USING DISCRETE AND CONTINUOUS LATENT VARIABLE MODELS TO EXPLORE INDUSTRY STATES

Rong Liu
HUID: 81369208
rongliu@g.harvard.edu

Yuting Kou
HUID: 91372452
yutingkou@g.harvard.edu

Yizhou Wang
HUID: 51373162
ywang@g.harvard.edu

May 13, 2019

ABSTRACT

We believe that the stock market price is determined by the market state of the corresponding industry. This paper wants to answer two questions by analyzing the daily closing prices of 9 industry indices: (1) What is the hidden state like – are there a fixed number of discrete states, or is it continuous? (2) What are the differences or connections among industries? We decide to explore the industry states through two latent variable models: discrete Hidden Markov Model (HMM) and continuous Variational Recurrent Auto-encoder (VAE) with Long Short-Term Memory (LSTM). First, we use a Kalman Filter to smooth out the daily noises without referring to future information. Under the discrete state hypothesis, we train an HMM and determine the optimal state number by maximizing posterior likelihood and mean squared error. Under the continuous state hypothesis, we relax the restrictions on emission and transition models, and we employ VAE with LSTM. With these two models, we compare the industries based on the means and variances of their hidden states. We further explore whether there exist unifying patterns in hidden states for different industries. These unifying patterns may be a result of industrial connections, such as sharing supply chains, or having a cyclic nature.

***K*eywords** Latent Variable Model · Hidden Markov Model · Variational Autoencoder · Kalman Filter · EM · LSTM

1 Introduction

Characterizing the dynamics of the stock market has always been a hot and difficult topic on Wall Street. We model stock market prices as a chaotic representation of underlying driven factors, such as industry performance, and we define these factors as different manifestations of hidden market states. So what are the hidden market states?

Terms like "bull market" and "bear market" are common ways of describing stock market states. Maheu and McCurdy [2000] use a two-state Markov-switching model to capture the nonlinear structure in conditional mean and variance of stock returns. De Angelis and Paas [2013] model the weekly prices of market index based on a hidden Markov model with four latent states, and Jiang and Fang [2015] attempt to detect more different states in a stock market according to marginal likelihoods. However, these models all assume hidden states to be discrete. In this paper, we experiment with continuous states. Furthermore, Pagan and Sossounov [2003] use predefined bull and bear market characteristics in different data generation processes to simulate the stock market, and they find the non-Gaussian process with Martingale difference to be successful. This motivates us to explore more complex emission and transition processes.

In this paper, we model the stock prices with discrete and continuous latent variable models, trying to answer the following questions:

1. What are the hidden market states – are there a fixed number of discrete states, or is it continuous?
2. What are the differences or connections among industries – do they exhibit cyclical changes or similarities due to industry relations such as supply chains?

To tackle these problems, we first use a Kalman Filter to smooth out the daily noises without involving future information (Section 2). Next, we implement a HMM model with the optimal number of states selected by maximizing the opportunity

gain in posterior likelihood and mean squared error (Section 3.1). The fact that the posterior likelihood increases with the number of hidden states motivates us to propose a variational autoencoder [Kingma and Welling, 2013] with recurrent LSTM [Gers et al., 1999] (Section 3.2). The association between HMM and VAE is discussed in Section 4.3. Lastly, in Section 4, we compare the relationship of latent market conditions in different industries with special attention to cyclical industries and industries in the same supply chain.

We believe that this article has the following contributions: First, we take an innovative approach to use a continuous latent variable model to analyze the states of the stock market, motivated by findings using the discrete latent variable model. Second, we discover that the latent space representation of HMM and VAE are related. The existing theories have not offered an explanation for this discovery. Third, we compare the industry hidden states. Such analysis can help us adjust the trading strategy accordingly, formulate corresponding industrial policies, and possibly predict the stock index returns.

2 Data

We download the daily closing price of 11390 individual stocks from 2010 to 2018 from CRSP[1] database. Then, we divide the stock into nine industries based on the SIC[2] code and calculate the average closing price of all stocks in each industry as the industry closing price. We focus on the finance industry first to build two type of latent variable models, and then compare the industry differences on the nine industries. All our models are fitted in the training dataset up to Jan. 1, 2017, and analyze the results on the test dataset (last two years).

2.1 Discovery

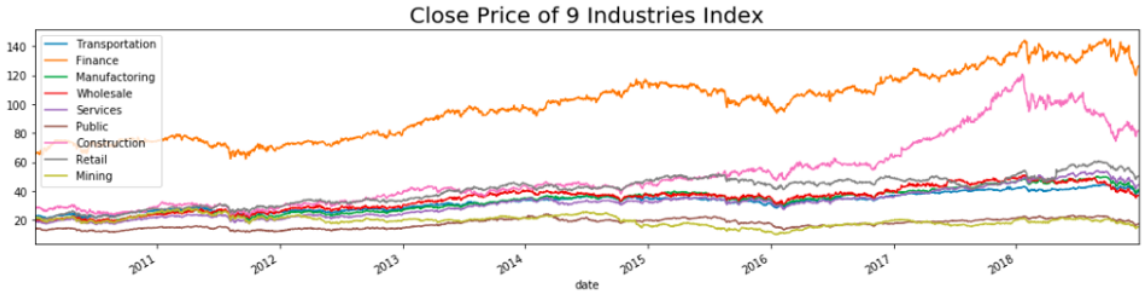


Figure 1: Nine industry closing prices.

Figure 1 shows the closing price of nine industry indices. We observe that almost every industry has experienced bull and bear markets, and sometimes the market may go through a period of stagnation as it tries to find direction. In addition, some industries show a common movement, some industries have periodicity, and some industries are different from other industries. We want to build a latent variable model and find the optimal number of hidden states. We try to explain the differences or connections of hidden states in different industries using two different latent variable models.

2.2 Kalman Filter

Since daily prices involve too much noise, we want to smooth them without involving future information. Kalman filter is essentially an algorithm that recover a more accurate estimate of the unknown variable for each time t based on a series of noise measurement observations (stock prices) using only the information up to time t.

We extend the dynamic transfer equations in traditional Kalman filters to ARiMA models.

$$\begin{cases} \Phi(L)(r_t) = \theta(L)\epsilon_t & \epsilon_t \sim N(0, \sigma_\epsilon^2) \\ y_t = r_t + \delta_t & \delta_t \sim N(0, \sigma_\delta^2) \end{cases} \Rightarrow \begin{cases} r_{t|t-1} = \sum_{i=1}^{5} \Phi_i r_{t-i|t-i} \\ \sigma_{t|t-1}^2 = \sum_{i=1}^{5} \Phi_i \sigma_{t-i|t-i} + \sigma_\epsilon^2 \sum_{i=1}^{5} \theta_i \\ \theta_{t|t} = r_{t-1|t-1} + \frac{\sigma_{t|t-1}^2}{\sigma_{t|t-1}^2 + \sigma_\delta^2}(y_t - r_{t-1|t-1}) \\ \sigma_{t|t}^2 = (\frac{1}{\sigma_{t|t-1}^2} + \frac{1}{\sigma_\delta^2})^{-1} \end{cases} \tag{1}$$

[1] the Center for Research in Security Prices maintains the most comprehensive collection of security price data for the NYSE, AMEX and NASDAQ stock markets.

[2] The Standard Industrial Classification (SIC) is a system for classifying industries by a four-digit code in the United States.

where L is a lag operator.

We find that the stock price is a unit-root process via Augmented Dickey-Fuller unit root test, so we model them in its one-order difference (return). Then by assuming that the coefficients in dynamic transfer equations are similar to the prices, we identify the orders of ARMA(1,1) from PCAF and ACF in the training set, and then estimate the coefficients using ARMA regression. Finally, we incorporate the fitted coefficients into our Kalman filters.

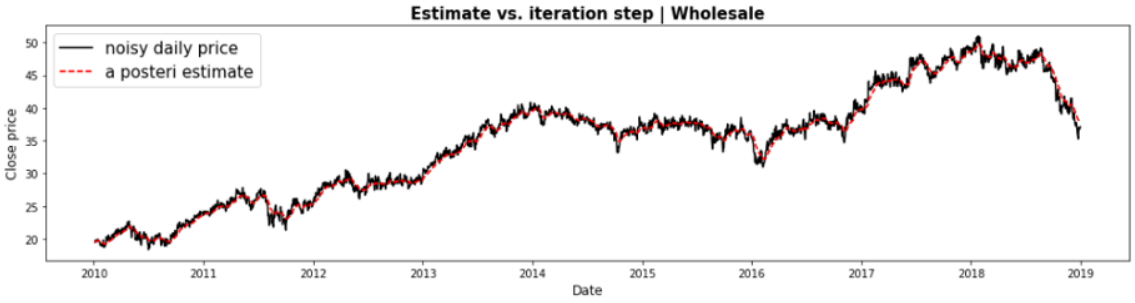


Figure 2: Kalman filter to smooth price.

Figure 2 shows the result of Kalman filter. It smooths the daily prices while maintaining the main pattern of stock movements. We build our models based on the filtered price, and the comparison of HMM results based on the original price and the filtered price further supports the usefulness of Kalman filter pre-processing.

Initially, we attempt to use the Kalman filter as a latent variable model to predict stock returns. However, we find that due to the Gaussian assumption on the emission function, the best prediction is simply a copy of the predicted hidden state, which is the weighted average of past returns. Therefore, we believe that the Kalman filter is not suitable for predicting stock prices, and we only use it as a denoising tool.

3 Method

We believe that there are a sequence of hidden states behind the stock market, and the prices are noisy realizations of the hidden states. Our initial question "What is the hidden state?" is now broken down into: how do we define the states, the transition model between consecutive states, and the emission model states to observations. In Hidden Markov Model, we assume that states are discrete, and the emission model has an underlying Gaussian distribution. However, we are afraid that we may be placing too much restriction on the nature of the stock market. That's why we introduce Variational Autoencoder, which models continuous states and potentially universal functions for the transition and emission processes.

We observe that the stock prices across industries are quite different. Thus, with the aim to combat such inconsistency, we train our models with return rates instead of raw stock prices:

$$r_{t+1} = \frac{p_{t+1} - p_t}{p_t}$$

where r_i denotes the return rate at day i, and p_i denotes the return rate at day i.

3.1 Hidden Markov Model

A Hidden Markov Model is defined by the following parameters:

1. $S = s_1, s_2, s_3...$: Set of possible states
2. $O = o_1, o_2, o_3...$: Set of possible observations
3. A: Transition probabilities p(st+1|st)
4. B: Emission probabilities $p(o_t|s_t)$
5. π: Initial state probabilities $p(s_1)$

In our case, we assume the possible observations are continuous, and the possible states are discrete. This is motivated by the general impression that the stock market has discrete states such as the bull market and the bear market. We further assume that the emission model follows a Gaussian distribution, and the prior initial state probability and transition probability are both uniform. The last hyperparameter we need to choose is the number of states.

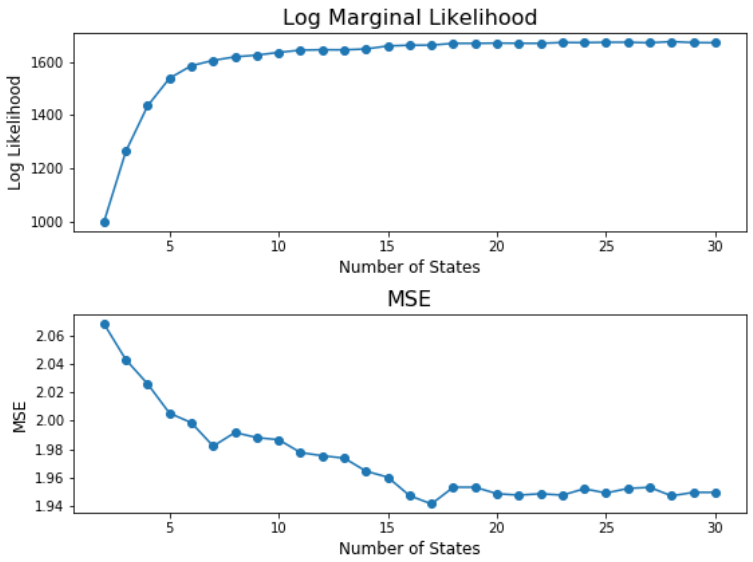


Figure 3: Posterior Log-likelihood and MSE of HMM with Different Numbers of States.

As plotted in Figure 3, we observe the general trend that the log-likelihood increases and MSE decreases as the number of states increases. This observation provides some intuition to model the market hidden states as continuous. However, as the number of states increases, the computational complexity also grows, and the amounts of improvement in the plotted metrics diminish by each step. In order to maximize the opportunity gain, we choose to have five hidden states.

We employ an "hmmlearn" package in Python to train our HMM model. It takes in the data series (the series of return rates in our case) and the hyperparameters that we select, and it learns the essential parameters with an Expectation-Maximization Algorithm. After that, it is able to estimate the optimal hidden state sequence given a data series, using a Forward Algorithm.

Usually the daily return rates for industry index are very small measures, and the difference among them is minimal. When we directly use return rates as the input to an HMM, relevant computation reaches the lower limit of computer precision, so the resulting states' means are nearly equal. Hence, we multiply the return rate data with 100 to expand the range and to exaggerate the difference among them.

3.2 Variational Autoencoder with LSTM

We consider using an autoencoder to learn the hidden states of the stock market because the encoder step maps the most important features in the input data to some latent variable. However, one problem with traditional autoencoders is that the latent space, where the encoded vectors lie, may not be continuous or allow easy interpolation. The variational autoencoder modifies the autoencoder architecture by replacing the deterministic function with a learned posterior recognition model $q_\phi(z|x)$. This model parametrizes an approximate posterior distribution over z with a neural network conditioned on x.

The latent variable z is generated from the normal distribution with mean μ and standard deviation σ, where μ and σ are hidden nodes in the encoder network. Now that z is not deterministic anymore, we need to use the reparameterization trick to enable gradient calculation in the back propagation. This is done by calculating z as $\mu + \sigma \odot \epsilon$ rather than generating from $\mathcal{N}(\mu, \sigma^2)$ directly. Then, we can take derivative with respect to μ and σ.

The VAE loss consists of two parts:

$$\mathcal{L}(x; \theta, \lambda) = D_{KL}(q(z|x; \lambda)\|p(z)) - \mathbb{E}_{z\sim q} \log p(x|z, \theta).$$

The KL-Divergence loss in the first part keeps the posterior distributions close to the prior $p(z)$, and the second part is a reconstruction loss. The specific steps of variational autoencoder are shown in Figure 4.

Since stock price is a time serie, we use some recurrent neural networks in the encoder part instead of the forward layers. Compared to normal RNN, LSTM can prevent vanishing gradients. Therefore, we use LSTM cells in the neural network.

The network structure of the VAE-LSTM model is shown in Figure 5. There are two layers in the encoder and three layers in the decoder. The latent dimension is set to be sixteen.

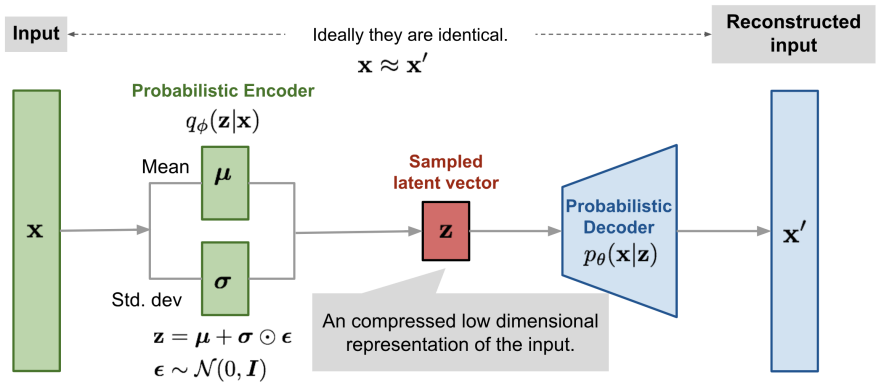


Figure 4: Whole process of Variational Autoencoder.

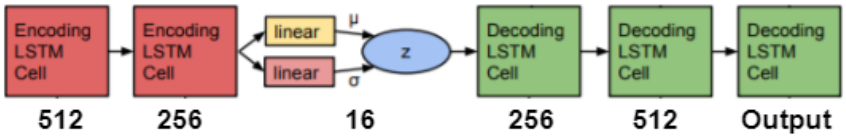


Figure 5: Network structure of the Variational Autoencoder with LSTM.

4 Result

4.1 Result of Hidden Markov Model

The states in the HMM models are specified by a mean and a variance, since the emission model is assumed to be Gaussian. The results for the Finance industry, measured in percentage return rate, are listed in the Table 1.

State Number	Mean (Filtered)	Var (Filtered)	Mean (Unfiltered)	Var (Unfiltered)
1	0.2648	0.0089	1.3437	1.5862
2	0.1224	0.0027	0.4492	0.4285
3	0.0148	0.0021	0.0891	0.3736
4	-0.1024	0.0052	-0.3016	0.5610
5	-0.2805	0.0290	-1.5195	1.6913

Table 1: Hidden States of Finance Industry

We observe that after filtering, the difference between any two consecutive days in the data is smaller, so the absolute values of the means and variances are much smaller than those of the model trained with unfiltered data. Moreover, the variance values are drastically larger than the rest when the corresponding mean values are the largest or the smallest. This is because given a limited number of hidden states, the model uses only two states to represent the extreme cases where the stock price rises or drops by an extensive amount. Such extreme change is reflected by the high variance.

We evaluate our HMM model with two methods. First, we use it to predict stock prices in 2017-2018. The expected return rate in the next day is computed with the current day's estimated state, the transition matrix, and the state means. The expected closing price is then calculated based on the expected return rate and the current day's true closing price. Second, we use our HMM model to estimate the sequence of hidden states. The hidden states are not ordered when they are learned, so we use their means to represent the states for clearer illustration.

In Figure 6, we plot price and state sequences predicted by HMM models that are trained with filtered and unfiltered data. We set the number of states to five as optimized before, and we also plot performances of two-state HMM models for comparison. It is not surprising to see that with unfiltered data, because of the daily noises, the hidden state is very unstable – it keeps altering between the two states that have relatively smaller absolute return rates. In contrast, when we fit our

model with filtered data, the state sequence is much smoother. This confirms that using the Kalman Filter helps us get a more accurate as well as interpretable understanding of the nature as well as changes of the hidden states.

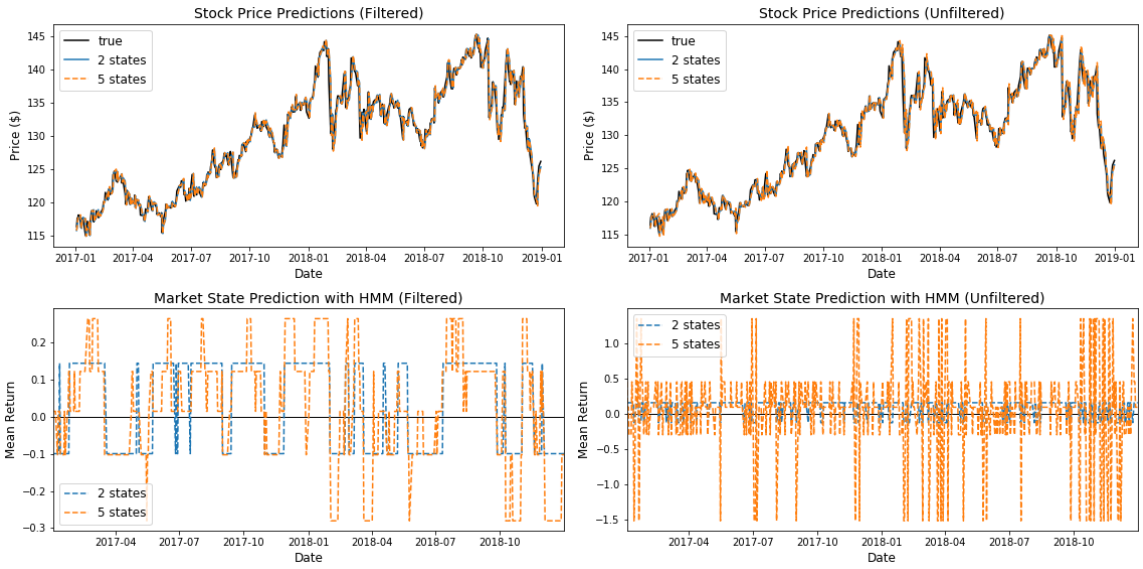


Figure 6: Price and State Predictions of HMM.

4.2 Result of VAE-LSTM

For LSTM cells, the time step is set to be two months, which is 60 days. Batch size is 32. We use Adam as the optimizer with learning of 0.0002. The model is trained for 15 epochs. We evaluate the model by predicting close price on the test set, which is from June 2017 to December 2018.

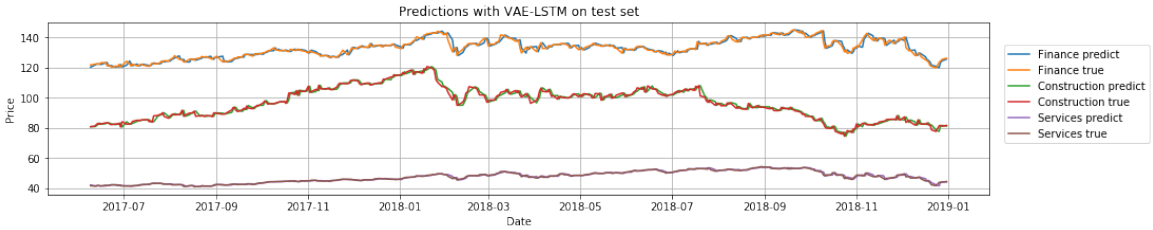


Figure 7: Predictions of close price for three industries with VAE-LSTM.

The result is shown in Figure 7. For clarity of the plot, we only select three industries (Finance, Construction, Services) to see the performance. It can be seen he predicted curve and true curve match quite well.

4.3 Connection between Hidden States of HMM and VAE-LSTM

The latent dimension of our Variational Autoencoder is 16, which means that there are 16 means and variances that are used to generate the latent variable z. To explore the latent states, we perform Partial Component Analysis (PCA) on the latent means and standard deviations, and extract the top 2 components. To compare the results of Hidden Markov Model and VAE-LSTM, we plot the training points in the latent space of VAE and color them using the states learned by HMM. The plot is shown in Figure 8. The points are colored in 5 different colors, corresponding to 5 hidden states in the Hidden Markov Model.

The plots indicate that the results of VAE and HMM are kind of consistent. From the plots, we can see that the points on each loop are classified as the same state by HMM. This means that the points of the same state do have some patterns in the latent space, which justifies that the hidden states learned by the HMM and the latent variables learned by VAE are reasonable.

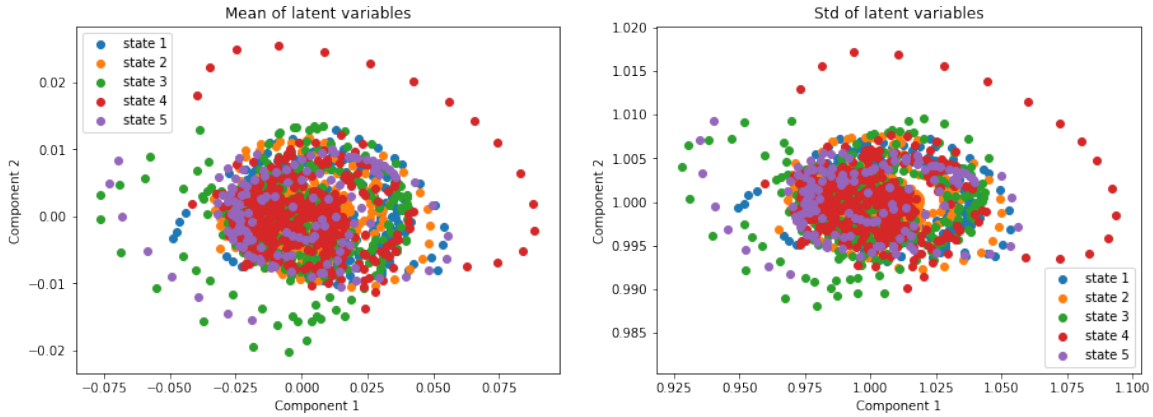


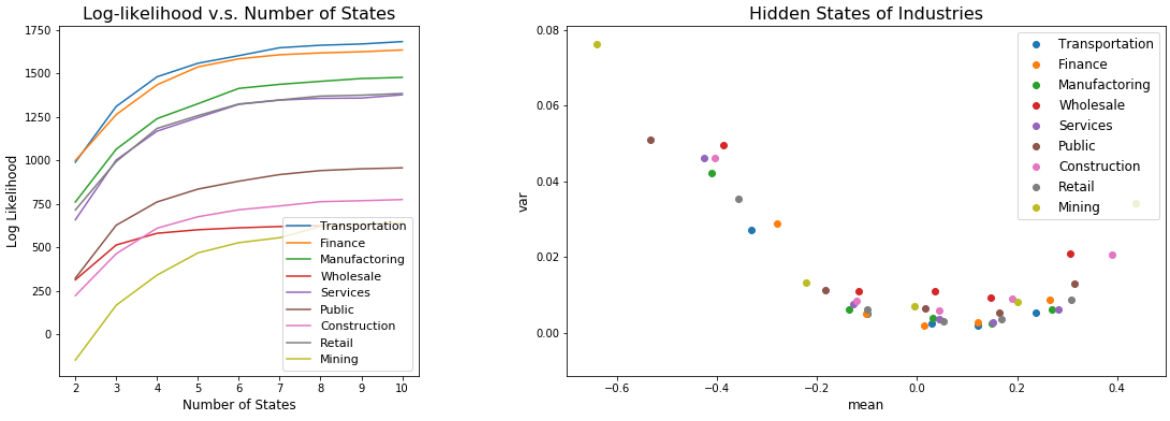
Figure 8: Scatter plot of construction industry index in VAE latent space, color-coded by hidden states from HMM.

5 Applications

5.1 Industry Hidden States

As shown in Figure 9a, most industries exhibit a drop in opportunity posterior likelihood gain when they reach five hidden states. For some industries, as many as five hidden states are not necessary, but it would not harm the performance either. For consistent comparison across industries, we model them with the same number of hidden states, which is five.

We plot in Figure 9b the information of the hidden states of all industries generated by HMM. Because each state is defined by a mean and a variance, we are able to plot them in a 2D scatter plot. We observe again the trend that the variances are higher when the means are relatively larger or smaller, in order to reflect extreme rises and drops in return. Mining, public service, and construction's hidden states have the largest absolute means and variances, indicating that they are among the most fluctuating industries. For the same reason, finance and transportation are relatively more stable.



(a) Posterior Log-likelihood of all Industries.

(b) Market States of Industries Generated by HMM.

Figure 10 shows the relationship between latent mean and volatility of two sample industries, finance and mining. In the VAE model, we have sixteen latent variables. For illustration's purpose, we perform principle component analysis on the sixteen means and sixteen variances, and we only plot the most principle components. The distribution of the results for finance and mining both exhibit an oval shape, but in opposite directions. We suspect the reason is that one single principle component is not enough to convey the whole story.

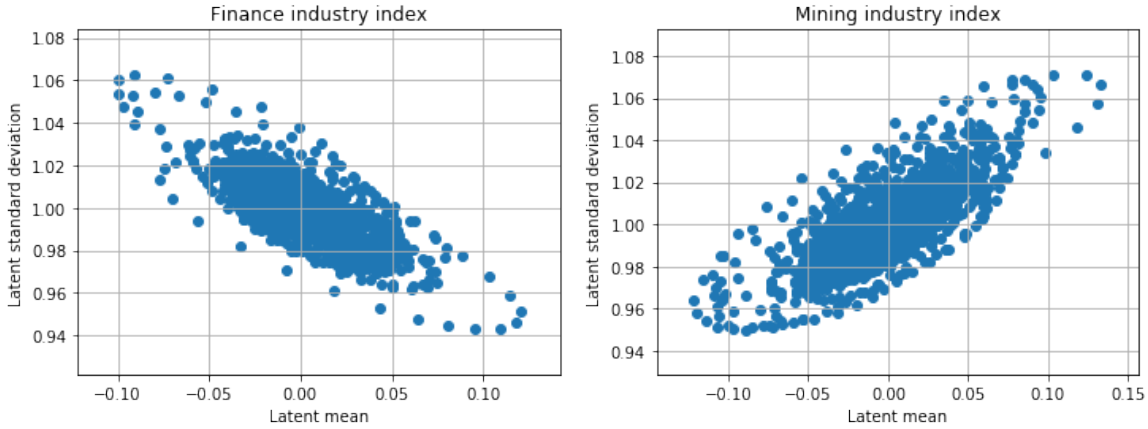


Figure 10: Relationship between latent mean and volatility among different industries.

5.2 Impact of Supply Chain

We select two sets of industries that consist supply chains, and we compare the results to see if their stock prices are correlated. In the first set, we compare manufacturing, wholesale, and retail; in the second set, we compare mining, construction, and public service. We observe that the state sequences in the first set resembles each other more than those in the second set. This is not surprising because manufacturing, wholesale and retail are more closed related in reality. It is also interesting to see that wholesale has higher variance than manufacturing and retail, even though wholesale locates at the middle of the supply chain. In the second set, it is arguably true that mining is more sensitive and thus more fluctuating than construction and public service.

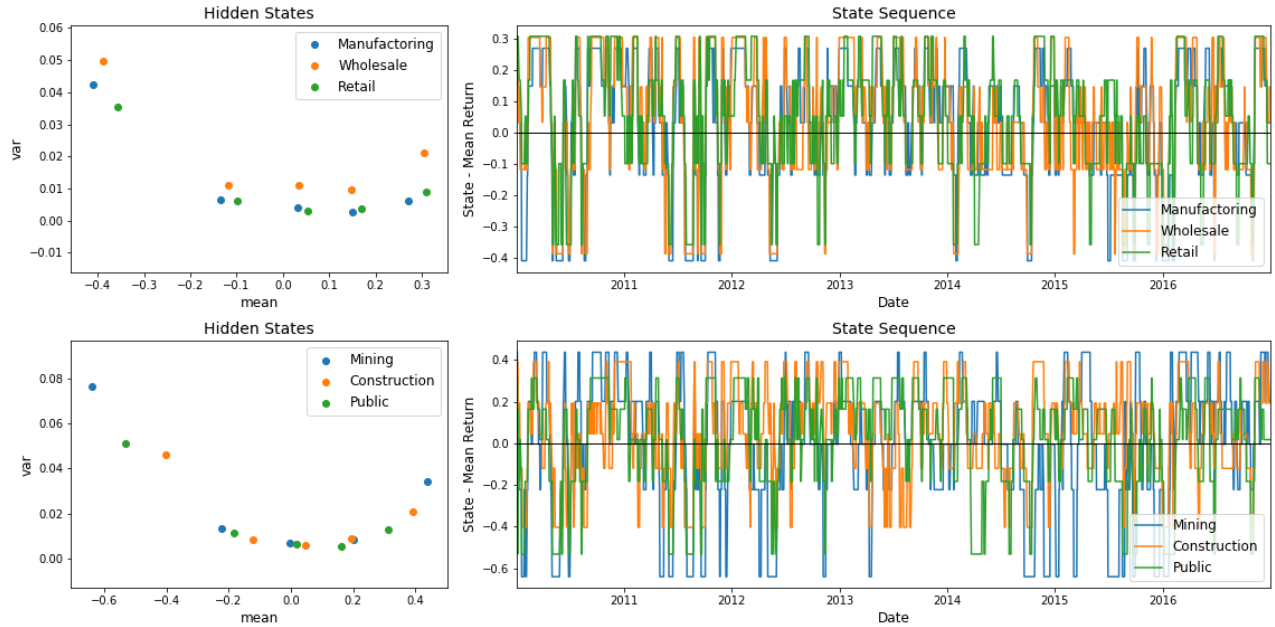


Figure 11: Comparison of Industries in the same Chain.

5.3 Impact of Cyclicity

Furthermore, we select industries that people generally perceive as cyclical, meaning that the industries experience a period of rising followed by a period of dropping, repeatedly. In the figure below, we plot the state sequences of Finance and

Mining, which indeed reflect such cyclicity. For example, the finance industry experiences positive return roughly in 2010, 2012, 2014, and 2016, and experiences negative return roughly in 2011, 2013, and 2015.

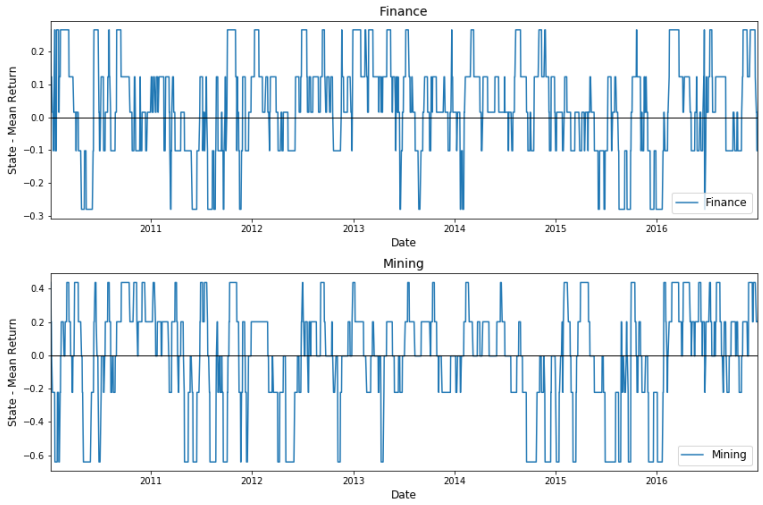


Figure 12: Performance of Cyclical Industries.

6 Conclusions

Recall that we raised two key questions at the beginning. Regarding the first question "What are the hidden market states", we attempt to model these states with HMM and VAE. We find that with HMM, it is optimal to have five discrete hidden states to explain the US stock market in the past 8 years. Moreover, the two models' results are consistent with each other, as observations identified as belonging to different states by the HMM have distinct patterns in the VAE latent space. Regarding the second question "What are the differences or connections among industries", we compare the industries states in terms of their means and variances generated by both HMM and VAE with LSTM. We further find that industries in the same supply chain have similar changes in hidden states, and cyclical industries exhibit cyclical trends in their state sequences.

Following our work presented in this paper, there are a few future tasks that we can perform to better understand the nature of the stock market hidden states. Currently, we are only using the time series of closing prices as our input. One can also include other features such as stock volume, GDP, and technical indicators such as Moving Average Convergence Divergence (MACD). We believe these features are not directly correlated with closing prices and will provide new perspectives in our analysis. For consistent comparison, we model all industries with five hidden states, even though for some particular industries, fewer hidden states are enough. In order to investigate any single industry other than finance, one can try with a different number of states. Lastly, the performance of VAE with LSTM in comparing industries requires more theoretical investigation.

References

Luca De Angelis and Leonard J Paas. A dynamic analysis of stock markets using a hidden markov model. *Journal of Applied Statistics*, 40(8):1682–1700, 2013.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

Yu Jiang and Xianming Fang. Bull, bear or any other states in us stock market? *Economic Modelling*, 44:54–58, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

John M Maheu and Thomas H McCurdy. Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics*, 18(1):100–112, 2000.

Adrian R Pagan and Kirill A Sossounov. A simple framework for analysing bull and bear markets. *Journal of applied econometrics*, 18(1):23–46, 2003.