Finding Hidden Market States with Kalman Filter, HMM and VAE

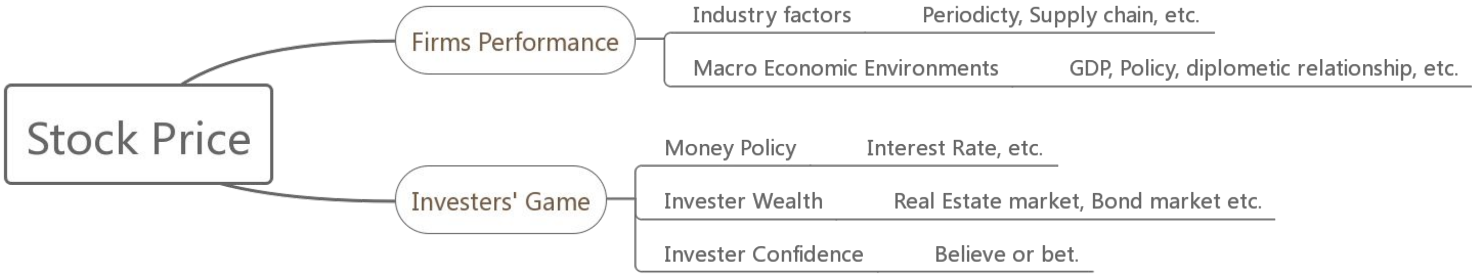
Yizhou Wang, Yuting Kou, Rong Liu

Overview

- Motivation
- Data
  - Preprocessing: denoise using Kalman Filter
- Models and Preliminary Results
  - Hidden Markov Model
  - Variational Recurrent Auto-encoder with LSTM
- Future Work

Motivation

- Data Generation Process: How does stock market work?

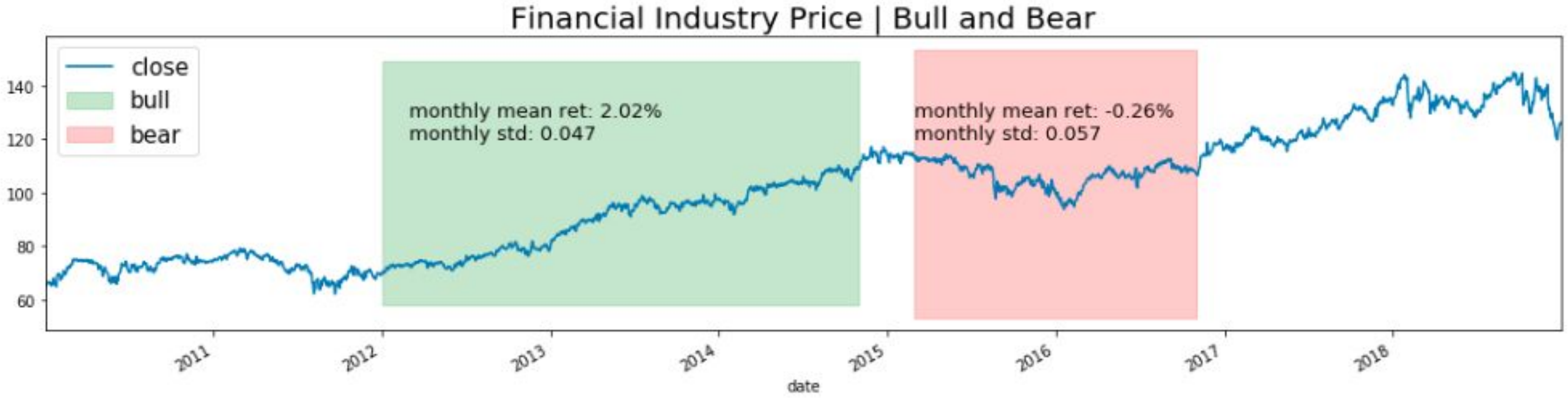


  - Market states drive the price
  - We want to recover the market states by prices

- Why we need to know market states?
  - Understand current states: predicting price, changing strategies, etc.
  - Know underlying connections: making contingent policy, etc.

Motivation

- 1. What is the Market States?
  - Bull and Bear markets?



Motivation

- 1. What is the Market States?
  - Bull and Bear markets? Or more?
  - Discrete or Continuous?

- 2. Is there any difference/connection among industries?
  - Periodicity because of industry periodical factors
  - Connections between industry because of supply chains

Motivation

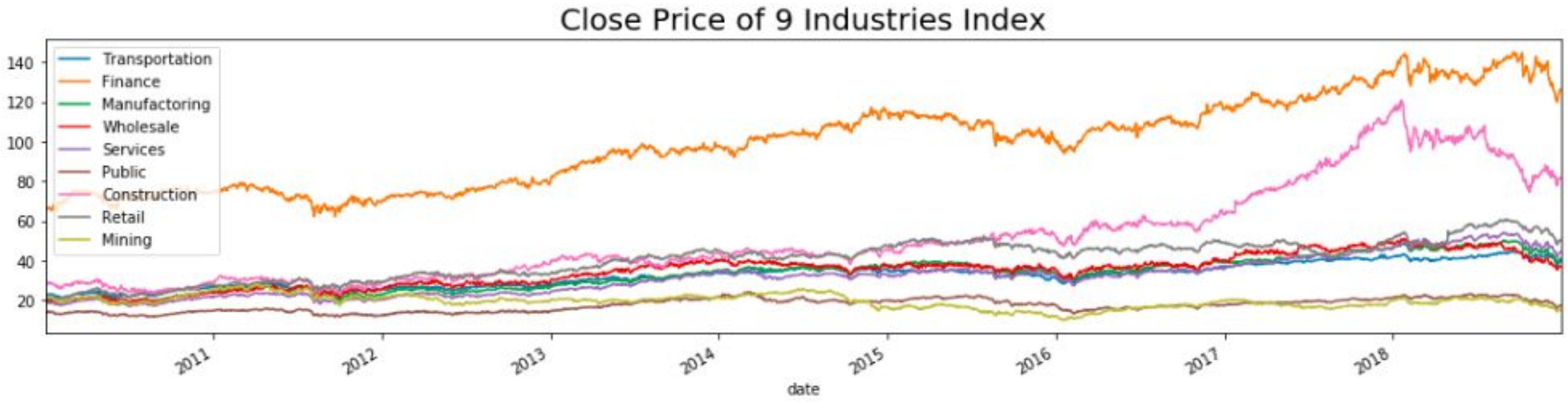
- 1. What is the Market States?
    - Bull and Bear markets? Or more?
    - Discrete or Continuous?

- 2. Is there any difference/connection between industries?
    - Periodicity of the industry
    - Connection like supply chain

- Hidden states → Latent Variable Models: HMM and VAE

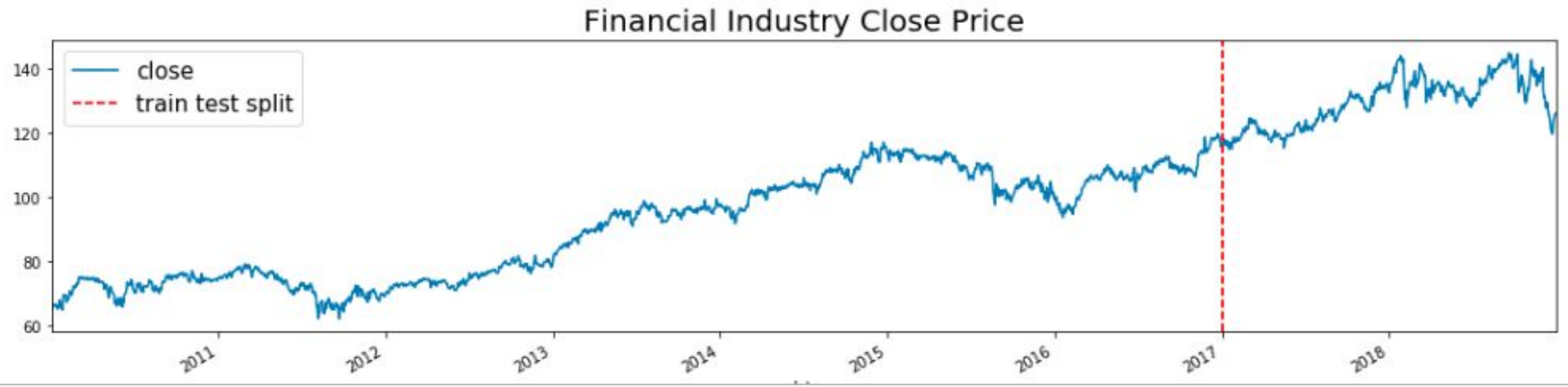
Data

- Daily closing prices of stocks in 9 industries, in 2010-2018
- CRSP - The Center for Research in Security Prices



Data

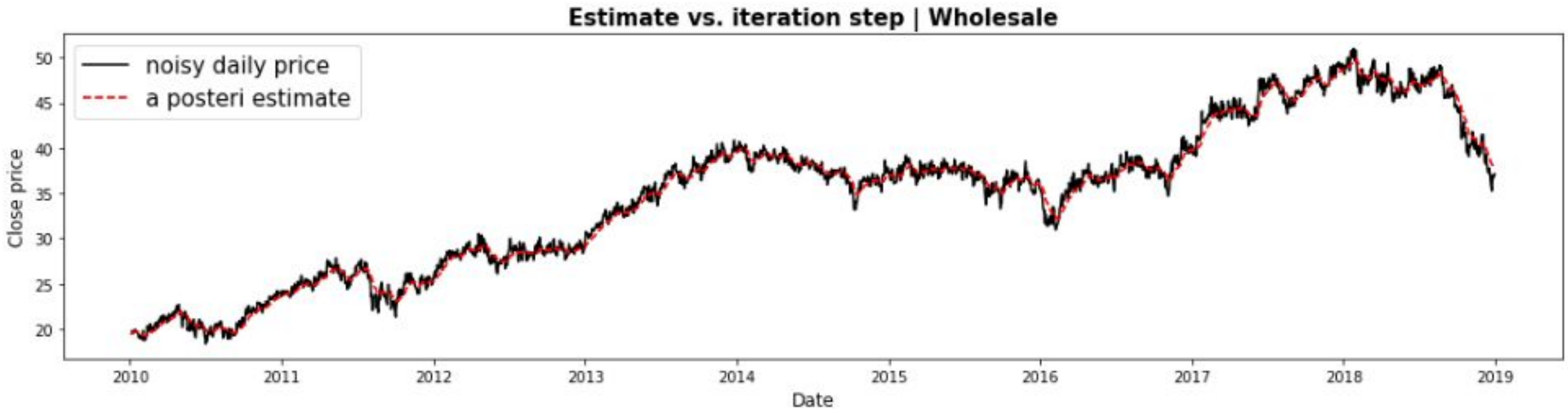
- Daily closing prices of stocks in 9 industries, in 2010-2018
  - We focus on Finance index first.



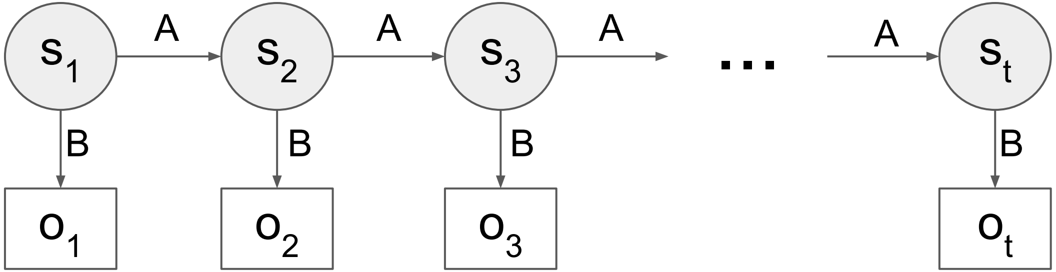
Kalman Filter

Denoise 9 Industry price using extended Kalman filter

$$
\begin{cases} \Phi(L)(p_t) = \theta(L)\epsilon_t & \epsilon_t \sim N(0, \sigma_\epsilon^2), \text{L is lag operator} \\ y_t = \theta_t + \delta_t & \delta_t \sim N(0, \sigma_\delta^2) \end{cases} \Rightarrow \begin{cases} p_{t|t-1} = \sum_{i=1}^5 \Phi_i p_{t-i|t-i} \\ \sigma^2_{t|t-1} = \sum_{i=1}^5 \Phi_i \sigma_{t-i|t-i} + \sigma_\epsilon^2 \sum_{i=1}^5 \theta_i \\ \theta_{t|t} = p_{t-1|t-1} + \frac{\sigma^2_{t|t-1}}{\sigma^2_{t|t-1} + \sigma^2_\delta}(y_t - p_{t-1|t-1}) \\ \sigma^2_{t|t} = (\frac{1}{\sigma^2_{t|t-1}} + \frac{1}{\sigma^2_\delta})^{-1} \end{cases}
$$

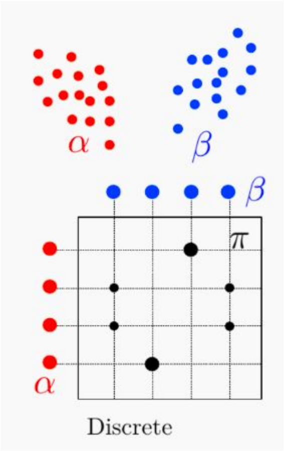


Model: General Approach

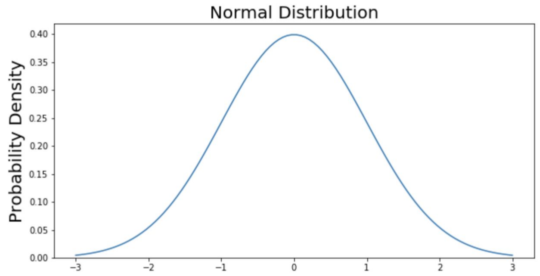


S: hidden state
O: observation
A: transition
B: emission

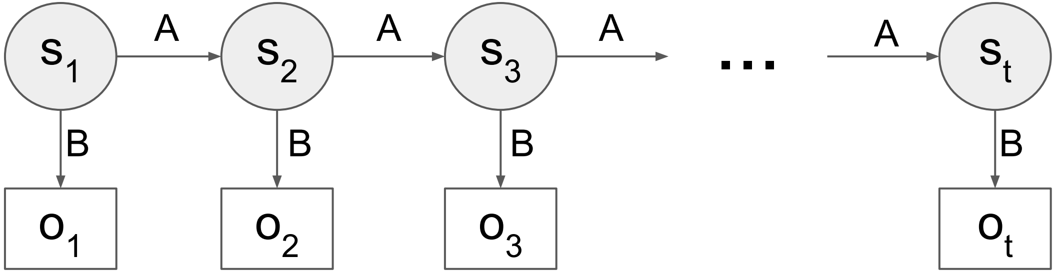
Transition



Emission

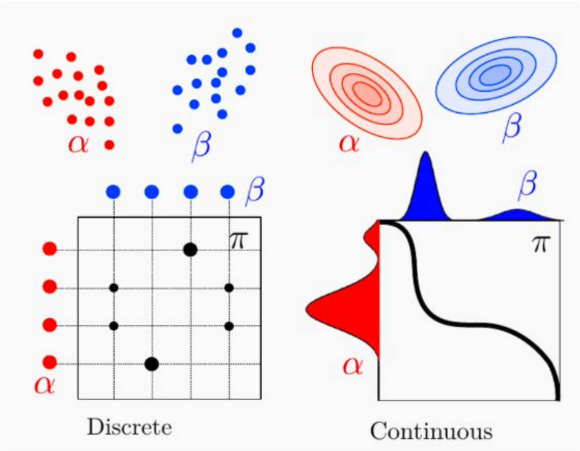


Model: General Approach

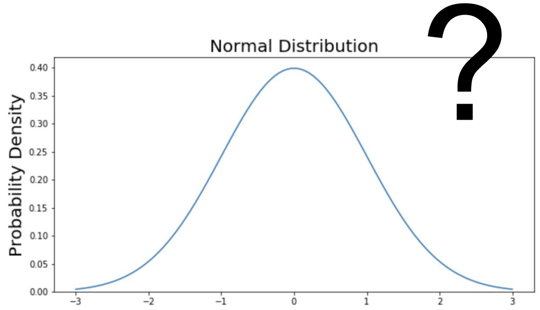


S: hidden state
O: observation
A: transition
B: emission

Transition

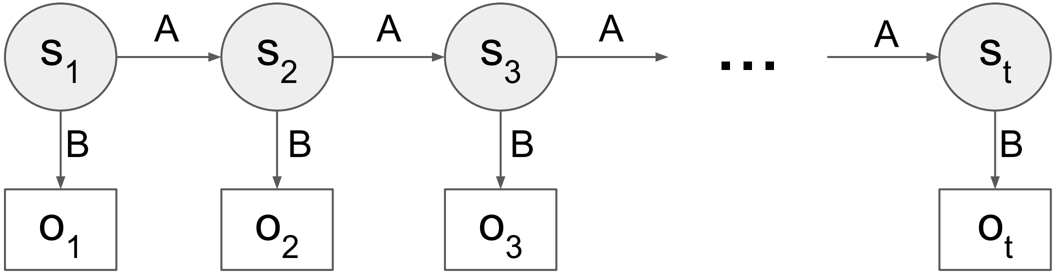


Emission



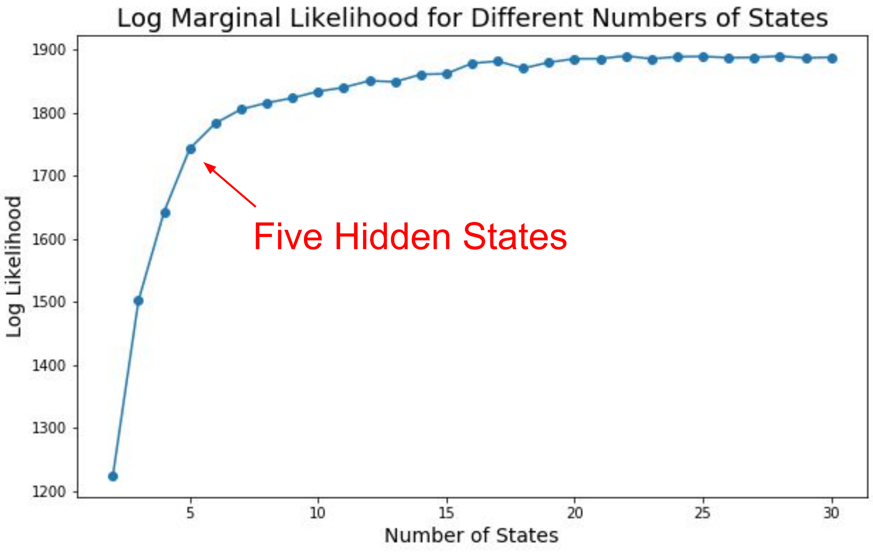
HMM

- Motivation:
  - It is generally assumed that there are a finite number of market states: bull, bear, etc.
- A Hidden Markov Model is defined by:
  - $S = \{s_1, s_2, s_3 ...\}$: Set of possible states
  - $O = \{o_1, o_2, o_3 ...\}$: Set of possible observations
  - A: Transition probabilities $p(s_{t+1}|s_t)$
  - B: Emission probabilities $p(o_t|s_t)$
  - π: Initial state probabilities $p(s_1)$



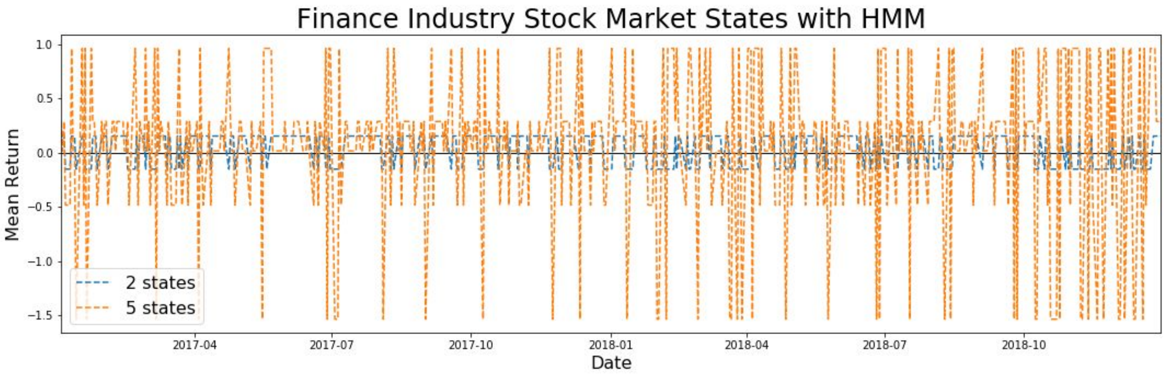
HMM - Training and Prediction

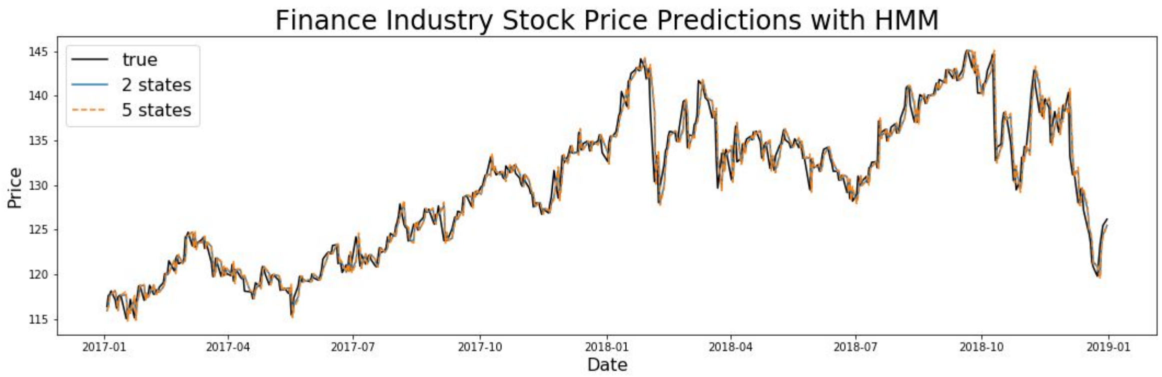
- Hyperparameters
  - Emission Model: **Gaussian**
  - Number of States: **Five**
  - Prior initial state probability and transition probability: **Uniform**
- Training
  - Input data sequence
  - Learns S, A, B, and π through iterative **Expectation-Maximization (EM)** to maximize posterior marginal likelihood
- Prediction
  - **Viterbi algorithm**: estimates optimal sequence of hidden states
  - **Forward algorithm**: calculate posterior marginal likelihood



HMM - Results (Unfiltered Input)

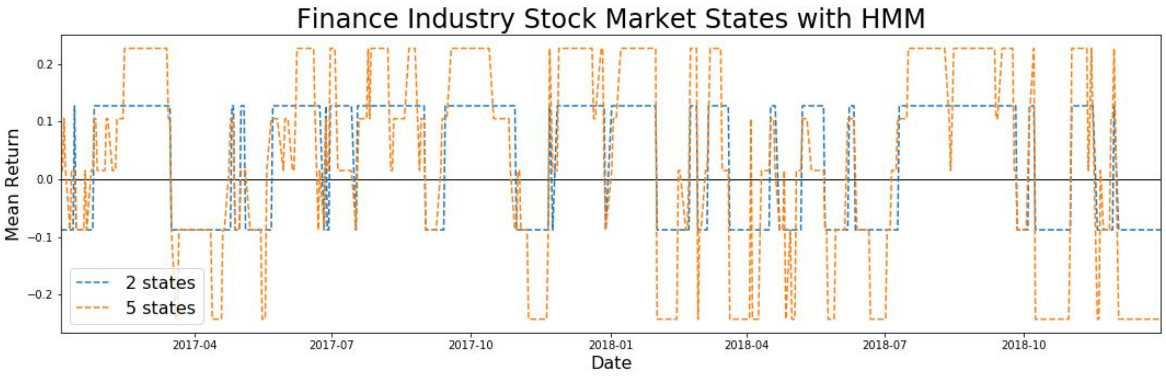
State No.	State (Mean, Variance)
1	(0.01606521, 0.29648902)
2	(0.96253467, 0.89453984)
3	(-1.53629294, 1.53925792)
4	(0.28834739, 0.3053085)
5	(-0.4837531, 0.57186103)

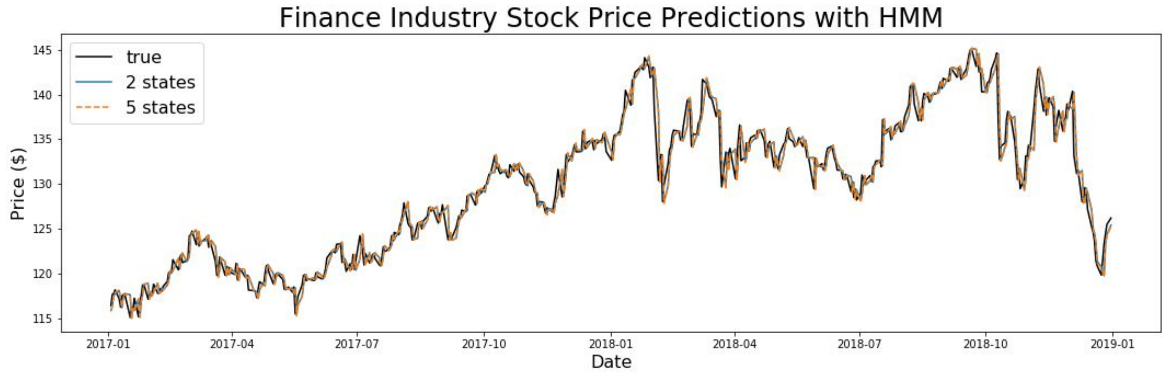




HMM - Results (Filtered Input)

State No.	State (Mean, Variance)
1	(0.07539053, 0.0007672)
2	(-0.00269605, 0.00034229)
3	(-0.04846012, 0.00064482)
4	(0.03338407, 0.00029805)
5	(-0.13193312, 0.00329104)

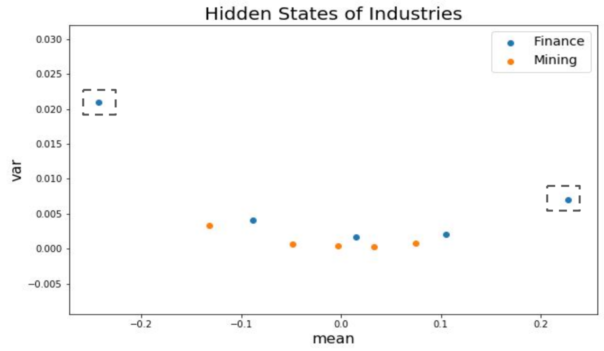


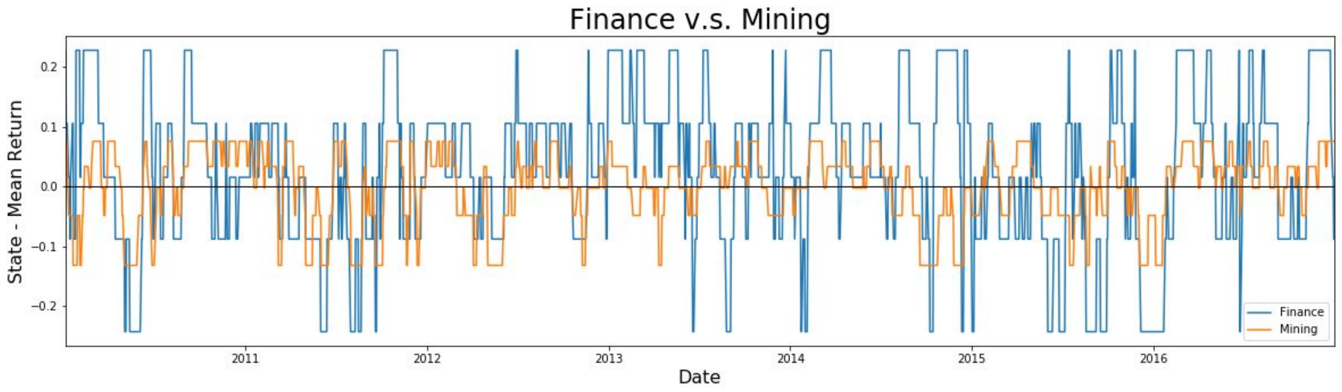


Application: Industry comparison

Case 1: Finance and Mining

- Finance fluctuates much more than mining
- No apparent correlation in sequence
- Finance: larger variance when falling

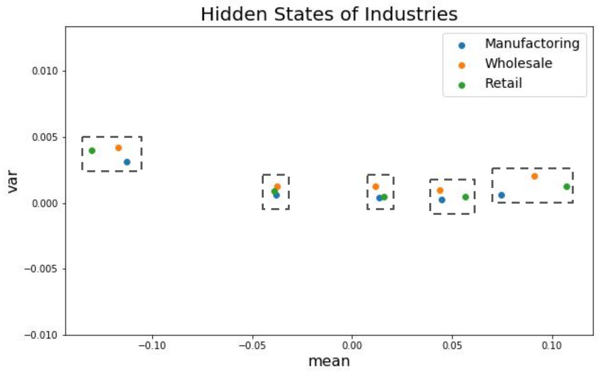


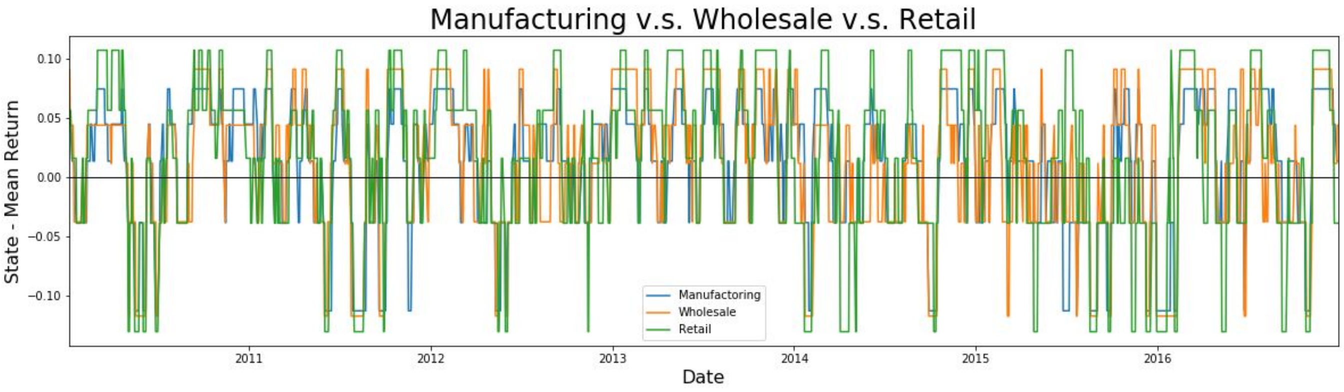


Application: Industry comparison

Case 2: Manufacturing, Wholesale, Retail

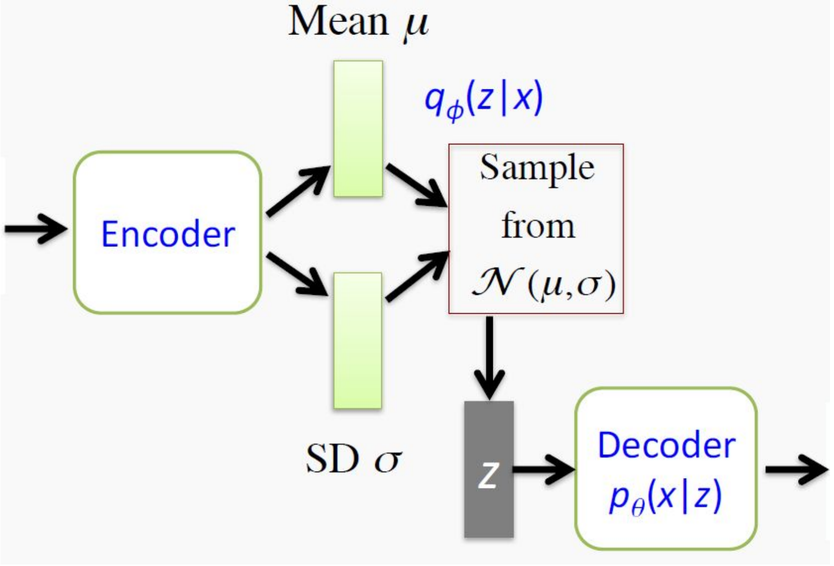
- Similar states
- Similar trend in sequence
- Fluctuation: Manufacturing < Wholesale < Retail





VAE with LSTM

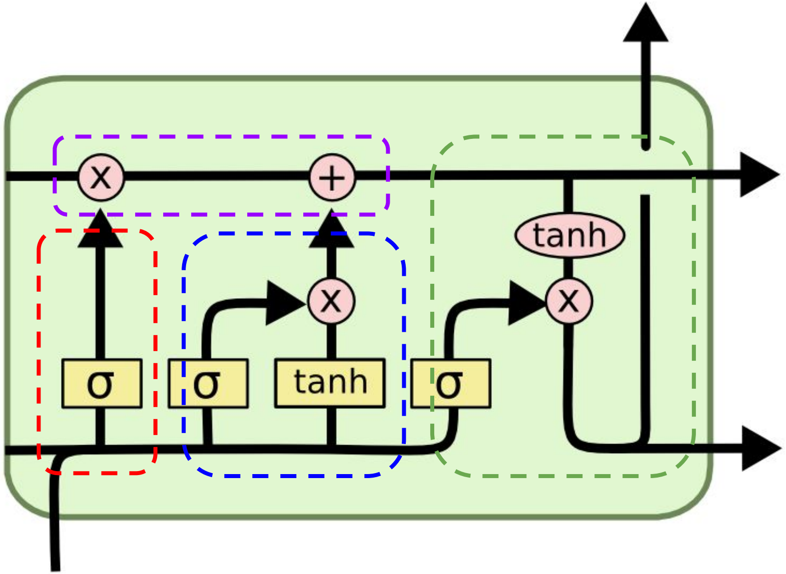
- Motivation
  - Continuous hidden state
  - More complicated emission probability distribution
- VAE
  - Replace the deterministic function
    with a learned posterior q(z|x).
  - Reparameterization trick to calculate gradient.
  - Total loss = KL Divergence loss
    + reconstruction loss



$$\mathcal{L}(\mathbf{x};\theta,\lambda) = D_{KL}\big(\underbrace{q(\mathbf{z}|\mathbf{x};\lambda)}_{\text{decoder}}\,||\,p(\mathbf{z})\big) - \mathbb{E}_{\mathbf{z}\sim q}\log \underbrace{p(\mathbf{x}|\mathbf{z};\theta)}_{\text{encoder}}$$

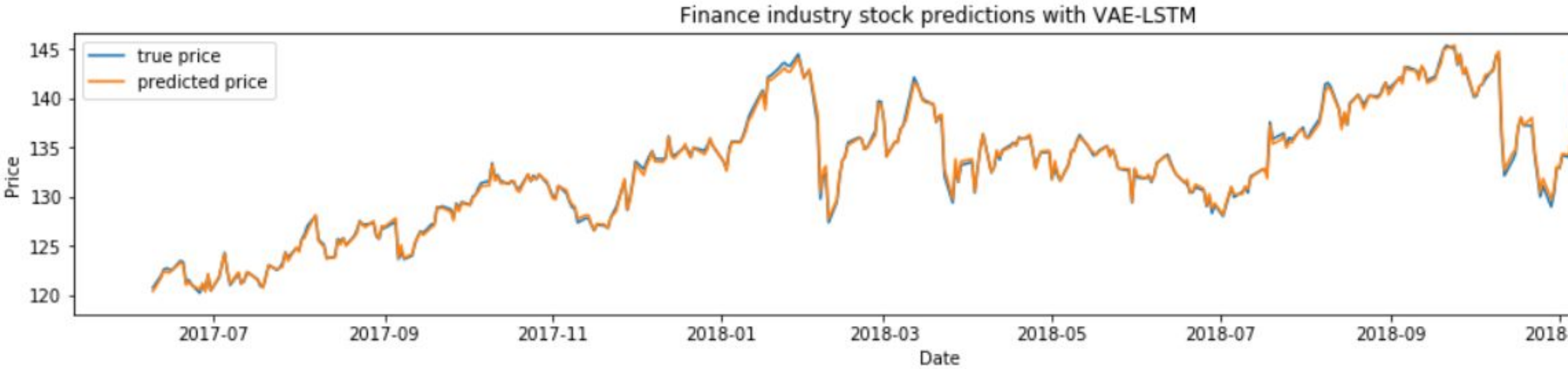
VAE with LSTM

- LSTM
  - Prevent vanishing gradients
  - Forget gate, input gate, output gate, cell gate
- Network Structure
  - Latent dim: 16
  - Time step: 60
  - Output: price sequence with 1 day forward



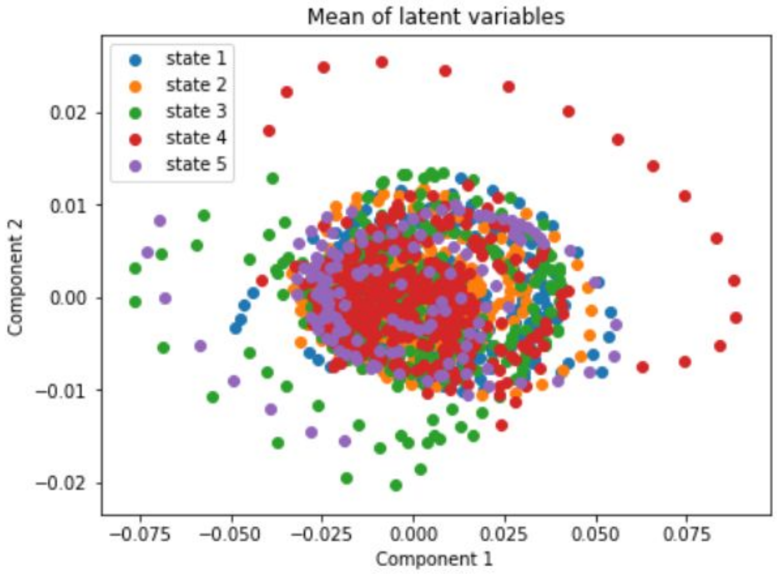
Enco
LSTI
Cell

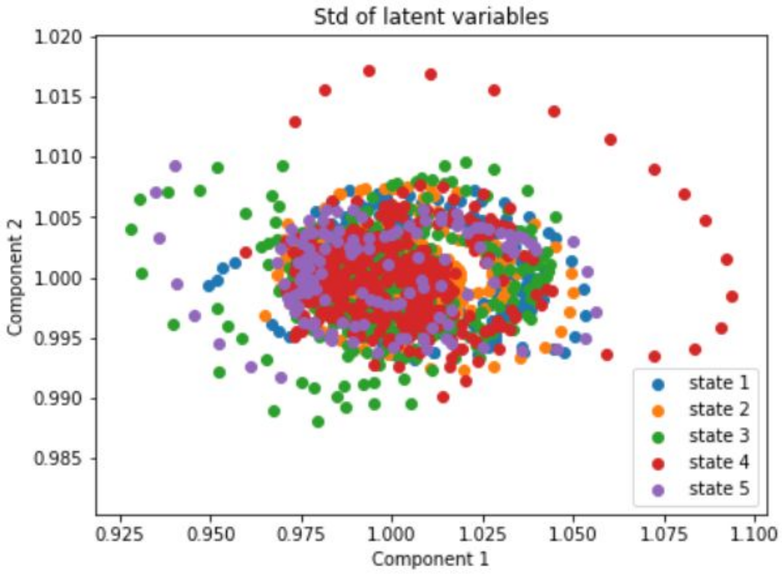
5



VAE-LSTM Results

- Perform PCA on the latent mean and variance and extract top 2 components.
- Color the points using the states achieved by HMM.





Future Work

- Improve model performance: involve more features
  - Volume
  - Technical indicators like MACD

- Model application: Dig deeper into industry comparison using HMM and VAE

Q&A