Robust Rolling Regime Detection (R2-RD): A Data-Driven Perspective of Financial Markets*

Ali Hirsa†, Sikun Xu‡, Satyan Malhotra§

February 16, 2024

Abstract

The nonstationary and high-dimensional nature of financial markets poses significant challenges for navigation. Temporally stable regime classification offers a perspective to manage these challenges. We propose the Robust Rolling Regime Detection (R2-RD) framework that adaptively retrains with streaming data and employs temporal ensemble, label assignment, and threshold policies to address temporal instability resulting from nonstationarity, model mismatches, etc. Further, the R2-RD framework's data-driven model selection procedure chooses the model that best describes the data from the wide variety of latent variable models. We demonstrate the application and ease of extensions of R2-RD via two different datasets: macroeconomic and futures markets. Numerical experiments also illustrate how different macroeconomic regimes separate the performance of mutual funds, allowing for regime-aware asset management. The findings make R2-RD an ideal support for data-driven decision models, and the implementations can be leveraged across investments, credit, risk, policies, etc.

1 Introduction

Financial markets are inherently challenging to navigate due to their pronounced non-stationarity and high dimensionality. The concept of *regimes* offers one way of bringing some order to these complexities. This perspective assumes the existence of a small number of distinct regimes, each marked by similar market behaviors. For instance, some funds or strategies may excel during inflationary times, yet falter in deflationary periods. Such variability underscores the importance for allocators to discern how to manage capital at risk. Where understanding the defining attributes of these regimes, as well as the dynamics of transitions between them, becomes paramount. Such insights are critical across all facets of the Investment Life Cycle, including security selection, asset allocation, portfolio construction, asset planning, and risk management. Implementations can also benefit other financial processes, including credit, risk, policies, etc.

Financial regimes, by their nature, are not directly observable. Hence, latent variable models become an ideal choice for their discovery. Among these, the family of Hidden Markov Model (HMM) stands out as particularly effective. At its core, a vanilla HMM presumes a finite number of hidden states (which, in our study, are termed as *regimes*) whose transition dynamics are Markovian, while the observations are conditionally independent given the hidden states. Variations of HMM relax the assumptions from the vanilla version, such as Markovian regimes and conditional independence of the observables, leading to models like the Hidden Semi-Markov Model (HSMM), Auto-Regressive HMM (ARHMM, also known as the Markov Switching Model), etc. Our goal is to delineate the unique features of each regime with these models and derive a time series that indicates the trajectory of realized regimes based on given observations.

When we apply a model to data, we are inherently relying on certain assumptions. However, since no assumption is perfectly valid in real-world scenarios, it's crucial to have a range of models at our disposal. This allows us to choose the one that most appropriately fits the specific situation. Selecting the *best* model is not

*Special thanks to Federico Kinkert, Miao Wang, Suraj Keshri and Ryan Holmes for their comments, suggestions, and implementations. Thanks also to the students at the Industrial Engineering and Operations Research Department at Columbia University for their contributions.

†Industrial Engineering & Operations Research Department, Columbia University, **ah2347@columbia.edu**, Chief Scientific Officer, Ask2.ai, **ali.hirsa@ask2.ai**

‡Olin Business School, Washington University in St. Louis, **sikun@wustl.edu** and **sikun.xu@ask2.ai**

§CEO of Ask2.ai, **satyan.malhotra@ask2.ai**

straightforward, especially for unsupervised tasks like regime detection. In this paper, we leverage a combination of classic statistical concepts, the marginal likelihood (or its approximation), and a geometric measure, the Silhouette score, to choose the model that best describes the data. The marginal likelihood of a model measures the probability of observing the data conditioning on the model, which helps us choose among different families of models, like the different variations of `HMM` mentioned above. The Silhouette score then determines the optimal number of regimes given a family of models to maximize inter-cluster separation and intra-cluster similarity.

Fitting a model once to all the time series data ignores the internal dynamics of how the market evolves across time. Therein, another vital emphasis of this paper is the need for rolling- or expanding-window retraining (or fine-tuning) with financial time series data to gain insights about temporal instability. By temporal instability, we refer to the unstable data generation process in financial markets, which summarizes concepts such as nonstationarity and non-Markovian. Temporal instability is even more challenging with optimization errors from solving a nonconvex model. A local optimum found from training the model in the first period could drastically differ from the one found from retraining the model in the second period. Such inconsistency is a result of both temporal instability and optimization error. We will illustrate the inconsistency issue in Section 4 using a popular regime model ([2]). To address this issue, `R2-RD` uses temporal ensemble techniques, which smooth the models fitted consecutively in order to alleviate the inconsistency issue from optimization error.

Another challenge `R2-RD` addresses is labeling. Like most unsupervised learning models, a regime detection model, say a vanilla `HMM`, generates several unlabeled clusters (regimes). The order of regimes in which the algorithm returns depends on the initialization and optimization processes, which are mostly stochastic by construction. Suppose we assign labels to regimes according to such orderings. In that case, we are effectively doing a random assignment, which is troublesome during retraining, as there will be mismatches between the previously found regimes and the newly refitted ones. For example, regime 1 from the previously fitted `HMM` describes a high gross domestic product (GDP) environment, while regime 1 from the newly fitted `HMM` could correspond to a low GDP environment. We introduce a label assignment model to resolve this problem without touching the internal optimization process of the regime models. The label assignment model is an integer programming that minimizes the total costs of assigning the labels from previously defined regimes to the newly found ones, where the costs are measured by statistical distances.

Lastly, we aim to tackle the task of determining the appropriate number of regimes. Although the model selection framework partially addresses this issue by comparing the Silhouette score, this approach primarily applies to the initial training period. For instance, suppose the first training session selects a `HMM` with 4 regimes. However, in subsequent periods, the score might suggest a `HMM` with only 3 regimes. Our objective is to maintain historical consistency: once a regime is established, it should remain present and not vanish, a premise we consider reasonable. Therefore, we introduce a threshold policy that relies on optimal label assignment costs to determine the number of regimes in ongoing recursive training sessions following the initial fit. This policy helps in deciding whether a new regime emerges in the market based on recent observations.

We demonstrate the efficacy of `R2-RD` using two distinct datasets: macroeconomic and futures markets. In the macroeconomic context, our method identified 4 unique regimes prior to 2008, and an additional one subsequently. Further analysis of regime characteristics, through rolling retraining, reveals shifting distributions over time for each regime. This shift underscores the necessity for dynamic decision models in financial markets. A similar pattern is observed in the futures markets. Moreover, our analysis indicates a pronounced distinction in mutual fund performances across different macroeconomic regimes, highlighting the substantial potential of regime-aware asset management strategies.

Our paper is organized as follows. Section 2 discusses the relevant literature on regime detection and hidden Markov models. In Section 3, we propose the `R2-RD`, including the label assignment problems and the threshold policy for determining the emergence of new regimes. In Section 4, we apply `R2-RD` to macroeconomic data, and in Section 5, we apply the `R2-RD` method to the futures market data. In Section 6, we compare different types of regimes and different datasets, and we examine the performance of mutual funds under the identified macroeconomic regimes (demonstrating the potential of regimes for many downstream tasks like security selection and portfolio construction). Finally, we conclude our paper in Section 7 and Section 8 with a discussion on enhancements underway.

2 Literature Review

Our paper is closely related to the literature on hidden Markov models. `HMM` has a wide range of applications, including biology [5], econometrics [7], computational linguistics [12], etc. Because of its wide applications,

researchers have designed different variations of the vanilla HMM. To capture the autoregressive nature of the observation process, [10] designed an Autoregressive HMM (ARHMM), where each observation is generated by both the hidden state and the previous observation. [6] proposed a Hierarchical HMM to model the scenarios with more than one layer of hidden states. That is, the hidden state in the HMM itself is generated by another deeper hidden state sequence. [18] discussed using the Hidden Semi-Markov Model (HSMM) to explicitly model the time the system has spent in a state as a factor in the transition dynamics.

The econometrics literature digs deep into the theoretical properties of different HMMs. This line of research focuses more on the consistency and asymptotic normality of the model. [14] proved consistency and asymptotic normality in a misspecified HMM with a finite state space. [4] extended the result to misspecified HMM with general hidden state spaces. [1] discussed the case where the hidden state transition also depends on the previous observations, namely covariates-dependent transition. They proved consistency of the model, assuming it is correctly specified. [15] then extended the results to misspecified models.

Our contribution to the literature is a comprehensive framework that utilizes the family of HMM for online regime detection. We emphasize the need for smart model selection and introduce a procedure based on statistical and geometric measures. R2-RD also adopts temporal ensemble, label assignment, and a threshold policy that alleviates the temporal instability issue and identifies the emergence of new regimes. In addition, we provide strong empirical evidence of nonstationary market dynamics that can only be found by our recursive framework.

3 Robust Rolling Regime Detection (R2-RD)

Consider a sample of multi-dimensional time series data $\{\mathbf{X}_t\}_{t=1}^T$. The observation in each period $\mathbf{X}_t$ is a D-dimensional column vector. We assume that there exists an unobserved discrete-time stochastic process $\{Z_t\}_{t=1}^T$ with a finite state space $\mathcal{Z}$. Consider the case of macroeconomic regimes; the D-dimensional time series data is the collection of macroeconomic indicators of Gross Domestic Product (GDP), unemployment rate, Consumer Price Index (CPI), etc. If we only look at GDP and CPI, Merrill Lynch's investment clock defines the state space as: $\mathcal{Z} = \{\text{Reflation}, \text{Recovery}, \text{Overheat}, \text{Stagflation}\}$. Using statistical models like HMM, we are able to consider more than just GDP and CPI and extract regime information from a much larger feature space.

The vanilla HMM with Gaussian observation model is a good benchmark latent variable model for regime detection. It assumes that the hidden state Z_t follows a Markov chain, and each observation $\mathbf{X}_t$ is conditionally independent given the hidden state. Gaussian observation model refers to the fact that $\mathbf{X}_t$ follows a D-dimensional Gaussian distribution, whose parameters depend on the hidden state Z_t. Variations of HMM are obtained by either relaxing the Markovian assumption or designing a more complicated observation model. For example, Gaussian mixture HMM uses the Gaussian mixture as the observation model. Autoregressive HMM relaxes the conditional independence assumption for observation and assumes that the parameters of $\mathbf{X}_t$'s distribution not only depend on Z_t but also $\mathbf{X}_{t-1}$. Each model makes certain assumptions about the underlying data generation process and we could never tell which one is closer to reality. Thus, in Section 3.1, we propose a model selection procedure for selecting the model that best describes the data statistically and also best separates the regimes geometrically.

HMM and its variations are single-stage models in that it is not designed for retraining or fine-tuning with streaming data. Thus, in later sections, we introduce the temporal ensemble, the label assignment model, and a threshold policy to ensure we can find a stable trajectory of regime dynamics. Altogether, this framework is called Robust Rolling Regime Detection (R2-RD).

3.1 Model selection

Model selection remains a crucial aspect in machine learning, as it enables us to critically evaluate and justify the underlying assumptions of various candidate models. Therefore, our approach emphasizes selecting models through data-driven methods, employing widely recognized techniques such as cross-validation. However, in many applications in unsupervised learning, social science, etc., cross-validation does not help much since we do not have the ground truth (e.g., label). Consequently, we incorporate a key statistical concept for model selection: marginal likelihood (see [3], [11], [16]). This metric, by its very nature, assists us in selecting the model that most closely aligns with the data. We will now elaborate on its underlying principles.

Let us first define marginal likelihood. Consider a set of models (hypothesis) $(M_1, \ldots, M_K)$, parameterized by $(\theta_1, \ldots, \theta_K)$ respectively. Given data $\mathbf{Y} = \{y_i\}_{i=1}^N$, the marginal likelihood of model M_i is defined as:

$$m(\mathbf{Y}|M_i) = \int_{\theta_i \in \Theta_i} \Pr(\mathbf{Y}|M_i, \theta_i)\pi(\theta_i|M_i)d\theta_i \tag{1}$$

where $\Pr(\mathbf{Y}|M_i, \theta_i)$ is the likelihood (or outcome model) and $\pi(\theta_i|M_i)$ is the prior distribution of parameters θ_i. This value quantifies the probability of observing the data set $\mathbf{Y}$ given model M_i. Next, we demonstrate why marginal likelihood can be used for model selection. Applying the Bayes formula, we have:

$$\Pr(M_i|\mathbf{Y}) = \frac{m(\mathbf{Y}|M_i)\Pr(M_i)}{\sum_k m(\mathbf{Y}|M_k)\Pr(M_k)}$$

where $\Pr(M_i)$ is the prior distribution for model i, and it is typically defined to be discrete uniform so that $\Pr(M_i) = \Pr(M_j)$ for all i and j. $\Pr(M_i|\mathbf{Y})$ is the posterior distribution and the value of interest, which describes the probability that M_i is the model generating the data given the observed data.

Finally, in order to select the model that best describes the data, we calculate the posterior odds:

$$\frac{\Pr(M_i|\mathbf{Y})}{\Pr(M_j|\mathbf{Y})} = \frac{m(\mathbf{Y}|M_i)}{m(\mathbf{Y}|M_j)}$$

The right-hand side is also called the Bayes factor. Notice that model comparison reduces to the problem of comparing marginal likelihood. That is, the higher the marginal likelihood, the better the model can describe the observed data. When the Bayes factor between two models, M_i and M_j is close to 1, indicating that the data does not significantly differentiate between the two models, we must consider alternative metrics for model selection, the silhouette score would be a good choice. In general, the calculation of the marginal likelihood is complicated. The authors in [3] discovered the Chib identity and conducted the Markov Chain Monte Carlo (`MCMC`) method to acquire this value. Other approximations, like the Laplace method, are also proposed for the calculation. A prominent approximation of the marginal likelihood is the Bayesian Information Criterion (`BIC`) score. Assuming unit information prior and using the central limit theorem for posterior distribution, the log of marginal likelihood reduces to the Schwarz criterion, minus twice of which is the `BIC` score:

$$\texttt{BIC}_i = -2\log\Pr(\mathbf{Y}|M_i, \theta_i) + d\log N$$

where d is the number of parameters of model M_i. Hence, a model with a larger marginal log-likelihood is approximately equivalent to a smaller `BIC` score. Then, if the precise marginal likelihood is unavailable, we could select models with a small `BIC` score.

However, marginal likelihood could have unsatisfying performance if the prior distribution is not chosen appropriately. Such issues would lead to ambiguous selection among models that are *close* to each other. `BIC` approximation has the same selection ambiguity issue. Thus, we only apply marginal likelihood-based selection among model families. For example, `HMM` with Gaussian observation model, `HMM` with Gaussian mixture observation model, `HSMM`, and `ARHMM`, etc. For comparison within a family of models, i.e., hyper-parameter selection, we resort to measures with geometric meanings: the Silhouette score. The Silhouette score measures the similarity of a data point to its own cluster against the dissimilarity of it to other clusters. It provides a more intuitive comparison for models with different numbers of clusters (regimes). Thus, our model selection procedure is summarized as the following process:

1. Compare model families using the marginal likelihood (or `BIC` approximation) and choose the one with the largest likelihood (or smallest `BIC` score)
2. Within the selected model family, choose the model with the best geometric separation (the highest Silhouette score)

Finally, we want to emphasize that the proposed model selection procedure is designed with the purpose of describing the data in the most statistically and geometrically meaningful way. If the focus is to improve the performance of some downstream tasks like security selection and portfolio construction, then it is better to customize a selection process that serves the best interest of the downstream tasks.

3.2 Temporal ensemble

Many latent variable models have nonconvex objective functions, solving which could lead to potentially large optimization errors (sub-optimality). Although the local optima might be close to the global one in the case of vanilla `HMM` due to the Baum-Welch algorithm ([17]), we do not have such nice guarantees for other more complicated variations. A direct consequence of optimization error is that when solving a model with rolling

window data, the sequence of locally optimal solutions could be far away from each other, leading to a highly volatile trajectory of regimes. Thus, to improve the robustness of our framework and find a stable trajectory of regimes, we need a way to mitigate the impact of optimization error.

Essentially, we want to smooth the sequence of models from rolling window fitting. Smoothing in the Hilbert space is complicated, so we introduce temporal ensemble, a set of heuristics for smoothing statistical models fitted across time. One of the techniques is temporal initialization, where we use the solution from a fitted model to initialize the subsequent model. Let us use `HMM` with the Gaussian observation model as an example. Suppose we fit the model on a sub-sequence $\{\mathbf{X}_t\}_{t=1}^{t_1}$ of data with 2 regimes. Denoted the regimes found as $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$. As we move forward to time t_2, we acquire new data $\{\mathbf{X}_t\}_{t=t_1+1}^{t_2}$ and retrain the model using all the data $\{\mathbf{X}_t\}_{t=1}^{t_2}$. Instead of random initialization or initializing with K-means, which is a common practice for vanilla `HMM`, we initialize the Expectation-Maximization algorithm at the previously found regimes $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$. If we are solving the model for the first time, we follow a simple optimization trick by randomly choosing multiple initializations and selecting the one that leads to the best objective. We only apply this trick to the first training because it could be time-consuming, and we do not have the consistency issue when we do not have fitted models to compare with.

Another useful temporal ensemble method is averaging model parameters across time. This is helpful, particularly when we are dealing with forecasting problems. That is, when we use the transition matrix from `HMM` type model to forecast future regime distributions, (exponentially) averaging the transition matrix from multiple training could alleviate the impact of temporal instability and optimization errors.

3.3 Label assignment

Consider a generalization of the previous example. When fitted to $\{\mathbf{X}_t\}_{t=1}^{t_k}$, we obtained a set of regimes denoted as I. When fitted to $\{\mathbf{X}_t\}_{t=1}^{t_{k+1}}$, we obtained a new set of regimes denoted as J. If we assign the label of an old regime $i \in I$ to a new regime $j \in J$, it will incur an assignment cost c_{ij}, measured by the statistical distance between the two regimes. In the case where the old and new regimes are modeled as two Gaussian distributions, an intuitive choice for the distance measure would be a weighted difference between the first two moments:

$$c_{ij} = \alpha\|\mu_i - \mu_j\|_2^2 + (1-\alpha)\|\Sigma_1 - \Sigma_2\|_F \tag{2}$$

α is a hyperparameter controlling the relative importance between the first and second moments. $\|\cdot\|_F$ denotes the Frobenius norm. Because the first two moments are sufficient statistics for a Gaussian distribution, (2) is general enough for quantifying the distance between Gaussian distributions.

However, in the case of more complicated distributions, we need to use other statistical distances. Distances from the information geometry literature (e.g., Kullback-Leibler distance) or optimal transportation costs (e.g., Wasserstein distance) are good choices for the assignment cost. Choosing among different cost functions is not trivial and should be carefully selected in practice, but we will not dive into the details of the choices in this study.

The decision variables of the label assignment problem is the assignment matrix with entries $x_{ij} \in \{0, 1\}$ indicating whether we are assigning the label of old regime i to new regime j. The assignment problem can then be formulated as follows:

$$\begin{aligned} \min_{x_{ij}} \quad & \frac{1}{|I|}\sum_{i\in I}\sum_{j\in J} x_{ij}c_{ij} \\ \text{s.t.} \quad & \sum_{i\in I} x_{ij} \le 1, \forall j \in J \\ & \sum_{j\in J} x_{ij} = 1, \forall i \in I \\ & x_{ij} \in \{0,1\},\ \forall i \in I, j \in J \end{aligned} \tag{3}$$

The first constraint specifies that, at most, one label is assigned to a new regime. If $|J| > |I|$ and a new regime does not get assigned, then it is defined as an emerging new regime. The second constraint specifies that each label of the old regime is assigned to exactly one new regime.

Although the normalization constant $|I|$ in the objective does not affect the solution, we keep it here because it makes the objective value an average assignment cost that is comparable under different numbers of regimes. This is important for the threshold policy described below.

3.4 Threshold policy for determining emergence of new regimes

One drawback of single-stage models like HMM is a constant state space. In practice, there could be new regimes emerging in the market, which cannot be modeled by traditional HMM. Thus, we use the optimal objective values from the label assignment problem as a criterion for determining if there is a new regime appearing. We do not treat this problem as a general hyper-parameter tuning because we are explicitly modeling an emerging new regime as a geometrically distant distribution from the existing ones. The concept of geometrically distant is nicely captured by the label assignment cost as it precisely measures the total distance from the old regimes to the new regimes. In addition, we only consider either one regime or no regime emerging. The reason is that with granular enough data (typically monthly frequency for macroeconomic indicators and minute-level frequency for futures data), it is almost impossible to have more than one new regime emerging.

Specifically, given $|I|$ regimes found in $\{\mathbf{X}_t\}_{t=1}^{t_k}$, we solve two models on data $\{\mathbf{X}_t\}_{t=1}^{t_{k+1}}$, one with the number of regimes equal to $|I|$ and the other with $|I|+1$. Then, we solve the label assignment problem for both of them and get the optimal objective c_0 and c_1, respectively. If the assignment cost of adding one more regime is smaller $c_0 > c_1$ and the assignment cost of maintaining the same number of regimes is higher than a certain threshold $c_0 > \bar{c}$, then we say there is a new regime in the market. The intuition is that when the distance between the old and new regimes is too large and adding one more regime to the new ones can decrease the distance, we consider it an emerging regime situation.

4 Application: Macroeconomic regime detection

In this section, we investigate the macroeconomic regimes using R2-RD. After introducing the data in Section 4.1, we go through the rolling retraining process of R2-RD in Section 4.2 and Section 4.3. Section 4.4 compares the results from our method with [2], a popular practice adopted by the industry. We will show that R2-RD alleviates the temporal instability issues and uncovers interesting patterns in the macroeconomic environment. Finally, we display the evolution dynamics of regime characteristics in Section 4.5.

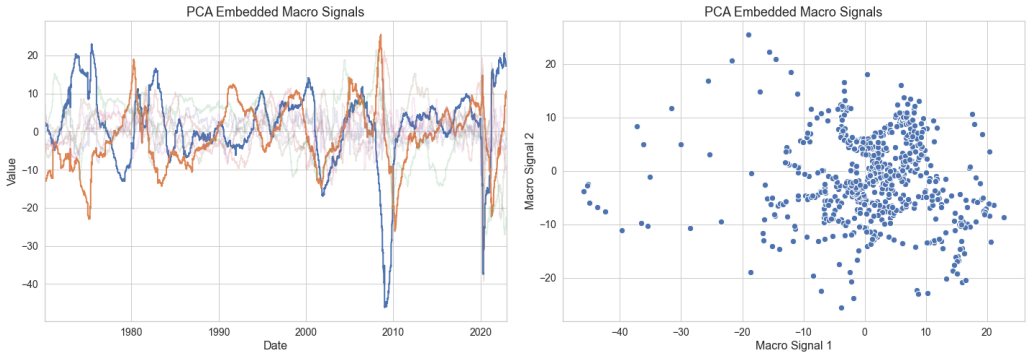
4.1 Macroeconomic Indicators

We selected over 650 macroeconomic indicators from the Federal Reserve Economic Data (FRED) from January 1970 to May 2022. The dataset covers a large variety of measures, including the GDP of various countries, rates with different maturities, Unemployment claims, treasury yields, etc. Figure 1 displays a snapshot of our database.

ask_id	fed_id	series_name	frequency	inception_da...	units
D062	SWPT	Assets: Central Bank Liquidity Swaps: Central B…	Weekly	2002-12-18	Millions
D063	WORAL	Assets: Other: Repurchase Agreements: Wedne…	Weekly	2002-12-18	Millions
D064	WSHOMCB	Assets: Securities Held Outright: Mortgage-Bac…	Weekly	2002-12-18	Millions
D065	TREAST	Assets: Securities Held Outright: U.S. Treasury…	Weekly	2002-12-18	Millions
D066	WALCL	Assets: Total Assets: Total Assets (Less Eliminat…	Weekly	2002-12-18	Millions
D067	AISRSA	Auto Inventory/Sales Ratio	Monthly	1993-01-01	Ratio
D068	CEU0500000003	Average Hourly Earnings of All Employees, Tota…	Monthly	2006-03-01	Dollars
D069	CES0500000003	Average Hourly Earnings of All Employees, Tota…	Monthly	2006-03-01	Dollars
D070	CES3000000008	Average Hourly Earnings of Production and Non…	Monthly	1939-01-01	Dollars
D071	CEU0500000008	Average Hourly Earnings of Production and Non…	Monthly	1964-01-01	Dollars
D072	ASPUS	Average Sales Price of Houses Sold for the Unit…	Quarterly	1963-01-01	Dollars
D073	SPTNSAWE	Average Sales Price of New Manufactured Hom…	Monthly	2014-01-01	Dollars
D074	AWHAEMAN	Average Weekly Hours of All Employees, Manuf…	Monthly	2006-03-01	Hours
D075	AWHAETP	Average Weekly Hours of All Employees, Total…	Monthly	2006-03-01	Hours
D076	AWHMAN	Average Weekly Hours of Production and Nons…	Monthly	1939-01-01	Hours
D077	UEMPMEAN	Average Weeks Unemployed	Monthly	1948-01-01	Weeks
D078	IEABCA	Balance on current account	Annual	1999-01-01	Millions
D079	QBPBSTAS	Balance Sheet: Total Assets	Quarterly	1984-01-01	Millons
D080	LOANINV	Bank Credit, All Commercial Banks SA	Monthly	1947-01-01	Billions
D081	H8B1001NCBCMG	Bank Credit, All Commercial Banks MoM Annual…	Monthly	1947-02-01	Percent
D082	LOANINVNSA	Bank Credit, All Commercial Banks NSA	Monthly	1947-01-01	Billions
D083	JPNASSETS	Bank of Japan: Total Assets for Japan	Monthly	1998-04-01	100 Millio…
D084	MPRIME	Bank Prime Loan Rate	Monthly	1949-01-01	Percent
D085	PRIME	Bank Prime Loan Rate Changes: Historical Dat…	Daily	1955-08-04	Percent
D086	EXBZUS	Brazilian Reals to U.S. Dollar Spot Exchange Rate	Monthly	1995-01-01	Brazilian…
D087	DEXCAUS	Canadian Dollars to U.S. Dollar Spot Exchange…	Monthly	1971-01-01	Canadian…
D088	TCU	Capacity Utilization: Total Index	Monthly	1967-01-01	Percent
D089	RKNANPUSA66…	Capital Stock at Constant National Prices for Un…	Annual	1950-01-01	Million of…
D090	H8B1048NCBCMG	Cash Assets, All Commercial Banks MoM Annu…	Monthly	1973-02-01	Percent
D091	CASACBM027S…	Cash Assets, All Commercial Banks	Monthly	1973-01-01	Billions
D092	FRGSHPUSM64…	Cass Freight Index: Shipments	Monthly	1990-01-01	Index Jan…

Figure 1: Sample macroeconomic indicators

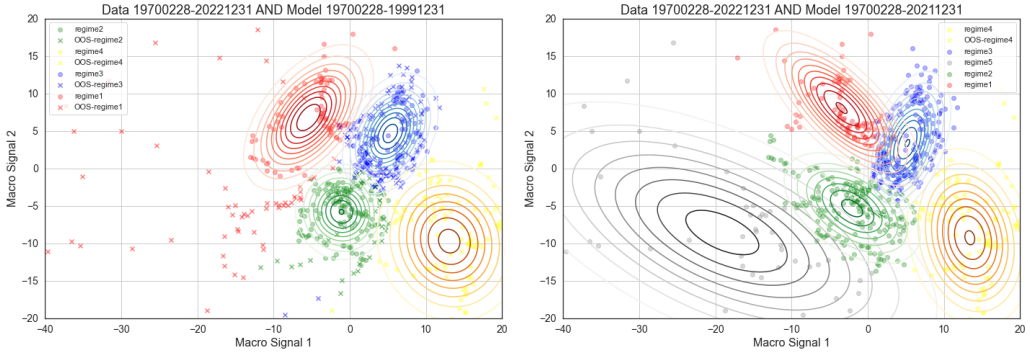
Notice that different indicators have different units, inception dates, and update frequencies, so we need careful preprocessing to align and transform them before further analysis. We adopt the robust rolling PCA (`R2-PCA`) method proposed in [9] to acquire the first eight principle components that explain over 80% of variance, as shown in Figure 2a. We highlight the first two principal components in blue and orange for better visualization. The scatter plot of the first two principal components is shown in Figure 2b.



(a) Embedded Macroeconomic Signals (b) First Two Principal Macro Signals

Figure 2: Macroeconomic Signals

Before going into the experiment details, we provide via visualization evidence of the temporal instability mentioned above. Latent variable models like `HMM` model nonstationarity as switching regimes (hidden states), where each regime is modeled as a fixed distribution across time. However, as shown in Figure 3a and Figure 3b, the distributions of regimes change across time. These two figures are generated using our `R2-RD` method with Figure 3a being the Gaussian `HMM` applied to data up to Jan. 2000 and Figure 3b being the Gaussian `HMM` retrained using data up to Jan. 2022. The dots in the figures represent in-sample data, and the crosses correspond to the out-of-sample ones. The regimes for the out-of-sample observations are identified via the trained regime detection model.



(a) Trained with data before 2000-01-01 (b) Trained with data before 2022-01-01

Figure 3: Regime distribution fitted using different input data

Clearly, the set of regimes fitted using data before January 2000 cannot describe the macroeconomic environment in a more modern era. It implies that the distribution depicting the same financial regime is evolving temporally. For example, a financial regime with average `CPI` of 2.25% is regarded as high inflation, but the 2.25% thresholds might change to 1.25% in a modern market condition, as we will show later in Section 4.5. A dynamic data-driven perspective of defining the regimes and describing their characteristics, as we are trying to accomplish with `R2-RD`, is more suitable than the common practice of fixed thresholding.

4.2 Initial training

When we train a regime detection model for the first time, we need to initialize the model with sufficient data so that the fitted regime distributions contain enough information. In our experiments of macroeconomic regime detection, we use the macro signals from January 1970 to December 1999 for initial training. First, we conduct the model selection process as described in Section 3.1. The candidate model families are `HMM` with Gaussian observation model and `HMM` with Gaussian mixture observation model. There is one hyper-parameter to decide within each family of models: the number of regimes, which ranges from 2 to 10. The comparison results are shown in Figure 4. The `BIC` score shows that for all choices of hyper-parameters, Gaussian `HMM` has a smaller `BIC`

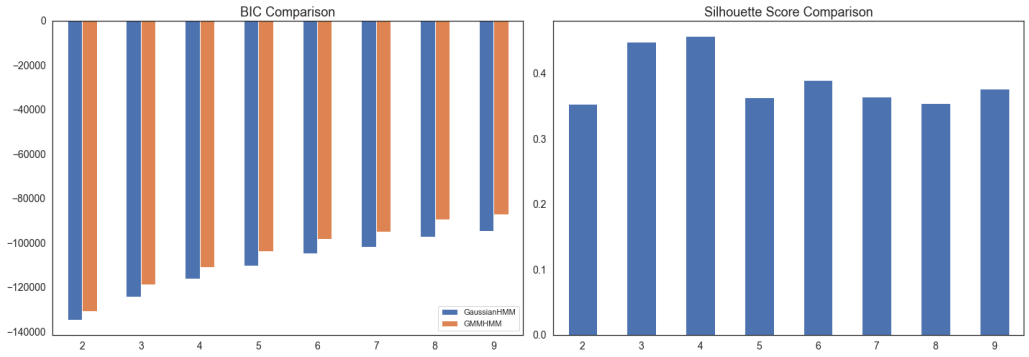
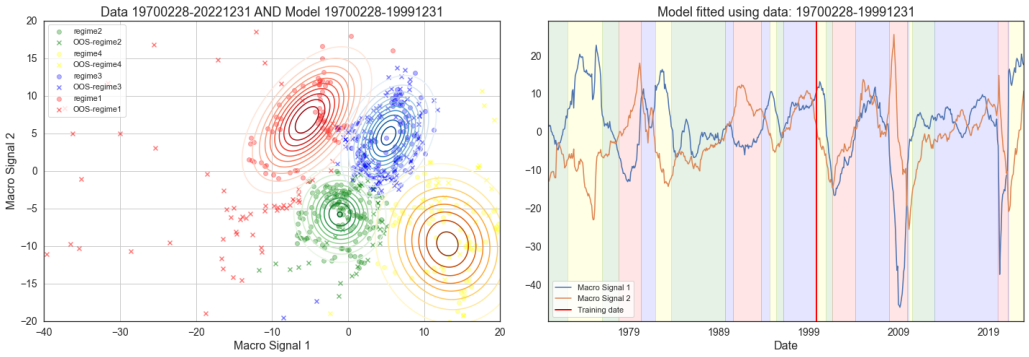


Figure 4: Model selection for macroeconomic regimes

score, hence a better model to describe the data. Then, four regimes have the highest Silhouette score within the Gaussian `HMM` family. Thus, the model selection procedure suggests that Gaussian `HMM` with 4 regimes is the best model for the initial training of macroeconomic regimes.

After training the selected model on the data for the first time, the regime distribution and transition plots are shown in Figure 5a and Figure 5b. The regime model utilizes the R2-KMeans [8] model which leads to consistent and stable centroids for initialization and rolling windows (Section 4.3).



(a) Regime distribution

(b) Regime transition

Figure 5: Initial training using macro signals from January 1970 to December 1999

Each color represents a different regime and is consistent in the two figures. For example, red indicates regime 1 in both Figure 5a and Figure 5b. Because regime detection models with unsupervised methods like `HMM` do not provide labeling for each regime, we need post-fitting analysis to gain insights and investigate the characteristics of the regimes. Table 1 summarizes four commonly used macroeconomic metrics in each regime, providing the unlabeled results with specific economic meanings.

	Real GDP (QoQ)	CPI (YoY)	Unemployment Rate	10-Yr Treasury Yield
Regime 1	2.20	6.67	6.69	8.42
Regime 2	4.01	4.25	6.37	9.20
Regime 3	3.06	4.64	6.00	9.41
Regime 4	3.09	6.20	7.50	9.67

Table 1: Summary statistics from initially-fitted `R2-RD`

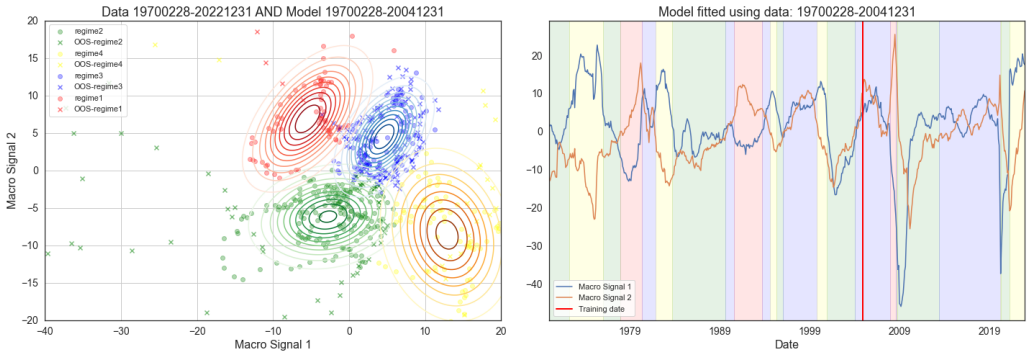
From Table 1, we see that the regimes are well-separated. Regime 1, colored red, describes the down period of the macroeconomic environment. It has low `GDP` growth, a high inflation rate, a high unemployment rate, and a low 10-year treasury yield. Regime 4, colored yellow, is similar to regime 1 but has a higher unemployment rate, possibly due to a high interest rate. On the contrary, regimes 2 and 3, colored green and blue respectively, correspond to the environments on the good side, where growth is strong, and inflation and unemployment rates are low. The difference between regime 2 and 3 are mainly in the `GDP` and `CPI`.

The out-of-sample performance of the model is also worth mentioning. The `R2-RD` model captures in the training data the bad economic conditions, including the oil crisis in the 1980s, the 1990 recession, etc. During the testing period (after Jan. 2000), the model identifies the 2008 financial crisis and the 2020 pandemic periods as similar bad economic conditions, as shown in Figure 5b.

4.3 Rolling retraining

In this section, we will refit the regime detection model under the `R2-RD` framework with an annual expanding window. It provides us with more in-depth insights into the dynamics of regime evolution.

As described in Section 3, when we have new data coming in, we refit a `HMM` model with the entire data set and use the regimes from the last fit as initialization. The labels for the regimes identified in the refitted `HMM` are assigned according to the label assignment model (3). The five refits from 2001 to 2005 are similar, signaling a stable market condition during the period. Thus, we only report the refit using data from January 1970 to January 2005. The regime distribution and transition with out-of-sample forecasts are shown in Figure 6a and Figure 6b.



(a) Regime distribution

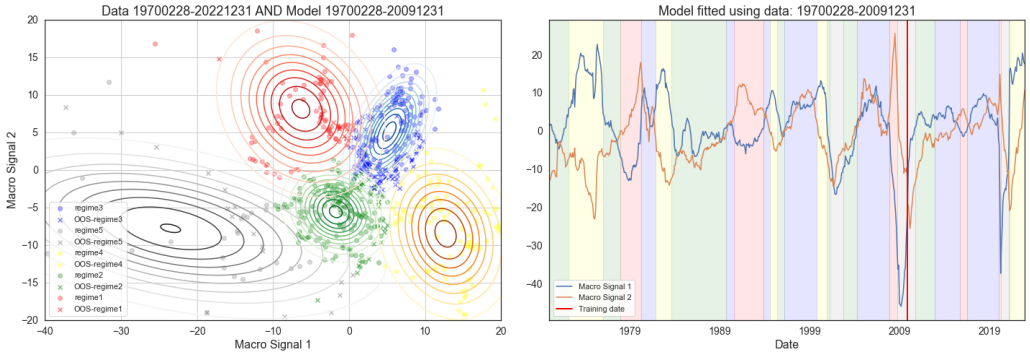
(b) Regime transition

Figure 6: Identified regimes using macro signals from January 1970 to December 2004

The major difference compared with the initial fitting is the covariance matrix of regime 2 (green). In the initial fitting, the marginal distributions for the two macro signals are almost independent under regime 2, while in the 2005 fit, they are positively correlated. But overall, the distribution and transition remain relatively stable, suggesting less need for frequent portfolio rebalancing and prediction model adjustment.

As we move past the 2008 financial crisis, data points on the lower left corner of the distribution plot (Figure 6a) realize, representing small values for both macro signals. Because they are (geometrically) far from the regimes identified in the previous periods, if we still set the number of regimes to be the same as before (3 regimes), at least one of the identified regime distributions will be distorted. Drastic distortion in the identified regimes leads to inconsistent characterization of regimes and hence should be avoided. The label assignment cost serves as a

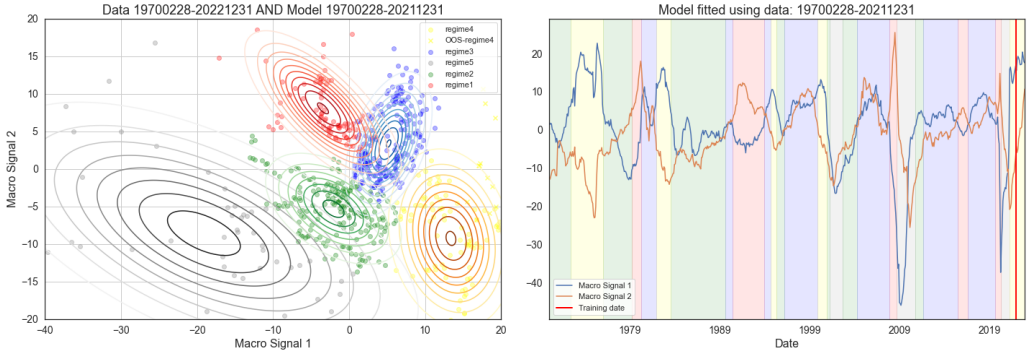
proxy indicator for such distortion. Thus, the threshold policy is triggered, suggesting that we should define a new regime for those newly appeared observations instead of distorting the identified ones. The resulting regime distribution and transition plots are shown in Figure 7a and Figure 7b.



(a) Regime distribution (2010-01-01) (b) Regime transition (2010-01-01)

Figure 7: Identified regimes using macro signals from January 1970 to December 2009

Finally, we jump to the last refitting time step on Jan. 1st 2022. As shown in Figures 8a and 8b, the regime characteristics do not exhibit much changes compared with the 2010 version. There are some data points whose assigned regimes changed, signaling a more suitable description of the market conditions at those times under a more modern lens.



(a) Regime distribution (2022-01-01) (b) Regime transition (2022-01-01)

Figure 8: Identified regimes using macro signals from January 1970 to December 2021

Lastly, we report some of the major macroeconomic indicators under each regime identified using the latest round of data in Table 2.

	Real GDP (QoQ)	CPI (YoY)	Unemployment Rate	10-Yr Treasury Yield
Regime 1	0.48	4.58	6.33	6.60
Regime 2	3.78	4.19	6.72	7.70
Regime 3	2.82	3.27	6.06	7.69
Regime 4	2.84	6.03	7.07	8.49
Regime 5	1.58	1.33	7.60	3.22

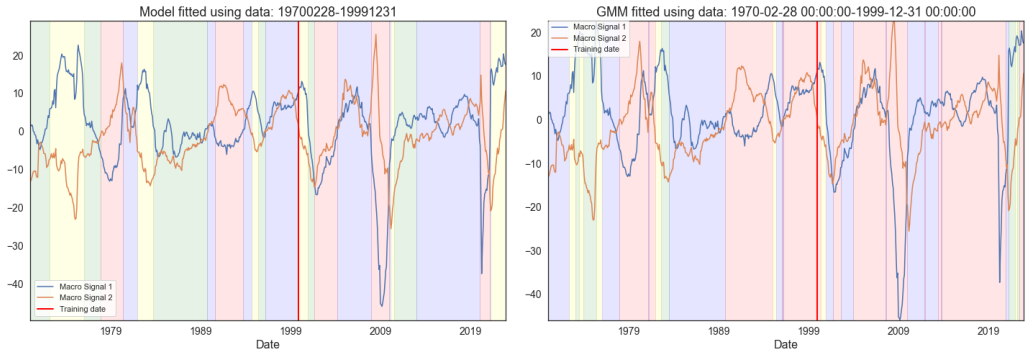
Table 2: Summary statistics from last-fitted `R2-RD`

First, notice that compared with the regime characteristics identified in the initial fit (Table 1), the `GDP` growth rates under all regimes are lower. Other characteristics also display obvious changes. It coincides with

our intuition that regime characteristics should be dynamic, and we cannot use fixed thresholds for determining macroeconomic conditions. Secondly, the new regime (regime 5 in gray) that appeared after the 2008 financial crisis represents another terrible market condition: extremely low GDP, CPI, and 10-year treasury yield, and the highest unemployment rate among all regimes. Powered by the recursive framework from R2-RD, we now have the dynamic view of how the regime characteristics evolve and a regime appears.

4.4 Comparison with Gaussian Mixture Model (GMM) for regime detection

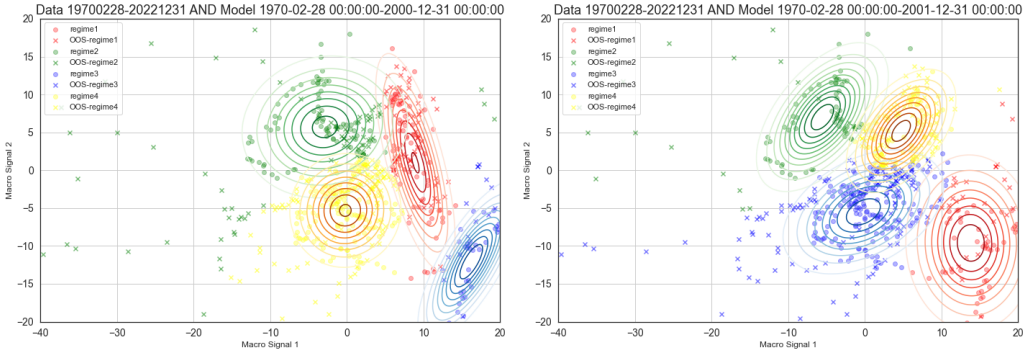
We assessed the performance of R2-RD versus other proposed Regime detection models[1]. Here, as an illustration, we compare the performance of R2-RD versus the Gaussian Mixture Model (GMM) ([2]). We find two weaknesses with GMM regime detection implementations. First, it does not consider the autocorrelation nature of time series data and falsely treats them as independent and identically distributed (i.i.d.) samples. As shown in Fig 9a and Fig 9b, GMM finds more frequent regime transitions indicated by the much thinner colored bars in the background.



(a) Regime transition from R2-RD (b) Regime transition from GMM

Figure 9: Identified regimes using macro signals from January 1970 to December 2000

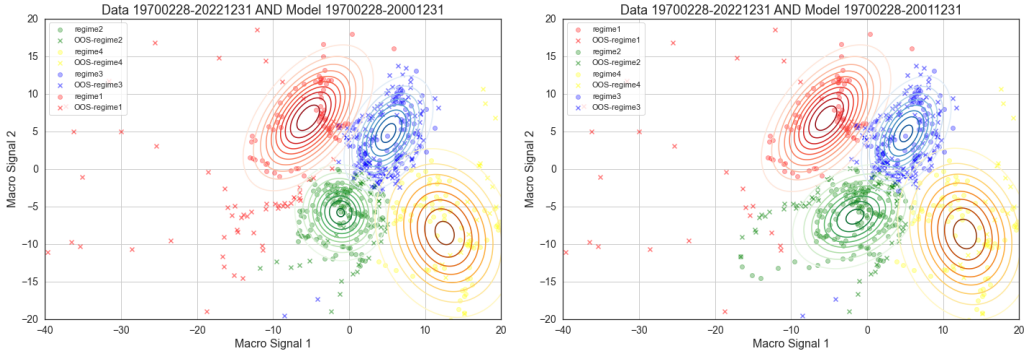
The other one is the inconsistent characterization of regimes during retraining. Fig 10a is the regime distribution identified by GMM on Jan. 1st, 2001, while Fig 10b shows the regime distribution fitted one year later on Jan. 1st, 2002. As we illustrate, the distributions change drastically. This may be because without proper initialization, optimizing the GMM model results in locally optimal solutions that are far from each other. Additionally, without label assignment, the default ordering of the regimes returned by GMM is different in each retraining. Hence, the labels are inconsistent. As a comparison, R2-RD identifies regimes more consistent across different retraining, as shown in Fig 11a and Fig 11b.



(a) Regime distribution (GMM) on Jan. 1st 2001 (b) Regime distribution (GMM) on Jan. 1st 2002

Figure 10: Distribution of GMM regimes differ significantly during retraining

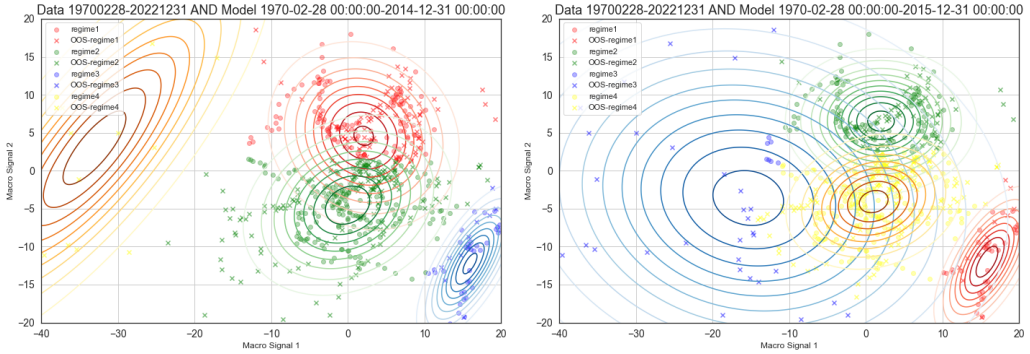
[1]See animation link: https://www.ask2.ai/r2rd/



(a) Regime distribution (R2-RD) on Jan. 1st 2001(b) Regime distribution (R2-RD) on Jan. 1st 2002

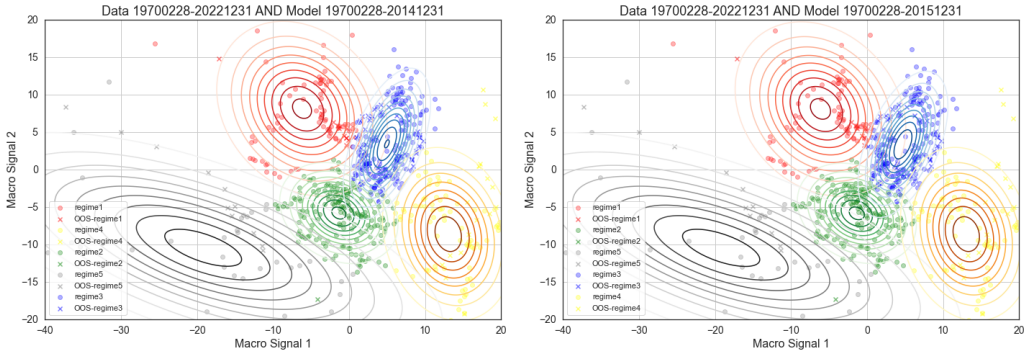
Figure 11: Distributions of R2-RD regimes show minimal variation during retraining

Similar results can be found in many other retraining periods. For example, Fig 12a and Fig 12b show the inconsistency of GMM from 2015 to 2016 while the R2-RD results in Fig 13a and Fig 13b are more consistent during the same period. Another reason for the consistent and stable results from R2-RD is its capability to identify new emerging regimes. The new regime can digest the new data patterns in the market so that the previously identified ones stay stable.



(a) Regime distribution (GMM) on Jan. 1st 2015 (b) Regime distribution (GMM) on Jan. 1st 2016

Figure 12: Distribution of GMM regimes differ significantly during retraining



(a) Regime distribution (R2-RD) on Jan. 1st 2015(b) Regime distribution (R2-RD) on Jan. 1st 2016

Figure 13: Distributions of R2-RD regimes show minimal variation during retraining

4.5 Regime evolution

After fitting the regime detection model rollingly, we acquire a trajectory of regime characteristics. This trajectory describes the regime distributions in each retraining, and hence, provides us with rich information regarding the evolution of some defining characteristics of the regimes. Thus, we could summarize the trajectory and examine the dynamics behind regime evolution. Let $\{\tau_1, \ldots, \tau_m\} \subset \{1, \ldots, T\}$ denote the set of time indices of retraining. Consider the τ-th retraining of Gaussian `HMM` model, which classifies all observations before τ as a sequence of regimes $\{r_1, \ldots, r_\tau\}$ where each regime r_t is an element of the set of regimes $\mathcal{J}$. With this sequence of regimes, we can summarize the performance of any measures under a specific regime. Let $\{W_t\}_{t=1}^{\tau}$ denote the time series of a measure of interest, say the Real GDP (QoQ). Then we can calculate the average Real GDP (QoQ) under regime j found by the τ-th Gaussian `HMM` as:

$$\bar{W}_{\tau j} = \frac{1}{\sum_{t=1}^{\tau} \mathbb{I}\{r_t = j\}} \sum_{t=1}^{\tau} \mathbb{I}\{r_t = j\} W_t \tag{4}$$

Finally, we can examine how the average Real GDP (QoQ) under regime j evolves across time by looking at the sequence $\{\bar{W}_{1j}, \ldots, \bar{W}_{\tau_m j}\}$. In Figure 14, we visualize such sequences $\{\bar{W}_{1j}, \ldots, \bar{W}_{\tau_m, j}\}$ for four prevalent macroeconomic indicators: Real GDP (QoQ), CPI (YoY), Unemployment Rate, and 10 Year Treasury Yield.

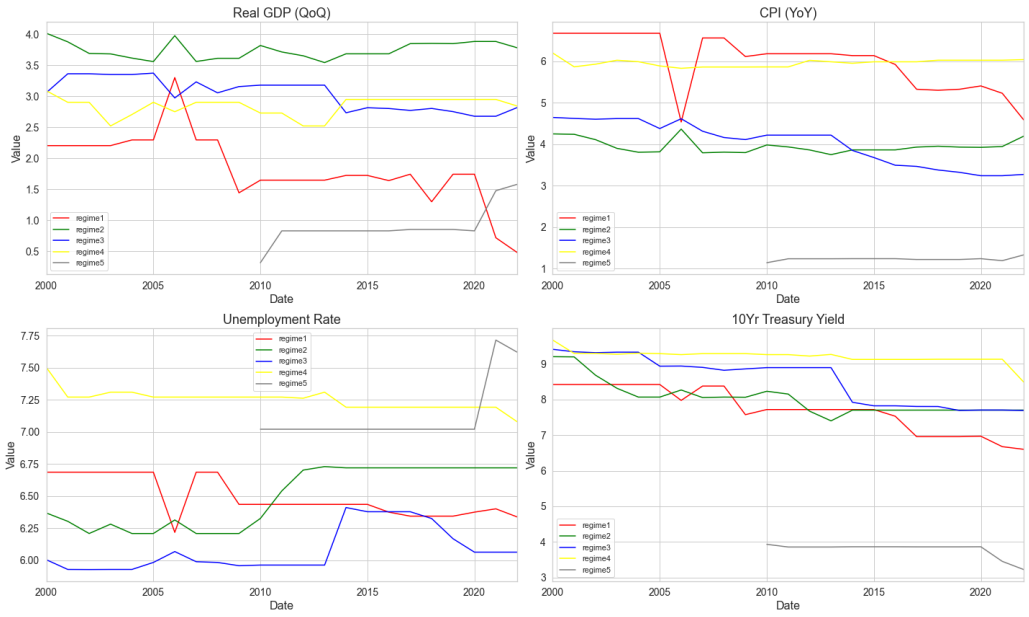


Figure 14: Evolving Dynamics of Some Macroeconomic Indicators under Different Regimes

Sequences constructed via (4) provide a brand new perspective on the nonstationary market and the distinction among different market conditions. We emphasize that some of the trajectories are not flat across time, addressing our claim that we should have a dynamic definition of financial and economic regimes. For example, the average CPI of regime 3 (blue) has a downward trend, while the average CPI of regime 4 (yellow) stays relatively stable, suggesting a growing distinction between the two. Another example is the average Real GDP of regime 1 (red), whose average value drops from 2.25% in the first decade of the 21st century to 1.25% in the second decade. Such nonstationarity suggests a need for dynamic policies (potentially regime-aware policies) for investment and policy making.

5 Application: Futures Market Regime Detection

To demonstrate the ease of extensions, in this section, we apply `R2-RD` to the futures market datasets. We focus on five major types of futures contracts with minute-level price series: Gold, U.S. Treasury (UST), e-mini (S&P 500), Crude oil, and EURUSD. The futures data covers the period from May 1998 to November 2023 with

minute-level granularity. In order to identify the regimes rather than simply capturing the long-term growth trend, we calculate 1-day rolling annual returns for each futures and pass it to `R2-RD`. We tested various rolling windows, ranging from minute level to hour level, and found that it does not affect the results much. Hence, we only report the 1-day rolling window results. The rolling annual return data is shown on the left of Figure 15.

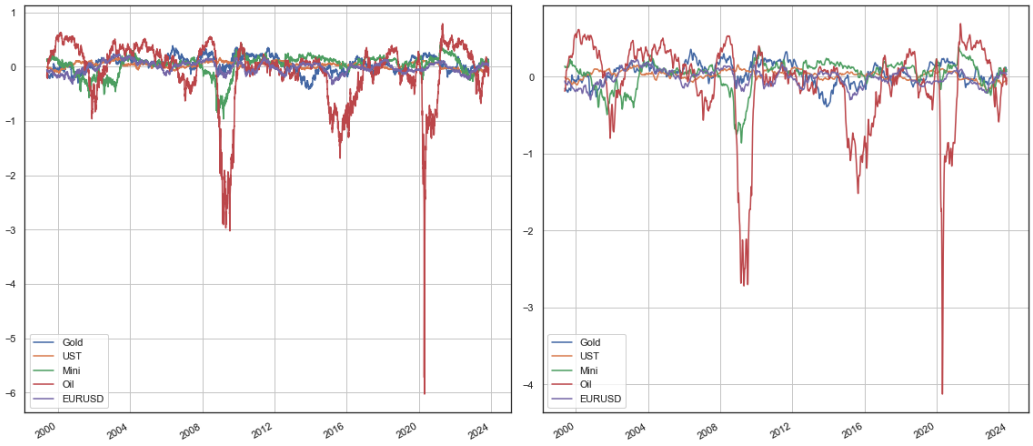


Figure 15: Original and smoothed rolling annual return

Because futures data are much more volatile than macroeconomic indicators, we also examined the impact of smoothing on regime identification. We adopt heat equation smoothing, which will not create lags in the data the way moving averages would. Compared with the original return data, the smoothed version is shown on the right of Figure 15.

5.1 Futures regimes

The initial training covers around 50% of the data in order to capture enough information about most of the regimes. First, Figure 16 shows model selection results. Candidate models are `HMM` with Gaussian observation models and `HMM` with Gaussian mixture observation models. The candidate number of regimes for both families of models ranges from 2 to 10. Because the metrics for the non-smoothed and smoothed data are very similar, we only report the ones for the non-smoothed data. The chart on the left shows the `BIC` score comparison between

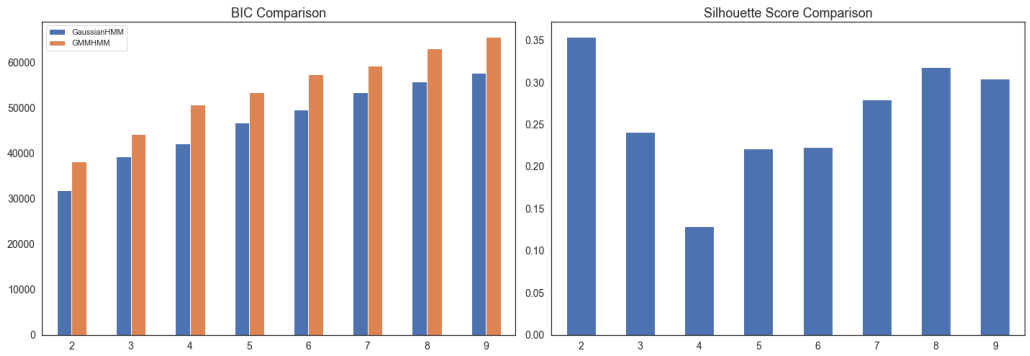


Figure 16: Model selection for futures regime

Gaussian `HMM` (blue) and Gaussian Mixture `HMM` (orange). We could see that Gaussian `HMM` has a lower `BIC` score across different number of regimes. Thus, it is a better model for describing futures data. With the Gaussian `HMM` family, we compare the Silhouette score based on the model selection procedure proposed in Section 3.1. Two regimes have the highest Silhouette score, indicating that it is the most geometrically reasonable hyper-parameter.

Thus, we train the futures regime detection model using Gaussian `HMM` with 2 regimes. The results are shown in Figure 17a and Figure 17b. Notice that there is almost no difference between the regimes fitted using

(a) Training with Non-Smoothed Data (b) Training with Smoothed Data

Figure 17: Initial training of futures regimes (up to Aug. 2011)

non-smoothed and smoothed data. The red regime (regime 1) corresponds to normal market conditions, while the green (regime 2) depicts the abnormal futures market.

As new data are being collected, the identified regimes stay relatively stable for several rounds of retraining. Figure 18a and Figure 18b show the last retraining of futures regimes with non-smoothed and smoothed data. Similar to the case in the initial training, smoothing does not make much difference to regime identification. Although most of the periods are identified as the same regime as the initial training, the regime for periods from 2004 to 2006 changes from regime 2 (green) to regime 1. The change is a result of regime characteristics calibration based on new observations. Although the drastic dive of the oil returns did not show up in the 2004 to 2006 period as in other regime 2 periods, the co-movement among different futures contracts plays a crucial role in the classification.

(a) Training with Non-Smoothed Data (b) Training with Smoothed Data

Figure 18: Last retraining of futures regimes (up to Aug. 2023)

The evolution dynamics of regime characteristics are reported in Figure 19. The dynamics chart shows that the two regimes can best separate UST, Mini, and Oil futures. Regime 1 (red) signals an active equity and oil market with a low interest rate. On the contrary, regime 2 (green) represents a worse equity market with a negative annual return on average and a high interest rate. In addition, notice that in more modern eras, the worse regime 2 is turning better compared with itself ten years ago. That is, the average return of Mini is higher in the 2020s compared with the 2010s.

Despite the interesting findings from the regime dynamics, we notice the similarity of dynamics between Gold and EURUSD futures contracts, especially under regime 2. It suggests that removing one of the two contracts could potentially improve the signal-to-noise ratio of input feature space. Figure 20a and Figure 20b show the regime transition of the smoothed 1-day rolling annual return series of Gold, UST, Mini, and Oil.

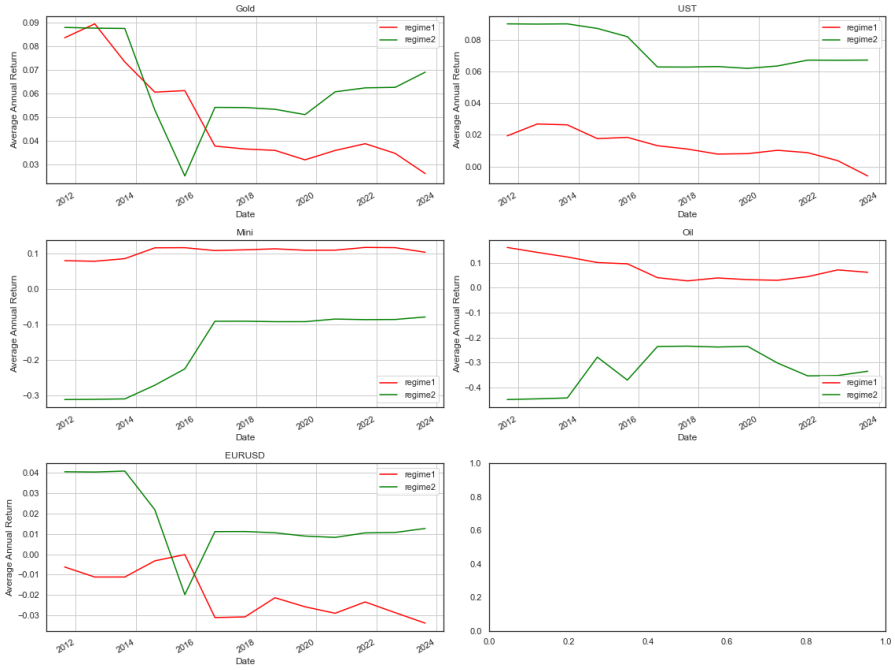


Figure 19: Evolution Dynamics of Futures Regimes

(a) Training with data before Aug. 2011 (b) Training with data before Aug. 2023

Figure 20: Futures regimes excluding EURUSD

Compared with regimes using all five contracts, regime transition excluding EURUSD is more stable in the sense that we have fewer frequent transitions, which are indicated by very thin green and red bars in the background. This suggests that rather than smoothing the input data with heat equation or other techniques, carefully selecting the input features plays a more significant role in regime detection performance.

Finally, we report the evolution dynamics of regime characteristics in Figure 21. We see similar patterns to the ones identified with five regimes in Figure 19. It reassures us about removing the EURUSD contracts and emphasizes again the need for careful feature selection for regime detection.

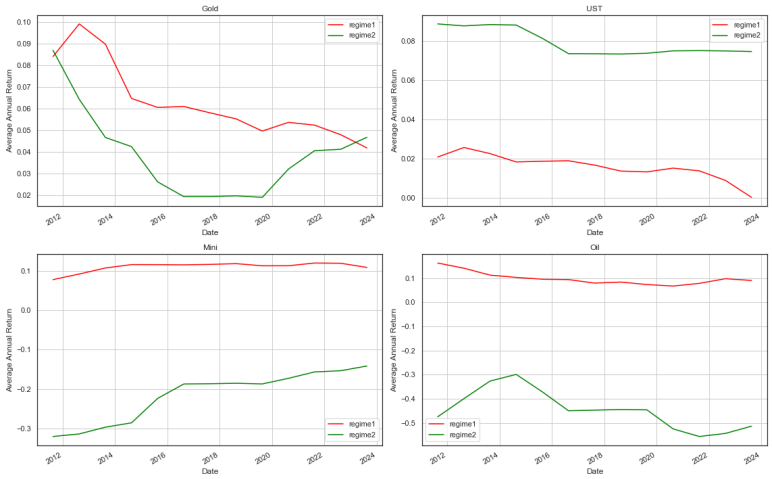


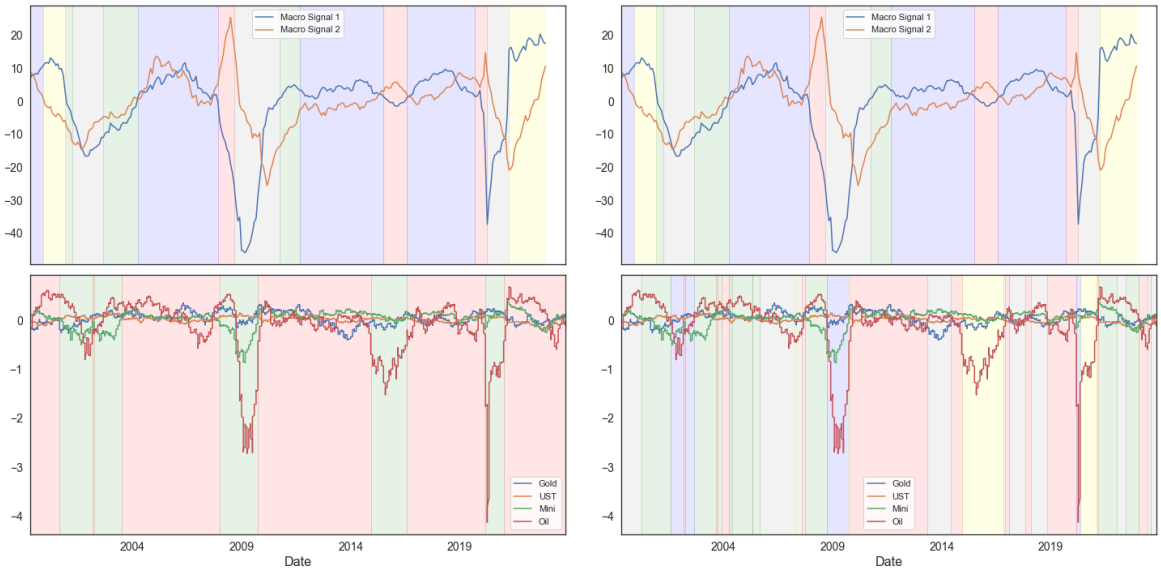
Figure 21: Evolution Dynamics of Futures Regimes (excluding EURUSD)

6 Regime Detection Assessments

In this section, we assess the macroeconomic regimes and futures regimes discovered by `R2-RD`. We begin by comparing the regimes in both markets to find similarities and cross-market predictability. Then, we apply `R2-RD` to another widely used dataset of macroeconomic indicators to examine the robustness of our method. Finally, we investigate the implication of macroeconomic regimes for mutual fund investment and discuss the potential of regime-aware asset management.

6.1 Futures versus Macroeconomic data

We compare the regimes identified in the futures market and in the macroeconomic environment. We compare the results trained using all available data in Figure 22a. The comparison covers periods from 1998 to 2023, where we have data for both markets.



(a) No. of Futures Regimes Chosen via Model Selection (b) No. of Futures Regimes Chosen the same as Macro's

Figure 22: Macroeconomic Regimes (above) v.s. Futures Regimes (below)

First, the number of regimes in the futures market is smaller than in the macroeconomic environment. On the one hand, this is a result of the Silhouette score-based model selection, which finds the hyperparameter that can best geometrically separate different regimes. On the other hand, such difference results from richer and more diverse information covered by macroeconomic indicators compared with the futures market. Hence, the macro environment requires more degrees of freedom to characterize. If we train the regime detection model in the futures market using five regimes, which is the one from the macroeconomic environment, we will get much noisier results, as shown in Figure 22b. This comparison emphasizes again the importance of careful model selection.

Second, the sort of comparisons as done in Figure 22a provides insights regarding the predictability of one regime towards another, leading to potential investment strategies. Notice that whenever the futures market enters a green regime, the crisis regime in the macroeconomic environment occurs not long after. Although this is an overly simplified observation, a more in-depth analysis yields interesting insights, which will be part of our forthcoming updates.

6.2 All macro indicators versus FED groupings

In this section, we report the regime detection results using the FRED-MD dataset proposed by [13]. They selected 134 indicators from FRED and grouped them into six categories: Output and Income, Labor Market, Consumption and Orders, Orders and Inventories, Money and Credit, Interest Rate and Exchange Rates, Prices, and Stock Market. Compared with applying `R2-PCA` on all indicators (more than 600 of them), applying dimension reduction within each group would yield more interpretable embeddings. However, inter-group correlation is significant, which would severely bias regime detection results. We expect orthogonality (or near-orthogonality) among the input multi-variate time series so latent variable models like `HMM` can accurately capture clusters in the latent space. Thus, using the groupings from [13] to explain the regimes identified from principal components of all macroeconomic indicators is better than using them directly for regime detection.

Our data covers the period from Jan. 1981 to Nov. 2023. First, we apply `R2-PCA` within each of the six categories of macroeconomic indicators and use the first principal component as the representing indicator. The result is shown in Figure 23.

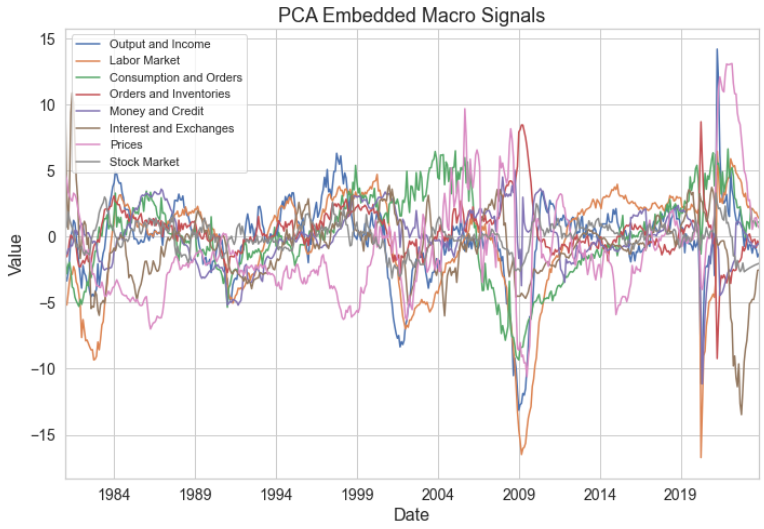


Figure 23: First Principal Macro Signal from each Category

The inter-group correlation is significant. We display the heatmap of correlation in Figure 24. For example, the labor market signal is positively correlated with the output and income signal, with a significant 0.75 correlation coefficient.

As mentioned, regime detection requires orthogonal (or near-orthogonal) input so that the model will not learn spurious correlations within each regime. Thus, we follow the practice in Section 4 and apply `R2-PCA` on all 134 indicators in the FRED-MD database. 6 principal components could explain 85% of the variance. The result is shown in Figure 25.

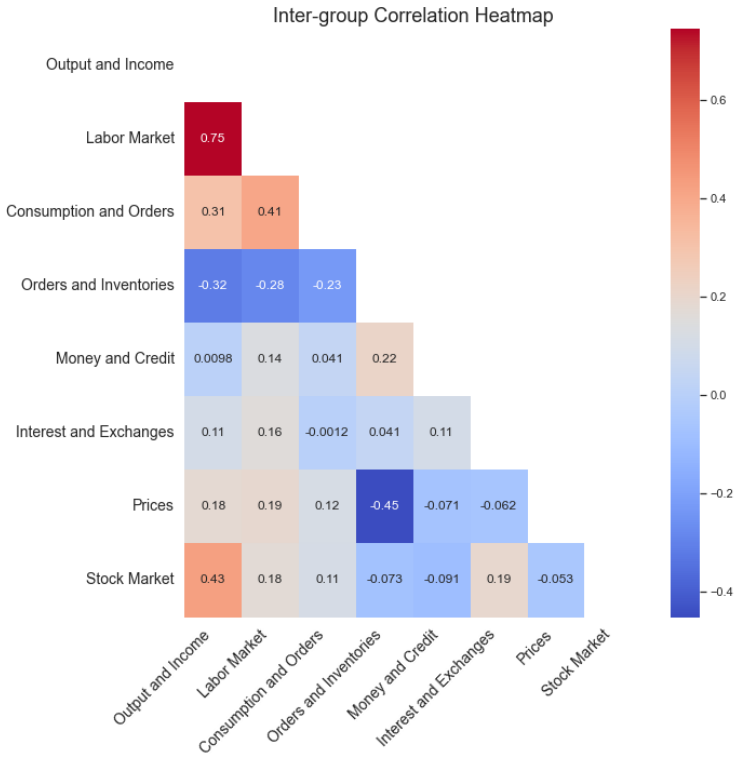


Figure 24: Correlations between Different Categories

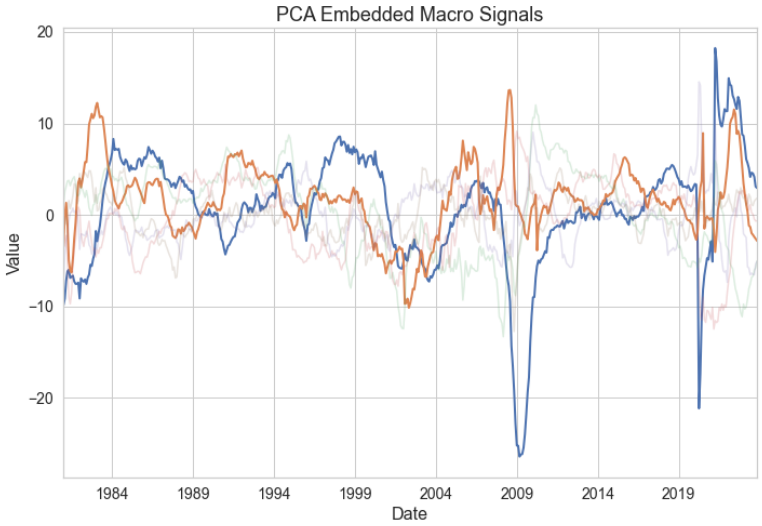


Figure 25: Principal Components of the 134 Macroeconomic Indicators from FRED-MD

After applying the rolling retraining process of `R2-RD`, we acquire the evolution trajectory of regime characteristics. We follow the calculation in Section 4.5 and report the results of the six representing principal macro signals in Figure 26.

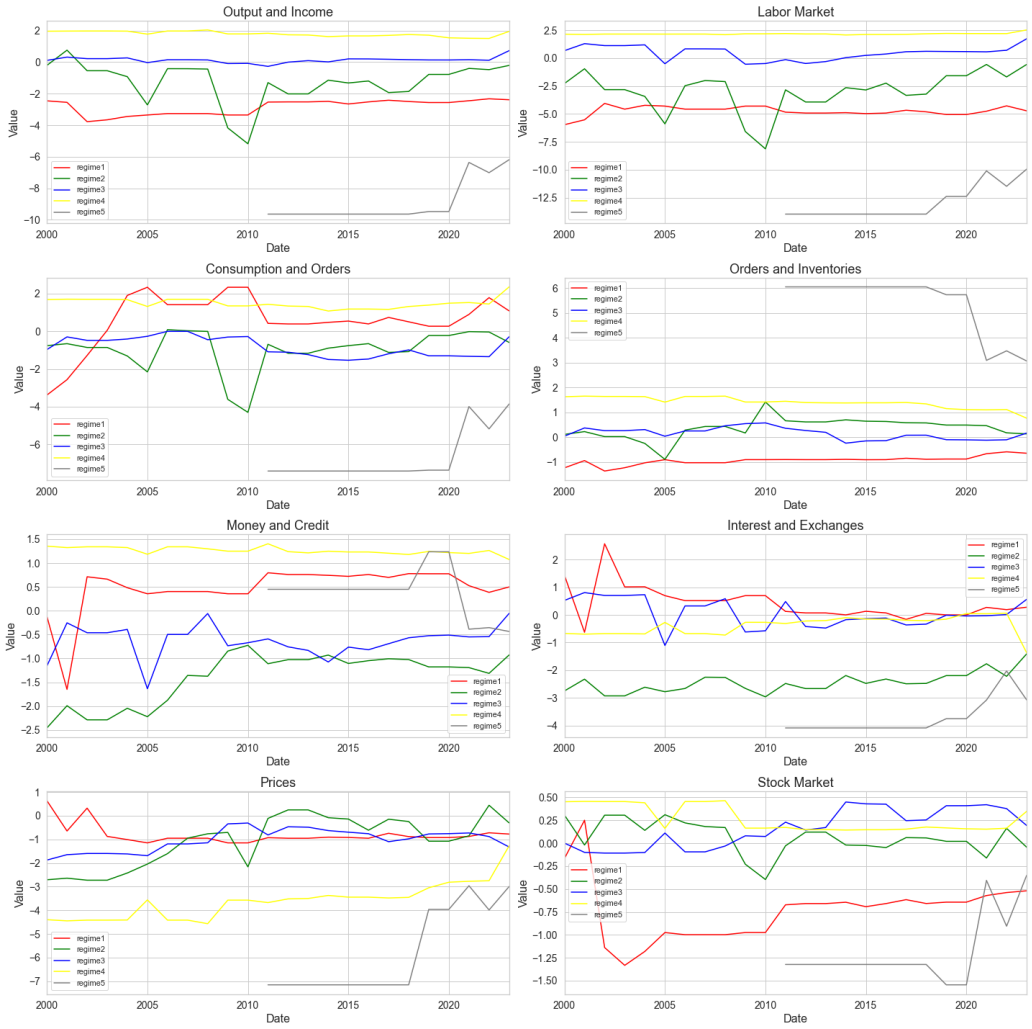


Figure 26: Evolving Dynamics of Six Categories of Macroeconomic Indicators under Different Regimes

First, notice that the inter-group correlation is also reflected in the trajectory. The highly correlated pair of the Labor Market signal and Output and Income signal has a very similar trajectory of regime evolution. The negatively correlated pair of Prices signal and Orders and Inventories signal appears to be an upside-down image of each other. Similar to the results in the main article, the dimension reduction and the regime detection models can compress a large number of input signals into a condensed latent variable model, based on which we can reconstruct some aspects from the input feature space. This validates the effectiveness of our method.

Second, although the regimes seem to be clustered under some categories, many are well separated under many categories. For example, regimes 1 (red), 2 (green), and 3 (blue) are not separated under the Prices signal. However, regimes 1 (red) and 3 (blue) exhibit different levels of Labor Market signals. The Interest and Exchange Rates signal nicely separate regime 1 (red) and 2 (green), as well as regime 2 (green) and 3 (blue). Such findings remind us again of the importance of using a large number of input features so that the model can identify more useful signals in the market.

6.3 Mutual Fund Performance under Macroeconomic Regimes

In this section, we analyze the performance of various mutual funds across different macroeconomic regimes. A distinct performance separation under these regimes emerges, underscoring the potential for regime-aware portfolios in these markets.

Figure 27 shows the monthly return distribution[2] of U.S. large-cap equity mutual funds under different regimes

[2]First, calculate the monthly return of all mutual funds within the specified asset category. Second, draw the Kernel Density

fitted in Jan. 2000 and Jan. 2022. The colors are consistent with the plots in the sections above. First, from the left plot, we see that the performance of large-cap funds is good under regimes 2 and 3, which is consistent with the macroeconomic conditions we analyzed in Table 1. Second, the performance of funds under the same category can shift significantly. For example, most large-cap funds under regime 1 have positive returns before 2000. However, under a modern definition of regime 1, more than half of the large-cap funds have negative returns. This finding suggests that a static view of financial regimes can lead to highly misleading results. Thirdly, the new regime that emerged after the 2008 financial crisis shows surprisingly positive skewness. It means that during the crisis periods, some large-cap mutual funds have significantly higher returns compared to non-crisis times.

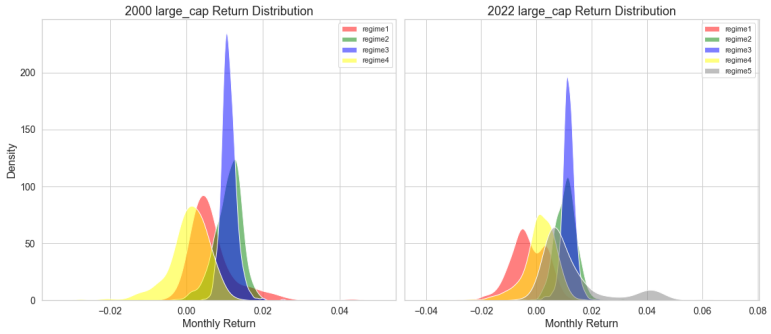
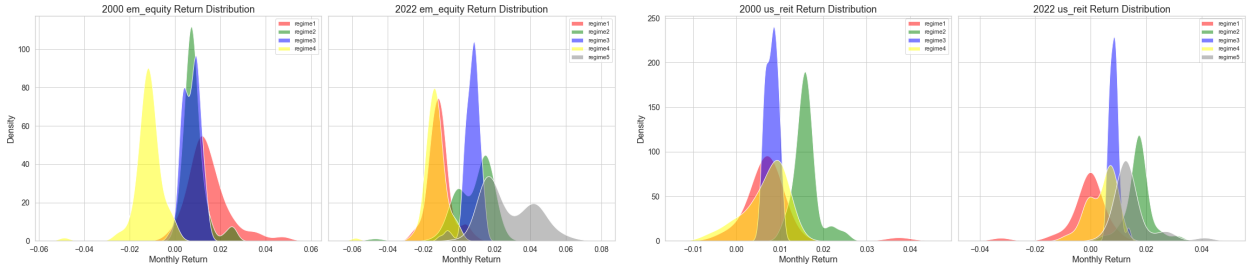


Figure 27: Returns of U.S. Large Cap Mutual Funds under Different Regimes

Similar findings can be found in other asset categories, like the emerging market equity funds[3] in Figure 28a and the U.S. REIT[4] funds in Figure 28b. Compared to the large-cap funds, we could also see different performance among different asset categories. For example, under the 2000-fitted regimes, the emerging market equity funds perform the best under regime 1, and the U.S. REIT funds perform much worse in regime 3 compared with regime 2. The highly diverse performance suggests that a regime-aware portfolio strategy can be significant. Lastly, we want to emphasize the fact that mutual funds under many asset categories during the so-called crisis regime (regime 5) displayed good performance compared with other regimes.



(a) U.S. Diversified Emerging Market

(b) U.S. Real Estate Mutual Funds

Figure 28: Returns of Mutual Fund Categories under Different Regimes

Next, let us compare the performance of mutual funds under the same regimes. The definition of regime 1 changes drastically from 2000 to 2022, as shown in Figure 3a and Figure 3b. Thus, the performance of mutual funds under regime 1 before 2000 and in 2022 differ. Figure 29 shows such performance changes. Noticeably, most mutual funds had positive monthly returns under the 2000 version of regime 1, while in the 2022 version, most had negative returns.

The performance comparison between regime 2 and 3 is similar, except for some minor differences from the U.S. REIT funds. Thus, we only report the KDE plot of regime 3 in Figure 30a. Recall from Table 2 regimes 2 and 3 correspond to good macroeconomic conditions. During these regimes, the performance of large-, mid- and small-cap U.S. equity funds perform the best, showing an average of about 1% monthly returns across all funds

Estimation (KDE) plots of the returns across the funds.

[3]U.S. mutual funds investing in emerging market equities

[4]U.S. Real Estate mutual funds

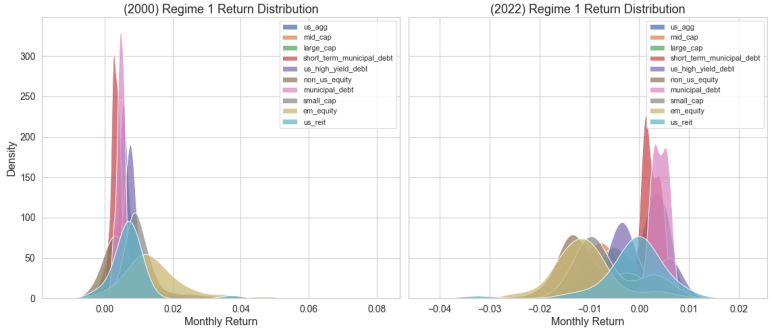
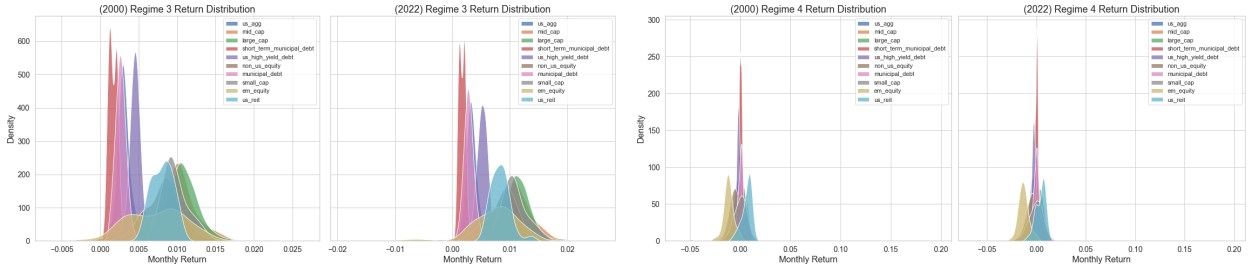


Figure 29: Performance Comparison of Mutual Fund Categories under Regime 1 (2000 and 2022)

within the categories. U.S. REIT funds also show good performance in these periods due to a good economic environment, especially in regime 2, where the U.S. REIT funds have the highest average returns. U.S. REIT funds also perform the best in regime 4, as shown in Figure 30b.



(a) Regime 3 (2000 and 2022)

(b) Regime 4 (2000 and 2022)

Figure 30: Performance Comparison among Different Mutual Fund Categories

As for regime 5, namely the crisis regime, the performance of different asset categories is reported in Figure 31. During this regime, the surprising finding is that most mutual funds exhibit positive monthly returns. In addition, the best-performing categories are U.S. REITs, emerging market funds, and small-cap equity funds.

Finally, we report the Sharpe ratios[5] of each mutual fund category under different regimes using the 2022 version of the regime model. We first calculate via (4) the regime-specific Sharpe ratios of all mutual funds. Then, for each regime $j \in \mathcal{J}$, we average the Sharpe ratios of all the mutual funds within each category. The results are shown in Table 3. From the regime's viewpoint, regime 3 and regime 5 are good times for investing overall. Surprisingly, they correspond to the two extremes of macroeconomic market conditions. From the mutual fund category's viewpoint, each mutual fund has different peak performance regimes. For example, U.S. large-cap equity funds perform the best in regimes 2 and 3, while EM equity funds perform the best in regime 5.

[5]We did not consider other factors like fees and taxes from mutual fund investment for clarity. The comparison extends easily to other performance measures

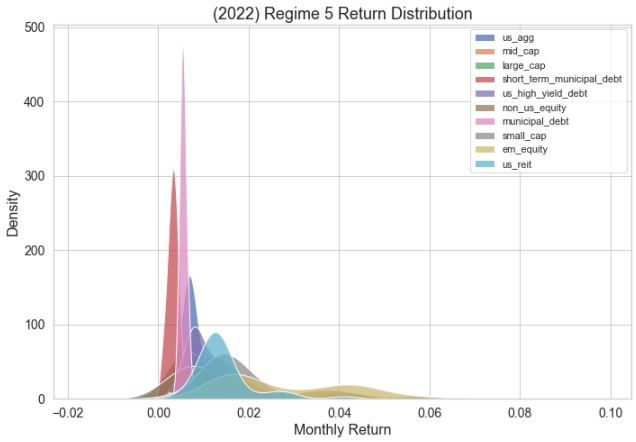


Figure 31: Performance Comparison of Mutual Fund Categories under Regime 5

	Regime 1	Regime 2	Regime 3	Regime 4	Regime 5
U.S. Aggregate	0.2364	0.5310	0.3512	-0.1100	0.4476
Mid Cap	-0.0790	-0.3355	0.2808	0.0264	0.2344
Large Cap	-0.0511	0.2549	0.3175	0.0303	0.2175
Short Term Municipal Debt	0.4475	0.6430	0.5135	0.0587	0.5899
U.S. High Yield Debt	0.0007	0.4991	0.3792	-0.0648	0.4067
EM Debt	-0.0252	0.4667	0.1586	-0.1412	0.4009
Non U.S. Equity	-0.1834	0.1534	0.2147	-0.0786	0.2817
Short Term U.S. Aggregate	0.2749	0.6490	0.7717	0.0550	0.6421
Municipal Debt	0.2752	0.3628	0.3137	0.0553	0.3702
Small Cap	-0.0969	-0.3866	0.2358	-0.0050	0.2500
EM Equity	-0.1661	0.1553	0.1466	-0.2407	0.5056
U.S. REIT	-0.0159	0.5385	0.2082	0.0771	0.2122

Table 3: Mutual fund monthly Sharpe Ratio under different regimes (identified in 2022)

7 Conclusion

In this paper, we proposed the `R2-RD` framework for dynamically identifying financial regimes. The statistical and geometric-based model selection framework automatically selects the model family that best describes the data and the model within the family that best separates the regimes. It mitigates the issue of falsely making unrealistic assumptions about the data generation process and lets the data speak for itself. By utilizing temporal ensemble, label assignment, and a threshold policy, `R2-RD` is able to identify a stable trajectory of evolving regime dynamics and the emergence of new regimes. Powered by `R2-RD`, we provide strong empirical evidence of nonstationary market dynamics in the macroeconomic environment and the futures market, which is difficult to identify before without the recursive training framework of `R2-RD`. The identified macroeconomic regimes also demonstrate separation among mutual fund performance, suggesting strong potential for regime-based asset management, including security selection and portfolio construction strategies. With detailed illustrations of numerical experiments, we also make suggestions about feature engineering for regime detection, making our methods more accessible for researchers and practitioners.

8 Future Work

The model selection procedure of `R2-RD` is more powerful when considering more candidate models. Thus, in addition to Gaussian `HMM` and Gaussian mixture `HMM`, which we tested in the main article, we are now focusing on incorporating more variations of `HMM` in the model portfolio. Some models under testing include `ARHMM` (also

known as the Markov switching model), Hidden Semi-Markov Model (`HSMM`), and neural `HMM`. Neural `HMM` is an important extension, as it unifies the assumptions behind many models like the `HSMM`, `ARHMM`, etc. It uses neural networks for the observation model with arbitrary lengths of past observations as inputs, hence relaxing any assumptions on the topology of the probabilistic graph. It also generalizes the hidden states' Markov or semi-Markov transition dynamics via a deep sequence model.

Another focus is forecasting future regimes. In addition to identifying the current regime, we want the model to tell us what regime we may be in six months. A benchmark method is the n-step transition matrix after temporal ensemble. However, we need much more careful analysis. Neural symbolic programming is an important direction as it unifies symbolic knowledge from statistical models like `HMM` and the powerful function approximation tool of neural networks. Another important direction is a mixture of experts.

References

[1] P. Ailliot and F. Pene. Consistency of the maximum likelihood estimate for non-homogeneous markov–switching models. *ESAIM: Probability and Statistics*, 19:268–292, 2015.

[2] A. Botte and D. Bao. A machine learning approach to regime modeling. *Two Sigma*, 2021.

[3] S. Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

[4] R. Douc and E. Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden markov models. 2012.

[5] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[6] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32:41–62, 1998.

[7] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384, 1989.

[8] A. Hirsa, R. Holmes, F. Klinkert, and S. Malhotra. Robust Rolling K-Means (R2K-Means): an updateable nonlinear k-means clustering methodology for financial time series. *Available at SSRN*, 2024.

[9] A. Hirsa, F. Klinkert, S. Malhotra, and R. Holmes. Robust Rolling PCA: Managing time series and multiple dimensions. *Available at SSRN*, 2023.

[10] B.-H. Juang and L. Rabiner. Mixture autoregressive hidden markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413, 1985.

[11] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[12] Y. Li, H. Duan, and C. Zhai. Cloudspeller: Spelling correction for search queries by using a unified hidden markov model with web-scale resources. In *Spelling Alteration for Web Search Workshop*, pages 10–14. Citeseer, 2011.

[13] M. W. McCracken and S. Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[14] L. Mevel and L. Finesso. Asymptotical statistics of misspecified hidden markov models. *IEEE Transactions on Automatic Control*, 49(7):1123–1132, 2004.

[15] D. Pouzo, Z. Psaradakis, and M. Sola. Maximum likelihood estimation in markov regime-switching models with covariate-dependent transition probabilities. *Econometrica*, 90(4):1681–1710, 2022.

[16] A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[17] F. Yang, S. Balakrishnan, and M. J. Wainwright. Statistical and computational guarantees for the baum-welch algorithm. *The Journal of Machine Learning Research*, 18(1):4528–4580, 2017.

[18] S.-Z. Yu. Hidden semi-markov models. *Artificial intelligence*, 174(2):215–243, 2010.