Uncovering Latent Regimes in the U.S. Housing Market: A Gaussian Hidden Markov Model Approach

Dennis Sun, Nick Pham, Noe Gonzalez, Wenrong Zheng

November 28, 2025

Abstract

This project utilizes the seasonally adjusted FHFA Purchase-Only House Price Index (HPI) from 1991 Q1 to 2025 Q2 to uncover latent price movement regimes in the U.S. housing market. We model the quarterly log-returns using a Gaussian Hidden Markov Model (HMM) and estimate parameters via the Baum-Welch (EM) algorithm. Comparing the HMM against a Gaussian Mixture Model (GMM) with the same components shows the HMM achieves a vastly superior fit ($\Delta BIC \approx 98$), confirming strong regime persistence in the housing market. The optimal model, selected using the Bayesian Information Criterion, uses three states, which are highly interpretable as Expansion, Neutral, and Downturn/Correction phases. Furthermore, the model's downturn state probability serves as a robust leading indicator, achieving an AUC of 0.998 for predicting downturns one quarter ahead, demonstrating its utility as an early warning system.

1 Problem Description

This project uses the FHFA House Price Index (HPI) to identify different price–movement regimes of the U.S. housing market and to analyze how these regimes evolve over time. In each quarter, the market is assumed to be in one latent state, such as:

- Expansion – high price growth, rising leverage and bubble risk;
- Neutral – mild growth or roughly flat prices;
- Adjustment/Downturn – low or negative returns and elevated default and wealth–loss risk.

Rather than defining these phases by ad hoc thresholds, we aim to learn the regimes and their transition probabilities directly from the HPI series, inspired by the regime-switching framework of Hamilton [Ham89]. A data–driven regime model can support quantitative assessments of housing affordability, mortgage default risk, and financial stability. This setting fits probabilistic reasoning and learning: we only observe price indices and their changes, while market regimes are latent variables. We use a Hidden Markov Model (HMM) with the EM algorithm to infer hidden states and estimate model parameters in an unsupervised way.

2 Data Sourcing and Processing

2.1 Data Source

We use the official FHFA House Price Index (HPI) quarterly data for the U.S. purchase–only index:

1. Main dataset page: FHFA House Price Index.[Fed25]
2. Original file used: U.S. Summary.xlsx.[Fed25]

The raw table reports both seasonally adjusted and non–seasonally–adjusted purchase–only indices, plus several percentage–change columns. From this file we construct four working datasets:

```
hpi_po_summary_cleaned.xlsx   hpi_po_summary_condensed.xlsx
index_observations.txt        adjusted_index_observations.txt
```

2.2 Preprocessing Steps

We utilized the FHFA Purchase-Only House Price Index (HPI), specifically the seasonally adjusted national series, to isolate economic trends from calendar effects. The dataset spans from 1991 Q1 to 2025 Q2. To ensure the stationarity required for HMM training, we transformed the raw index levels P_t into quarterly log-returns:

$$\gamma_t = \ln(P_t) - \ln(P_{t-1})$$

The initial observation (1991 Q1) was dropped due to the differencing step, resulting in a continuous time series of $T = 134$ observations. No further discretization was applied; we model the continuous returns directly using Gaussian emissions to preserve magnitude information.

3 Modeling and Inference

3.1 Mathematical Formulation

Formally, the system is defined by a set of parameters $\lambda = (\pi, A, \mu, \sigma^2)$. Let q_t denote the state at time t and O_t denote the observation (log-return). The model assumes that the observation O_t is drawn from a Gaussian distribution conditional on the state k:

$$P(O_t \mid q_t = k) = \mathcal{N}(O_t; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(O_t - \mu_k)^2}{2\sigma_k^2}\right)$$

The transition between hidden states follows a first-order Markov process governed by the transition matrix A, where $A_{ij} = P(q_{t+1} = j \mid q_t = i)$. The joint probability of a sequence of observations O and states Q is given by:

$$P(O, Q \mid \lambda) = \pi_{q_1} b_{q_1}(O_1) \prod_{t=2}^{T} A_{q_{t-1} q_t} b_{q_t}(O_t)$$

where $b_k(O_t)$ is the emission probability defined above. We maximize the likelihood $L = \sum_Q P(O, Q \mid \lambda)$ via the EM algorithm.

3.2 Algorithm Implementation

We model the housing market using a Gaussian Hidden Markov Model (HMM). The latent state sequence $S_1, S_2, \ldots, S_T$ represents unobserved market regimes (e.g., Expansion, Contraction), and observations correspond to quarterly log-returns.

- Inference: We implemented the Forward-Backward algorithm to compute the posterior probabilities of hidden states, $\gamma_t(k) = P(S_t = k \mid O_{1:T})$.
- Learning: Parameters $(\pi,\ A, \mu, \delta^2)$ were estimated using the Baum-Welch (EM) algorithm [Rab89]. To mitigate the risk of local optima—a common issue in EM training—we performed random restarts with 5 different initialization seeds and selected the model with the highest final log-likelihood.
- Decoding: The Viterbi algorithm was used to find the single most likely sequence of regimes.
- Model Selection: We determined the optimal number of states (K) by minimizing the Bayesian Information Criterion (BIC), striking a balance between model fit and complexity.

4 Results and Discussion

4.1 Stability Analysis

We first assessed the numerical stability of the training process. Figure 1 illustrates the EM log-likelihood trajectories for Gaussian HMMs with $K = 2, 3, 4$ states. For all values of K, the log-likelihood increases monotonically and stabilizes quickly without oscillation, indicating that the Baum-Welch algorithm is numerically stable on this dataset.

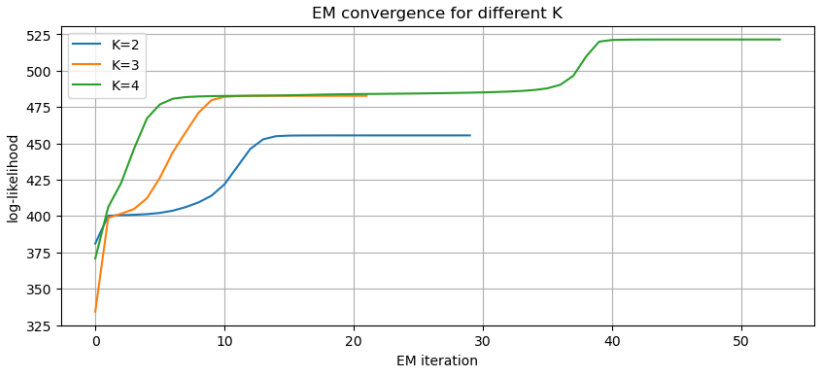

Figure 1: EM convergence trajectories for K=2, 3, 4

To ensure that our identified regimes are structural features of the data rather than artifacts of random initialization, we performed 5 random restarts for each K. Figure 2 displays the pairwise agreement matrix for the Viterbi-decoded state sequences across these seeds. The high agreement ($> 95\%$ for convergent runs) confirms that the model consistently identifies the same underlying regime structure regardless of the starting parameters.

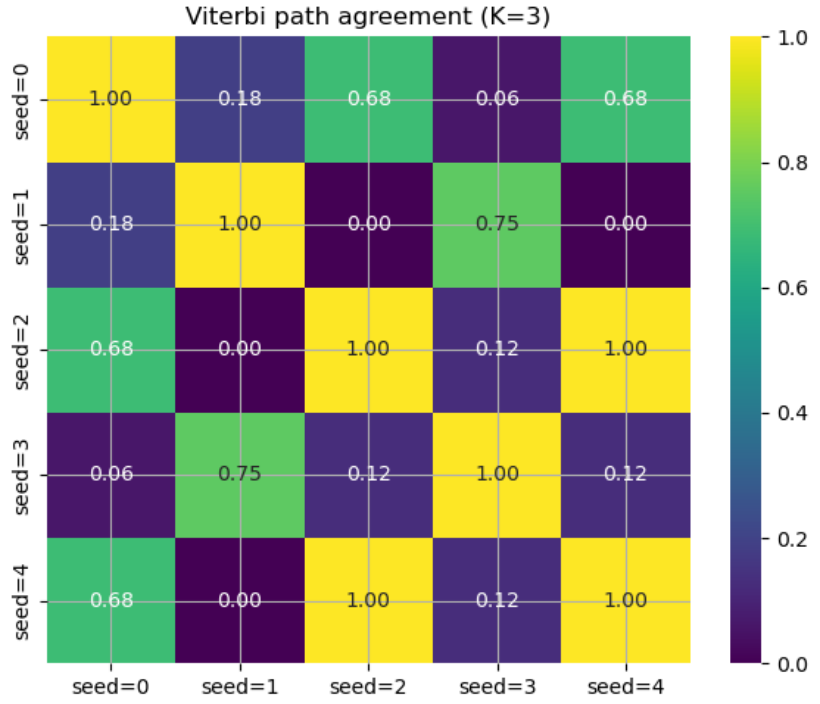

Figure 2: Pairwise Viterbi path agreement across random seeds

4.2 Baseline Comparison: HMM vs. GMM

To validate the necessity of modeling temporal dependencies (the "Markov" assumption), we compared our HMM against a Gaussian Mixture Model (GMM) with the same number of components ($K = 3$)[MP00]. While a GMM also models the data as a mixture of Gaussian distributions, it assumes observations are independent and identically distributed (i.i.d.), effectively ignoring the time-series nature of housing cycles.

Table 1 compares the Bayesian Information Criterion (BIC) for both models (lower is better). The HMM achieves a significantly lower BIC ($\Delta \approx 98$), indicating a vastly superior fit.

Model Type	Assumptions	BIC
Gaussian Mixture Model (GMM)	Independence (No temporal dependency)	-798.57
Hidden Markov Model (HMM)	Markov Dynamics (Regime persistence)	**-896.55**

Table 1: Baseline Model Comparison (BIC)

Theoretical Analysis of Temporal Dependency: The fundamental difference lies in how the two models treat time. The GMM assumes that each quarterly return is an independent and identically distributed (i.i.d.) sample from a mixture distribution. Mathematically, this implies that the probability of being in a regime depends only on the global mixture weights, ignoring the immediate past: $P(S_t = j \mid S_{t-1} = i) = P(S_t = j)$

In contrast, the HMM explicitly parameterizes the state persistence through the diagonal elements of the transition matrix A (i.e.,A_{ii}). Financial and real estate time series exhibit "volatility clustering," where high-volatility periods tend to cluster together. By allowing A_{ii} to be close to 1, the HMM captures the "stickiness" of economic cycles, explaining why it fits the trajectory of housing prices significantly better than the memoryless GMM.

Result confirms that housing market returns exhibit strong regime persistence—a boom today increases the probability of a boom tomorrow (Table 4.2). The GMM treats returns as random draws, failing to capture the "volatility clustering" and sequential logic inherent in economic cycles.

4.3 Model Selection

We determined the optimal number of latent states by training HMMs with $K = 2, 3, 4$ and comparing their information criteria.

K	Log-Likelihood	AIC	BIC
2	455.49	-896.97	-876.53
3	482.71	-937.43	-896.55
4	521.48	-996.96	-929.80

Table 2: HMM Model Selection Metrics

As K increases, the log-likelihood improves substantially. While yields the lowest BIC, the improvement from $K = 3$ to $K = 4$ is marginal compared to the jump from $K = 2$ to $K = 3$. We selected $K = 3$ as our final model to balance goodness-of-fit with interpretability, as three states map naturally to standard economic phases (Expansion, Neutral, Downturn).

4.4 Regime Interpretation

Using the model, we decoded the most likely sequence of market regimes via the Viterbi algorithm. Figure 3 overlays these regimes on the quarterly log-returns, with macroeconomic periods marked for context. The identified states are highly interpretable:

- Regime 0 (Neutral): Characterized by low variance and moderate positive returns, corresponding to stable growth periods.
- Regime 1 (Expansion): Associated with high means and moderate volatility, appearing prominently during the mid-2000s housing boom and the post-2020 recovery.
- Regime 2 (Downturn/Correction): Captures periods of negative returns and high volatility. This regime clusters tightly around known crisis points, including the early-2000s recession and the 2007-2009 Global Financial Crisis.

Historical Context Alignment: The decoded regimes align remarkably well with known macroeconomic history.

- 2007-2009 Subprime Crisis: The model correctly identifies the onset of the housing crash (Regime 2) almost immediately as prices began to decelerate in 2007, capturing the high volatility and sharp negative returns of the Great Recession.

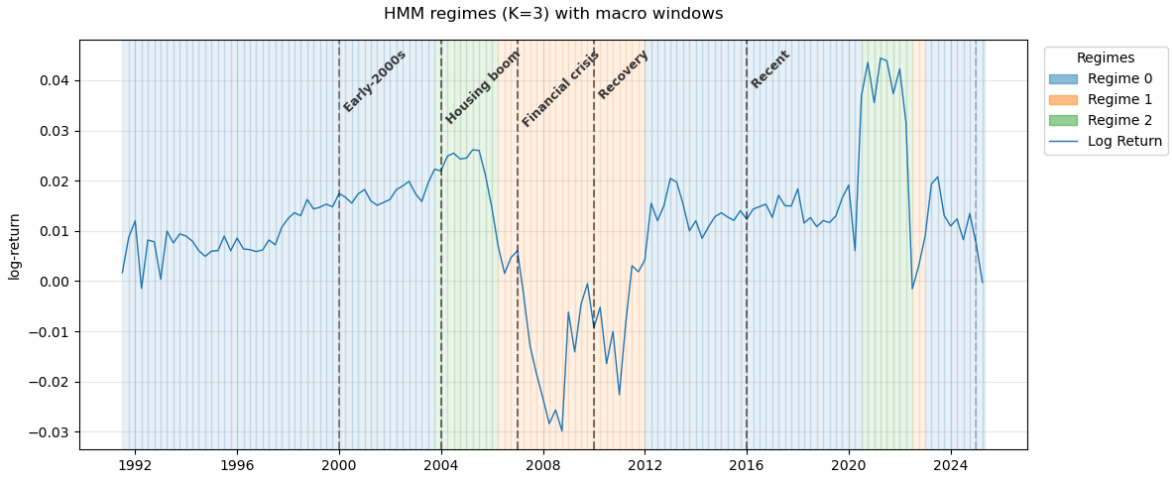

Figure 3: Viterbi-decoded regimes overlaid on log-returns with macro windows

- COVID-19 Pandemic: Interestingly, the model briefly flags the uncertainty at the start of 2020 but quickly transitions back to Regime 1 (Expansion) as the housing market experienced an unexpected boom driven by remote work trends and low interest rates.
- 2022-2023 Rate Hikes: The transition out of the "Expansion" regime in late 2022 coincides precisely with the Federal Reserve's aggressive interest rate hikes, which cooled demand and slowed price appreciation, moving the market toward a Neutral or Correction phase.

This alignment with historical macroeconomic events demonstrates that the unsupervised HMM successfully extracts meaningful economic signals solely from price data.

4.5 Downturn Detection & Predictive Power

Finally, we evaluated the model's utility as an early warning system. We defined a binary "Downturn Label" for periods where the rolling 3-month cumulative return was negative. Figure 4 compares the Receiver Operating Characteristic (ROC) curves for two tasks:

1. Same-period Detection: Classifying the current quarter's downturn status.
2. Lead-1 Prediction: Predicting a downturn one quarter ahead using the current posterior probability.

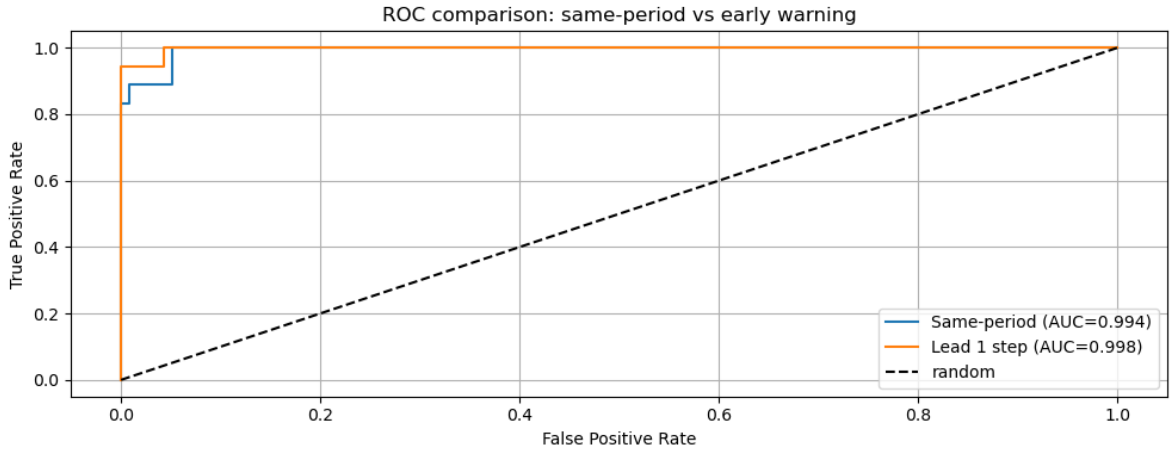

Figure 4: ROC comparison for same-period vs. early warning downturn detection

The model achieved an AUC of 0.998 for the 1-step ahead prediction. This indicates that the "Downturn State" probability is a robust leading indicator, capable of signaling negative market momentum before it is fully reflected in the lagging aggregate statistics.

5 Conclusion

Using the FHFA seasonally adjusted purchase-only HPI, we build quarterly log returns for 1992 Q1–2025 Q2 and fit a 1-D Gaussian HMM to capture latent housing-market regimes. The EM algorithm converges monotonically for $K = 2, 3, 4$, indicating stable training. As K increases, the log-likelihood improves, and BIC reaches its minimum at $K = 4$, while $K = 3$ offers a good trade-off between fit and interpretability. In the three-state model, the regimes roughly correspond to "high-return expansion," "neutral," and "low/negative-return downturn," with the downturn state used more often during crisis and correction periods. This suggests that meaningful market phases can be extracted from prices alone.

Limitations and Future Extensions: While effective, our univariate model has limitations. It relies solely on price dynamics, ignoring exogenous drivers such as 30-year mortgage rates, unemployment rates, and housing inventory levels.

Future work could extend this framework to a Multivariate HMM, where the emission probability becomes a multivariate Gaussian distribution over a vector of economic indicators. Additionally, moving from a discrete-time HMM to a Hidden Semi-Markov Model (HSMM) could explicitly model the duration of each economic cycle, potentially improving long-term forecasting accuracy for regime switches.

6 Reflections & Contributions

6.1 Team Contributions

- Dennis Sun: Developed the core HMM framework code and established the visualization framework for the Jupyter notebooks. Led the final integration of the report, ensuring logic coherence and adherence to formatting guidelines.

6.2 Individual Reflections

- Dennis Sun: Through this project, I gained a higher-dimensional perspective on the application of Hidden Markov Models to real-world data. It bridged the gap between theoretical probabilistic reasoning and practical economic analysis, specifically showing how latent variables can effectively model complex market behaviors.

6.3 Generative AI Statement

We acknowledged the use of AI assistance (Gemini 3) in this project. The AI was utilized to debug Python plotting code within our Jupyter notebooks to ensure accurate visualization. Additionally, the AI provided the conceptual idea for the "pairwise Viterbi path agreement across random seeds" plot (Figure 4) to rigorously test model stability. All core modeling logic, code implementation, and final analysis were performed and verified by the human team members.

References

[Fed25] Federal Housing Finance Agency (FHFA). HPI Technical Description. https://www.fhfa.gov/data/hpi/datasets/, 2025. Accessed [11/20/2025].

[Ham89] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[MP00] G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[Rab89] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.