:** Consider if explicitly modeling regime persistence and transition probabilities is valuable for the intended application (e.g., forecasting, dynamic allocation).41
1. **Distance Metric Selection:** Abandon standard Euclidean distance.
* **Mahalanobis Distance:** Experiment with this metric to account for the expected correlations among the input variables.4 This requires estimating covariance matrices, perhaps iteratively within clusters or using rolling windows.
* **DTW:** Test if accommodating potential lead-lag relationships between indicators (e.g., LEI vs. yield curve slope) improves regime identification.58
* **Correlation-Based Distance:** Evaluate if similarity in the *pattern* of variable movements, rather than levels, better defines the regimes.22
1. **Variable Weighting / Feature Importance:** Address the equal weighting limitation.
* **PCA:** Apply PCA to the standardized input variables. Analyze the loadings to understand variable contributions and co-movements.68 Cluster on the first few significant PCs (e.g., those explaining >85% cumulative variance) as a primary alternative strategy.75
* **Feature-Based Clustering:** Consider extracting dynamic features (e.g., 6-month change in LEI, 3-month volatility of yield slope, trend in CAPE) and clustering based on these characteristics to capture regime dynamics more directly.20
1. **Validation:** Rigorous validation is essential for unsupervised learning tasks like regime identification.
* **Internal Metrics:** Use metrics like Silhouette score, Davies-Bouldin index, or Calinski-Harabasz index to compare the mathematical quality (separation, cohesion) of clusterings produced by different methods *using the same distance metric*.40 However, recognize that optimal scores on these metrics do not guarantee economic meaningfulness.
* **Stability Analysis:** Assess the robustness of the identified regimes by:
* Running algorithms with different initializations (for k-means, GMM).
* Applying methods to different sub-periods of the data.23
* Using bootstrapping or subsampling techniques. Stable regimes should emerge consistently.
* **External Validation / Economic Plausibility:** This is the most critical validation step. Evaluate whether the identified clusters correspond to recognizable and distinct economic regimes:
* *Characterize Regimes:* Calculate average values, volatilities, and correlations of input variables within each identified cluster/regime. Do these characteristics align with the intended labels (Recession, Expansion, etc.)?
* *Historical Alignment:* Compare the timing of identified regimes with known economic events and NBER recession dates.32
* *Asset Behavior:* Analyze the performance (mean return, volatility, correlation) of different asset classes (equities, bonds, commodities) conditional on the identified regimes. Meaningful regimes should exhibit distinct and ideally predictable asset behavior patterns.32
* **Method Comparison:** Systematically compare the results (regime assignments, characteristics, stability) obtained from different combinations of algorithms, distance metrics, and weighting strategies to understand sensitivities and build confidence in the final chosen methodology.

**5.4 Addressing the 5 Regimes**

The pre-specification of five regimes (Recession, Expansion, Slowdown, Recovery, High Inflation) should be data-driven rather than assumed.

* **Hierarchical clustering** naturally allows exploring different numbers of clusters by cutting the dendrogram at various levels.21
* For **GMM**, use information criteria like BIC or AIC to guide the selection of the optimal number of components (k).29 These criteria balance model fit with complexity.
* For **k-means** (or variants), use methods like the Elbow method (plotting inertia vs. k) or average Silhouette scores to suggest an appropriate k.94

Combine these statistical suggestions with economic interpretability. Does a 4-regime solution provide more distinct and meaningful economic states than a 5-regime solution? Does a 6-regime solution reveal a significant nuance missed by 5? The data itself, analyzed through appropriate clustering techniques and validation, should guide the determination of the most representative number of distinct macroeconomic regimes present in the historical period under study. Some studies using similar approaches have identified four distinct regimes.32

**6.0 Conclusion**

The task of identifying macroeconomic regimes using clustering techniques is essential for informed financial decision-making, but the standard approach of k-means with Euclidean distance often falls short due to the complex nature of economic data. This report has detailed a range of advanced methodologies that offer potential improvements in robustness, flexibility, and economic interpretability.

Exploring alternative algorithms such as Hierarchical Clustering, Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM) allows for capturing hierarchical relationships, probabilistic uncertainties, and dynamic transitions, respectively, which are often characteristic of economic cycles. Moving beyond Euclidean distance to metrics like Correlation-based distances, Dynamic Time Warping (DTW), or Mahalanobis distance enables the incorporation of co-movement patterns, time lags, and variable interdependencies crucial for understanding macroeconomic relationships. Furthermore, addressing the limitation of equal variable weighting through techniques like Principal Component Analysis (PCA), feature-based clustering, or explicit weighting algorithms allows the clustering process to focus on the most salient indicators driving regime shifts.

Ultimately, the selection of the "best" methodology is not universal but depends critically on the specific characteristics of the input data (yield curve factors, LEI, CAPE, payout ratio) and the precise objectives of the analysis. A rigorous, iterative process involving experimentation with different combinations of algorithms, distance metrics, and weighting strategies, coupled with robust validation focusing on both statistical cluster quality and, most importantly, economic plausibility and interpretability, is paramount. By carefully considering these advanced techniques and validation approaches, practitioners can significantly enhance their ability to identify meaningful macroeconomic regimes and leverage this understanding for improved forecasting, risk management, and asset allocation.

**Works cited**

1. Gaussian Mixture Model: A Comprehensive Guide, accessed May 1, 2025, <https://www.pickl.ai/blog/gaussian-mixture-model/>
2. Classifying market regimes | Macrosynergy, accessed May 1, 2025, <https://macrosynergy.com/research/classifying-market-regimes/>
3. 5 Key Steps to Master Gaussian Mixture Model Challenges, accessed May 1, 2025, <https://www.numberanalytics.com/blog/gaussian-mixture-model-master-steps>
4. A NEW INDEX OF THE BUSINESS CYCLE - State Street Global Markets, accessed May 1, 2025, <https://globalmarkets.statestreet.com/research/service/public/v1/article/insights/pdf/v2/5b4b47fa-8256-4e4a-8991-afe13469268b/joim_a_new_index_of_the_business_cycle.pdf>
5. Financial development and growth: A clustering and causality analysis, accessed May 1, 2025, <https://www.tandfonline.com/doi/abs/10.1080/09638199.2011.570364>
6. Asset Allocation Perspective Research Articles - R Discovery, accessed May 1, 2025, [https://discovery.researcher.life/topic/asset-allocation-perspective/22236894?page=1&topic\_name=Asset%20Allocation%20Perspective](https://discovery.researcher.life/topic/asset-allocation-perspective/22236894?page=1&topic_name=Asset+Allocation+Perspective)
7. MULTI-ASSET STRATEGIES AND ASSET ALLOCATION - Portfolio Management Research, accessed May 1, 2025, <https://www.pm-research.com/content/iijpormgmt/51/5/local/complete-issue.pdf>
8. Optimizing target-date fund rebalancing through threshold-based strategies, accessed May 1, 2025, <https://www.vanguardmexico.com/content/dam/intl/americas/documents/latam/en/2024/12/the-rebalancing-edge.pdf>
9. Optimizing target-date fund rebalancing through threshold-based strategies - Vanguard, accessed May 1, 2025, <https://corporate.vanguard.com/content/dam/corp/research/pdf/the_rebalancing_edge_optimizing_target_date_fund_rebalancing_through_threshold_based_strategies.pdf>
10. Rational rebalancing: An analytical approach to multiasset portfolio rebalancing decisions and insights - Vanguard, accessed May 1, 2025, <https://corporate.vanguard.com/content/dam/corp/research/pdf/rational_rebalancing_analytical_approach_to_multiasset_portfolio_rebalancing.pdf>
11. MULTI-ASSET STRATEGIES AND ASSET ALLOCATION - Portfolio Management Research, accessed May 1, 2025, <https://www.pm-research.com/content/iijpormgmt/50/5/local/complete-issue.pdf>
12. Vanguard's portfolio construction framework, accessed May 1, 2025, <https://corporate.vanguard.com/content/dam/corp/research/pdf/vanguards_portfolio_construction_framework.pdf>
13. 2.3. Clustering — scikit-learn 1.6.1 documentation, accessed May 1, 2025, <https://scikit-learn.org/stable/modules/clustering.html>
14. Grouping and clustering methods in econometrics - Institute of Economic Research, Seoul National University, accessed May 1, 2025, <https://ier.snu.ac.kr/journals/kej/korean-economic-journal?md=download&articleidx=1242>
15. An Overview of Clustering Methods in The Financial World - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/363359772_An_Overview_of_Clustering_Methods_in_The_Financial_World>
16. In cluster analysis, how does Gaussian mixture model differ from K Means when we know the clusters are spherical? - Cross Validated, accessed May 1, 2025, <https://stats.stackexchange.com/questions/489459/in-cluster-analysis-how-does-gaussian-mixture-model-differ-from-k-means-when-we>
17. What Is Principal Component Analysis (PCA)? - IBM, accessed May 1, 2025, <https://www.ibm.com/think/topics/principal-component-analysis>
18. Principal Component Analysis (PCA) Explained | Built In, accessed May 1, 2025, <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
19. Clustering Financial Time Series - DSpace, accessed May 1, 2025, <https://dspace.ut.ee/bitstreams/238ca2e1-d755-4147-af52-2fb08219bda2/download>
20. Characteristic-based Clustering for Time Series Data - Rob J Hyndman, accessed May 1, 2025, <https://robjhyndman.com/papers/DMKD.pdf>
21. Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods | PLOS One, accessed May 1, 2025, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116201>
22. Correlation Based Hierarchical Clustering in Financial Time Series - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/239831442_Correlation_Based_Hierarchical_Clustering_in_Financial_Time_Series>
23. www.econstor.eu, accessed May 1, 2025, <https://www.econstor.eu/bitstream/10419/180670/1/clustering.pdf>
24. Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/262841110_Relation_between_Financial_Market_Structure_and_the_Real_Economy_Comparison_between_Clustering_Methods>
25. Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods - PMC, accessed May 1, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4365074/>
26. Clustering Financial Time Series: How Long Is Enough? - IJCAI, accessed May 1, 2025, <https://www.ijcai.org/Proceedings/16/Papers/367.pdf>
27. Hierarchical clustering applied to our economic state distance matrix Ω... | Download Scientific Diagram - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/figure/Hierarchical-clustering-applied-to-our-economic-state-distance-matrix-O-S-The-algorithm_fig6_359616422>
28. (PDF) Ensemble Financial Time Series Clustering - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/382455348_Ensemble_Financial_Time_Series_Clustering>
29. 2.1. Gaussian mixture models - Scikit-learn, accessed May 1, 2025, <https://scikit-learn.org/stable/modules/mixture.html>
30. Market regime detection using Statistical and ML based approaches | Devportal, accessed May 1, 2025, <https://developers.lseg.com/en/article-catalog/article/market-regime-detection>
31. Clustering performance comparison using K-means and expectation maximization algorithms - PMC, accessed May 1, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4433949/>
32. Data-driven Approach to Clustering Similar Macroeconomic Regimes - Alpha Architect, accessed May 1, 2025, <https://alphaarchitect.com/clustering-macroeconomic-regimes/>
33. A Machine Learning Approach to Regime Modeling - Two Sigma, accessed May 1, 2025, <https://www.twosigma.com/articles/a-machine-learning-approach-to-regime-modeling/>
34. Full article: Evaluation of Value-at-Risk (VaR) using the Gaussian Mixture Models, accessed May 1, 2025, <https://www.tandfonline.com/doi/full/10.1080/27684520.2024.2346075>
35. Portfolio construction with Gaussian mixture returns and exponential utility via convex optimization - Stanford University, accessed May 1, 2025, <https://web.stanford.edu/~boyd/papers/pdf/exp_util_gm_ret_portfolio.pdf>
36. Using a “Human + Machine” Approach to Determine Financial Market Regimes - S&P Global, accessed May 1, 2025, [https://cdn.ihs.com/www/pdf/Using%20a%20'human%20and%20machine'%20approach%20to%20determine%20financial%20market%20regimes.pdf](https://cdn.ihs.com/www/pdf/Using%20a%20%27human%20and%20machine%27%20approach%20to%20determine%20financial%20market%20regimes.pdf)
37. Tactical Asset Allocation with Macroeconomic Regime Detection - arXiv, accessed May 1, 2025, <https://arxiv.org/html/2503.11499v2>
38. Tactical Asset Allocation with Macroeconomic Regime Detection - arXiv, accessed May 1, 2025, <https://arxiv.org/html/2503.11499v1>
39. High-Dimensional Markov-Switching Vector Autoregression Spillover Networks - Erasmus University Thesis Repository, accessed May 1, 2025, <https://thesis.eur.nl/pub/72919/Thesis-Pronk-501149-Final.pdf>
40. Characteristics of time series clustering. | Download Scientific Diagram - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/figure/Characteristics-of-time-series-clustering_fig1_335762113>
41. Asset Allocation Using a Markov Process of Clustered Efficient Frontier Coefficients States - Newswise, accessed May 1, 2025, <https://www.newswise.com/pdf_docs/170108570199650_1-s2.0-S2405918823000260-main.pdf>
42. Macroeconomic Patterns in System-Wide Liquidity Regimes - NYU Stern, accessed May 1, 2025, <https://www.stern.nyu.edu/sites/default/files/assets/documents/FloodLiechtyPiontek_MacroPatternsinSystemWideLiquidityRegimes_Nov2014.pdf>
43. NBER WORKING PAPER SERIES REGIME-SWITCHING AND THE ESTIMATION OF MULTIFRACTAL PROCESSES Laurent Calvet Adlai Fisher Working Pape, accessed May 1, 2025, <https://www.nber.org/system/files/working_papers/w9839/w9839.pdf>
44. (PDF) Machine Learning in Macro-Economic Series Forecasting - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/321053452_Machine_Learning_in_Macro-Economic_Series_Forecasting>
45. THE DETERMINANTS OF INFLATION - Journal of Investment Managment, accessed May 1, 2025, <https://www.joim.com/wp-content/uploads/emember/downloads/p0733.pdf>
46. THE DETERMINANTS OF INFLATION | State Street, accessed May 1, 2025, <https://www.statestreet.com/web/insights/articles/documents/ssa-the-determinants-of-inflation-061322_4787363.1.1.GBL.pdf>
47. A Regime-Switching Model of the Yield Curve at the Zero Bound - San Francisco Fed, accessed May 1, 2025, <https://www.frbsf.org/research-and-insights/publications/working-papers/2016/01/a-regime-switching-model-of-the-yield-curve-at-the-zero-bound>
48. Investigating Emerging Market Economies Reverse REIT-Bond Yield Gap Anomalies - WIReDSpace, accessed May 1, 2025, <https://wiredspace.wits.ac.za/bitstreams/0d1943a3-bba5-40c3-87fc-89c5a16cfd4c/download>
49. (PDF) Tactical Asset Allocation with Macroeconomic Regime Detection - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/389894855_Tactical_Asset_Allocation_with_Macroeconomic_Regime_Detection>
50. Relation between financial market structure and the real economy: comparison between clustering methods - IDEAS/RePEc, accessed May 1, 2025, <https://ideas.repec.org/p/ehl/lserod/61644.html>
51. Full article: Clustering financial institutions in countries based on a hybrid random forest and induced ordered weighted averaging - Taylor & Francis Online, accessed May 1, 2025, <https://www.tandfonline.com/doi/full/10.1080/23270012.2025.2454674>
52. Clustering of financial time series - IDEAS/RePEc, accessed May 1, 2025, <https://ideas.repec.org/a/eee/phsmap/v392y2013i9p2114-2129.html>
53. (PDF) Ensemble Clustering of Financial Time Series - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/378797814_Ensemble_Clustering_of_Financial_Time_Series>
54. Dynamic Time Warping (DTW) and time series clustering - RPubs, accessed May 1, 2025, <https://rpubs.com/esobolewska/dtw-time-series>
55. Principal Component Analysis(PCA) | GeeksforGeeks, accessed May 1, 2025, <https://www.geeksforgeeks.org/principal-component-analysis-pca/>
56. Dynamic Time Warping Algorithm in Modeling Systemic Risk in the European Insurance Sector - MDPI, accessed May 1, 2025, <https://www.mdpi.com/1099-4300/23/8/1022>
57. An introduction to Dynamic Time Warping - Romain Tavenard, accessed May 1, 2025, <https://rtavenar.github.io/blog/dtw.html>
58. (PDF) Clustering Macroeconomic Time Series - ResearchGate, accessed May 1, 2025, <https://www.researchgate.net/publication/325876743_Clustering_Macroeconomic_Time_Series>
59. Intertemporal Similarity of Economic Time Series: An Application of Dynamic Time Warping, accessed May 1, 2025, <https://ideas.repec.org/a/kap/compec/v56y2020i1d10.1007_s10614-020-09986-0.html>
60. Yield Curve and Predicted GDP Growth - Federal Reserve Bank of Cleveland, accessed May 1, 2025, <https://www.clevelandfed.org/indicators-and-data/yield-curve-and-predicted-gdp-growth>
61. Yielding Clues About Recessions: The Yield Curve as a Forecasting Tool, accessed May 1, 2025, <https://www.stlouisfed.org/publications/regional-economist/october-1997/yielding-clues-about-recessions-the-yield-curve-as-a-forecasting-tool>
62. Predicting Recessions Using the Yield Curve: The Role of the Stance of Monetary Policy, accessed May 1, 2025, <https://www.bostonfed.org/publications/current-policy-perspectives/2020/predicting-recessions-using-the-yield-curve.aspx>
63. Is This Time Different? LEI And Yield Curve Inversion as Recession Predictors - PGIM, accessed May 1, 2025, <https://www.pgim.com/ucits/de/article/time-different-lei-and-yield-curve-inversion-recession-predictors>
64. Is This Time Different? LEI and Yield Curve Inversion as Recession Predictors, accessed May 1, 2025, <https://www.pgimquantitativesolutions.com/article/time-different-lei-and-yield-curve-inversion-recession-predictors>
65. The use of the dynamic time warping (DTW) method to describe the COVID-19 dynamics in Poland | Oeconomia Copernicana - Economic Publishing Platform, accessed May 1, 2025, <https://journals.economic-research.pl/oc/article/view/1900>
66. Bridging the Gap: A Decade Review of Time-Series Clustering Methods - arXiv, accessed May 1, 2025, <https://arxiv.org/html/2412.20582v1>
67. Skulls, Financial Turbulence, and Risk Management - Top1000funds.com, accessed May 1, 2025, <https://www.top1000funds.com/wp-content/uploads/2010/11/FAJskulls.pdf>
68. Interpret all statistics and graphs for Principal Components Analysis - Support - Minitab, accessed May 1, 2025, <https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/multivariate/how-to/principal-components/interpret-the-results/all-statistics-and-graphs/>
69. Detecting multivariate market regimes via clustering algorithms - Imperial College London, accessed May 1, 2025, <https://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/212236006---James-Mc-Greevy---MCGREEVY_JAMES_01075416.pdf>
70. Clustering market regimes using the Wasserstein distance - Journal of Computational Finance - Risk.net, accessed May 1, 2025, <https://www.risk.net/journal-of-computational-finance/7959937/clustering-market-regimes-using-the-wasserstein-distance>
71. Principal Component Analysis (PCA) - Analytics Vidhya, accessed May 1, 2025, <https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/>
72. Applications of Principal Components Analysis in Finance - Aptech, accessed May 1, 2025, <https://www.aptech.com/blog/applications-of-principal-components-analysis-in-finance/>
73. Principal component analysis - Wikipedia, accessed May 1, 2025, <https://en.wikipedia.org/wiki/Principal_component_analysis>
74. Principal Component Analysis | Trading Interview, accessed May 1, 2025, <https://www.tradinginterview.com/courses/linear-algebra/lessons/principal-component-analysis/>
75. Principal Component Analysis: PCA: PCA: How to Reduce the Dimensionality of Your Investment Data - FasterCapital, accessed May 1, 2025, <https://fastercapital.com/content/Principal-Component-Analysis--PCA---PCA--How-to-Reduce-the-Dimensionality-of-Your-Investment-Data.html>
76. Principal Component Analysis (PCA) for Portfolio Risk - QuestDB, accessed May 1, 2025, <https://questdb.com/glossary/principal-component-analysis-pca-for-portfolio-risk/>
77. Portfolio Optimization under Time-Varying Economic Regimes - CS229, accessed May 1, 2025, <https://cs229.stanford.edu/proj2009/HanTurner.pdf>
78. Clustering on Principal Component Analysis - Charles Holbert, accessed May 1, 2025, <https://www.cfholbert.com/blog/cluster-pca/>
79. Using principal components to construct macro trading signals - Macrosynergy, accessed May 1, 2025, <https://macrosynergy.com/research/using-principal-components-to-construct-macro-trading-signals/>
80. Principal Component Analysis Using R - SOA, accessed May 1, 2025, <https://www.soa.org/digital-publishing-platform/emerging-topics/principal-component-analysis-using-r/>
81. 11.4 - Interpretation of the Principal Components | STAT 505, accessed May 1, 2025, <https://online.stat.psu.edu/stat505/lesson/11/11.4>
82. Feature Importance in PCA: Analyzing Loadings and Biplots | GeeksforGeeks, accessed May 1, 2025, <https://www.geeksforgeeks.org/feature-importance-in-pca-analyzing-loadings-and-biplots/>
83. Full article: A Gentle Introduction and Application of Feature-Based ..., accessed May 1, 2025, <https://www.tandfonline.com/doi/full/10.1080/00273171.2024.2432918>
84. Feature-driven Time Series Clustering - GitHub Pages, accessed May 1, 2025, <https://edbt2021proceedings.github.io/docs/p270.pdf>
85. Forecasting With Feature-Based Time Series Clustering - DiVA portal, accessed May 1, 2025, [https://www.diva-portal.org/smash/get/diva2:1779968/FULLTEXT01.pdf](https://www.diva-portal.org/smash/get/diva2%3A1779968/FULLTEXT01.pdf)
86. Imaging feature-based clustering of financial time series | PLOS One, accessed May 1, 2025, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0288836>
87. Temporal clustering of time series via threshold autoregressive models: application to commodity prices - IDEAS/RePEc, accessed May 1, 2025, <https://ideas.repec.org/a/spr/annopr/v260y2018i1d10.1007_s10479-017-2659-0.html>
88. Time-series clustering – A decade review, accessed May 1, 2025, <https://wiki.smu.edu.sg/18191isss608g1/img_auth.php/f/fd/Time_Series_Clustering_A_Decade_Review.pdf>
89. Central Clustering of Categorical Data with Automated Feature Weighting - IJCAI, accessed May 1, 2025, <https://www.ijcai.org/Proceedings/13/Papers/190.pdf>
90. Regularised Jump Models for Regime Identification and Feature Selection - OpenReview, accessed May 1, 2025, <https://openreview.net/forum?id=JXvEzl8YkS>
91. Evaluating Feature Selection Methods for Macro-Economic Forecasting, Applied for Iran's macro-economic variables Mahdi Goldani - arXiv, accessed May 1, 2025, <https://arxiv.org/pdf/2406.03742>
92. Macroeconomic conditions, corporate default, and default clustering - CentAUR, accessed May 1, 2025, <https://centaur.reading.ac.uk/109180/1/EM%20final%282%29.pdf>
93. Editor's Introduction for the 2023 Special Issue on Investing in Non-US Financial Markets, accessed May 1, 2025, <https://www.pm-research.com/content/iijpormgmt/49/6/local/complete-issue.pdf>
94. Comparison of K-means, Gaussian Mixture, and Hierarchical Clustering Models on Countries' Economic Freedom Index - ResearchGate, accessed May 1, 2025, [https://www.researchgate.net/publication/389646198\_Comparison\_of\_K-means\_Gaussian\_Mixture\_and\_Hierarchical\_Clustering\_Models\_on\_Countries'\_Economic\_Freedom\_Index](https://www.researchgate.net/publication/389646198_Comparison_of_K-means_Gaussian_Mixture_and_Hierarchical_Clustering_Models_on_Countries%27_Economic_Freedom_Index)