Uncovering Market Regime Structure in Financial Data

Dimitris Markopoulos
Statistics Department

Columbia University
New York, United States

djm2269@columbia.edu

Abstract— **Financial markets exhibit non-stationary behavior characterized by recurring yet evolving market regimes. Identifying such regimes is challenging due to the absence of ground-truth labels and the high dimensionality of modern financial data. This project studies unsupervised market regime discovery using a large set of macroeconomic and financial time series. Rather than committing to a single modeling choice, we adopt a systematic workflow that evaluates regime structure across a grid of data preprocessing choices, dimensionality reduction techniques, and clustering algorithms. Regime quality is assessed using quantitative, label-free criteria that emphasize robustness and consistency across modeling assumptions. This framework allows us to distinguish regime structures that are stable across alternative representations of the data from those that are sensitive to specific analytical choices. The contribution of this work is methodological: it demonstrates how stability-driven evaluation can be used to justify unsupervised regime discovery in high-dimensional financial settings, providing a reproducible and interpretable alternative to ad-hoc regime definitions.**

I. Introduction

A. Motivation

Financial markets exhibit persistent yet evolving patterns of behavior, commonly referred to as *market regimes*. These regimes reflect underlying changes in macroeconomic conditions, risk sentiment, liquidity, and policy environments, and they play a central role in asset pricing, portfolio construction, and risk management. Empirically, markets do not behave in a stationary manner: periods of relative calm are punctuated by episodes of heightened volatility, stress, or structural transition. Models calibrated in one regime often degrade or fail entirely when the market shifts into another.

Despite their importance, regimes are not directly observable. Traditional approaches frequently rely on ad-hoc or economically motivated labels, such as "risk-on / risk-off," recession indicators, or volatility thresholds. While intuitive, these definitions are inherently subjective, often depend on a small number of variables, and may fail to capture the full multivariate structure of modern financial systems. Moreover, such labels are typically imposed *ex ante*, rather than inferred from the data, making them fragile to changing market dynamics and structural breaks.

From a modeling perspective, the absence of reliable labels motivates the use of unsupervised learning. In contrast to supervised or rule-based approaches, unsupervised methods allow regimes to emerge directly from the joint behavior of high-dimensional financial variables, without imposing restrictive assumptions about their number, form, or economic interpretation. This is particularly well-suited to regime analysis, where regimes are latent, evolving, and unlabeled by construction.

A data-driven approach to regime discovery provides several advantages. First, it enables the identification of recurring market structures that may not align cleanly with predefined economic narratives. Second, it allows regimes to be defined by collective behavior across assets and macro variables, rather than isolated indicators. Finally, unsupervised methods provide a flexible framework for studying regime stability, transitions, and robustness under alternative representations of the data.

Taken together, these considerations motivate a systematic, unsupervised investigation of market regimes that emphasizes robustness, stability, and interpretability, rather than reliance on a single model or predefined economic interpretation.

B. Related Work

The identification of market regimes has been extensively studied in both the academic literature and practitioner research. Traditional approaches [1] include rule-based economic classifications, volatility- or drawdown-based thresholds, and parametric regime-switching models such as Markov-switching or hidden Markov models. While these methods offer interpretability, they typically rely on strong structural assumptions, predefined state dynamics, or a small set of conditioning variables, which can limit their flexibility in high-dimensional financial settings.

More recently, unsupervised learning methods have been adopted as a data-driven alternative for regime discovery. These approaches aim to infer latent market states directly from the joint behavior of assets or macroeconomic factors, without requiring predefined labels. A prominent practitioner example is the work by Two Sigma [2], who apply a Gaussian Mixture Model (GMM) to a fixed set of macro and style factors in order to identify a small number of latent market regimes. Their framework demonstrates that unsupervised clustering can recover economically interpretable regimes such as crisis, inflationary, and steady-state environments, and highlights the practical relevance of data-driven regime modeling.

Despite their empirical success, most existing unsupervised regime-detection approaches, including mixture-model-based methods, share several limitations. First, they typically commit to a single modeling choice, such as a specific clustering family or data representation, without examining sensitivity to preprocessing, scaling, or dimensionality reduction. Second, regime quality is often assessed qualitatively, through ex post economic interpretation or visual inspection, rather than through quantitative validation. As a result, it is unclear whether the identified regimes reflect stable structural features of the data or are artifacts of particular modeling assumptions.

More broadly, while unsupervised learning is frequently motivated by the absence of labels in regime detection, there is limited emphasis in the literature on systematically justifying unsupervised regimes using quantitative metrics. Measures such

Columbia University

as clustering stability, robustness across resampling, or consistency across alternative representations are rarely used as primary selection criteria. This creates a gap between the conceptual appeal of unsupervised regime discovery and its empirical validation. Recent work has emphasized the need for structured and reproducible workflows in unsupervised learning, particularly in settings where ground truth labels are unavailable. Allen et al. [3] propose a model-agnostic framework for unsupervised scientific discovery that explicitly stresses exploring multiple preprocessing choices, modeling techniques, and hyperparameter settings, followed by rigorous validation based on stability and generalizability metrics. Rather than treating unsupervised learning as a single-model exercise, their approach frames discovery as a systematic evaluation across a grid of reasonable analytical decisions.

This paper adopts that workflow philosophy and applies it in a financial regime-detection setting. In contrast to prior regime-modeling work that typically fixes a single preprocessing pipeline and clustering method, we explicitly evaluate how regime structure varies across alternative data transformations, dimensionality reduction techniques, and clustering algorithms, using quantitative stability-based metrics to guide model selection.

C. Contributions

The contributions of this work are primarily methodological rather than algorithmic. Specifically, this paper:

- Formulates market regime detection as a model-selection problem in an unsupervised setting, rather than committing to a single clustering method or data representation.
- Implements a systematic workflow that evaluates regime structure across multiple data preprocessing strategies, dimensionality reduction techniques, and clustering algorithms.
- Emphasizes quantitative, label-free evaluation criteria to assess regime robustness and sensitivity to modeling assumptions.
- Applies this workflow to a high-dimensional macro-financial dataset, demonstrating how stability-driven evaluation can justify unsupervised regime discovery in practice.

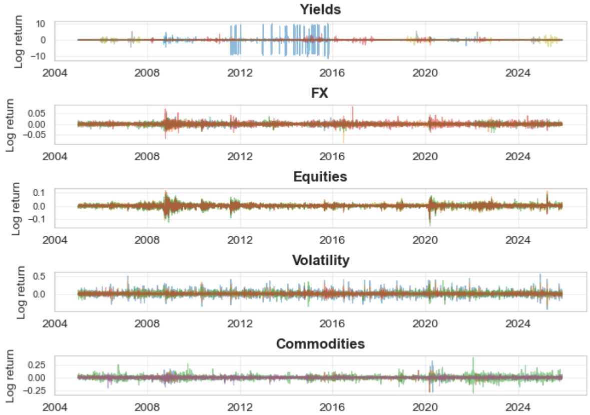
II. Data and Preprocessing

A. Data Sources

The dataset consists of a high-dimensional panel of macroeconomic and financial time series obtained from Bloomberg. All series are observed at a daily frequency and span 2005–2025, resulting in approximately 5,000 observations per feature after alignment.

The final dataset contains 37 features selected to capture broad macro-financial conditions across multiple asset classes and risk dimensions. Specifically, the dataset includes U.S. Treasury yields across the maturity spectrum, yield curve slope measures, and breakeven inflation rates, which collectively reflect interest rate expectations, term structure dynamics, and inflation compensation. Equity market conditions are represented using major U.S. and global equity indices, while volatility and tail-risk measures include implied volatility indices, skewness, and options-based indicators. Currency markets are captured through major G10 and emerging market exchange rates, providing information on global risk sentiment and dollar strength. Finally, commodity prices and shipping indices are included to proxy for real economic activity, inflation pressures, and supply-side conditions.

For visualization and exploratory analysis, features are grouped by asset class [yields, foreign exchange, equities, volatility, and commodities] and plotted across the full time horizon.



B. Data Alignment and Cleaning

All time-series are aligned to a common daily date index to ensure consistent cross-sectional observations across assets. Since the underlying series originate from markets with different trading calendars and holiday schedules, alignment is performed prior to any transformation or modeling to avoid artificial sparsity in the data.

Missing observations arising from non-trading days or asynchronous market closures are handled using forward-filling, such that the most recently observed value is carried forward until a new observation becomes available. This approach preserves information continuity while remaining consistent with standard practice in financial time-series analysis, particularly for macroeconomic and market-level indicators that do not update continuously.

Finally, one yield-related series contains intermittent zero values that are economically meaningful and correspond to periods in which the underlying rate was effectively constrained at zero. However, these values prevent the computation of log returns. To enable consistent return-based transformations, zero observations are replaced with a small positive constant ($\epsilon = 10^{-6}$) prior to transformation. This adjustment preserves the qualitative behavior of the series while avoiding undefined operations, and accounts for the presence of occasional large return fluctuations observed in the yield group visualizations.

C. Feature Transformations

To examine the sensitivity of regime discovery to data representation, three feature transformation schemes are considered. These transformations are applied consistently

Columbia University

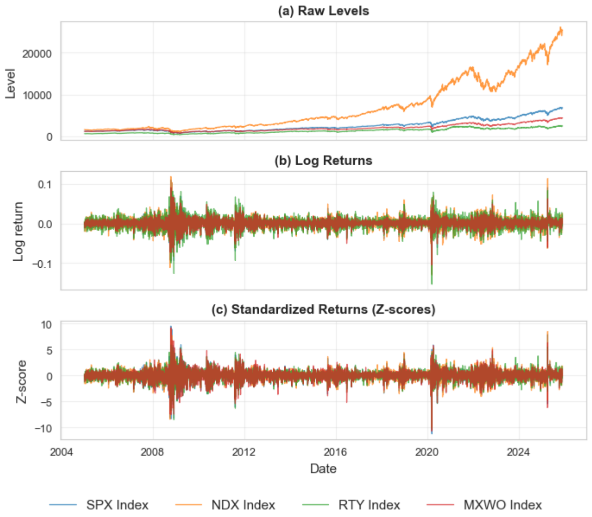
across all assets and serve as alternative inputs to the downstream dimensionality reduction and clustering procedures.

First, the **(a) raw time series** are used directly, preserving level information such as absolute interest rate levels, price indices, and volatility magnitudes. This representation retains long-run structural information but may be sensitive to non-stationarity and scale differences across features.

Second, **(b) log returns** are computed to emphasize short-term dynamics and relative changes rather than absolute levels. Return-based representations are commonly used in financial modeling to mitigate non-stationarity and to place features on a more comparable scale across asset classes. This transformation highlights periods of rapid market adjustment and stress, which may be informative for regime identification.

Third, features are standardized using a **(c) z-score transformation**, defined as $z_t = (x_t - \mu)/\sigma$, where μ and σ denote the sample mean and standard deviation of each series. Standardization removes differences in scale and variance across features, allowing clustering algorithms to focus on relative deviations rather than magnitude. This representation emphasizes cross-sectional structure and co-movement patterns in the data.

For visualization the equity features have been plotted across every transformation. As is evident from the transformed series, the three representations operate on markedly different scales and emphasize distinct statistical properties. Raw time series preserve level information and exhibit long-run trends and compounding effects, whereas log returns are centered near zero and more closely resemble approximately stationary, symmetric distributions. The z-score transformation further normalizes variation across features, highlighting relative deviations and co-movement rather than absolute magnitude. **These differences in statistical representation can materially influence the regime structure identified in subsequent analysis.**



Rather than selecting a single transformation a priori, all three representations are evaluated as part of the broader unsupervised workflow. This allows regime structure to be compared across alternative views of the data and supports a stability-driven approach to model selection.

III. Methods and Theory

The methodology below defines a systematic framework for unsupervised regime discovery across a broad set of modeling choices. The following section presents empirical results obtained from applying this workflow and evaluates regime quality across alternative configurations.

A. *Workflow Overview*

The overall workflow consists of three stages: (i) data preprocessing and transformation, (ii) dimensionality reduction to obtain low-dimensional representations, and (iii) clustering of the embedded observations to infer latent regimes. Each combination of modeling choices is evaluated quantitatively, and regime quality is compared across the full grid.

B. *Grid-Based Model Selection*

Given the absence of ground-truth labels for market regimes, no single data representation or clustering method can be assumed optimal a priori.

Three feature transformation schemes are considered: raw time series, log returns, and standardized (z-score) representations. These transformations emphasize different statistical properties of the data, ranging from long-run level information to short-term dynamics and relative deviations. Evaluating all three representations enables comparison of regime structure under substantially different scaling and stationarity characteristics.

For each transformed dataset, dimensionality reduction is applied to obtain low-dimensional embeddings suitable for clustering. Both linear and nonlinear techniques are included. Principal Component Analysis (PCA) is used as a linear baseline, with the number of components varied over a grid. Nonlinear structure is explored using Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE). For UMAP, the embedding dimension, neighborhood size, and minimum distance parameters are varied to capture alternative manifold assumptions. For t-SNE, both embedding dimension and perplexity are varied to assess sensitivity to local versus global structure.

Clustering is then performed on each low-dimensional embedding using three commonly used unsupervised algorithms: k-means, Gaussian mixture models (GMMs), and hierarchical clustering with ward linkage. The number of clusters is varied over $\{3, 4, 5\}$, reflecting common assumptions about the number of macro-financial regimes while avoiding excessive fragmentation. This combination of clustering families allows both centroid-based and distribution-based regime definitions to be examined.

Taken together, the Cartesian product of preprocessing transformations, embedding methods, clustering algorithms, and hyperparameters defines a comprehensive model grid.

Columbia University

C. Evaluation Metrics

Since regime labels are unobserved, clustering performance is evaluated using label-free, quantitative criteria. Two complementary metrics are computed for each candidate model:

- **Silhouette score**, which measures cluster separation and cohesion within the embedded space.
- **Stability**, measured via the Adjusted Rand Index (ARI) across repeated perturbations of the data and clustering procedure, capturing the robustness of regime assignments to sampling variability.

Together, these metrics assess both the quality of the clustering and the consistency of the inferred regimes.

D. Model Scoring and Selection

To enable comparison across modeling choices, both evaluation metrics are normalized to the unit interval using min–max scaling across the full grid of candidate models. A composite score is then constructed as a weighted average of the normalized silhouette and stability scores. This scoring rule reflects a balanced preference for cluster separation and robustness, while avoiding reliance on a single evaluation criterion.

Mathematically this is expressed as follows: for model configuration $m \in \mathcal{M}$, let S_m denote its silhouette score and A_m its stability score. Define min–max normalized versions:

$$\tilde{S}_m = \frac{S_m - \min\limits_{j \in \mathcal{M}} S_j}{\max\limits_{j \in \mathcal{M}} S_j - \min\limits_{j \in \mathcal{M}} S_j} \qquad (1)$$

$\tilde{A}_m$ is defined equivalently as in equation (1). The composite score used for model selection is given by:

$$\text{Score}(m) = w\tilde{S}_m + (1 - w)\tilde{A}_m, \qquad w \in [0,1] \qquad (2)$$

Models are ranked according to this composite score, and the highest-scoring configurations are examined in greater detail in the empirical analysis.

IV. Empirical Studies

A. Grid Search Results

The full grid-based evaluation consists of 999 model configurations spanning preprocessing strategies, dimensionality reduction methods, clustering algorithms, and hyperparameter choices. Each configuration is evaluated using the composite score defined in Section III.D, which balances cluster separation and stability. The resulting score distribution reveals strong concentration among a narrow subset of modeling choices, with clear implications for both interpretation and computational cost.

Table I reports the three highest-scoring configurations. These results exhibit limited diversity, as all top-ranked models correspond to return-based preprocessing combined with PCA and a small number of clusters. While these configurations achieve near-perfect stability and strong silhouette values, their dominance suggests that global geometric criteria favor linear embeddings under return-based representations. As a result, a naïve ranking by composite score alone provides limited insight into how regime structure varies across alternative representations.

TABLE I. Top-Performing Regime Configurations

Preprocessing	Embedding	K	Clustering	Score
Returns	PCA	3	Hierarchial	1.000
Returns	PCA	3	Kmeans	1.000
Returns	PCA	5	Kmeans	0.994

While the return-based PCA configurations dominate the top of the ranking, this dominance should be interpreted with care. In particular, the high composite scores achieved by these models appear to arise in a degenerate regime-discovery sense, where the clustering solution is characterized by one overwhelmingly dominant regime and a small number of extreme, short-lived regimes corresponding to large market shocks. Such solutions naturally yield near-perfect stability under resampling and strong global separation metrics, as most observations are consistently assigned to a single cluster while periods of acute market stress are sharply isolated. In this sense, PCA on returns favors a coarse partition that emphasizes abrupt macroeconomic dislocations rather than a richer segmentation of intermediate market states. While this behavior is not undesirable, and indeed can be advantageous for identifying major structural breaks, it suggests that the top-ranked configurations reflect a specific notion of regime structure that prioritizes robustness to perturbations over diversity of latent states. Consequently, composite score maximization alone may over-select such degenerate solutions, motivating a broader analysis of regime quality beyond the highest-scoring models.

To obtain a more informative and diverse view of regime robustness, results are further analyzed by aggregating performance across preprocessing and dimensionality reduction pairs. For each pair, composite scores are averaged across clustering methods, cluster counts, and hyperparameter settings. This aggregation reveals greater insights in regime quality than the top-ranked configurations alone. While PCA remains competitive across preprocessing schemes, nonlinear embeddings such as UMAP and t-SNE exhibit comparable average performance in several settings, particularly when paired with standardized or raw representations.

TABLE II. Top Configurations Grouped By Preprocessing & Embedding

Preprocessing	Embedding	K	Silhouette	Stability	Score
Returns	PCA	3	0.946	1.000	1.000
Raw	PCA	5	0.545	0.999	0.792
Raw	UMAP	5	0.491	0.999	0.764
Raw	t-SNE	3	0.397	1.000	0.717
Zscore	UMAP	5	0.395	0.989	0.709
Returns	t-SNE	4	0.400	0.981	0.708
Zscore	t-SNE	5	0.386	0.980	0.700
Returns	UMAP	3	0.360	0.995	0.695
Zscore	PCA	3	0.317	0.958	0.653

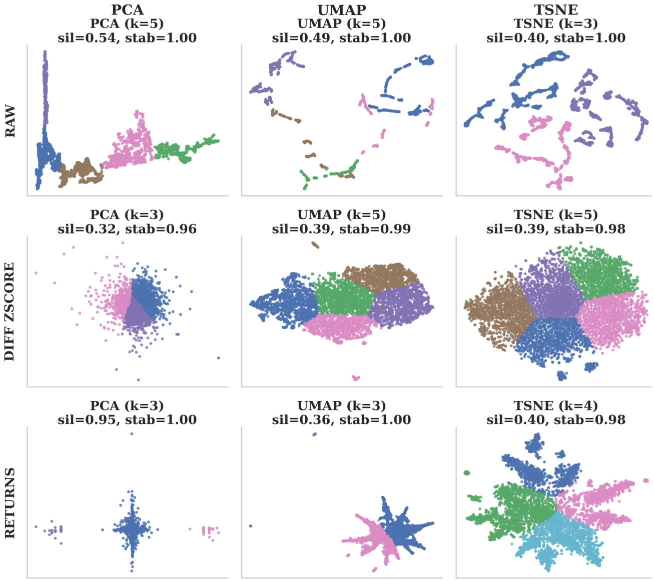
Together, these results demonstrate that regime discovery outcomes are highly sensitive to data representation and embedding choice. Although return-based PCA embeddings maximize stability and separation metrics, alternative

Columbia University

preprocessing, embedding combinations yield distinct and robust regime structures that are not captured by extreme score-based rankings alone.

B. *Effect of Preprocessing on Regime Structure*

To further illustrate the impact of preprocessing on regime discovery, representative two-dimensional embeddings from top-performing configurations are visualized for each preprocessing strategy. These projections reveal qualitatively distinct regime geometries despite similar clustering procedures.



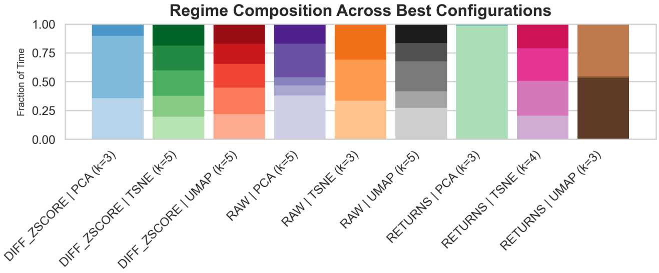
Raw time series tend to produce elongated clusters, reflecting long-run level variation and gradual transitions across observations. Differenced and standardized inputs yield more compact clusters that emphasize relative deviations and volatility-driven structure. In contrast, return-based preprocessing produces more clearly separated clusters under linear embeddings, with visually higher stability across repeated samples.

Together, these visual results demonstrate that preprocessing choices materially affect the geometry and separability of inferred regimes, even when downstream clustering methods are held fixed.

C. *Regimes: Composition Across Model Configurations*

Having identified the most stable clustering configurations through the grid search, we now examine the structure of the discovered regimes across preprocessing and dimensionality-reduction choices. Rather than interpreting regimes temporally or economically at this stage, we focus on how different modeling decisions partition the data and the relative prevalence of the resulting regimes.

The figure below summarizes the fraction of observations assigned to each regime for the best-performing configurations. Each stacked bar corresponds to a single combination of preprocessing method, dimensionality reduction technique, and number of clusters. Within each bar, distinct color shades indicate different regimes discovered under that specific configuration. Importantly, colors are treated as configuration-specific and are not aligned across bars, as cluster labels in unsupervised learning are inherently non-identifiable across different model specifications. Consequently, the figure should be interpreted in terms of within-configuration regime composition, rather than as a direct comparison of regime identities across models.



Several informative patterns emerge from this comparison. Configurations based on return preprocessing, most notably returns with PCA and $k = 3$, exhibit a highly concentrated regime composition, with the vast majority of observations assigned to a single dominant regime and only brief excursions into alternative regimes. This concentration likely explains the exceptionally high stability scores observed for these configurations: when most observations are assigned to a single cluster, resampling and perturbations of the data are unlikely to alter the clustering outcome, resulting in near-degenerate but highly stable solutions. In contrast, raw-level and differenced representations distribute observations more evenly across multiple regimes, indicating a richer partitioning of the data at the cost of lower, but still substantial, stability. These results highlight an important trade-off in unsupervised regime discovery: high stability may reflect either robust structural separation or, in some cases, a collapse toward a dominant regime, underscoring the need to jointly consider stability metrics and regime composition when selecting models.

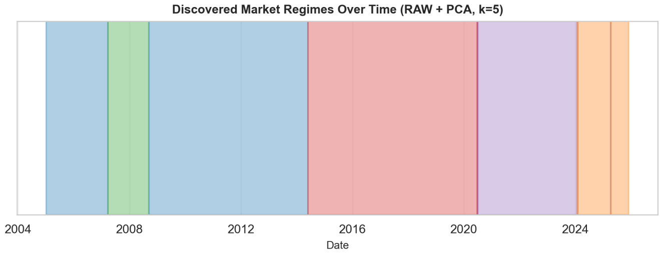
V. Discussion

A. *Interpretation*

While the empirical analysis focuses on stability and structural properties of the discovered regimes, temporal and economic interpretation is essential for assessing whether these regimes correspond to meaningful market states rather than purely statistical artifacts. Interpretable regime structure provides validation that the unsupervised workflow is capturing economically relevant variation in the data, particularly around known periods of market stress and transition.

The figure below presents the temporal evolution of discovered regimes for one of the most interpretable and stable configurations, raw-level preprocessing combined with PCA and $k = 5$ clusters. Rather than displaying pointwise regime assignments, the figure emphasizes regime block structure, highlighting contiguous periods during which the model assigns observations to the same latent regime. This visualization facilitates interpretation by revealing persistent market states and discrete transitions over time.

Columbia University



A particularly striking feature of this decomposition is the emergence of a distinct and temporally localized regime coinciding with the 2008–2009 Global Financial Crisis**.** This regime appears as a short-lived but clearly separated block, indicating that the model identifies the crisis period as fundamentally different from both preceding and subsequent market conditions. Importantly, this regime does not recur in later periods, suggesting that the clustering is capturing a genuine structural break rather than elevated volatility alone. The isolation of the Global Financial Crisis as its own regime provides strong evidence that raw-level preprocessing preserves information about large-scale market dislocations that may be suppressed under return-based or differenced transformations.

Beyond this canonical example, other high-performing configurations also exhibit interpretable temporal structure, though with differing emphases. Some specifications produce regimes characterized by long periods of stability punctuated by brief transitional states, while others fragment the sample into more frequent but shorter-lived regimes. These differences reinforce a central finding of the study: preprocessing and representation choices materially affect not only the stability of discovered regimes, but also the economic phenomena they emphasize**.** Together, these results underscore the importance of combining quantitative stability metrics with targeted temporal interpretation when evaluating unsupervised regime discovery methods.

B. Limitations

While the proposed framework provides a systematic and stable approach to unsupervised regime discovery, several limitations should be acknowledged.

- **Dependence on preprocessing choices**: As demonstrated in the empirical results, inferred regime structure is highly sensitive to the data representation. Even when clustering algorithms and evaluation metrics are held fixed, alternative preprocessing strategies can yield materially different regime partitions. This sensitivity complicates the interpretation of regimes as intrinsic economic states.
- **Substantial computational cost:** The Cartesian product of preprocessing choices, embedding methods, clustering algorithms, and hyperparameters results in a large combinatorial search space. Even with a deliberately constrained grid of fewer than 1,000 model configurations, the full evaluation required approximately **60 minutes** on a multi-core CPU, limiting the feasibility of broader hyperparameter exploration. As a result, the reported grid represents a pragmatic rather than exhaustive search of the model space.
- **Lack of ground-truth regime labels:** In the absence of observed regime assignments, evaluation relies on internal criteria such as silhouette score and clustering stability. While these metrics quantify geometric separation and robustness, they do not directly measure economic validity.
- **Static clustering assumption:** The clustering procedures applied treat observations as exchangeable in time, ignoring temporal dependence and transition dynamics. As a result, regime persistence and switching behavior are inferred post hoc rather than explicitly modeled.

Despite these limitations, examining the temporal evolution of the highest-scoring regime configuration provides useful insight into how the inferred regimes align with well-known macroeconomic episodes.

VI Acknowledgment

The author thanks the instructor of STAT GR 5244 at Columbia University for providing the academic framework for this study.

VII References

[1] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," Econometrica, vol. 57, no. 2, pp. 357–384, Mar. 1989. [Online]. Available: https://www.jstor.org/stable/1912559

[2] A. Botte and D. Bao, "A machine learning approach to regime modeling," Two Sigma Insights, Oct. 2021. [Online]. Available: https://www.twosigma.com/articles/a-machine-learning-approach-to-regime-modeling

[3] G. I. Allen, A. Chang, T. M. Tang, and T. M. Zikry, "Unsupervised machine learning for scientific discovery: Workflow and best practices," arXiv:2506.04553, Jun. 2025. [Online]. Available: https://arxiv.org/abs/2506.04553