ALERT: Anomaly Learning via Eigen-Residual Regime Tracking

Introduction

We developed an early warning system for financial crises and anomalous economic activity using a Karhunen-Loeve (KL) and Hidden Markov Model (HMM) machine learning pipeline. It is by now well known that the stock market can be unpredictable and that returns can have heavier tails, meaning that extreme gains or losses happen more frequently than the standard normal distribution suggests. However, the textbook literature and traditional financial models, such as Modern Portfolio Theory, rely on the assumption of normality, a significant flaw.

To overcome this, we focus on the Option-Adjusted Spread (OAS). By comparing against a "risk-free" benchmark, such as corporate bonds, the OAS quantifies the yield premium investors require to hold more volatile bonds. Thus, this metric isolates the market's perception of potential defaulting or liquidity risk, making it a sensitive indicator of overall systemic financial stress. By analyzing the dynamics of the OAS across different credit qualities, we have a robust method for identifying which indicators they are and predicting future indicators.

We therefore decided to use a Karhunen-Loève expansion to develop probabilistic baselines for anomaly detection, which are then classified into different market regimes using a Gaussian Hidden Markov Model (GHMM). The benefits of the KL Expansion are that no prior distribution or parameters are assumed, and subtle market regimes and structures are used instead to extract interesting information from the data. This renders our model quite versatile.

Methodology

Data Collection and Preprocessing

We collected Index Option-Adjusted Spread (OAS) data for nine corporate bond rating categories from the Federal Reserve Bank of St. Louis (FRED). These spreads measure the compensation that investors demand, above risk-free Treasury yields, for holding corporate bonds of a given credit quality. OAS reflects the market's pricing of credit, liquidity, and overall risk sentiment within each rating class. Because these spreads respond directly to changes in perceived default risk and financial stress, they provide a natural proxy for systemic risk in the corporate credit market.

We acknowledge that this may not be the whole picture, but for demonstration purposes, we decided to take this direction. We also know that bonds with different ratings, security, and yields can exhibit different response times or general behavior, so we decided to use indices for nine categories/ratings: *High-Yield, Corporate, AAA, AA, A, BBB, BB, B, and CCC_Lower*. Having a diverse dataset of indices with different ratings can represent various market segments, providing the model with more information for detecting specific regimes.

Karhunen-Loeve Expansion and Residual Calculation

Using this data, we conducted the KL expansion to construct a baseline for normal market movement. For background, the KL expansion is mathematically analogous to Principal Component Analysis in the discrete case, but we use it for a different purpose. Rather than reducing dimensionality for prediction or visualization, it is used to define what "normal" multivariate market behavior looks like and to measure deviations from that standard structure. To do this, the model first estimates the covariance matrix of credit spread changes during a historically calm period. This covariance matrix summarizes how the different parts of the credit curve typically move together under stable market conditions. Applying PCA to this covariance matrix produces a set of orthogonal components that represent the dominant patterns of joint movement across all credit spreads. The leading components capture the bulk of normal market variation, while the remaining components capture minor, less critical fluctuations.

Once this normal covariance structure is learned, each new day of data is projected onto the space spanned by these dominant principal components. This produces a reconstruction of what that day would look like if the market were still behaving according to its historical normal relationships. The difference between the observed data and this reconstruction is called the residual. These residuals represent the portion of market movement that the usual correlation structure cannot explain. In standard PCA, these low-variance residuals are often discarded, but here we treat the discarded portion as the primary signal.

The magnitude of the residual is summarized into a single anomaly measure, the KL score. When the KL score is small, the market is moving in line with its typical covariance structure. When the KL score becomes large, it indicates that the relationships between credit spreads are breaking down and that the market is entering an abnormal or stressed regime. Because this method is based on covariance and cross-sectional structure rather than on individual price levels, it can detect early structural changes even before significant absolute price moves occur.

Our KL expansion driver was written to allow us to select the percent variance we would like explained in the model. Keeping the variance low (around 80%) results in the function using fewer principal components, forcing the model to retain only the strongest market trends. The remaining variance is discarded and ultimately ends up in the residuals. On the other hand, keeping the variance high (around 99%) results in the model including almost everything, including noise and minor fluctuations that provide little to no information about actual market conditions. In finding the balance between these two results, we found that our best results yielded 95% variance explained in the decomposition. Surprisingly, using PC1 and PC2 captures 95.43% of the data's variance, providing a solid baseline for training the HMM against volatile regions.

Hidden Markov Model Design

Now, the cross-section at each time step is compressed into an anomaly score, which is classified into three main regimes: Calm, Stress, and Crisis using a Gaussian HMM (GHMM).

The Gaussian aspect of the model is an upgrade, as it uses a normal distribution to detect anomalies rather than static cutoffs. We varied the number of hidden states in the HMM to determine which would be most appropriate. While two states are sufficient for splitting into a normal market state and a stressed crisis state, we tended to achieve better results with three states, separating not just a stressed state but an additional third crisis state for extreme volatility. Splitting further into four states was less valuable, as only a few dozen (less than 0.5% of our data) were classified into this more extreme state. Additionally, we added a hyperparameter to preprocess the inputs via a log transformation to reduce data skew and achieve a more normal distribution. The data is also normalized using sklearn's StandardScaler before training.

Training of the GHMM is optimal directly after the chosen window for performing the KL expansion. For example, for evaluation, we selected time windows when the market was relatively normal to compute the KL, and trained the GHMM in a more volatile region. The higher residuals in the volatile region are represented in the GHMM as crisis/stressed states, while those resembling the training region are classified as calm. When using the model for inference, it's ideal to follow a similar process, even if that means using earlier data to avoid high-variance market trends during the expansion.

Evaluation

KL Reconstruction Scores

We found our best results when running the expansion with a window from January 1, 2004, to May 31, 2007, representing our calm market. Plotting the KL score over time shows a maximum of about 9, with most other scores around 2-4.

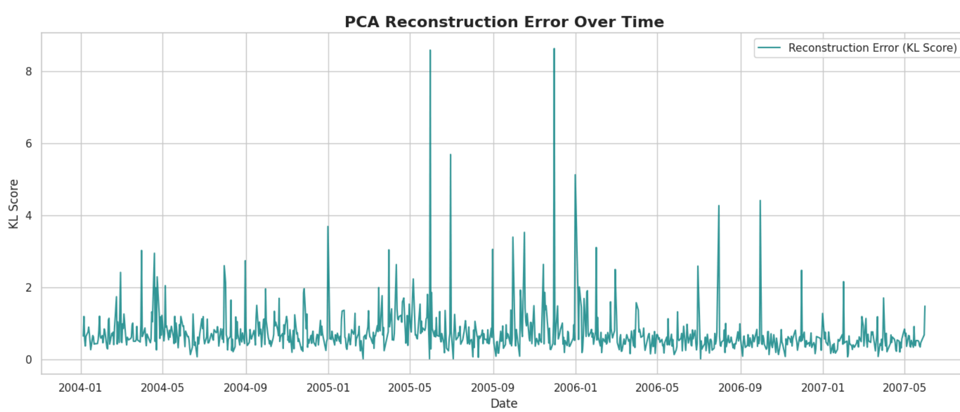


Figure 1: PCA Reconstruction Error (2004-2007).

However, when reconstructing the entire dataset, the KL scores reach much higher, upwards of 60-90. With the full dataset reconstruction, we get a much better picture of how the market is doing in terms of our expansion. We see that in 1998-2004, the market was quite volatile, but became much calmer in our training window (by design, since we chose that as our "normal" market). Following our training, the model accurately captures major crises through the KL score, with peaks during the Great Recession in 2008 and the COVID-19 pandemic in 2020.

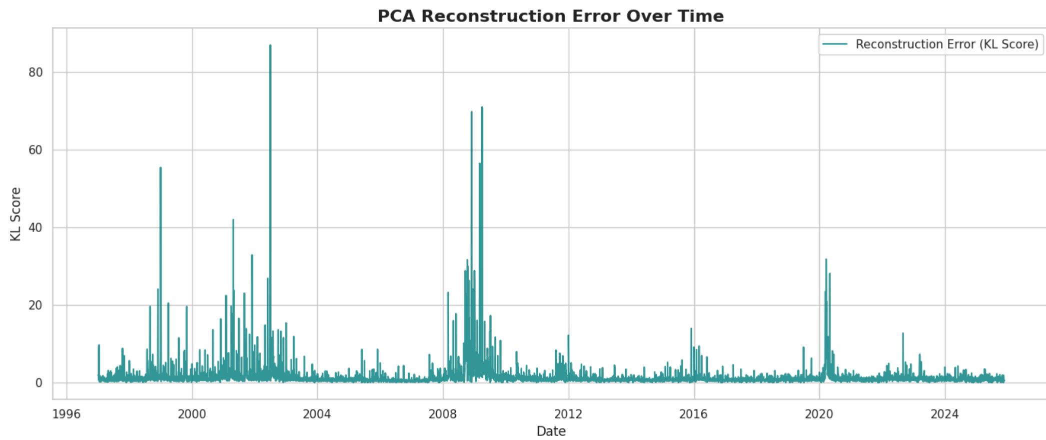
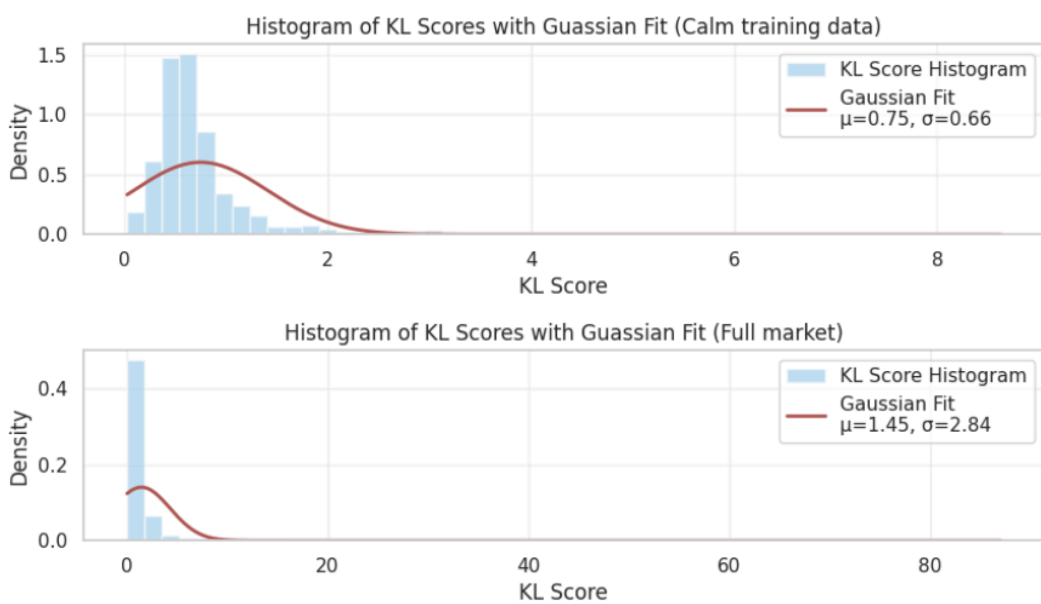


Figure 2: PCA Reconstruction Error (1996-2025).

As expected, the distribution of KL scores during the calm training period is tightly concentrated at low values, with a modest rightward skew. This shape is desirable, as the KL score is constructed as a nonnegative reconstruction error and therefore naturally exhibits asymmetry even under normal market conditions. In contrast, the distribution across the whole sample shows a dramatically heavier right tail, reflecting rare but extreme stress events. The magnitude of this tail is an order of magnitude larger than that observed during the calm period, highlighting the strong separation between normal and crisis regimes.



Figures 3 & 4: Histogram of KL Scores.

GHMM Model Training Variations

Using the same KL expansion, we created two primary GHMMs using (1) the reconstructed full dataset as a multivariate input and (2) the univariate KL scores as input. The goal was to determine whether compiling the reconstructed data as a magnitude of difference from the original data (the KL score) would perform better or worse at predicting the market state.

	GHMM with KL Score Input			GHMM with Reconstructed Data Input		
	Count	**Mean**	**Max**	**Count**	**Mean**	**Max**
Calm	3438	0.67	1.73	6022	0.84	2.37
Stressed	3397	1.29	4.02	1236	2.57	7.30
Crisis	615	6.71	87.02	192	13.50	87.02

Figure 5: Summary Statistics of the two GHMMs.

The summary statistics above describe the distribution of KL scores for both model training strategies. We can see that the KL Score split the counts between calm and stressed states almost evenly, whereas the reconstructed data was mainly labeled as calm. Clearly, the KL score is more sensitive to market changes than the reconstructed one, with about three times as many labels labeled as crisis in the KL score, and the mean KL score in the reconstructed data is twice that in the KL Score.

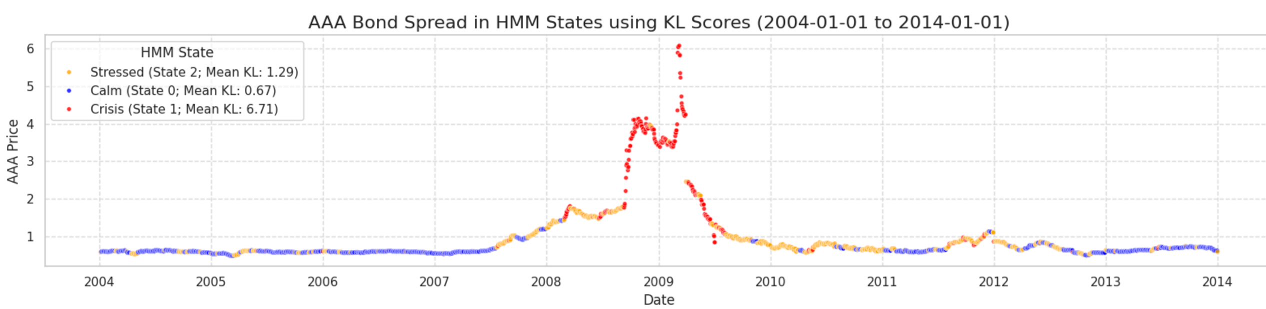


Figure 6: Labeled AAA Bond Spreads using KL Scores.

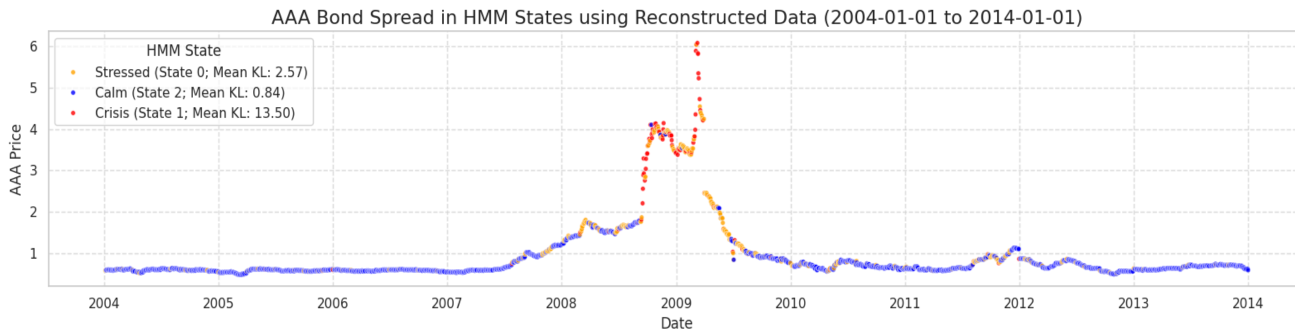
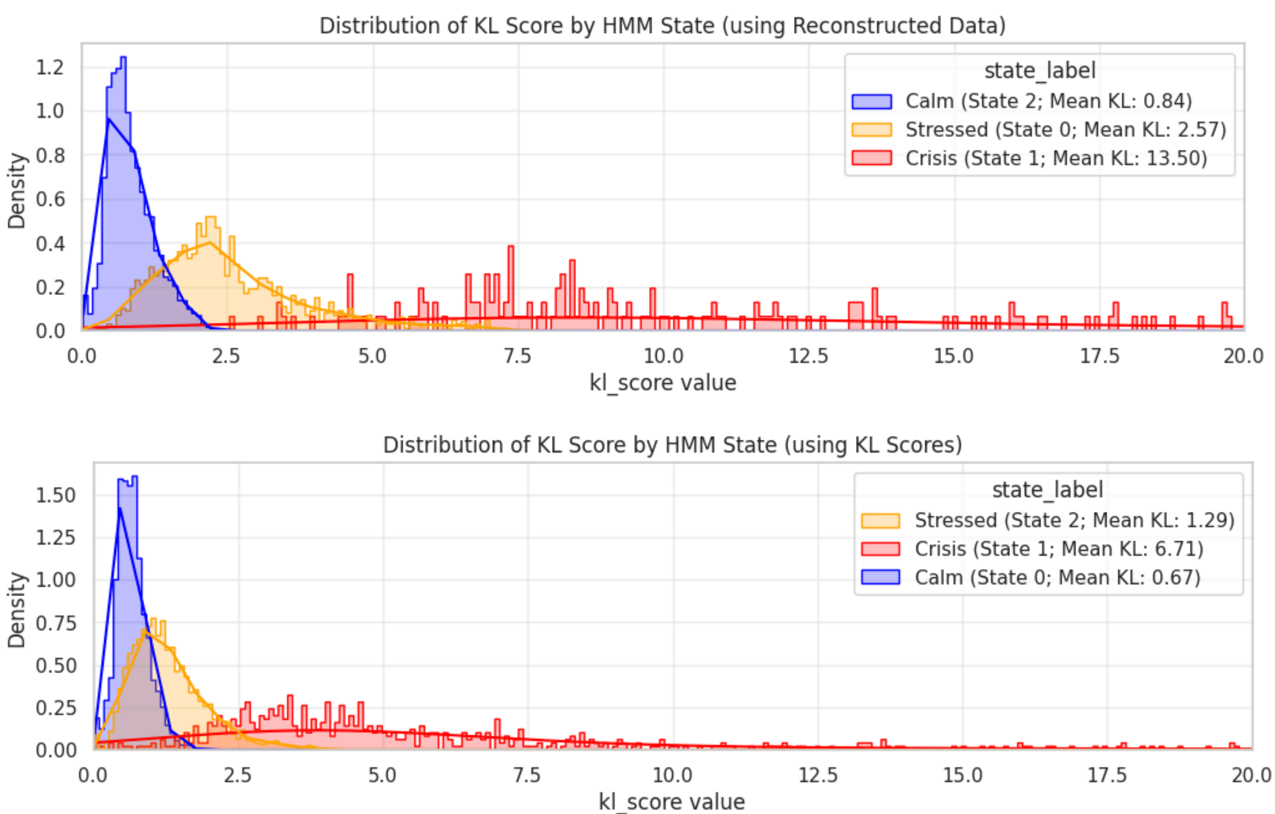


Figure 7: Labeled AAA Bond Spread using Reconstructed Data

The above time trends support the idea that the KL Score model is more sensitive to minor changes in the market, with no calm labels during the height of the Great Recession. In contrast, the Reconstructed Data model even has a few calm labels at that time. The KL score's ability to consistently predict the crisis state during the recession comes at the cost of more stressed labels appearing in data with relatively low variance (such as between 2004 and 2006).



Figures 8 & 9: Distribution of KL Scores of both KL Score HMM and Reconstructed Data HMM

Note that the tails of both graphs actually extend to a KL score of 80, but are reduced to 25% of the total for visual purposes. Again, we can see that the distributions for the reconstructed data extend further to the right, suggesting more sensitive predictions. However, it is crucial to note that the model using reconstructed data achieves better class separation than the KL Score model. Overall, the model choice may depend on the use case, and the user should evaluate the trade-off between false positives in the KL model and false negatives in the reconstructed-data model.

Experiments

Trading Game

To evaluate our model's performance quantitatively, we set up a simple trading game. We tested how we can leverage its signals and “early warnings” to gain an edge in the stock market. Success would mean we outperform the S&P 500's average 8% annual return. Due to our limited background in algorithmic trading and to truly test our model, we developed a naive trading strategy that invests, holds, or divests based on our model's signals. The trading game begins with a portfolio value of $10,000 invested in the S&P 500. In the calm period, the portfolio maintains its investment. In the crisis regime, the portfolio moves to cash. In the stressed regime, the portfolio remains defensive and stays out of the market. However, to account for frequent state fluctuations, we allowed the algorithm to be aware of duration. If the unstable regime persists longer than its typical historical duration, the strategy gradually re-enters the market with partial exposure, reflecting improving conditions.

We benchmark this strategy's performance against buying and holding $10,000 in the S&P 500. The first game was played from 2022 to 2025, with the training window running from July 2020 to 2021. This training window was selected for its relatively consistent KL scores, which are indicative of the calm period following the COVID-19 recession.

Training: 2020-07-01 to 2021-12-31
Trading: 2022-2025

Figure 10: Trading Game 1. Portfolio Value of the Duration-Aware Strategy vs. Buy and Hold.

As shown, the model performs very well in the short run, achieving a higher portfolio value in the first 3 years of deployment. However, this success is short-lived and dwindles in later years. Recognizing the need for more frequent updates to the KL structures, we also implemented a rolling KL window that consistently recomputes the KL scores. This allowed a much more conservative yet consistent strategy that also outperformed the S&P 500 in the short run. Figure 11 below shows the overall comparison of all three approaches.

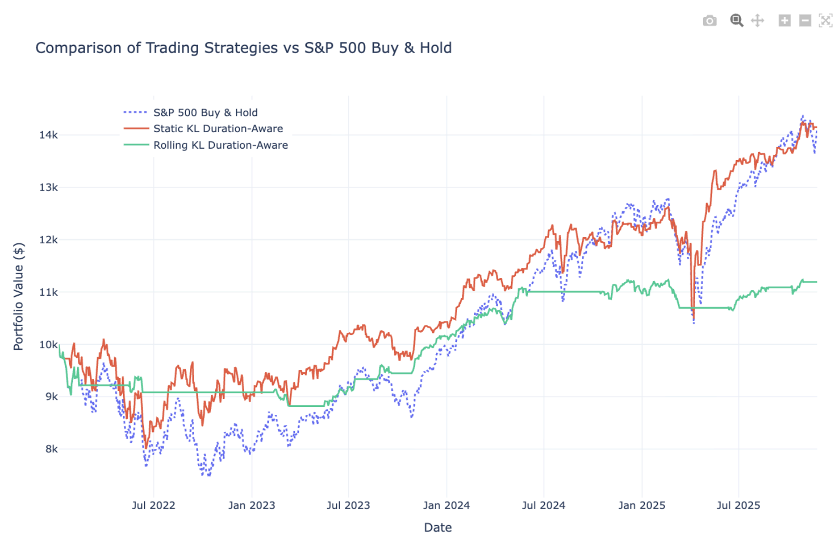
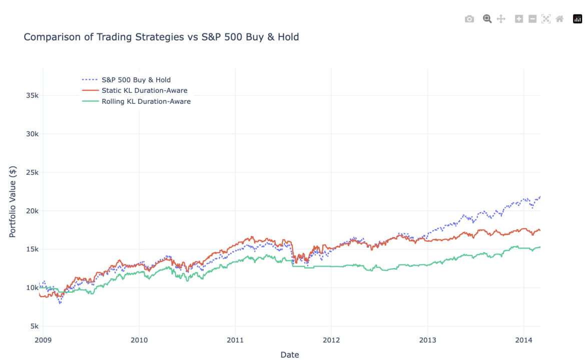


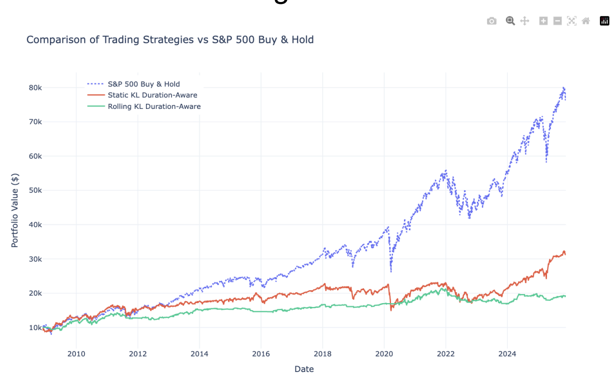
Figure 11: Performance of all three strategies.

Training: 2004-01-01 to 2006-12-31
Trading: 2008-11-04 to 2014-03-31

Figures 12 & 13: Trading Game 2. Duration-Aware Strategy vs. Buy and Hold.

To ensure that this short-run performance is not just a one-time success, we tried a training window from 2004 to 2006, then conducted a trading game from 2008 to 2014. As shown, we achieved relative success in the short run, but were beaten in the long run.





Figures 14 & 15: Comparing Trading Strategies

In this scenario, the rolling KL window performed as well as the previous trading game. This can most likely be attributed to the fact that this is a more conservative strategy in a long, calm period, resulting in fewer opportunities to outshine the other methods.

Results

As the above graphs show, our models perform relatively well in the short run when compared with the historical returns of the S&P 500. However, as the structures rapidly change over time, the predictions tend to decline, necessitating more sophisticated parameter updates.

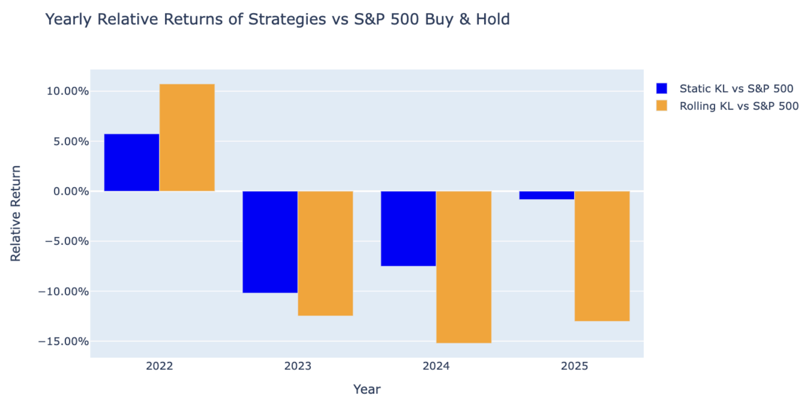


Figure 16: Trading Game 1. Comparing returns of both strategies against the S&P 500.

In Figure 16, we see that both strategies outperform the S&P 500 significantly in the short run but eventually do not fare well in later years.

	Year	Strategy	Start Balance	End Balance	Annual Return
0	2022	Static KL Duration-Aware	$10,000.00	$8,583.93	-14.16%
1	2023	Static KL Duration-Aware	$8,583.93	$9,789.52	14.04%
2	2024	Static KL Duration-Aware	$9,789.52	$11,336.39	15.80%
3	2025	Static KL Duration-Aware	$11,336.39	$12,945.72	14.20%
0	2022	Rolling KL Duration-Aware	$10,000.00	$9,082.78	-9.17%
1	2023	Rolling KL Duration-Aware	$9,082.78	$10,150.33	11.75%
2	2024	Rolling KL Duration-Aware	$10,150.33	$10,972.18	8.10%
3	2025	Rolling KL Duration-Aware	$10,972.18	$11,192.96	2.01%
0	2022	S&P 500 Buy & Hold	$9,993.54	$8,004.63	-19.90%
1	2023	S&P 500 Buy & Hold	$8,004.63	$9,944.19	24.23%
2	2024	S&P 500 Buy & Hold	$9,944.19	$12,262.08	23.31%
3	2025	S&P 500 Buy & Hold	$12,262.08	$14,105.57	15.03%

Figure 17: Annual returns for all three strategies.

Again, in terms of our model, we get our best results when emitting three hidden states in the GHMM and through capturing 95% of the data for our KL Expansion reconstruction.

Evaluating on a more general level, both the KL Score trained HMM and the Reconstructed Data HMM perform sufficiently. Still, there is a clear tradeoff between false positives in the KL score HMM and false negatives in the reconstructed Data HMM. Acknowledging this is essential to how well the trading game performs, as it may be making suboptimal decisions due to the trade-off between precision (where having too many false positives is bad) and recall (where having too many false negatives is bad).

Furthermore, we wish to continue developing our model to notice when the market spread is decreasing. Right now, our model predicts stressed markets when volatility is high (no matter if bond spreads are increasing or decreasing). By noting the direction of volatility, we can also make more optimal decisions about when to trade.

Conclusions

Our system leverages OAS bond spreads to find early warning signals of anomalous economic activity. The process is relatively straightforward and has the benefits of extremely low data requirements and training time.

Overall, we have shown the model's success in the short run, but relatively lackluster performance in the long run. While the model is not perfect, we recognize areas for future development that could yield dramatic improvements. This project serves as an example that even low-computational, low-data-intensive models still hold significant value in today's world of large-scale deep learning. The inclusion of an updating method can be a potential extension. Our model's modest success in the naive trading game suggests it could be applied to other contexts.